HUMANA MENTE

ISSUE 20 - FEBRUARY 2012

Philosophy of Self-Deception

EDITED BY
Patrizia Pedrini

Edizioni ETS

Editorial Board

Editorial Staff

TABLE OF CONTENTS

COMMENTARIES

BOOK REVIEWS

INTERVIEWS

Obituary

On January 14th died Paolo Rossi Monti, philosopher of science and scholar of philosophy and of evolution of sciences. He studied Philosophy with Eugenio Garin in Florence 1947 and was research fellow with Antonio Banfi in Milan.

Combining the study of history of science, technology and philosophy, he expounded the interdependence of scientific thought and practice on one hand and technical developments on the other with great lucidity. Paolo Rossi' work on the history of science has been groundbreaking and is still read all over the world. He received many international acknowledgements and prizes, including the Balzan Prize in 2009.

He was a member of our editorial board.

Introduction
Philosophy of Self-Deception

Patrizia Pedrini[†]
patpedrini@gmail.com

The phenomenon of self-deception is one of those topics that, perhaps more than others, is capable of intriguing and fascinating those who decide to devote to it a part of their studies and research. It is also a topic that, once encountered and reflected upon, does not leave us the same as before, in our relationships either with ourselves or with others. This can happen because we get in touch with the psychological event, which is pervasive and complex, and which we feel may have been crucial, for better or for worse, or at least insidious, at many junctures of our own existence. We sense that perhaps many decisions we made – maybe even more than we would be willing to acknowledge - have been made upon one variety or the other of self-deception – that is, upon beliefs that are false, that we additionally may, at times, have the sense that are false, and yet are strongly, sometimes even irresistibly wanted, or desired. Its disconcerting hallmark lies in the fact that we somehow seem to come to believe a proposition that we should at least doubt is likely to be true, and that we seem to do that because of a strong motivation to acquire that false belief. That is why self-deception is included among the so-called "motivated irrationality" phenomena, to which other phenomena also belong, e.g., wishful thinking, cases of precipitate believing under the influence of strong emotions, and so on.

It is thus easy to get caught up in the attempt to analyse it as to the best of our ability, so as to have a coherent description of it, and also a convincing explanation as to why human beings embark on it at all. It is also tempting to believe that, if we can come up with such a description, and such an explanation, we might perhaps be better equipped to identify its occurrence in

[†] Senior Research Fellow in Philosophy funded by the Mensa Society; Fixed-Term Professor, Dept of Philosophy, University of Rijeka; Assistant Fellow, College of Letters and Philosophy, University of Modena & Reggio-Emilia, and Dept. of Philosophy, University of Florence.

ourselves and others, and so, possibly, also try to overcome it. This may be the hope we might want to ascribe to those who believe that self-deception is not a good thing. Other people, however, consider self-deception bliss, by virtue of its allegedly evolutionary, or simply individual, advantages.

Although it was notably described by Donald Davidson[1], in the early days of the debate, as an *intentional* attempt at deceiving oneself, in the hope, among other things, of distinguishing it from other, non-intentional forms of motivated irrationality, many people subscribed later on to the *anti-intentional* view of self-deception promoted by Al Mele (2001), now also referred to as "motivationalism", as Mele replaces the explanatory hypothesis of an intention to deceive oneself with a more palatable, paradox-free explanatory account in which a motivational state, mainly a desire, triggers self-deception and explains it convincingly. After Mele's seminal work, the debate has flourished greatly, and many other related, and vital questions, the way to which was fully paved by Mele's research and the subsequent discussion, have been tackled.

Many of these questions have been brilliantly addressed anew by the authors who have contributed to this issue, but other, brand-new ones have also been posed and argued for.

In his article "When Are We Self-Deceived?", Al Mele provides a sketch of his view about how self-deception happens and, interestingly, he returns to the proposed set of jointly sufficient conditions for entering into self-deception and offers a couple of amendments.

Dion Scott-Kakures gets back critically to the traditional question of intentionalism, in his article: "Can You Succeed in Intentionally Deceiving Yourself?", and argues that if we take the model of interpersonal intentional deception seriously, we ought to conclude that a self-deceiver, so regarded, deceives herself *unintentionally.*

Anna Elisabetta Galeotti ("Self-Deception: Intentional Plan or Mental Event?") also addresses the issue of whether self-deception is an intentional plan or a mental event, and argues that self-deception 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.

[1] See Davidson 1985.

José Eduardo Porcher, in his "Against the Deflationary Account of Self-Deception", critically examines the anti-intentional, deflationary strategy, where the theorist attributes to a subject just one belief – the false belief – as opposed to two beliefs, the true one and the false one, as supposed by intentionalists. He captivatingly suggests that the deflationary view contains a failure that support the neglected view that the self-deceived are not accurately describable as believing either of the relevant propositions.

Eric Funkhouser breaks into new territory, that of "Practical Self-Deception", as his article is titled. He argues that, in the very same sense that we can be self-deceived about belief, we can be self-deceived about matters that concern our practical identities – e.g., our desires, emotions, values, and lifestyles –, and he offers an striking account of where practical self-deception is accommodated.

The thread of the practical issues concerning self-deception is also taken up by Carla Bagnoli, in her "Self-Deception and Agential Authority", and by Dana Kay Nelkin in her "Responsibility and Self-Deception: A Framework". Both of them go on to touch directly on specific moral questions raised by self-deception.

Bagnoli adopts a constitutivist approach to self-deception, which has the merit of explaining the selective nature of self-deception, as well as its being subject to moral sanction, while also describing it as a pragmatic strategy for maintaining the stability of the self, hence being continuous with other rational activities of self-constitution. However, she argues, its success is limited, and its costs are high: it protects the agent's self by undermining the authority she has on her mental life.

Dana Kay Nelkin focuses instead directly on the question of whether and, if so, when people can be responsible for their self-deception and its consequences. In particular, she argues that a particular motivationist account, the "Desire-to-Believe" account, together with other resources, best explains how there can be culpable self-deception, and that self-deception is a good test case for deciding important questions about the nature of moral responsibility.

The "Desire-to-Believe" account is the target of my own contribution, "What Does the Self-Deceiver Want?", where I argue that it is unlikely that the self-deceiver's primary want to believe, or interest in believing that p, occurs as the result of a merely contingent interest in p being true, as one version of such

general account wants us to agree. I also assess various consequences of the view I favour, regarding the self-deceiver's avoidance behaviour, "twisted" self-deception, and whether we should provide a unifying explanation of "straight" and "twisted" self-deception, as we are encouraged to do by the Desire-To-Believe" account defenders.

Julie Kirsch, in her "Narrative and Self-Deception in *La Symphonie Pastorale*", addresses the ever-lasting sceptical issue of whether forging a personal narrative is always at risk of self-deception. She looks at the ways narratives can actually contribute to self-deception, but she argues that not all narratives are invariably self-deceptive. Rather, when they are not, they can make a very positive contribution to self-knowledge and moral understanding.

Mark Young ("The Therapeutic Value of Intellectual Virtue") argues that the development of intellectual character has necessary therapeutic value with regard to self-deception. A motivational/dispositional account of self-deception is offered and linked to a predominant psychological theory of virtuous character worked out by contemporary virtue ethicists and virtue epistemologists.

Lisa Bortolotti and Matteo Mameli ("Self-Deception, Self-Delusion, and the Boundaries of Folk-Psychology") lead us directly into the domain of philosophical psychopathology as well as back to vital and more general philosophical issues, such as the psychological vocabulary we should use to capture and explain some specific mental phenomena, and argue that both self-deception and delusions can be understood in folk-psychological terms. They suggest that there is continuity between the epistemic irrationality manifested in self-deception and in delusion.

Massimo Marraffa ("Remnants of Psychoanalysis. Rethinking the Psychodynamic Approach to Self-Deception") gets back to how self-deception fits the crucial psychoanalytic topic of defence mechanisms. Building on Giovanni Jervis' criticism of psychoanalysis, he sets out to integrate that psychodynamic approach to defence mechanisms fully into the neurocognitive sciences.

In the "Commentaries" section, Clancy Martin and Alan Strudler focus on two texts: Kierkegaard's *Diary of the Seducer* and Shakespeare's *Much Ado About Nothing*, and use the phenomenon of seduction to explore the complicated philosophical and psychological terrain of how truth, trust,

deception and self-deception may interact in a process with which we are all intimately familiar.

Mark A. Wrathall offers an analysis of Sartrean "bad faith" and claims that it amounts to a motivated failure to apprehend the state of dis-integration that exists between one's facticity and transcendence. This "failure to see" is explained by drawing on Merleau-Ponty's account of perceptual ambiguity and existential opacity.

In the "Book Reviews" section, the reader will find Elisabetta Sirgiovanni reviewing Lisa Bortolotti's *Delusions and Other Irrational Beliefs* (OUP, 2010), and Brad Bolman assessing Clancy Martin's collection *The Philosophy of Deception* (OUP, 2009).

Last but not least, we have a "Interview" section, where Professor Amélie O. Rorty agreed to be interviewed by me and generously answered questions on how the self must be to be capable of self-deception, the adaptive fitness of self-deception, its motivational content, the failures of self-knowledge involved in self-deception, and confabulation, and on the lines of research on which she encourages self-deception theorists to embark.

The idea of compiling this issue dates back to July 2010, when I received the invitation to suggest a topic and a team of contributors. The help and encouragement I have had from the members of the editorial board from the outset has been incalculable; the enthusiasm I have encountered in all the contributors who agreed to write a paper and have subsequently been so generously ready to discuss their views with me and other referees unforgettable and immensely instructive. I thank each of the authors warmly for making this issue a busy "virtual lab" that has enabled me to reflect further on the topic. I also thank my diligent assistant, Alice Giuliani, for her decisive help in getting me into, and especially *out of*, the final editing.

REFERENCES

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

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

When Are We Self-Deceived?*

Alfred R. Mele †
almele@fsu.edu

ABSTRACT

This article's point of departure is a proto-analysis that I have suggested of entering self-deception in acquiring a belief and an associated set of jointly sufficient conditions for self-deception that I have proposed. Partly with the aim of fleshing out an important member of the proposed set of conditions, I provide a sketch of my view about how self-deception happens. I then return to the proposed set of jointly sufficient conditions and offer a pair of amendments.

Introduction

In *Self-Deception Unmasked* (Mele 2001) and in earlier work, I tried to show that self-deception is masked by traditional models of the phenomenon that treat it as an intrapersonal analogue of stereotypical interpersonal deception.[1] According to these models, self-deceivers intentionally deceive themselves into believing that *p*, and there is a time at which they believe that *p* is false while also believing that *p* is true. In Mele 2001, I offered an alternative model of self-deception and, drawing heavily on empirical work, I developed a detailed explanation of how garden-variety self-deception happens.

The contributors to this issue have been asked to focus on philosophical aspects of self-deception. I focus here on a question about conceptually sufficient conditions for self-deception. In section 1, I review a proto-analysis that I have suggested of entering self-deception in acquiring a belief and an

* In parts of this article, I draw on Mele 2001 and 2009. This article was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this article are my own and do not necessarily reflect the views of the John Templeton Foundation. I am grateful to an anonymous referee for comments on a draft of this article.

† Florida State University, USA.

[1] For citations of this tradition in philosophy, psychology, psychiatry, and biology, see Mele 2001, p. 125, n. 1. Stereotypical interpersonal deception does not exhaust interpersonal deception.

associated set of jointly sufficient conditions for self-deception that I have proposed. In section 2, partly with the aim of fleshing out an important member of the proposed set of conditions, I provide a sketch of my view about how self-deception happens. In section 3, I return to the proposed set of jointly sufficient conditions and offer two amendments.

1. A Proto-Analysis and Proposed Sufficient Conditions

Although I have never offered a conceptual analysis of self-deception, I have suggested the following proto-analysis of entering self-deception in acquiring a belief: people enter self-deception in acquiring a belief that *p* if and only if *p* is false and they acquire the belief in «a suitably biased way» (Mele 2001, p. 120). The suitability at issue is a matter of kind of bias, degree of bias, and the nondeviance of causal connections between biasing processes (or events) and the acquisition of the belief that *p*. My suggestion is that someone interested in constructing a conceptual analysis of entering self-deception in acquiring a belief can start here and try to work out an account of suitable bias. Of course, an analysis of entering self-deception in acquiring a belief will not be a complete analysis of self-deception if there are other ways of entering self-deception; and, as I have explained elsewhere, people sometimes enter self-deception in *retaining* a belief (Mele, 2001, pp. 56-59). Someone who faultlessly acquires the belief that *p* may later enter self-deception in persisting in believing that *p*. It may be suggested that if a complete analysis of self-deception is constructable, it is constructable out of analyses of these two ways of entering self-deception.[2]

I have also proposed a set of conceptually sufficient conditions for self-deception, as follows:

> *S* enters self-deception in acquiring a belief that *p* if:
>
> 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

[2] Some theorists would definitely reject this suggestion. Robert Audi, for example, contends that no one who is self-deceived about p has a false belief that p; rather, self-deceived people have an unconscious true belief that ~p and – in the absence of a belief that p – sincerely avow that p (1982, 1985, 1997). I criticize Audi's attempted analysis of self-deception in Mele 1982 and 2010; I will not do so again here.

that p, and

4. The body of data possessed by S at the time provides greater warrant for $\sim p$ than for p. (Mele 2001, pp. 50-51; see Mele 1997, p. 95)

I comment briefly on each condition and then forge ahead.

Condition 1 captures a purely lexical point. A person is, by definition, *deceived in* believing that p only if p is *false*; the same is true of being *self-deceived in* believing that p. The condition does not imply that the falsity of p has special importance for the *dynamics* of self-deception. Motivationally biased treatment of data may sometimes result in someone's believing an improbable proposition, p, that happens to be *true*. There may be self-deception in such a case, but the person is not self-deceived in believing that p nor in acquiring the belief that p.

People may be deceived *into* believing something that they are not deceived *in* believing (see Mele 1987, pp. 127-28). Ann might execute a complicated strategy for deceiving Alan into believing something that, unbeknownst to her, is true. And she might thereby cause him to believe this proposition, p. Since p is true, Alan is not deceived *in* believing it. Even so, it is plausible that Ann deceived him *into* believing it, if she caused him to believe that p partly by deceiving him into believing some false propositions suggestive of p.

My discussion of motivated bias and various ways of entering self-deception in the following section puts some flesh on the bones of condition 2. An interpretation of condition 2 will emerge from that section.

My inclusion of the term "nondeviant" in condition 3 is motivated by a familiar problem for causal characterizations of phenomena in any sphere. Specifying the precise nature of nondeviant causation of a belief by motivationally biased treatment of data is a difficult technical task. Mele 2001 provides guidance on the issue.

The thrust of condition 4 is that self-deceivers believe against the weight of the evidence they possess. I do not view 4 as a *necessary* condition of self-deception. In some instances of motivationally biased evidence-gathering, for example, people may bring it about that they believe a falsehood, p, when $\sim p$ is much better supported by evidence readily available to them, even though, owing to the selectivity of the evidence-gathering process, the evidence that they themselves actually *possess* at the time favors p over $\sim p$. In my view, such people are naturally deemed self-deceived, other things being equal. However, some philosophers require that a condition like 4 be satisfied (Davidson 1985, McLaughlin 1988, Szabados 1985), and I do not object to including 4 in a list

of jointly *sufficient* conditions. Of course, in some cases, whether the weight of a person's evidence lies on the side of *p* or of ~*p* (or equally supports each) is subject to legitimate disagreement.

2. Explaining Self-Deception

Elsewhere, I have distinguished between what I call *straight* and *twisted* cases of self-deception (Mele 1999, 2001). In straight cases, which have dominated the literature, people are self-deceived in believing something that they want to be true – for example, that their spouse is not having an affair. In twisted cases, people are self-deceived in believing something that they want to be false (and do not also want to be true). For example, an insecure, jealous husband may believe that his wife is having an affair despite having only thin evidence of infidelity and despite wanting it to be false that she is so engaged (and not also wanting it to be true that she is). In cases of both kinds, as I have explained in Mele 2001 and briefly explain below, self-deceivers have motivationally biased beliefs.

Some illustrations of ways in which our desiring that *p* can contribute to our believing that *p* in instances of straight self-deception will be useful (see Mele 2001, pp. 26–27). Often, two or more of the phenomena I describe are involved in an instance of self-deception.

> 1) *Negative Misinterpretation.* Our desiring that *p* may lead us to misinterpret as not counting (or not counting strongly) against *p* data that we would easily recognize to count (or count strongly) against *p* in the desire's absence. For example, Rex just received a rejection notice on a journal submission. He hopes that the rejection was unwarranted, and he reads through the referees' comments. Rex decides that the referees misunderstood two important but complex points and that their objections consequently do not justify the rejection. However, the referees' criticisms were correct, and a few days later, when Rex rereads his paper and the comments in a more impartial frame of mind, it is clear to him that this is so.
>
> 2) *Positive Misinterpretation.* Our desiring that *p* may lead us to interpret as *supporting p* data that we would easily recognize to count against *p* in the desire's absence. For example, Sid is very fond of Roz, a college classmate with whom he often studies. Because he wants it to be true that Roz loves him, he may interpret her declining his invitations to

various social events and reminding him that she has a steady boyfriend as an effort on her part to "play hard to get" in order to encourage Sid to continue to pursue her and prove that his love for her approximates hers for him. As Sid interprets Roz's behavior, not only does it fail to count against the hypothesis that she loves him, it is evidence that she does love him. This contributes to his believing, falsely, that Roz loves him.

3) *Selective Focusing/Attending*: Our desiring that *p* may lead us to fail to focus attention on evidence that counts against *p* and to focus instead on evidence suggestive of *p*. Beth is a twelve-year-old whose father died recently. Owing partly to her desire that she was her father's favorite, she finds it comforting to attend to memories and photographs that place her in the spotlight of her father's affection and unpleasant to attend to memories and photographs that place a sibling in that spotlight. Accordingly, she focuses her attention on the former and is inattentive to the latter. This contributes to Beth's coming to believe – falsely – that she was her father's favorite child. In fact, Beth's father much preferred the company of her brothers, a fact that the family photo albums amply substantiate.

4) *Selective Evidence-Gathering*: Our desiring that *p* may lead us both to overlook easily obtainable evidence for $\sim p$ and to find evidence for *p* that is much less accessible. For example, Betty, a political campaign staffer who thinks the world of her candidate, has heard rumors from the opposition that he is sexist, but she hopes he is not. That hope motivates her to scour his past voting record for evidence of his political correctness on gender issues and to consult people in her own campaign office about his personal behavior. Betty may miss some obvious, weighty evidence that her boss is sexist – which he in fact is – even though she succeeds in finding less obvious and less weighty evidence for her favored view. As a result, she may come to believe that her boss is not sexist. Selective evidence-gathering may be analyzed as a combination of hyper-sensitivity to evidence (and sources of evidence) for the desired state of affairs and blindness – of which there are, of course, degrees – to contrary evidence (and sources thereof).

In none of these examples does the person hold the true belief that $\sim p$ and then intentionally bring it about that he or she believes that *p*. Yet, if we assume that these people acquire relevant false, unwarranted beliefs in the

ways described, these are garden-variety instances of self-deception; or so I have argued elsewhere.[3] Rex is self-deceived in believing that his article was wrongly rejected, Sid is self-deceived in believing certain things about Roz, and so on.

We can understand why, owing to her desire that her father loved her most, Beth finds it pleasant to attend to photographs and memories featuring her as the object of her father's affection and painful to attend to photographs and memories that put others in the place she prizes. But how do desires that *p* trigger and sustain the two kinds of misinterpretation and selective evidence-gathering? It is not as though these activities are intrinsically pleasant, as attending to pleasant memories, for example, is intrinsically pleasant.

Attention to some sources of *unmotivated* biased belief sheds light on this issue. Several such sources have been identified (Mele 2001, pp. 28–31), including the following three:

> (a) *Vividness of information.* A datum's vividness for us often is a function of such things as its concreteness and its sensory, temporal, or spatial proximity. Vivid data are more likely to be recognized, attended to, and recalled than pallid data. Consequently, vivid data tend to have a disproportional influence on the formation and retention of beliefs (Nisbett and Ross 1980).
>
> (b) *The availability heuristic.* When we form beliefs about the frequency, likelihood, or causes of an event, 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 and Ross, 1980, p. 18). For example, we may mistakenly believe that the number of English words beginning with 'r' greatly outstrips the number having 'r' in the third position, because we find it much easier to produce words on the basis of a search for their first letter (Tversky & Kahnemann, 1973). Similarly, attempts to locate the cause(s) of an event are significantly influenced by manipulations that focus one's attention on a specific potential cause (Nisbett and Ross, 1980, p. 22; Taylor & Fiske, 1975, 1978).

[3] If, in the way I described, Betty acquires or retains the false belief that her boss is not sexist, it is natural to count her as self-deceived. This is so even if, owing to her motivationally biased evidence-gathering, the evidence that she actually has does not weigh more heavily in support of the proposition that her boss is sexist than against it.

(c) *The confirmation bias*. People testing a hypothesis tend to search (in memory and the world) more often for confirming than for disconfirming instances and to recognize the former more readily (Baron, 1988, pp. 259–265). This is true even when the hypothesis is only a tentative one (and not a belief one has). People also tend to interpret relatively neutral data as supporting a hypothesis they are testing (Trope, Gervey, & Liberman, 1997, p. 115).

Although sources of biased belief apparently can function independently of motivation, they also may be triggered and sustained by desires in the production of *motivationally* biased beliefs.[4] For example, desires can enhance the vividness or salience of data. Data that count in favor of the truth of a proposition that one hopes is true may be rendered more vivid or salient by one's recognition that they so count; and vivid or salient data, given that they are more likely to be recognized and recalled, tend to be more "available" than pallid counterparts. Similarly, desires can influence which hypotheses occur to one and affect the salience of available hypotheses, thereby setting the stage for the confirmation bias.[5] Owing to a desire that *p*, one may test the hypothesis that *p* is true rather than the contrary hypothesis. In these ways and others, a desire that *p* may help produce an unwarranted belief that *p*.

An interesting theory of lay hypothesis testing is designed, in part, to accommodate self-deception. I explore it in Mele 2001, where I offer grounds for caution and moderation and argue that a qualified version is plausible.[6] I call it the *FTL theory*, after the authors of the two articles on which I primarily drew, Friedrich 1993 and Trope & Liberman 1996. Here, I offer a sketch of the theory.

The basic idea of the FTL theory is that a concern to minimize costly errors drives lay hypothesis testing. The *errors* on which the theory focuses are false beliefs. The *cost* of a false belief is the cost, including missed opportunities for gains, that it would be reasonable for the person to expect the belief – if false – to have, given his desires and beliefs, if he were to have expectations about such things. A central element of the FTL theory is a "confidence threshold" –

[4] I develop this idea in Mele 1987, ch. 10 and 2001. Kunda 1990 develops the same theme, concentrating on evidence that motivation sometimes primes the confirmation bias. Also see Kunda 1999, ch. 6.

[5] For motivational interpretations of the confirmation bias, see Friedrich 1993 and Trope and Liberman 1996, pp. 252–265.

[6] See Mele 2001, pp. 31–49, 63–70, 90–91, 96–98, 112–18.

or a "threshold," for short. The lower the threshold, the thinner the evidence sufficient for reaching it. Two thresholds are relevant to each hypothesis: «The acceptance threshold is the minimum confidence in the truth of a hypothesis,» *p*, sufficient for acquiring a belief that *p* «rather than continuing to test [the hypothesis], and the rejection threshold is the minimum confidence in the untruth of a hypothesis,» *p*, sufficient for acquiring a belief that ~*p* «and discontinuing the test» (Trope & Liberman, 1996, p. 253). The two thresholds often are not equally demanding, and acceptance and rejection thresholds respectively depend «primarily» on «the cost of false acceptance relative to the cost of information» and «the cost of false rejection relative to the cost of information». The "cost of information" is simply the «resources and effort» required for gathering and processing «hypothesis-relevant information» (p. 252).

Confidence thresholds are determined by the strength of aversions to specific costly errors together with information costs. Setting aside the latter, the stronger one's aversion to falsely believing that *p*, the higher one's threshold for belief that *p*. These aversions influence belief in a pair of related ways. First, because, other things being equal, lower thresholds are easier to reach than higher ones, belief that ~*p* is a more likely outcome than belief that *p*, other things being equal, in a hypothesis tester who has a higher acceptance threshold for *p* than for ~*p*. Second, the aversions influence *how* we test hypotheses – for example, whether we exhibit the confirmation bias – and *when we stop* testing them (owing to our having reached a relevant threshold).[7]

Friedrich claims that desires to avoid specific errors can trigger and sustain «automatic test strategies» (1993, p. 313), which supposedly happens in roughly the nonintentional way in which a desire that *p* results in the enhanced vividness of evidence for *p*. In Mele 2001 (pp. 41–49, 61–67), I argue that a person's being more strongly averse to falsely believing that ~*p* than to falsely believing that *p* may have the effect that he primarily seeks evidence for *p*, is more attentive to such evidence than to evidence for ~*p*, and interprets relatively neutral data as supporting *p*, without this effect's being mediated by a belief that such behavior is conducive to avoiding the former error. The stronger aversion may simply frame the topic in a way that triggers and sustains

[7] Whether and to what extent subjects display the confirmation bias depends on such factors as whether they are given a neutral perspective on a hypothesis or, instead, the perspective of someone whose job it is to detect cheaters. See Gigerenzer & Hug 1992.

these manifestations of the confirmation bias without the assistance of a belief that behavior of this kind is a means of avoiding particular errors. Similarly, having a stronger aversion that runs in the opposite direction may result in a skeptical approach to hypothesis testing that in no way depends on a belief to the effect that an approach of this kind will increase the probability of avoiding the costlier error. Given the aversion, skeptical testing is predictable independently of the agent's believing that a particular testing style will decrease the probability of making a certain error.

The FTL theory applies straightforwardly to both straight and twisted self-deception. Friedrich writes:

> a prime candidate for primary error of concern is believing as true something that leads [one] to mistakenly criticize [oneself] or lower [one's] self-esteem. Such costs are generally highly salient and are paid for immediately in terms of psychological discomfort. When there are few costs associated with errors of self-deception (incorrectly preserving or enhancing one's self-image), mistakenly revising one's self-image downward or failing to boost it appropriately should be the focal error. (1993, p. 314)

Here, he has straight self-deception in mind, but he should not stop there. Whereas for many people it may be more important to avoid acquiring the false belief that their spouses are having affairs than to avoid acquiring the false belief that they are not so engaged, the converse may well be true of some insecure, jealous people. The belief that one's spouse is unfaithful tends to cause significant psychological discomfort. Even so, avoiding falsely believing that their spouses are faithful may be so important to some people that they test relevant hypotheses in ways that, other things being equal, are less likely to lead to a false belief in their spouses' fidelity than to a false belief in their spouses' infidelity. Furthermore, data suggestive of infidelity may be especially salient for these people and contrary data quite pallid by comparison. Don Sharpsteen and Lee Kirkpatrick observe that «the jealousy complex» – that is, «the thoughts, feelings, and behavior typically associated with jealousy episodes» – is interpretable as a mechanism «for maintaining close relationships» and appears to be «triggered by separation, or the threat of separation, from attachment figures» (1997, p. 627). It certainly is conceivable that, given a certain psychological profile, a strong desire to maintain one's relationship with one's spouse plays a role in rendering the potential error of falsely believing one's spouse to be innocent of infidelity a "costly" error, in the FTL sense, and more costly than the error of falsely

believing one's spouse to be guilty. After all, the former error may reduce the probability that one takes steps to protect the relationship against an intruder. The FTL theory provides a basis for an account of both straight and twisted self-deception (Mele 2001, ch. 5).

3. Proposed Sufficient Conditions Revisited

I return to my proposed set of jointly sufficient conditions for entering self-deception in acquiring a belief. Some philosophers have argued that my four conditions fall short of collective sufficiency because they do not capture a kind of *tension* that is necessary for self-deception. According to Robert Audi, this tension «is ordinarily represented [...] by an avowal of *p* [...] *coexisting* with knowledge or at least true belief that ~*p*» (1997, p. 104). Eric Funkhouser claims that self-deception requires tension between some of the agent's behavior and certain of her sincere avowals (2005, p. 304). Michael Losonsky contends that self-deceivers have the unwarranted, false belief that *p*, lack the true belief that ~*p*, and have evidence for ~*p* that is «active» in their «cognitive architecture» (1997, p. 122). This activity, he claims, is manifested in such indications of tension as recurrent or nagging doubts, and he uses the contention that self-deception conceptually requires such conflict to support a distinction between self-deception and instances of "prejudice" or "bias" that satisfy the quartet of conditions I offered as conceptually sufficient for entering self-deception. Mike W. Martin mentions a similar tension, «a cognitive conflict» such as «suspecting *p* and believing ~*p*» (1997, p. 123). And Kent Bach maintains that self-deception requires actively avoiding or suppressing certain thoughts, or ridding oneself of these thoughts when they occur (1997; also see Bach 1998, pp. 167–168).

The quartet of conditions I offered certainly does not entail that there is no tension in self-deception. Nor do I claim that self-deception normally is tension-free. Significant tension may be present in most people who satisfy my four conditions. But the issue raised by the authors mentioned in the preceding paragraph is whether the alleged kinds of tension are conceptually *necessary* for entering self-deception. And my answer has been *no*. As I see it, given the details of Rex's story, even if he is tension-free during the process of acquiring the belief that his article was wrongly rejected and while that belief is in place, he is self-deceived and he enters self-deception in acquiring that

belief. In my view, the same is true of bigots who, without psychic conflict, satisfy my four conditions in acquiring a bigoted belief that *p*.

The primary topic of the present section is conceptually sufficient conditions for entering self-deception in acquiring a belief – not individually necessary and jointly sufficient conditions for this. Different philosophers require different kinds of tension for self-deception, as the first paragraph of this section suggests; and I have argued that tension of the various kinds at issue is not required for self-deception (Mele 2001). But even if I am right in keeping tension off a list of necessary conditions of self-deception, it may appear on a useful list of jointly sufficient conditions. The following addition to my proposed quartet of jointly sufficient conditions (which resembles Martin's condition of suspecting that the pertinent proposition one believes is false [1997, p. 123)]) would result in a less latitudinarian proposal about sufficient conditions for entering self-deception: (5) *S* consciously believes at the time that there is a significant chance that $\sim p$ (see Mele 2001, pp. 71–73 and 2010, p. 749). For example, the resulting proposal would not entail that tension-free Rex enters self-deception in acquiring the belief that his submission was wrongly rejected.

The second and third conditions in my proposed set of sufficient conditions include the expressions "*S* treats data" and "This biased treatment." I intended my discussion (in Mele 1997 and 2001) of various ways of entering self-deception in acquiring a belief that *p* to provide guidance on the interpretation of "treats" and "treatment" in these conditions. But if, strictly speaking, relatively simple motivationally biased misperception counts as motivationally biased *treatment* of data (given the standard meaning of "treats data"), trouble is brewing. Imagine that a hungry cat misperceives a noise as the sound of her food being shaken into a bowl and runs into the room from which the noise is emanating (Scott-Kakures 2002, pp. 578–580). Those who are happy to attribute beliefs to cats may be happy to say that the cat has a belief to the effect that food is available, and that belief may be a relatively direct product or a constituent of her motivationally biased misperception of the noise. If feline self-deception is out of the question and if "treats data" has a broader sense than I intended, then something should be done about "treats" in condition 2 or a useful condition should be added. How should this be handled?

Dion Scott-Kakures argues that «reflective, critical reasoning is essential to the process of self-deception» (2002, p. 577) and that «the error of self-

knowledge that makes [...] self-deception possible is a misconception about what animates [the believer's] doxastic or cognitive activities. Like any reflective reasoner, she will regard her investigations as directed by [...] her grasp upon what reason recommends,» but she is wrong about this (p. 599). «Her investigations are directionally driven by desire or interest» (p. 599), in ways featured in my account of how self-deception happens. If Scott-Kakures is right in requiring these things for self-deception, something like the following condition should be added to my proposed sufficient conditions for *S*'s entering self-deception in acquiring a belief that *p*: (6) *S*'s acquiring the belief that *p* is a product of "reflective, critical reasoning," and *S* is wrong in regarding that reasoning as properly directed.[8] I have no objection to including condition 6 in a list of jointly sufficient conditions for entering self-deception in acquiring a belief that *p*.

Putting things together, I arrive at the following statement of proposed jointly sufficient conditions for entering self-deception in acquiring a belief:

> *S* enters self-deception in acquiring a belief that *p* if:
>
> 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 $\sim p$ than for *p*
> 5. *S* consciously believes at the time that there is a significant chance that $\sim p$
> 6. *S*'s acquiring the belief that *p* is a product of "reflective, critical reasoning," and *S* is wrong in regarding that reasoning as properly directed.

My primary aim in previous work on self-deception has been to explain how it happens. The explanation I developed elsewhere and sketched in section 2 applies straightforwardly to cases in which these six conditions are satisfied.

[8] Scott-Kakures motivates a condition of this kind not only by means of reflection on the case of the hungry cat, but also by means of reflection on «"precipitate cases" of motivated believing» in human beings (2002, p. 587), cases in which a person leaps to a motivationally biased conclusion in the absence of reflective reasoning.

REFERENCES

Audi, R. (1982). Believing and Affirming. *Mind*, *91*(361), 115–120.

Audi, R. (1985). Self Deception and Rationality. In M. Martin (Ed.), *Self Deception and Self Understanding*. Lawrence: University of Kansas Press, 169–194.

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

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

Bach, K. (1998). (Apparent) Paradoxes of Self-Deception and Decision. In J. Dupuy (Ed.), *Self-Deception and Paradoxes of Rationality*. Cambridge: Cambridge University Press, 163–189.

Baron, J. (1988). *Thinking and Deciding*. Cambridge: Cambridge University Press.

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

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

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

Gigerenzer, G., & Hug, K. (1992). Domain-Specific Reasoning: Social Contracts, Cheating, and Perspective Change. *Cognition*, *43*, 127–171.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin 108*(3), 480–498.

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

Losonsky, M. (1997). Self-Deceivers' Intentions and Possessions. *Behavioral and Brain Sciences*, *20*(1), 121–122.

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

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.

Mele, A.R. (1982). Self-Deception, Action, and Will: Comments. *Erkenntnis*, *18*(2), 159–164.

Mele, A.R. (1987). *Irrationality*. New York: Oxford University Press.

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

Mele, A.R. (1999). Twisted Self-Deception. *Philosophical Psychology*, *12*(2), 117–137.

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.

Mele, A.R. (2010). Approaching Self-Deception: How Robert Audi and I Part Company. *Consciousness and Cognition*, *19*(3), 745–750.

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

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

Sharpsteen, D, & Kirkpatrick, L. (1997). Romantic Jealousy and Adult Romantic Attachment. *Journal of Personality and Social Psychology*, *72*(3), 627–640.

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

Taylor, S., & Fiske, S. (1975). Point of View and Perceptions of Causality. *Journal of Personality and Social Psychology*, *32*(3), 439–445.

Taylor, S., & Fiske, S. (1978). Salience, Attention and Attribution: Top of the Head Phenomena. In Leonard Berkowitz (Ed.), *Advances in Experimental Social Psychology*, vol. 11. New York: Academic Press, 249-288.

Trope, Y., & Liberman, A. (1996). Social Hypothesis Testing: Cognitive and Motivational Mechanisms. In E. Higgins & A.W. Kruglanski (Eds.), *Social Psychology: Handbook of Basic Principles*. New York: Guilford Press, 239–270.

Trope, Y., Gervey, B., & Liberman, N. (1997). Wishful Thinking from a Pragmatic Hypothesis-Testing Perspective. In M. Myslobodsky (Ed.) *The Mythomanias: The Nature of Deception and Self-Deception*. Mahwah, NJ: Lawrence Erlbaum, 105–131.

Tversky, A., & Kahnemann, D. (1973). Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, *5*(2), 207–232.

Can You Succeed in Intentionally Deceiving Yourself? *

Dion Scott-Kakures †
dion.scott-kakures@scrippscollege.edu

ABSTRACT

According to intentionalists, self-deceivers exercise the sort of control over their belief-forming processes that, in standard cases of interpersonal deception, the deceiver exercises over the deceived's belief forming processes – they intentionally deceive themselves. I'll argue here that interpersonal deception is not an available model for the sort of putatively distinctive control the self-deceiver exercises over her belief-forming processes and beliefs. I concentrate attention on a kind of case in which an agent allegedly intentionally causes herself to come to have a false belief. I hope to show that contrary to appearances, the agents in such cases do not intentionally cause themselves to have false beliefs – do not intentionally deceive themselves. Indeed, if we take the model of interpersonal intentional deception seriously, we ought to conclude that a self-deceiver, so regarded, deceives herself *unintentionally*.

1. Introduction

We are all familiar with the unhappy fact that we frequently deceive ourselves – cause ourselves to have false beliefs. If this sounds hyperbolic or alarming, it should be recalled that, typically, we cause ourselves to have false beliefs in unintentional fashion.[1] Aiming to settle a question of the form "p or not-p?", I may, for example, decide to consult a friend knowledgeable about such

* I owe thanks to Yuval Avnur, Paul Hurley, and Rivka Weinberg for helpful discussion.
† Scripps College, Claremont, CA, USA.
[1] For a discussion of lexical considerations relating to the use of "deception" and "deceive," see Mele 2001, Chapter 1.

matters. I ask and she answers: "p." I think, "She's always right," and I come to believe that p. Alas, she's mistaken. I've caused myself to have the false belief that p and, so, I deceive myself – but not intentionally so. Deflationists about self-deception point out that such unintentional causings of ourselves to have false beliefs can frequently have a motivational or affective basis. In a simple sort of case, I may, as a result of my desire that p, spend much time thinking about p, and this may well make data in support of p very salient. I may come, in unwarranted and motivationally biased fashion, to believe that p.[2] If p is false, I've unintentionally caused myself to come to have a false belief. A fundamental issue raised in the investigation of self-deception is whether appeal to such depressingly familiar features of our cognitive lives is sufficient for the explanation of the phenomenon. Those who reject the explanatory sufficiency of such spare resources very often insist that self-deception requires more. Intentionalists insist that *real* self-deception demands that a subject *intentionally* deceive herself – that is, a self-deceiver must intentionally cause herself to come to have a false belief.

It's no doubt the case that some intentionalists find comfort in the intuition that any phenomenon worth calling "self-deception" (as distinguished from, say, "wishful thinking") must follow the contours of the processes underlying prototypical cases of interpersonal deception. Even so, it's worth noting that there's additional powerful intuitive basis for such a view. Self-deceivers very frequently believe in the teeth of the evidence and often regard as evidence for p just what the rest of us take to be – and obviously so – evidence for not-p. A self-deceiver's doxastic behavior is sometimes so striking that we are tempted to ask, "How can you possibly *believe* that?"[3] It's natural, then, to entertain the suspicion that some distinctive explanation of the self-deceiver's doxastic behavior is required. A seductive diagnosis is that self-deceivers display the light-fingered and strategic behavior characteristic of means-end rationality and, so, of intentional activity. Self-deceivers explain away just what must be explained away in order to embrace some favored proposition; they search for

[2] See Mele 1997 and 2001 for a defense of the deflationist account of self-deception and a very influential characterization of the dispute between the intentionalist and deflationist.

[3] Of course, the very nature of the phenomenon, self-deception, is increasingly disputed. In particular, it has been denied that those we describe as "self-deceived" believe what they are self-deceived about. See, for example, Gendler 2007 and Elga 2009. Here I'll just take it for granted that those who are self-deceived *believe* what we take them to be self-deceived about.

evidence favoring their focal hypotheses; they do not consider just what must not be considered. Self-deceivers, it seems, aren't trying to settle a question of the form "p or not-p?"[4] Rather, they are trying to come to believe a particular proposition. They intend to deceive themselves – they try to cause themselves to hold a false belief or try to come to believe that p regardless of the truth of p.[5] Moreover, it seems, they sometimes succeed in coming to believe that p and, so, succeed in intentionally deceiving themselves. The powerful suspicion, then, is that self-deceivers intelligently and intentionally direct, control or guide their belief-forming processes in ways that truth-oriented (or at least, non-self-deceptive) hypothesis testers do not. Their belief-forming processes are sensitive and responsive to their practical (and non-epistemic) aims in the way our intentional behavior is, generally, so sensitive and responsive. Only something like this could explain the remarkable doxastic behavior of self-deceivers. That, at least, is the intuition.[6]

[4] One way of characterizing the different aims of the intentional self-deceiver and the normal hypothesis tester is to note that the self-conscious inquirer, in aiming to settle a question, turns to the world, seeking considerations that bear on her question. An upshot of this is that, if things go well, her evidentiary or reasons condition will become determinative for belief is the following sense:
She'll come to believe that p, if by her then current lights she has sufficient reason to believe that p; or
She'll come to believe that not-p, if by her then current lights she has sufficient reason to believe that not-p.
In self-deception, as imagined by the intentionalist, things are different. The self-deceiver isn't interested in settling a question. She doesn't aim to turn to the world to seek considerations that bear on her question. Her explicit aim is precisely *not* to be doxastically open to alternatives (1) and (2). In this way, the aims or intentions of the putative intentional self-deceiver and the subject engaged in settling a question are at odds with each other. They are inconsistent aims.

[5] William Talbott (1994) characterizes the goal of self-deception so: «It [...] involves intentionally biasing one's cognitive processes to favor belief in p, due to a desire to believe that p regardless of whether p is true» (p. 30). Talbott rejects a contradictory beliefs requirement. In rejecting such a requirement, he aims, thereby to avoid a strong "divisionist" or partitioning account of the self in self-deception (p. 29). Jose Bermúdez makes note of three distinct ways – «in ascending order of strength» – in which we might characterize "core episodes" of self-deception à la intentionalism: (1) as involving «the intention to bring it about that one acquires a certain belief»; (2) as involving «holding a belief to be false and yet intending to bring it about that one acquires that belief»; and, (3), as involving «intending to bring it about that one acquires a false belief» (2000, p. 310). Nothing I say hinges on a contradictory belief requirement.

[6] This is, I think, one way of putting the perennial appeal of traditional accounts of self-deception. In familiar fashion, a traditionalist about self-deception will hold of a self-deceiver that:

1) He believes some proposition, not-p – or believes that, given the evidence, he ought to believe that not-p).
2) He engages in intentional activity the aim of which is his acquisition of the belief that p.

In this paper I focus a critical eye directly upon intentionalism about self-deception. Needless to say, intentionalism has been the object of intensive critical scrutiny by skeptics (Haight, 1980) and by deflationists (Mele, 2001) about self-deception. Much of this work has focused upon the difficulties implicated in the effort to carry out the intention to deceive oneself.[7] I am less interested in the allegedly self-defeating nature of such an effort *per se*, than I am in trying to get a grip upon the nature of the control over their belief-forming processes that self-deceivers, à la intentionalist accounts, exercise. The intentionalist holds that a self-deceiver

i. intentionally deceives herself; that is, she
ii. intentionally causes herself to come to have a false belief. [8]

Thus, the self-deceiver exercises the sort of control over her belief-forming processes that, in standard cases of interpersonal deception, the deceiver exercises over the deceived's belief forming processes and that we, more generally exercise in intentionally altering states of affairs in the broader world in non-basic action. This is the distinctive form of control over her beliefs that the self-deceiver exercises.

I'll argue here that interpersonal deception is not an available model for the sort of allegedly distinctive control the self-deceiver exercises over her belief-forming processes and beliefs. Such a view can seem plausible only by failing to recognize the real limits on our capacity to exert intentional or agential control

3) He believes, at least for a time, both that not-p and that p.

In a much cited passage, Donald Davidson appears to have embraced these three elements of a traditionalist conception of the phenomenon; he puts it so: «The acquisition of a belief will make for self-deception only under the following conditions: A has evidence on the basis of which he believes that p is more apt to be true than its negation; the thought that p, or the thought that he ought rationally to believe that p, motivates A to act in such a way as to cause himself to believe the negation of p. The action involved may be no more than an intentional turning away from the evidence in favor of p, or it may involve the active search for evidence against p. All that self-deception demands of the action is that the motive originate in a belief that p is true [...] and that the action be performed with the intention of producing belief in the negation of p. Finally, and this is what makes self-deception a problem, the state that motivates the self-deception and the state that produces it co-exist.» (1985, p. 145)

[7] See, for example, Alfred Mele's discussion of the "dynamic puzzle" (2001, pp. 13-14).

[8] The deflationist holds that self-deception is, rather, a matter of a subject

iii. non- or unintentionally deceiving herself; that is, she
iv. non- or unintentionally causes herself to come to have a false belief.

over our doxastic states.[9] In this essay, I concentrate attention on a kind of case in which an agent allegedly intentionally causes herself to come to have a false belief – a kind of case that has long-figured in discussions of self-deception but whose significance, if I am right, has not been fully appreciated. I hope to show that contrary to appearances, the agents in such cases do not intentionally cause themselves to have false beliefs – do not intentionally deceive themselves. Indeed, if we take the model of interpersonal intentional deception seriously, we ought to conclude that a self-deceiver, so regarded, deceives herself *unintentionally*. I conclude that the failure of intentionalism – or at least an intentionalism that looks to the sort of control a deceiver exercises over the deceived in interpersonal deception – in such cases constitutes indirect support for deflationist accounts of self-deception.

2. Intentional Self-Deception?

An instance of the sort of case I have in mind is this:

Happy Days : Sammy is a talented, youngish mathematician. Since his youth he's devoted himself to his career and he has enjoyed some not inconsiderable professional success and acclaim. Still, his devotion to mathematics has taken a toll on other areas of his life. He has no real friends, no lovers, no hobbies or other avocations. Sammy knows that colleagues and acquaintances derive great satisfaction from these things. He understands that there is joy attached to human intimacy but, he thinks, so long as he can do and appreciate good mathematics, he is satisfied, indeed delighted, with the trajectory of his life. Even so, there is a problem: Sammy knows that one's ability to do creative and original mathematics ebbs dramatically as one ages. Worse, still, is the fact that Sammy's family has a depressingly systematic history of early on-set Alzheimer's disease. So, not only is there reason to believe that at a certain point in his life he will be unable to gain satisfaction from the pursuit of mathematics, there is reason for believing that he will be unable to reflect backwards upon his past achievements or to take delight in the work of younger mathematicians. Sammy does believe, however, that he might

[9] My argument here, it's worth noting, is, by my lights, a development of a suggestion made by Jon Elster that beliefs are instances of states that are essentially "by-products" – states that cannot be «brought about intelligently and intentionally» (1983, p. 43). Elster also notes that such states are such as to resist or thwart «indirect as well as direct attempts to bring them about» (1983, p. 57).

well gain satisfaction, even after the on-set of illness, from reflecting backwards on a life devoted to less intellectually demanding pursuits. Of course, Sammy might change his ways now and seek out human companionship and intimacy. But why should he? The pursuit of mathematics is what now offers him the greatest satisfaction. It is far from obvious that discounting the future in this way is irrational. It seems as if Sammy is leading his life up a blind alley. But there is a solution. He now embarks upon a complex strategy designed to bring it about that he come, later in life, to believe that he has led a life rich in human connections. He fills many notebooks detailing imagined friendships, love-affairs and travels. He offers a bounty to those he engages via social media who send photographs, postcards, and letters, and other memorabilia detailing imagined intimacies with him. He secures the services of a trustee who will make certain that the relevant materials are delivered when likely to prove effective. There's no real barrier to our imagining that this strategy could succeed in the way Sammy foresees. We can imagine that, many years later, as he sits in bed at an Alzheimer's center, he's asked by an inquisitive volunteer if he has many friends or has traveled to exotic places. Sammy may say "I don't remember." Seeing the many boxes marked "friends" and "travels," the volunteer may ask, "Perhaps we should look in those? And Sammy may reply, "Yes, let's do that." He is delighted to discover that, as he now comes to believe, he has led a life that touched (and was touched by) so many others.

Such cases have been regarded by some as obvious cases of intentionally causing oneself to come to have a false belief, by others as obviously *not* cases of self-deception and by, still, others as unclear cases of self-deception. Mark Johnston (1988), Alfred Mele (2001), Brian McLaughlin (1988) and Donald Davidson (1985) all consider structurally similar cases.

Davidson writes of such a case that it «is not a pure case of self-deception, since the intended belief is not *sustained* by the intention that produced it, and there is not necessarily anything irrational about it» (1985, p. 145, n. 5). The chief source of Davidson's worry about counting such a case as a case of self-deception is his conviction that robust self-deception involves a continuing form of internal irrationality that requires the subject, a least for a time, to have contradictory beliefs. As he puts it, «the state that motivates the self-deception and the state that produces it co-exist» (1985, p. 145). Since I am concerned here solely with the demand of the intentionalist that self-deception requires that the agent intentionally cause herself to have a false belief, I cannot rely upon this sort of worry.

Brian McLaughlin (1996, p. 41) notes that one basis for holding that such cases do not count as self-deception (but are, rather, instances of what he usefully terms «self-induced deception») is that, e.g., Sammy's belief, given his evidence, is not epistemically unwarranted but being self-deceived with respect to p requires that one's belief that p be epistemically unwarranted.[10] While I agree that this is a symptom of the fact that that Sammy's isn't a case of intentional self-deception, I can imagine an interlocutor insisting that this is, rather, a mark of *really* successful intentional self-deception. After all, in successful cases of interpersonal deception, the belief the deceived individual comes to have is typically warranted. Needless to say, this reply is all the more plausible if we jettison the contradictory beliefs requirement for self-deception.

While Mele notes that such cases are «remote» from «garden variety self-deception» (2001, p. 16), he does conclude that such cases make clear that «[i]ntentionally deceiving oneself is unproblematically possible» (2001, p. 16). After all, if intentional deception is a matter of intentionally causing a subject to believe what is false then, e.g., Sammy's causing himself to believe what is false seems no less intentional than if, say, he had perpetrated the ruse on his aged father. Sammy has a plan for bringing it about that he comes to believe as he does. Events transpire as he foresees. Surely, in such circumstances he intentionally deceives himself – intentionally causes himself to come to have a false belief.

So, even if, as Mele rightly notes, Sammy's case is very different from typical cases of self-deception, such cases apparently display the fact that self-deception *can* be modeled on interpersonal deception. And this is a fact – if it is a fact – that the intentionalist might hope to exploit.[11]

Mark Johnston makes dialectical use of such cases: his aim is to show that cases like Sammy's make essential use of an «autonomous means» – a means to an end the operation of which does not require and, sometimes, does not permit agential attention to them «under the description «means of producing

[10] In this regard, it's worth noting that Sammy in *Happy Days* would appear not to satisfy Mele's "impartial observer test" for self-deception. See Mele 2000 (pp. 106-110) and 2003 (p. 164).

[11] For example, Bermúdez (2000) might well be understood to exploit this fact in his defense of intentionalism.

in me the desired belief'» (1988, p. 77).[12] This is in aid of showing that cases of self-deception that do *not* involve such means involve sub-intentional mechanisms rather than intentional activity. In his consideration of cases like *Happy Days*, Johnston writes:

> [I]t is important that one does not intend or monitor the process throughout. But, then, the operation of the means, though intended to occur, is not an intentional act and neither is the outcome produced by the means, although it is an intended outcome of a means one set in motion. [...] One intended to deceive oneself by arranging misleading evidence and taking the amnestic drug. But what one did in arranging the evidence and taking the amnestic drug did not itself constitute self-deception. Only the cooperation of future events made what one did deserve the name of deceiving oneself by arranging misleading evidence and taking the drug. So: [...] nothing that itself constitutes motivated believing or motivated cessation of (conscious) belief is an intentional act. In cases of self-deception and repression in which autonomous means are employed, the motivated believing and accompanying repression are constituted by the intentional acts of setting the means in motion plus the brute operation of the means culminating in the belief and the forgetting. [...] Even where there is a self-deceptive or repressive action plan, no intentional act is intrinsically a self-deception or a forgetting. (1988, p. 78)

I'm in sympathy with these remarks. Still, if we are modeling self-deception on interpersonal deception, it is not apparent why Sammy's actions (generating the false evidence, arranging to have it delivered) are any less "intrinsically" (or otherwise) acts of intentional deception than various acts that constitute interpersonal deception. In interpersonal deception, in the simplest sort of case, if there is an act that *is* an act of deception, it is presumably my act of saying to you that q (when we both regard $q \rightarrow p$ to be obvious), with the aim of getting you to believe falsely that p. Issues of causal deviance aside, if my act causes you to come to believe that p, I've intentionally deceived you. If to intentionally deceive is to intentionally *cause* another (or oneself) to believe

[12] In Sammy's case there are various autonomous means: his anticipated cognitive decline together with his arranging of the materials to be delivered to him at the appropriate time, etc. Autonomous means figure in various practical contexts, of course. The Soviets' doomsday device in Stanley Kubrick's *Dr. Strangelove* is an autonomous means. Autonomous means, in more familiar contexts, operate to produce ends in the face of, for example, the anticipated failure of attention or a short-term change of preference.

what is false, then, the act which is intentionally performed (with an eye to producing false belief) is the act which is the act of deception.[13]

The mere fact that Sammy (and others like him) no longer consciously intends to deceive himself for some period of time prior to coming to believe that p should, by itself, be no obstacle to our viewing Sammy as intentionally deceiving himself. Certainly, in a case of interpersonal deception, as with other such cases of non-basic actions, once I do whatever I do – for example, assert that q – to initiate the casual chain that results in your coming to believe that p, my contribution is over. I need no longer actively intend or monitor the situation. Indeed, as deceiver I could *die* during the temporal interstice between my act and the deceived's coming to believe and, yet, I would, nonetheless, count as having deceived you.[14] In familiar cases of non-basic action, say, sinking a putt, my contribution is over – body English aside – once I strike the ball. Yet I sink the putt, if acting as I do, I cause the ball to drop into the cup. So it can't be the fact that, in Sammy's case, there's a point at which he can't or doesn't intervene in his deception that makes it the case that his self-deception is not intentional.

So, should we conclude that *Happy Days* and other similar cases are cases of intentional self-deception? We should resist such a conclusion. In this regard, we do well, I think, to ask how a subject might *try to bring it about that he unintentionally deceives himself that p*. (We can imagine that something important – a large wager or his life – hinges upon his coming to believe that p and upon his doing so in unintentional fashion.) It seems to me that he might do this via an effort to arrange evidence in such a way that, at some later point, he comes to believe that p and that he does so as a result of his, then, good-faith effort to settle the question "p or not-p?" If this is so, Sammy's effort to deceive himself intentionally and our current subject's effort to unintentionally deceive himself look to be no different.

It might, I suppose, be suggested that someone who aims to bring it about that he non- or unintentionally deceives himself that p must resort to other sorts of maneuvers. Perhaps, what such a subject must do is, e.g., to seek out experts on p-related matters and simply ask "p or not-p?", believing that they are experts but *hoping*, somehow, that they will offer erroneous counsel. In

[13] Presumably, whatever we mean by an act that *is* an act of self-deception we cannot mean an act that is somehow constituted by the *coming* to believe what is false.

[14] Such a view is not mandatory, of course; see Sorensen 1985.

such circumstances, if an expert says "p" and the subject, believing the expert is always right, comes to believe that p, she'll have deceived herself via her asking the expert.[15] But this hardly seems like a way of *trying* to deceive oneself unintentionally. Indeed, such a "plan" for bringing it about that one unintentionally deceives oneself seems no different than trying to settle the question "p or not-p?" but hoping, somehow, that one gets it wrong.

But if this is right, then, we seem to be in a position of concluding either that what one does when one's trying to intentionally deceive oneself and what one does when one's trying to unintentionally deceive oneself are no different or, perhaps, worse, that when one intentionally deceives oneself one also unintentionally deceives oneself. Needless to say, it may be claimed that the effort to unintentionally deceive myself is, somehow, essentially self-defeating. There would, of course, be an irony here since we've become used to thinking that it's, rather, the effort to bring it about that I intentionally deceive myself that is essentially self-defeating.

3. Who Deceives Whom?

Since we are considering a potentially puzzling consequence of the effort to model intentional self-deception on prototypical cases of interpersonal deception we would do well to consider, in brief, the nature of the activity of the deceiver and the deceived in interpersonal deception. If to deceive another is to cause her to believe falsely that p, we should be clear that what the deceiver's action causes is an event - presumably the event of the deceived's coming to believe that p. I take it that this is so in virtue of the fact that a deceiver alters the evidence or epistemic reasons of the deceived and this results in the latter's coming to believe that p. In this way, if all goes well (for the deceiver, that is), the deceived's belief-forming processes are sensitive and responsive to the deceiver's intentions and practical reasons in the way that the path of the ball is sensitive to my aims when I sink a putt.[16] In this way, we are

[15] Thus, the expert, as well, will have unintentionally deceived the subject.

[16] It's important that the control I exercise over the deceived in cases of intentional deception is not merely causal. Consider the following: I assert that p to you, believing it false and thinking you regard my testimony as trustworthy. As it happens, you aren't at all inclined to believe on the basis of my assertion alone. Still, you've just emerged from a session with your much esteemed psychic. You've consulted him, as you're consumed with the question "p or not-p?" since you desperately desire that

right to regard deceiving another as treating another as a mechanism. In familiar fashion, we exploit machinery and the causal structure of the world, more generally, in order to extend the range of our control and, so, to secure our ends. In this way the deceiver acts upon and through the deceived.

I take it that Iago is remarkably successful in this regard with respect to Othello in the matter of the question of Desdemona's fidelity. Iago, in this sense, treats Othello as a mechanism in order to secure his aims. He exercises control over Othello's reasons and belief-forming process. And that is why we say he deceives Othello – causes Othello to come to believe that Desdemona is unfaithful. Iago accomplishes his deception via the alteration of Othello's evidentiary or reasons condition. Believing p false and wanting Othello to come to believe that p, Iago arranges things such that Othello comes to possess evidence in favor of p; his reasons condition becomes determinative for p and he comes to believe it. Iago intentionally deceives Othello – causes him to believe something false. Presumably, this is something Iago does. One agent intentionally deceives another via the first agent's pursuit of a deceptive project that exploits the second agent's pursuit of the project of settling a question. So, there are two projects simultaneously at play – two projects traceable to two agents and to two constellations of practical reasons.

This is, of course, one reason why it is nonsense for a deceiver to say to a deceived: "Don't look at me. Your coming to believe that p is something *you* did, not *me*. You came to believe for your own reasons." This is nonsense even though Othello's coming to believe or forming the belief that p is not something *Iago* does. Othello *does* that and for his reasons. In this way, Othello is not a passive by-stander to his deception. But this should be no surprise. Deceptive projects in the interpersonal arena exploit the rational

not-p. He's just told you that if you can get to midnight without hearing a typically trustworthy speaker assert that p, you can be assured that not-p is true – otherwise, p is true. You immediately try to make your way home to seek seclusion, when you encounter me. Now, my assertion certainly plays a causal role in your coming to believe that p; yet, if p is false, I don't intentionally deceive you. We have a case of consequential waywardness or deviance. But it's important to point out that this is so because what I do fails to exert the sort of control that I aim to exercise over the direction your cognition. My intention to cause you to believe that p is, of course, not appropriately related to how it is you are caused or come to believe that p. Here it seems to me, were I to come to learn why it is you came to believe that p, I might well reasonably say: "Your coming to believe that p is something you did or brought about, not me!" In such a case, the deceiver may certainly be said to cause the deceived to come to believe as he does, but he does not intentionally deceive. I lack the appropriate sort of control over your being caused to come to believe as you do.

agency of another. Iago is trying to deceive Othello. Othello is trying to settle a question. Othello (with the assistance of Iago, to be sure) takes certain data to constitute powerful evidence in favor of the view that Desdemona is unfaithful. In focusing upon various data he causes himself to come to believe this. So he causes himself to have a false belief. In this way, Othello deceives himself – but unintentionally, of course – and in the manner that we all often deceive ourselves in unintentional fashion. Unless Iago could somehow directly implant the belief that Desdemona is unfaithful into Othello, it's hard to see how this result is avoidable.

Iago deceives Othello. But he does something else: he intentionally causes Othello to deceive himself unintentionally. Thus, typical cases of interpersonal deception require the presence of intentional deception (on the part of the deceiver) and unintentional deception (on the part of the deceived). With this as a model for intentional self-deception, we may want to say of Sammy that he:

(1) He intentionally causes himself to (come to) have a false belief; that is,
(2) He intentionally deceives himself.

But, as well, when Sammy comes to believe as he does, he does so in the aftermath of his effort to settle a question. He takes various data to constitute sufficient reason for settling his question. In this way Sammy, like Othello,

(3) Unintentionally causes himself to (come to) have a false belief; so, he
(4) Unintentionally deceives himself.

And this, of course, because Sammy, like Iago in his deception of Othello,

(5) Intentionally causes himself to deceive himself unintentionally.

This is the source of the puzzle at the end of the last section. If interpersonal deception is our preferred model, then we must conclude that if Sammy were to aim to deceive himself *unintentionally* he could do no better than to do precisely as he does in the case as described in *Happy Days* – a case in which he, of course, allegedly intentionally deceives himself; and this, because, as we now see, Sammy *does* unintentionally deceive himself in that case. Indeed, there is an additional puzzle since Sammy both, (2), intentionally deceives himself and, (4), unintentionally deceives himself. Thus, the very same doxastic alteration, at the very same time, by the very same agent must be counted an instance of both intentional deception and unintentional deception. Of course,

we might insist that if Sammy in *Happy Days* unintentionally deceives himself he does not, as well, intentionally deceive himself.

It is precisely because of the presence of two agents with two distinct projects in cases of familiar interpersonal deception that there is no puzzle attached to conceiving of Iago's deception of Othello as involving both intentional and unintentional deception – and this, of course, because Iago intentionally deceives Othello, while Othello deceives himself unintentionally. So, the presence of two agents with two distinct projects is crucial to our understanding of interpersonal deception and, in particular to the way in which one agent may intentionally cause another to come to believe falsely that p and, in this way, to control or manipulate the belief-forming processes of another agent via her (i.e., the deceiver's) deceptive intentions and intentional activities. The deceptive agent counts upon or exploits the fact that the deceived is engaged in and pursuing her own project: settling a question or trying to get things right. Iago intentionally causes Othello to come to have a false belief via his pursuit of his deceptive project. Othello deceives himself unintentionally via his effort to settle a question. So, again, on this interpersonal model, we say of Sammy that he intentionally causes himself to have a false belief via his pursuit of his deceptive project while he also unintentionally deceives himself via his effort to settle a question. At the least, we're compelled to view Sammy as possesses two competing and contrary projects.

But there's just one Sammy. Now, this might be disputed, of course. In obvious ways we can claim that it is the earlier time-slice of Sammy who succeeds in intentionally deceiving the later time-slice of Sammy, while the later time-slice of Sammy unintentionally deceives himself in the midst of his trying to settle a question. To be clear, though, Sammy comes to believe that p at a particular time; so, at that time Sammy's deceptive project succeeds *and* Sammy unintentionally deceives himself. But this is to treat Sammy not merely as if he were like two distinct agents but, rather, as if he were, in fact, two distinct agents. And the cost here, it seems to me, is very great.

If self-deception literally implicated two agents or two independent centers of rational activity, I take it that it would be foolhardy to gainsay the possibility of intentional self-deception. Needless to say, there are accounts on offer that appear to involve something like this strategy (Pears, 1984; Rorty, 1988). Still, I take it that there's something profoundly unsatisfying about such radical homuncular accounts. If, we explicate intentional self-deception by appeal to

two independent centers of rational activity, we will have failed to come to grips with the phenomenon and what we find puzzling about it. We would have failed to come to grips with the phenomenon because we would have turned a case of self-deception into a case of interpersonal deception. And we would have failed to explain what we find puzzling about the phenomenon ("How can you possibly *believe* that?") since there's nothing puzzling about how or why one comes to believe as a result of the activity of a deceiver. (At best we would have explained away our puzzlement.) Rather, my point is that if self-deception, à la intentionalism, is to be compellingly defended and explained, the phenomenon must be realized, as William Talbott puts it in a single, coherent self (1996). If what we call "self-deception" involved one center of rational activity or agent controlling the epistemic reasons possessed by another independent center of rational activity in precisely the way Iago controls Othello's reasons, there is, it seems to me a straightforward way in which we would have to conclude that there is no self-deception.

What should we say about Sammy in *Happy Days*? I think we should say, (5), that Sammy intentionally causes himself to deceive himself unintentionally,[17] but that we should resist saying that he intentionally deceives himself. Sammy tries to bring it about that he unintentionally deceives himself. He does unintentionally deceive himself. Of course, one imagines the immediate rejoinder: but then he also must, (2), intentionally deceive himself. If he intentionally causes himself to unintentionally deceive himself *then* he intentionally deceives himself. Indeed, Sammy, we will say, *intentionally* deceives himself by *unintentionally* deceiving himself.[18] My own view is that we can say this only if Sammy is treated precisely like Iago and Othello – as two distinct agents with two distinct projects and constellations of practical reasons. In the next section, I aim to consider why, in the case at hand, we should reject the suggestion that, in these circumstances, Sammy intentionally deceives himself by unintentionally deceiving himself.

[17] More felicitously we can say that Sammy intentionally brings about conditions in which he unintentionally deceives himself.

[18] Needless to say, an agent can intentionally ϕ by unintentionally ψ-ing. For example, I can intentionally amuse the children by intentionally causing myself, unintentionally, to trip down the stairs. But in this case, the intentional causing (an action) produces my unintentional tripping which then produces a distinct event: the children's merriment. In the case of self-deception, though, it is the causing to believe what is false that is both intentionally and unintentionally produced.

Before turning to that task, I want to note that the modest conclusion that Sammy intentionally deceives himself via unintentionally deceiving himself would, itself, appear to have awkward consequences for intentionalists. For while it may be insisted that it is clear how, in Sammy's case, intentional self-deception succeeds, it is far from clear how, without similar improbable contrivances (e.g., Alzheimer's-induced forgetfulness together with fabricated, but compelling, evidence delivered by a trustee, etc.) intentional self-deception could succeed. Indeed, as Mele notes, such cases as *Happy Days* are remote from typical cases of self-deception; and they are so in part precisely by virtue of the presence of such fanciful elements. Those fanciful elements are, of course, critical to Sammy's coming to believe as he does. He comes to believe as he does, in the midst of settling a question because he comes to have sufficient reason so to believe. But, then, we must ask, how without such contrivances is intentional self-deception to succeed?

Here it should be pointed out that instances of intentional self-deception either involve intentionally causing oneself unintentionally to deceive oneself or they do not. If they do, then, in the absence of the baroque elements critical to success in *Happy Days* some other mechanisms and processes must be at work which result in a subject's unintentionally deceiving herself. If success hinges upon intentionally causing myself to deceive myself unintentionally, it is not at all easy to see what these other mechanisms and processes could be if not the non-intentional motivational and affective mechanisms described by deflationists. After all, the self-deceiver must be moved to regard her data as sufficient reason for belief.

Of course, it may be that the intentional self-deceiver does not succeed in intentionally deceiving herself by unintentionally deceiving herself while in the midst of trying to settle a question. That is, it may be that there are not two projects – the deceptive project and the effort to settle a question – at work. A natural way of developing this suggestion is to appeal to unconscious deceptive intentions and projects (Talbott, 1994; Bermúdez, 2000). While it is certainly the case that I cannot take up this challenge with the attention it deserves, one consequence of this view should be noted. Appeals to unconscious deceptive projects and intentions are very often accompanied by an insistence that the requisite sort of unconscious is a familiar one – an innocent or minimal unconscious (Talbott, 1994; Bermúdez, 2000). William Talbott insists, for example, that the sort of unconscious upon which his account relies «requires no more division of the self then does explaining ordinary communication, or

explaining such activities as singing a duet, or painting a house together [...]» (1994, pp. 36-37).[19]

Thus, the claim is that in intentional self-deception there are not two competing projects or intentions, there is just one: the self-deceptive project of intending to come to believe that p (regardless of its truth.) Still, there is the stubborn fact that self-deceivers – in the midst of deceiving themselves – do take themselves to be doing whatever they are doing when they, in fact, are trying to settle a question. So, at the very least, in such cases, an agent who intends to deceive herself, and whose activities through time are presumably organized and directed toward that end, also takes herself to be settling a question. Moreover these are projects or intentions that are at odds with each other. On such a view, the agent isn't merely ignorant of the project she's really engaged in and of the intentions and reasons animating it; she is positively mistaken about what she is doing; in particular, she is mistaken about why, when, for example, she rejects a datum as probative, she is rejecting that datum as relevant. Such an agent takes herself to be trying to settle a question, takes herself to be organizing her activities toward that end, but she is not. She is, in fact, engaged in the contrary project of trying to deceive herself. But this seems less a familiar and innocent appeal to an unconscious of the sort present in communicative activity or to "innocent" divisions of the self, than it does an appeal to a robustly psychodynamic conception according to which our conscious projects and aims are epiphenomena floating powerlessly above of our unconscious intentions, aims, and reasons.

4. Occluded Reasons

The challenge to which I now return is this: to intentionally deceive is to intentionally cause to believe falsely. Sammy, I have suggested, intentionally causes himself to deceive himself unintentionally. That is to say (rebarbatively): Sammy intentionally causes himself to cause himself unintentionally to come to have a false belief; or (somewhat less rebarbatively), Sammy intentionally brings about conditions in which he unintentionally deceives himself. But, again, if Sammy intentionally causes himself to deceive himself unintentionally,

[19] Talbott appeals to Grice on communicative intentions and to the intentions that figure in Bratman's theory of shared or joint activity as analogues to the unconscious intentions implicated in self-deception.

then it seems that he intentionally deceives himself (*by* getting himself to deceive himself unintentionally). Moreover, the same conclusion seems to result when we make note of the fact (apparent in the rebarbative formulation above) that an intentional causing of a causing surely collapses into an intentional causing; that is, if Sammy intentionally causes himself to cause himself unintentionally to come to believe falsely that p, then, he intentionally deceives himself.

Why, then, deny that Sammy (and others) intentionally deceives himself in circumstances in which he intentionally causes himself to deceive himself unintentionally? I will argue – too briefly here – that Sammy's earlier intention and practical reasons are occluded or screened off from playing an intentional or rationalizing explanatory role in his deceiving of himself.

To see how this is so, consider a case, from the strictly practical sphere, described by Alfred Mele. In the case, Ann is offered $10,000 if she offends Bob unintentionally. «Ann,» Mele writes

> will be inclined, in some measure to bring it about that she offends Bob unintentionally. In one relevant scenario, she knows that she tends to offend Bob unintentionally when she is extremely busy: when she is preoccupied with her work, for example, she tends, without then realizing it, to speak more tersely than she ordinarily does to people who phone her at the office; and, when Bob calls, her terse speech tends to offend him. Knowing this, Ann may undertake an engrossing project [...] with the hope that her involvement in it will render her telephone conversation at the office sufficiently terse that should Bob call (as he frequently does), she will unintentionally offend him. This is a coherent attempt [...]. (1995, p. 414)

That seems right. When Ann offends Bob by speaking tersely to him that evening, she does so for considerations then salient to her and not in virtue of the considerations salient to her when she formulated her plan. She acts thoughtlessly and unintentionally. She does not offend Bob intelligently and intentionally. What about the fact or state affairs <Bob's being unintentionally offended by her>? She does intentionally bring about or cause that *state of affairs*; but this is to say that she intentionally brings about conditions in which she insults Bob unintentionally. And this is consistent, of course, with her exerting no intentional control or guidance via her earlier practical reasons over her current treatment of Bob. Luckily for Ann, those have come to be screened off by the interposition of her current motivational and cognitive constitution. Of course, in the aftermath of her success, Ann may think:

"Yahoo! I've done just what I wanted to do – the $10,000 is mine." But for all that, she does not exert (in virtue of her practical reasons at the time she formulated the plan) intentional control over her offending of Bob. At the time she formulates her plan, she foresees that she will offend Bob but that, too, is consistent with the claim that she unintentionally offends Bob when she does. Of course, it's clear that what she has done at an earlier time as well the practical reasons then animating her activities are causally relevant to her later unintentional offending of Bob. But it is not in virtue of those that she does what she does when she offends Bob.

How, then, does Ann succeed in bringing it about that she unintentionally offends Bob? Well, what she must do is to arrange things such that she will come to have a different constellation of reasons and a different aim from those that give rise to the original aim or project. It is, of course, the temporally later set of reasons that produces her action whereby the state of affairs <Bob's being unintentionally offended by her> is realized – the state of affairs that Ann aimed to bring about, given her earlier reasons. In short, what she does when she offends Bob is explained by the reasons she has come to acquire: she's working hard in the evening, doesn't have time for a meandering conversation and wants to get off the phone. She answers Bob's question tersely wanting to get off the phone and he is thereby offended. The reasons from which her earlier aim (i.e., offending Bob unintentionally) emerged explain – in the rationalizing way – not why she acts as she does when she offends Bob, but rather why she comes to have the reasons that, at the later time, explain her acting as she does. So, while it's certainly the case that her earlier reasons and intention figure in the causal explanation of her later activity, they do not figure in the intentional or rationalizing explanation of her later activity. What she does then is explained by the reasons she has come to possess at the later time. What's crucial here, again, is that the practical reasons which generate her project are screened off – in ways she hopes will occur – from those which generate her later behavior.

Thus, in Ann's case, we will say that she intentionally causes herself to insult Bob unintentionally.[20] Let me be clear about the relationship of this case to that of intentionally deceiving oneself: since to deceive oneself is to *cause* oneself to have a false belief, the structural analogue in the case of alleged

[20] Or we may say, a bit more felicitously, that she intentionally brings about conditions in which she unintentionally insults Bob.

intentional self-deception is this: Sammy intentionally causes himself to cause himself unintentionally to come to hold a false belief. In more familiar settings, the intentional causing of a causing will collapse into an intentional causing. But not so in these cases, since the means by which one brings about the state of affairs one wants to bring about entails that one's reasons-condition and intention be altered in order that the desired state of affairs be the upshot of a distinct reasons condition and intention.

Sammy comes to believe as he does because he's motivated to settle a question and he comes to settle his question in virtue of the epistemic reasons he comes to possess – that is what explains his coming to believe as he does. But, as well, his causing himself to come to have a false belief is something explained by his then current aim and reasons. In the midst of settling a question, he asks to see what's in the boxes and, as a result, comes to believe that he's led a life rich in human connections, and, he thereby deceives himself unintentionally. His earlier reasons are, like Ann's, occluded or screened off from providing a rationalizing account of his deception of himself. Something like this point is noted by Jonathan Bennett; he argues that agents can be appropriately said to act through «long, complicated causal chains but not ones whose whole effectiveness runs through the will of an agent» (Bennett, 1988, p. 227). He writes that

> at noon I set up a delayed-action mechanism, knowing that when it kicks into action at dusk it will irresistibly tempt me to close the gate. In that case, what qualifies me as the one who closes the gate is what I do at dusk not what I do at noon. (Bennett, 1988, p. 227).

Sammy at the time he comes to believe he has led a life rich with human intimacy comes to believe as he does as a result of his effort to settle a question and the epistemic reasons he comes, then, to possess. In this way, his earlier plan and intention to bring it about that he deceives himself is one whose, as Bennett puts it, "whole effectiveness runs through the will of an agent."[21] The parallel between Bennett's gate-closer and *Happy Days* case might appear to be vitiated by the fact that Bennett does, of course and rightly, want to speak of

[21] I am certainly not presuming that there is a "doxastic will." I am presuming that trying to settle a question is an intentional activity and, as well, that settling a question – i.e., coming to believe, as it may be, that p – is an instance of rational activity. That I come to believe as I do is something I do because of my apprehension of reasons. See, for example Raz 1999 (Chapter 1) and Moran 2002.

the agent in this case as (by virtue of what he does at noon) causing himself to close the gate at dusk (1988, p. 227). But, of course, Sammy is trying to deceive himself, which is just to try to *cause* himself to come to have a false belief. So, I agree that Sammy intentionally causes himself to deceive himself. This is what I have been arguing: Sammy does not intentionally cause himself to come to have a false belief – what he does is to intentionally cause himself unintentionally to deceive himself. Less awkwardly, he intentionally brings about conditions in which he unintentionally causes himself to have a false belief.

Thus, when I intentionally cause my own action, when that action is *already* a causing, as with deception, then, after Bennett, we should say that Sammy counts as deceiving himself in virtue of what he does while in his bed at the Alzheimer's center, rather than in virtue of what he does as a young mathematician. As with Ann, his earlier reasons and intention are occluded or screened off from providing a rationalizing explanation of his deceiving of himself. In this case, then, the intentional causing of an unintentional deception does not collapse into an intentional deception and this because, like Ann, Sammy now has another aim and constellation of reasons, and these provide the rationale for his coming to believe as he does and, so, for his deception. Of course, in virtue of what he does as a young mathematician, Sammy counts as causing his later deception; but, again, this is not to say that he intentionally deceives himself – intentionally causes himself to come to have a false belief. There is no act which is an act of intentional deception.

This is why, if Sammy wanted to deceive himself unintentionally, he could do no better than to arrange things such that at some later time, while in the midst of trying to settle a question, he would come to take himself to have sufficient reason for coming to believe that p. His earlier reasons and intentions are occluded from playing a rationalizing explanatory role in his deception, as Ann's are from her offending of Bob. He comes to have a different aim, settling a question; as a result, he comes to have various epistemic reasons his apprehension of which constitutes by his lights sufficient reason for coming to believe as he does. His coming to believe as he does is explained by appeal to these propositional attitudes and by his rational activity at that time. As a result of his current activities – his inquiry – he causes himself to come to have a false belief and, so, to deceive himself unintentionally. By virtue of what he did as a young mathematician, he

intentionally caused himself to deceive himself unintentionally. He does not intentionally deceive himself

5. Conclusion

I have argued that Sammy does not succeed in intentionally deceiving himself. I have, as well, pointed out that if intentional deception, in fact, requires that the agent unintentionally deceives herself, intentionalism faces serious challenges.

The interpersonal model of intentional deception is no model for self-deception because, since I am a single agent, once my evidentiary or reasons condition is altered – the condition of success of my project – I have altered the reasons condition of the actor and, in fact, have abandoned the intention to deceive prior to coming to believe. Indeed, that aim to deceive myself has been replaced by another contrary aim: the aim of settling a question. Iago's act of successful deception requires for its success the rational activity of another agent. In self-deception, there is no distinct agent to whom the deceptive project can be traced. To treat such a case as a case of intentional deception is to treat a single agent precisely as two agents. Moreover, since there is but one agent in self-deception, there is no other agent whose activity or aims could be the source of the deception. In self-consciously aiming to alter my reasons condition and my aims in order to bring it about that I come to have a false belief as a result of settling a question, I guarantee that what I do is to intentionally bring about conditions in which I cause myself unintentionally to deceive myself.

REFERENCES

Bennett, J. (1988). *Events and Their Names*. Indianapolis: Hackett.

Bermúdez, J. (2000). Self-Deception, Intentions, and Contradictory Beliefs. *Analysis*, *60*(4), 309–319.

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

Elga, A. (2009). Imagination, Delusion, and Self-Deception. In T. Bayne, & J. Fernández (Eds.), *Delusion and Self-Deception*. New York: Psychology Press, 263–280.

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

Gendler, T. (2007). Self-Deception as Pretense. In J. Hawthorne (Ed.), *Philosophical Perspectives 21: Philosophy of Mind*. New York: Wiley Interscience, 231–258.

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

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

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

Mele, A. (1995). Motivation: Essentially Motivation-Constituting Attitudes. *The Philosophical Review*, *104*(3), 387–423.

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

Mele, A. (2003). Emotion and Desire in Self-Deception. In A. Hatzimoysis (Ed.), *Philosophy and the Emotions*. Cambridge: Cambridge University Press, 163–179.

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

Moran, R. (2002). Frankfurt on Identification: Ambiguities of Activity in Mental Life. In S. Buss, & L. Overton (Eds.), *Contours of Agency: Essays on Themes form Harry Frankfurt*. Cambridge, Mass.: The MIT Press, 189–217.

Oksenberg Rorty, A. (1988). The Deceptive Self: Liars, Layers, Lairs. In A. Rorty, & B.P. McLaughlin (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 11–28.

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

Sorensen, R. (1985). Self-Deception and Scattered Events. *Mind*, *94*(373), 64–69.

Raz, J. (1999). *Engaging Reason*. Oxford: Oxford University Press.

Talbott, W. (1995). Intentional Self-Deception in a Single Coherent Self. *Philosophy and Phenomenological Research, 55*(1), 27–74.

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.

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.

Practical Self-Deception

Eric Funkhouser *
efunkho@uark.edu

ABSTRACT

Philosophical accounts of self-deception almost invariably treat it as a phenomenon concerning belief. But this article argues that, in the very same sense that we can be self-deceived about belief, we can be self-deceived about matters that concern our practical identities – e.g., our desires, emotions, values, and lifestyles. Given that our practical identities are at least as important to us as are our beliefs, philosophical accounts of self-deception should accommodate such practical self-deception.

1.

The philosophical literature on self-deception has, by and large, treated it as a phenomenon concerning belief.[1] That is, the self-deceived are almost always described, defined, or theorized as being deceived with respect to a belief. This is probably because philosophers who discuss psychological matters tend to have a heavy bias toward belief in general, perhaps because it has clear connections to theoretical reasoning. Self-deception is supposed to be a type of irrationality, and beliefs are particularly well-suited for rational evaluation in terms of their standing with respect to evidence and other epistemic norms. Since epistemology is squarely within the field of philosophy, such a characterization of self-deception makes it appropriate for philosophical investigation as well.

However, I think that we should expand our conception of self-deception and our corresponding philosophical theories so that they cover a wider

* University of Arkansas, USA.

[1] Mele 2001, one of the most prominent book-length treatments of self-deception in recent years, does not consider self-deception about anything but belief-like attitudes. Almost all other philosophical treatments of self-deception have been similarly narrow in their focus.

assortment of attitudes, states, and actions with respect to which one can be deceived. In particular, I offer our desires, emotions, values, and lifestyles as additional respects in which we can be self-deceived. Of course, we could be deceived about these things in virtue of being self-deceived about beliefs that mislead us into acquiring the wrong desires, emotions, values, or lifestyles. For example, I might desire to give a public speech on a topic – or simply make a blog post – because I have deceived myself into believing that I am an expert on that topic. But this is not what I have in mind. Here the desire is an effect of the deceptive belief, and this desire would be appropriate were the belief accurate. Rather, my claim is that we can be directly self-deceived about these things in the same sense that we can be directly self-deceived when it comes to our beliefs. Further, philosophical attention should be given to this broader range of self-deception. In contrast with the theoretical nature of belief, I call such cases *practical self-deception* because of their close connections to action.

2.

I will begin by providing some justification for this expansive understanding of self-deception. These are reasons for thinking that a theory of self-deception should concern itself with more than just the psychological state of belief.

1) Scope.

Other things being equal, or at least for some explanatory purposes, explanations and theories with wider scope are to be preferred over those with more limited scope. For example, a theory of self-deception that covers deception about both self-affirming and self-negating beliefs should, other things being equal, be preferred over a theory that covers only one of these categories. Likewise, a theory of self-deception that covers deception about our beliefs, desires, emotions, values, and lifestyles, should be given precedence over a theory that covers only one of these categories.

2) There are interpersonal analogues to practical self-deception.

Puzzles about self-deception are often introduced by comparing it to interpersonal deception. In interpersonal cases, it is often said, we want to deceive someone into believing some falsehood and we take steps so as to trick them into believing that falsehood. This is often true. But there is nothing about the notion of deception that tethers it to belief. We also

deceive or trick people into acquiring certain desires, emotions, values, and lifestyles. It is perhaps a stretch to call the resultant desires, emotions, values, or lifestyles *false* (even though some people do commonly speak this way). But there is some negative term – such as *mistaken*, *inappropriate*, or *inauthentic* – that applies to such cases. We then have some reason to look for parallel varieties when it comes to self-deception.

3) Self-deception about belief often uses means that can be applied to acquiring desires, emotions, values, and lifestyles as well.

Some philosophers might think that we should limit our understanding of self-deception to belief because, as philosophers, we advocate certain standards (e.g., truth) and we advance norms of good reasoning that are violated in self-deception about belief. Further, those who self-deceive about belief are often seen as engaging in an activity that is paradoxical. Consider a man who deceives himself with respect to his wife's infidelity. How can he know or suspect that his spouse is having an affair (as seems required for pulling off the trick of consistently avoiding the decisive evidence in favor of her infidelity), but also *not* know or suspect the truth (which seems required for the trick to succeed, so that he nevertheless believes that she is faithful)?

However, when we look at many of the psychological mechanisms employed in self-deception about belief, they also can apply to self-deception about desires, emotions, values, and lifestyles. In all these areas we can ignore alternatives, suppress doubts, be motivated to misinterpret contrary considerations, or simply remain unreflective. Also, there are practical paradoxes that parallel the paradoxes about self-deceptive belief – e.g., How can one know or suspect the value of some career (as seems required for pulling off the trick of consistently avoiding the considerations that make that career appealing), but also *not* know or suspect the value of that career (which seems required for the trick to succeed, so that one values an alternative career instead)?

4) Practical self-deception is of fundamental importance.

Philosophers are often passionate about truth, even for its own sake. But whatever importance there is in having true beliefs – and as a corollary, whatever importance there is in avoiding self-deception about belief – is at least equaled by the importance of getting our desires, emotions, values, and lifestyles right. Of course there is some uncertainty as to what "getting it right" means in these cases or if there even is such a standard. Regardless,

our desires, emotions, values, and lifestyles are at least as important to us in practice as are our beliefs. As such, we should be at least as concerned about practical self-deception as we are about our beliefs.

5) Practical self-deception occurs.

It is a simple fact that we sometimes are self-deceptive with respect to our desires, emotions, values, or lifestyles. As such, and given the previous reasons, we should be interested in theories that cover practical self-deception. In the next section I will make a case, through examples and distinctions, for self-deception of these varieties.

3.

Let us start with a case of self-deception with respect to desire. Suppose that a young man finds himself naturally inclined to have sexual thoughts – desires – about other men. For whatever reason, he is motivated not to have these desires and to have heterosexual desires in their place. The motivation here is not simply for a belief. It is true that he does not want to believe that he is a homosexual, but this is because at a more fundamental level he does not want it to be true that he is a homosexual (i.e., has a certain set of sexual and otherwise intimate desires directed at men). The primary motivation in this case is for certain kinds of desires. Of course, there are other cases in which people are motivated simply to repress or hide their desires rather than replace them. This is not the case, however, with our young man. He deceives himself into having his thoughts involving sexuality and intimacy directed at women. As a consequence the bulk of his conscious thoughts about sex do not mesh with his more brute, biological desires for men. This is, unsurprisingly, not a great success, and he remains celibate.

Next imagine someone who genuinely is not happy, perhaps for good reason. Her mother recently died, say, and she also lost her job. But she wants to be happy. She forces smiles. She repeats to herself that things are fine. She focuses on the more pleasant parts of her life, however minor they are. She is trying to be happy. But this is only a partial success. She will be doing fine for a while, but then she suddenly breaks down into tears seemingly out of nowhere.

Or consider a boy who was passionately attracted to the arts and found great value in them. He loved drawing and painting more than anything else. But his father taught him that the arts were impractical and feminine, and that there is nothing of value (at least for a man) in pursuing or appreciating them. So the

boy focused his attention on more practical business interests and suppressed his value judgments about the arts. He strained to value practical matters instead. Echoing his father, he now claims that the arts are a waste of time. But he still catches himself engaged in extensive doodling from time to time as well as being moved, against his declarations, by the arts.

Finally, consider a case of self-deception with respect to lifestyle. Suppose that a girl is raised by parents who have pushed her from an early age to be a medical doctor. She grows up, goes to medical school, and becomes a medical doctor. She works at her profession with care and great competence, but she lacks real passion for her work. She knows that other people find much pleasure and fulfillment in their work, or that a certain job was "meant to be" for them, but she experiences no such feelings herself. She has doubts from time to time, doubts that first started back in her teenage years, about whether a medical career is for her. But the influence of her parents and the years of schooling carry great weight. She suppresses any thoughts about a change in career – the will of her parents as well as a great educational investment have provided her with reason to do this. She instead focuses on the objective value of helping the sick. She deceives herself into accepting this lifestyle, this career.

I offer each of these as an example of practical self-deception. I think that each of these four types can exist independent of the others, though they often will come bundled together – e.g., those who are self-deceived about their lifestyles often engage in desire self-deception as well. Some might challenge the independence claim by arguing, for example, that whenever there is lifestyle self-deception there is also desire self-deception. But I do not believe that is correct. Our doctor need not deceive herself into desiring to be a doctor – she just continues to go to work and go through the motions. Our doctor also need not deceive herself into losing her desires for a change of career – these desires could persist, but she simply discounts or ignores them. Because desire and lifestyle can come apart, it is worthwhile to consider these as two different categories of self-deception. The repressed homosexual deceives himself with respect to his desires, but this need not result in a heterosexual lifestyle. The doctor deceives herself with respect to her lifestyle, but this need not result in a change in her career desires. And each of our 4 examples is different from the standard philosophical examples of self-deception about belief, at least in its motivation and target state.

Some might object that even in these cases of practical self-deception belief still plays a privileged role. Namely, one might argue that in order for practical self-deception to succeed one needs to have the right kinds of beliefs. For example, in order for effective practical self-deception about his sexual desires he must believe that he is heterosexual.[2] If true, this could justify the focus on belief in philosophical discussions of self-deception. I do not think that belief plays such a privileged role, however. First note that the motivation and target state of practical self-deception is not belief, but some practical identity instead. At best, then, self-deception about belief would be a necessary step toward acquiring this practical identity. But acquiring such beliefs is not necessary for practical self-deception. Rather than being a means to practical self-deception, such beliefs are often a consequence of practical self-deception. This is clear in some of our examples. Someone can deceive themselves into being happy not by means of believing this, but by doing things like forcing smiles and selectively attending to the evidence as described in our example. She engages in these activities while not yet believing that she is happy. In fact, she engages in these activities in part *because* she does not believe that she is happy. If she eventually succeeds to some extent in making herself happy, it is true that she will likely believe herself to be happy as a consequence. But that belief is not self-deceptive; it reflects the actual success of her practical self-deception in bringing about some happiness.

4.

Practical self-deception is similar to other types of practical irrationality that have received philosophical treatment, such as Mill's notion of a *life of custom*. In writing of custom Mill had in mind those who are unreflective and dogmatic with respect to their desires and lifestyles. To live a life of custom is to passively accept a set of desires or a manner of living without any rational scrutiny.

> [...] though the customs be both good as customs, and suitable to him, yet to conform to custom, merely *as* custom, does not educate or develop in him any of the qualities which are the distinctive endowment of a human being. The human faculties of perception, judgment, discriminative feeling, mental activity, and even moral preference, are exercised only in making a choice. He who does anything because it is the custom, makes no choice. (Mill, 1993, p. 67)

[2] I would like to thank Patrizia Pedrini for raising this objection.

Such people are *lazy*, at least when it comes to their practical reasoning. They are irrational in virtue of not even reflecting or trying. They do not question their desires or lifestyle, nor do they consider alternatives to them. They are passive, either out of pure laziness or out of a false belief that they have been assigned a role to play in life (e.g., the feminine role or the waiter role).[3]

But one can also be actively irrational with respect to one's desires, lifestyle, emotions or values. I here have in mind those who are *perverse*, rather than merely lazy, when it comes to their practical rationality. This perversion is a motivated misuse of reasons or reasoning, rather than simply a failure to engage with reasons or reasoning. Thus, it is a perversion of rationality. This lazy/perverse distinction can also be found in the theoretical realm. The lazy believe (if they believe at all) dogmatically, passively accepting some belief as if it has been assigned to them by the press, their peers, their parents, or nature itself. The perverse, on the other hand, are the self-deceivers who are motivated to misuse reasons or reasoning. They ignore (due to their motivation) reasons for one belief, and they sometimes actively abuse reasoning by selectively attending to the evidence or rationalizing their favored alternative. The examples of practical self-deception in the previous section are all supposed to involve perversions of rationality in this sense. They are motivated to have heterosexual desires, be happy, value the practical life of business, or be a doctor. But they have good reasons for being otherwise. Their body pushes them to desire the same sex; their situation is anything but a happy one; they recognize little value in business and find much value in the arts; or they feel alienated from their career. But by suppressing these reasons and putting a positive spin on the alternatives, they push for the alternatives they desire. This active engagement with reasons and reasoning makes them perverse practical reasoners.

Those who merely live a life of custom are often wholehearted in their desires, emotions, values, or lifestyles. They need not feel any tension or uncertainty about how they live or how they want to live. Perhaps they *should* feel seem tension or uncertainty, but they do not since they are unreflective or simply inactive in this regard. Practical self-deceivers, in contrast, often experience tension or uncertainty. Tension results from the recognition, or simply the fact, that things are not as they want them to be. And while self-

[3] Bad faith is similar to a life of custom, but I will avoid discussing it as Sartre represents it as too psychologically sophisticated and metaphysically loaded for my simpler purposes here.

deceivers sometimes are fully successful at eventually bringing about the desire, emotion, value, or lifestyle that they want, they frequently are only half-successful in this regard. Like a self-deceiver about belief who "half-believes" both that her husband is faithful and that he is having an affair, practical self-deceivers will frequently "half-desire" something, be "half-happy", "half-value" an activity, or "half-identify" with their career. The ambivalence here is a result of reasons conflicting with motives, and the ambivalence remains because many of us cannot overcome the force of these reasons no matter how much we may want to.

There is a large literature discussing the reasons for such disconnect – between our reasons and motives – when it comes to belief. Because belief aims at the truth, it has been argued, it is impossible to believe at will.[4] More generally, reasons for belief have a tendency to prevail over, or at least frustrate, our reason-independent motives. But one might be skeptical about there being such built-in obstacles to our motives for particular practical identities. That is, one might claim that there is nothing analogous to the built-in norm of truth when it comes to desire, emotion, value, or lifestyle. This would undermine the comparison of practical self-deception to theoretical self-deception, as well as the necessity of engaging in any kind of *deception* in order to satisfy our practical aims.

I will not argue for the claim that desire or value aims at the good, or that our emotions and lifestyles have their own constitutive aims. However, such a strong claim is not necessary to establish a conflict between reasons and motives for these practical identities, nor for the necessity of engaging in deception to satisfy these practical aims. All that is needed is that there are reasons for or against these practical identities and, at least as an empirical fact, there is some difficulty in flatly discounting (consciously or not) the force of these reasons. The difficulty in simply avoiding the force of these reasons would then explain the need to resort to deceptive measures. And I think it is manifest that there are such obstacles, aptly described as reasons, to satisfying our motives to desire, feel, value, or live in a particular way. His natural inclinations provide him with reasons to desire men, reasons that cannot be dismissed at will. One might object that these inclinations *constitute* his homosexual desires, rather than serve as *reasons* for these desires. Even so, they at least are reasons that speak against him desiring, and attempting to

[4] See, for example, Williams 1973 and Velleman 2000.

desire, to have sex with women. His motivation is for an all-things-considered preference (desire) to be heterosexual. But his natural inclinations provide reasons against such an all-things-considered preference. And this point generalizes. Recall our unhappy woman, whose unpleasant circumstances provide her with reasons to be unhappy. These are reasons that she cannot dismiss at will. Such reasons are forceful even if we do not consciously reflect on them. Our self-deceivers then need to resort to deceptive measures to overcome their force – e.g., they suppress their natural inclinations or focus on the (few) positives.[5]

5.

We are now in a position to consider the conditions that are characteristic of both theoretical and practical self-deception. I offer the following 5 conditions that are at least close to being necessary and jointly sufficient for either theoretical or practical self-deception with respect to some psychological state or behavior X.

1) Motivation: A is motivated to X.
2) Frustration: A is in a state that directly conflicts with X.
3) Insufficient Rational Support: A does not have adequate reason to X.
4) Deception: A employs some deceptive strategies, often involving perversions of rationality, to further X.
5) Success: A has some success in furthering X.

Let us discuss each of these 5 conditions, with special consideration given to their application to practical self-deception.

1) Motivation.

Here 'X' can be one of a variety of mental states or behaviors concerning which an agent A can be self-deceived. As previously discussed, the motivation can be for belief, desire, emotion, value, or lifestyle, and this list is not intended to be exhaustive. Certainly most people do have motives with respect to each of these categories from time to time – e.g., people want to be hopeful or they want to be a lawyer. This motivation itself often has its own psychological explanation, and such explanations can be quite varied. A woman might want to be hopeful

[5] Millgram 1997, Ch. 2, argues that desires possess such backward-looking commitments (i.e., reasons) that make it impossible to desire at will.

for its own sake, for example, but want to be a doctor simply in order to please her father. As the latter case is supposed to show, the motivation here might not reflect what we would naturally describe as what the agent "really desires". She desires to be a doctor, but she prefers that her father were not so overbearing or that he at least favored a career path more in line with her temperament. In that sense, while she does have a motive to be a doctor, it is not what she "really desires". The motivation for self-deceptive belief can similarly have varied psychological explanations, these differences accounting for the distinction between straight and twisted self-deception for example.[6] Straight self-deceivers typically desire a belief for its own sake or for the peace of mind that comes with it, but twisted self-deceivers – whose motives do not accord with what they want to be true – often have a more complex motivation.

2) Frustration.

The agent is in a state that conflicts with their motivation. This means that they do not have what they want. But more than this, they are in a state that frustrates their desires. He desires to be heterosexual, but he finds himself with homosexual desires. She wants to be a writer, but she is a doctor. He wants to believe that the ship is seaworthy, but he has doubts or outright believes that it is not seaworthy. In cases like these the conflict is obvious and direct. In order to prompt deception, the conflict should be straightforward and obvious enough to cause psychic tension or be evident to a neutral observer. The existence of this conflict is largely due to condition 3.

3) Insufficient Rational Support.

Deception results from a conflict between motivation and reasons. While A is motivated to X, the reasons available to her do not support X or they support a state that straightforwardly conflicts with X. The most well-developed accounts of rational support apply to belief, which likely explains why discussions of self-deception have focused on belief. Skeptics about practical self-deception will probably attend to this condition, arguing against the applicability of rational support to desires, emotions, and the like. But we often are capable, if pressed, of justifying such states by citing considerations on their behalf. I view these considerations as reasons, though some will likely insist on a division between genuine reasons (such as for belief) and aptness conditions or the like (such as

[6] See Mele (2001) for a characterization of the distinction between these two different kinds of theoretical self-deception.

for desires). Regardless, I take it that it is practically undeniable that there are conditions that speak to the appropriateness of a desire or lifestyle. The fact that you have no interest in a particular career or that you have no aptitude for it, for example, are considerations that speak to the inappropriateness of that career for you. Such conditions are likely not produced by a faculty of reasoning, but they still are considerations for or against these states. Further, these considerations, like epistemic considerations that count as reasons for belief, cannot be resisted at will. We cannot simply decide to have a career for which we have neither interest nor aptitude.[7] The fact that the considerations speaking against this career also prompt deceptive tactics further suggests that they are reasons, as such tactics are employed to manipulate their rational force.

4) Deception.

Reasons have force that often cannot be straightforwardly denied. This is particularly clear with belief, with some arguing that it is a conceptual or psychological necessity that we cannot ignore such reasons and simply will to believe. Hence, theoretical self-deceivers must employ tactics like suppression, biased evidence gathering, rationalization, and the like. These same tactics are employed when it comes to our practical identities as well. Our unhappy woman supresses her unhappy thoughts and feelings. She selectively attends to the meager evidence that shows things are going well for her. She attempts to rationalize away her unhappy thoughts and feelings – e.g., they are merely the product of a bad night's sleep or indigestion. Such efforts, aimed at acquiring the emotion of happiness, clearly amount to a deception. The fact that she has to deceive in and of itself strongly supports the claim that the considerations she manipulates are reasons for, and not merely causes of, her unhappiness. She is not merely addressing an impediment to her happiness; she is doing so in a way that amounts to a perversion of rationality.

5) Success.

Some degree of success is required to be self-deceived, rather than merely self-deceiving. Full success, however, is not required. That is, the self-deceived do

[7] Some will think that lifestyle is different from belief in that it is conceptually impossible to believe at will, whereas at best it is psychologically impossible to pursue a certain lifestyle at will. In Funkhouser (2003) I argued that our inability to believe at will is similarly a mere psychological impossibility, at best. Regardless, there are reasons that provide psychological obstacles to our ability to acquire desired practical identities at will.

not have to fully acquire the belief, desire, emotion, value, or lifestyle that they desire. A partial success can be good enough to count as self-deception. Such is often the case, as when the subject remains ambivalent and continues the self-deceptive enterprise because the rational force behind the contrary belief, desire, etc. remains. Our woman must keep thinking happy thoughts, as the reasons for her unhappiness intrude every now and then and cause her to cry. Outright delusion, in which the agent fully satisfies his motivation, is the extreme that terminates the process of self-deception.[8] But in some cases it might not even be possible for the agent to fully satisfy his desire through a process of self-deception. Such might be the case for the homosexual who wants to have heterosexual desires.

6.

I have argued that there are cases of practical self-deception that share the same structural features, the 5 conditions discussed in the previous section, with the common examples of theoretical self-deception. Theorists of self-deception should investigate and treat these practical cases as well. Practical self-deception deserves treatment because it exists and is of importance. Whatever virtue there is in getting our beliefs right is likely matched, if not exceeded, by getting our desires, emotions, values, and lifestyles right. Considering such cases can also shed further light on the nature of rationality itself, as they show us the diversity of reasons and, on the perverse side, the diversity of deception.

REFERENCES

Funkhouser, E. (2003). Willing Belief and the Norm of Truth. *Philosophical Studies*, *115*(2), 179–195.

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

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

[8] Funkhouser (2009) and Noordhof (2009) mark this distinction between ambivalent self-deception and full-fledged self-delusion. Both accounts of self-deception focus more on the cases involving ambivalence and instability, in which self-deceptive pressures persist.

Mill, J.S. (1993). *On Liberty and Utilitarianism*. New York, NY: Bantam Books.

Millgram, E. (1997). *Practical Induction*. Cambridge, MA: Harvard University Press.

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

Velleman, J. D. (2000). On the Aim of Belief. In J.D. Velleman, *The Possibility of Practical Reason*. New York, NY: Oxford University Press, 244–281.

Williams, B. (1973). Deciding to Believe. In B. Williams, *Problems of the Self*. Cambridge: Cambridge University Press, 136–151.

Self-Deception and Agential Authority. A Constitutivist Account

Carla Bagnoli †
carla.bagnoli@unimore.it

ABSTRACT

This paper takes a constitutivist approach to self-deception, and argues that this phenomenon should be evaluated under several dimensions of rationality. The constitutivist approach has the merit of explaining the selective nature of self-deception as well as its being subject to moral sanction. Self-deception is a pragmatic strategy for maintaining the stability of the self, hence continuous with other rational activities of self-constitution. However, its success is limited, and it costs are high: it protects the agent's self by undermining the authority she has on her mental life. To this extent, self-deception is akin to alienation and estrangement. Its morally disturbing feature is its self-serving partiality. The self-deceptive agent settles on standards of justification that are lower than any rational agent would adopt, and thus loses grip on her agency. To capture the moral dimension of self-deception, I defend a Kantian account of the constraints that bear on self-constitution, and argue that it warrants more discriminating standards of agential autonomy than other contemporary minimalist views of self-government.

1. Introduction

There is empirical evidence that self-deception is a quite pervasive phenomenon, even though some would prefer to believe that it is not. Here is an example. Amy knows that her teenager daughter Bea is visibly too thin, does not eat properly, is always concerned with her weight, and selects obsessively

† University of Modena and Reggio Emilia, Italy.

her food; but Amy does not believe that Bea is anorexic. The evidence is accessible and available to her but does not count as a reason for believing that her daughter is anorexic. In fact, she avoids discussing and investigating Bea's eating habits, and other related matters. What does prevent Amy from acquiring the belief that Bea is anorexic? Is Amy irrational, and in what sense? Does her epistemic state bear moral implications, and if so which ones?

Philosophers have given rather different answers to these questions. As a preliminary step, I take self-deception to be the acquisition and retention of a belief despite overwhelming evidence to the contrary. On some views, self-deception is the case where one holds a false belief *p*, possesses evidence that ~*p*, and has some desire or emotion that favor *p*. Intentionalists take self-deception to result from an intention to deceive (Davidson, 1986; Bermudez 2000). Instead, Motivationalists hold that the deception depends on some interfering motivational state, typically a desire or an emotion (Mele 2001; Funkhouser 2003). The self-deceptive agent discounts evidence which one normally would find sufficient to warrant ~*p*, and yet believes *p* instead because of the interference of some desire, emotion or other motivational state favoring *p*.

Self-deception is regarded a case of irrationality, and in extreme cases a pathology that impedes self-knowledge and it is subject to moral sanction. Indeed, this is partly the reason why it is paradoxical. On the one hand, self-deception is a moral charge, which applies to something one does. On the other hand, it implies lack of the relevant sort of self-knowledge that the moral imputability and the applicability of moral sanctions imply. The moral implications of self-deception are to some important extent similar to (interpersonal cases of) deception. But there are some morally relevant aspects of the phenomenon that are absent in interpersonal deception. Arguably, (the charge of) deception implies intentionality, while (the charge of) self-deception does not. For this very reason, lying to oneself is more threatening than being lied to by others, since in the former case it becomes unclear how to protect oneself from deception. We ordinarily assume self-transparency, even though we know that there are large areas of our mental processes and operations that remain inaccessible. One solution is to treat self-deception as a case where our mind is opaque, as it happens for many mental sub-personal processes and

operations.[1] But the interesting aspect of self-deception is that it concerns beliefs and mental states that are normally accessible. Hence, the selective character posits an obstacle to reducing self-deception to a general case of the opacity of the mind because it appears to exhibit some sort of finality. That is, it concerns a selected cluster of beliefs that whose knowledge the agent has an interest in blocking, even though she may not intend to block it.

The question I would like to address is what kind of irrationality self-deception represents, and what moral consequences it carries for the self-deceptive agent. I will argue that self-deception is not merely a pathological phenomenon, but a defensive strategy that is functional to maintaining the stability of the self. As such, the phenomenon of self-deception can be evaluated under different dimensions of rationality. My starting point is to take self-deception as a practical rather than a theoretical phenomenon. Its philosophical relevance resides in the relation the agent bears to her reasons to believe, rather than in the issue of whether she accurately represents the world and her mind as independent objects. I speculate that self-deception is more similar to alienation than to interpersonal deception in this regard. In both cases, self-opacity undermines agential authority, that is, the authority that the agent claims on her action. In focusing on agential authority, I am driven by the conviction that the relevant source of interference in the production and formation process of self-deceptive beliefs is neither an intention to deceive nor a desire, but a concern with one's self-representation. The constitutivist account I am proposing has the merit of explaining the selective nature of self-deception as well as its being subject to moral judgment and sanction. These are very important features of self-deception that elude traditional accounts of self-deception. The motivational explanation fails to fully capture them, and the intentionalist approach treats them inadequately and generates well-known paradoxes about how the agent holds intentionally contradictory beliefs. The proposal is to adopt a constitutivist account of self-knowledge, where agents are responsible for making up their mind, and they are also responsible for self-deception. This is not because self-deception is analogous to deception in that it is brought about by the intention to deceive. Rather, it is because of the special relation (of authority) the agent bears to her own mind and agency.

[1] For a treatment of self-deception that relinquishes the claim about intentionality and say that it self-deception is operated at subintentional level, see Johnston 1988. White (1988) discusses the case of self-deception as an argument for homuncular theories of identity.

Focus on the special responsibility that agents have for their own beliefs helps us see what is wrong with self-deception and why it can be evaluated morally.

This practical account of self-deception has some important consequences about theories of self-knowledge. It belongs to a broadly constitutivist view that takes self-knowledge to be a practical rather than a theoretical matter. A canonical objection against the constitutivist approach to self-knowledge is that it makes the formation and retention of beliefs arbitrary, insofar as it holds that agents make up their mind. Clearly, self-deception would be impossible to describe in any voluntarist view of self-knowledge on which the agent simply decides at whim what to believe, without being subject to any constraint. Constitutivism avoids this problem of arbitrariness by arguing that the formation and retention of beliefs is indeed constrained, hence there are right and wrong ways of constituting beliefs. But constitutivists differ as to what the relevant constraints and their rationale are.[2] The constitutivist account I defend attempts to separate the issue of stability from the issue of autonomy, which is crucial to genuine agential authority. It points out that stability and autonomy are both issues that can be evaluated rationally, but they call into play different dimensions of rationality. My argument is that while self-deception works as a pragmatic strategy to improve or guarantee the stability of the self, it nonetheless undermines its autonomy, hence its authority over action. The self-deceptive agent can even be more stable than the autonomous agent, but it loses authority on her actions. These are all matters of degrees, of course; and one interesting question concerns the scope of self-deception. I will argue that the selective and circumscribed nature of self-deception is crucial to its success as a pragmatic strategy for maintaining stability of the self, which is of limited sustainability.

Furthermore, this argument has some bearing against those theories of agency that either take stability as a property of autonomous agents or take stability equivalent to autonomy. It reveals that the constitutivist views of agency, which hold that agents make up their mind in action, need to lay down some stricter criteria than stability for authorship on mental life. Such criteria

[2] Notably, the main difference concerns the nature of constraints. According to Kantians, such constraints are moral and necessary; for others, they are contingent and their nature is not moral. See Korsgaard 2008, Velleman 2009.

should include moral constraints about how to relate to self and others, and my suggestion is that they be grounded on respect. [3]

2. Self-Deception, Negligence, and Ignorance

It is tempting to think of self-deception as a peculiar case of deception, as if deceiving oneself is analogous to deceiving others. [4] Is not Amy lying to herself, after all? The analogy with deception helps us distinguish self-deception from mere error. The case of Amy who refuses to believe her daughter Bea to be anorexic, despite all the evidence to the contrary, is different from the case in which Greta fails to realize that her son Phil is a drug addict because she fails to recognize and properly collect the evidence, or because she ignores the symptoms of heroin addiction and thus she does not know what counts as evidence in this particular case. Greta may fail to collect the evidence or to adequately interpret the evidence through no faults of her own. Or, she may be utterly negligent. She may simply not care about the whereabouts of her son, and thus refrain from inquiring about his state of health. Or else she may voluntarily disengage from such investigations not out of negligence, but because she does not think it is right of her to intrude and interfere with her son's life, even when his health and prospects are at stake. In these three scenarios, Greta may be holding false beliefs or lacking beliefs about the state of health of her son, but she is not self-deceptive. Self-deception is more similar to deception than to culpable and not culpable ignorance, or error, in this respect.

The selectivity of self-deception is a very important aspect of it. [5] Self-deception concerns only a very specific set of beliefs. In this case, it is only about the mother's beliefs about anorexia, rather than say all beliefs concerning the general state of health of Bea. It is not uncommon that the self-deceptive agent is attentive to all signs but those that matter for the belief she resists. Amy may be quite perceptive of all other aspects of her daughter health, and worried about seasonal cold, while disregarding only the signs of anorexia. The analogy with interpersonal cases of deception helps us see that the self-

[3] This suggestions shows that I tend to side with Kantian forms of constitutivism, but I will not argue directly for any Kantian claim in this paper.

[4] On the moral dimension of the similarity between deception and self-deception, see Baron 1988.

[5] On the so-called selectivity problem, see Bermúdez 1997, 2000.

deceptive agent is deceptive about some particular beliefs, even though she is largely reliable on all other matters. The scope of self-deception is very specific. A massive, global, and self-consistent delusion such as Don Quixote's imaginary world is not self-deception. It requires no internal struggle, no split of the self, and no effort to reach unity. By contrast, the world is always about to intrude in the self-deceptive's existence, and it threatens disaster. There is always a moment where mothers like Amy have to confront reality. Such moments are experienced rather differently than the acquisition of new shattering information about the world. They are likely to be experience of failures, as well as experiences of having failed others. A scenario where Amy eventually realizes that her daughter is anorexic differs significantly from the scenario where she suddenly learns that her daughter is affected by a life-threatening disease.

This asymmetry is often signaled in moral terms. The coming out of self-deception is typically accompanied, or rather, partially constituted by emotions that are appropriate also in the case of moral failure.6 For instance, it is appropriate for Amy to feel guilty for having disregarded the evidence that pointed to Bea's anorexia. Correspondingly, the self-deceptive agent is the target of moral judgments of condemnation or pity, and she is expected to feel guilty upon realization. Perhaps, the moral judgment addressed to the self-deceptive agent is not as strongly negative as the one addressed to the liar, but it is certainly not positive. It is an open question whether the self-deceptive agent deserves to be blamed, and this partly depends on the conditions for being the appropriate target of moral judgment. But it seems largely agreed that the self-deceptive agent morally differs from the one faultlessly lacking relevant information. If the negligent is culpable, the self-deceptive agent is not completely innocent. When she is excused, it is because she is considered a pathological case, less than a fully morally competent agent.

In the case of Amy, the relevant moral implication is that she failed Bea, that is, she failed to pay attention to and take care of her. These are also failures to value Bea as worthy of attention and care, or as I will show next, as failing to recognize Bea as legitimately claiming attention and care. Other cases of self-

[6] I take the category of feelings of guilt to be rather inclusive, and not linked to intentionality. That is, such feelings are appropriate even when the agent did not intentionally cause any harm or violate any moral claims.

deception may not have direct victims as in this case, but there is always something morally objectionable involved. The question is what that is.

3. Feelings of guilt, blame, and pity: the moral relevance of self-deception

It is hard to pinpoint exactly the moral offense of which the self-deceptive agent is guilty. The case is not clear-cut as deception. The deceptive agent manipulates others in order to pursue her own interests or plans. Unlike the deceiver, it is not obvious that the self-deceptive agent intends to deceive herself to further her own interests or plans. In fact, this sounds paradoxical. What further plans and interests does the self-deceptive agent try to pursue despite herself? And how could she pursue some plans in the ignorance of what she herself knows? These are puzzling questions that arise because the analogy with deception leads to thinking of self-deception in terms of intentions. But the analogy can be taken to highlight aspects of self-deception other than its alleged intentionality.

As I take it, the analogy points out that self-deception is a moral charge, associated to some kind of moral sanction. This association, however, does not imply that self-deception is a thoroughly intentional affair. In fact, moral reproach takes different forms whether it is directed to the deceiver or to the self-deceptive agent. It is morally appropriate to blame people who manipulate others in order to get what they want, while pity is a more appropriate moral attitude to address the self-deceptive agent. The moral grammar of feelings of guilt is compatible with the claim that self-deception is not fully intentionally deceptive; and so is the grammar of pity. Nonetheless, self-deception is a case of moral relevance.

Here is the interesting asymmetry, though. In the case of deception, it is apparent who the victim of the moral crime is. In the case of self-deception, instead, things are not simple. The difficulty does not reside only in the fact that it is not obvious whether the self-deceptive agent is culpable of any moral crime, since it is questionable that she intends to deceive. Rather, the difficulty is that it is not clear how to describe the moral offence of the self-deceptive agent. I will argue that there is something morally objectionable about self-deception, even when we put the issue of intentionality aside. There is some self-serving partiality involved in disregarding evidence selectively, which makes the self-deceptive agent look more like the liar than either the negligent

or the ignorant. The reluctance that motivates self-deception is self-concerned. But what does the self-deceptive agent promote, protect, or express?

4. Self-deception as a Practical Phenomenon: a Normative Account

Traditionally, self-deception is seen as an epistemic and theoretical case of irrationality: the self-deceptive agent holds contradictory beliefs about the world. On this description, the problem with Amy is that she believes that Bea is too thin and does not eat properly *and* she also believes that Bea is not anorexic, where these are two contradictory beliefs. The literature on self-deception abounds with strategies to avoid such paradoxical condition (Rorty 1988a, I-II). Countenance of incoherence can take different forms. For instance, some adopt a strategy of temporal partitioning (Bermúdez, 2000). Others, instead, favor the strategy of psychological division, where the self is partitioned into psychological parts that play the role of the deceiver and deceived respectively (Pears, 1984; Davidson, 1985; Rorty, 1988b).[7]

In contrast to these interpretations, I suggest that we describe the case of Amy more like a case where the agent does not take the available evidence to *count as* reasons. This description points to a different aspect of Amy's activity of belief formation. Amy's problem is not that she holds contradictory beliefs, but that she has reason to believe something that she does not in fact believe. Why? The selective nature of self-deception and the analogy with deception discussed above naturally invite us to find answers by investigating further aims of the agent. What does the self-deceptive agent want? What does she try to obtain by lying to herself? Motivationalists respond to these questions by invoking an interfering desire, and treat self-deception as a case of desire-biased belief (Mele, 2001; Nelkin, 2002; Funkhouser, 2005). It is because Amy does not want to be the case that Bea is anorexic. An interesting and illuminating suggestion is that the interfering desire may be not concerned with some state of affairs (a world in which Bea is anorexic), but a self-focused desire (Funkhouser, 2005).[8] What is interesting about this suggestion is that it connects the bias involved in self-deception to the self.

[7] Among the strategies that for avoiding these paradoxes, there are more moderate views about how to draw the division within the self, such as Pears 1984, 1986, 1991, and Davidson 1982, 1985.

[8] Funkhouser (2005) accepts motivationalism, but he interestingly distinguishes between self-focused and world-focused desires, and defends the former account versus the latter.

To further develop this suggestion, however, we need to abandon the talk of desires. I propose a normative model, where the interfering force is neither desire nor an emotion, but a broad normative concern with the agent's own self-representation. I contend that this concern is not reducible to second-order desires about the selves, but it involves appeal to normative ideals of agency, to which the agent holds herself accountable. In the interesting cases of self-deception, this normative concern plays a role in blocking the normative value and weight of beliefs about the agent's not being up to such standards. That is, the self-deceptive agent defends herself against the charge of not being up to her own standards of agency, by blocking the normative force of reasons that support such a judgment. For instance, self-deceptive Amy bracketed or suspended the normative power of reasons for believing that Bea is anorexic. Amy's resistance to form the belief that Bea is anorexic has certainly to do with her desire that Bea be healthy, and with her emotional discomfort of confronting a world where the child is sick, but it has also some more profound connections to how Amy thinks of herself in relation Bea. She knows that Bea is too thin and shows worrisome eating habits, but these considerations have little normative weight in her overall epistemic system. The point is not that the Amy does not access her most intimate thoughts, or that she misses strong evidence about some states of affairs, and thus forms false beliefs or disbelieves what is true. Rather, the key philosophical point in self-deception is that the self-deceptive agent does not take the evidence available to her as reasons. I want to argue that this is a practical mistake, not a theoretical one.

The (practical) problem of how Amy forms her self-deceptive beliefs does not get resolved by endorsing some coherence-driven strategies. More importantly, it is a problem that only Amy can resolve by engaging in practical reflection. In contrast to theoretical reflection about how the world is, practical reflection is driven by the agent's practical concerns. It does not aim at establishing the truth about the world, even though it is constrained by concerns of accuracy and truthfulness. Its purpose is for the agent to determine what she has reason to believe, and this is something that pertains to the context of deliberation.

Here I am invoking a distinction that stands in the background of constitutivist accounts of self-knowledge. Richard Moran has thus formulated the distinction:

> Roughly, a theoretical question is one that is answered by discovery of the fact about oneself of which one was ignorant, whereas a practical question is answered by a decision, and does not arise from ignorance of some antecedent fact about oneself. (Moran, 1988, p. 141).

Accordingly, what it takes to Amy to realize that Bea is anorexic is not some new information about the state of the world, but a change in her practical attitude toward the evidence she already has about Bea. The change is prompted by practical reflection, which does not aim at accuracy in the representation of the world, but it is itself productive of such representations, and driven by a practical concern about what to believe about the world.

To treat self-deception as a practical rather than theoretical issue is not to discount the fact that it is an epistemic condition. Self-deception raises issues about knowledge of oneself. But to capture its philosophical import we should focus on the special relation that the agent bears to her own states of mind. That relation is of authorship. This is the basic claim of constitutivist accounts of self-knowledge. In such accounts, the agent is responsible for what she believes. Hence, she is also responsible for her self-deception. The agent engaged in self-knowledge does not discover some truths about herself through the course of an introspective theoretical investigation; rather, she engages in deliberative activities that are productive of epistemic states.

The epistemic stories that agents elaborate in deliberation are not epistemic stories about themselves as independent objects of knowledge. Such stories are constitutive of self-knowledge. Claiming authorship for what the agent believes of herself is to take responsibility for herself as an agent. Because of its focus on the responsibility for belief, the constitutivist account seems suitable to make sense of two important aspects of self-deception: its selective nature and its moral status. The self-deceptive agent is entitled to feel guilty because she is responsible for her self-deceptive condition. That she is responsible for her beliefs also explains why self-deception is never global or random, but it concerns some beliefs that bear a particular relevance for the self.

5. Self-Deception as a Pragmatic Strategy

It may seem that by making the agent responsible for belief formation the constitutivist account actually dissolves the very problem of self-deception. If it is up to the agent what to believe, how can one discriminate between genuine cases of self-knowledge and self-deception? This is a special case of a general objection against constitutivist views of self-knowledge, which is based on some misunderstanding. Constitutivism does not claim that the agent simply decides what to believe. The claim is not intended to be causal. It is not that the agent brings self-deception about insofar as she decides what to believe. To this extent, the intentionality of the belief is not the relevant philosophical issue. On the constitutivist view, agents are self-interpreting animals. What to believe is something they determine in the first-person, as part of the activities by which they take responsibility for themselves. While belief formation is a practical matter, there are, indeed, norms that constrain and guide its processes.

According to Richard Moran, for instance, such process should respond to criteria of theoretical transparency. One should make up one's mind about *p* on the basis of reasons related to the truth or falsity of *p*. The criteria of theoretical transparency constrain also the formation of attitudes and emotions of fear and love.[9] It is exactly because such constrains hold that we can rationally assess believes, emotions, and attitudes. When such criteria are violated, then the agent makes up her mind for the sake of reasons that are merely *pragmatic*. It seems plausible to treat self-deception as a case where pragmatic reasons prevail, and the agent comes to form and retain beliefs for reasons that are not constrained by criteria of theoretical transparency.

It may seem that self-deception still counts as a case of theoretical irrationality, under this description. My point here is that self-deception is a complex phenomenon and should be assessed according to different dimension of rationality. It is easy to fill in a story where Amy holds very strong pragmatic reasons to discount the evidence she has that her daughter is

[9] «One answers the question of whether to feel hopeful or ashamed by determining whether something is actually hopeful or shameful. Similarly, a practical question about what I want will often be transparent to an impersonal theoretical question about what is good, desirable or useful. It is essential to the rationality of belief that practical questions about it should be transparent in this way» (Moran, 1988, p. 145).

anorexic. Understandably, this is no welcome news. It is normal for parents to think of their kids as safe and perfect, and to resist evidence about their vulnerability. Moreover, suppose that medical research relates anorexia to some deep emotional instability of the anorexic, due to problematic family relations. Perhaps, Amy does not want to confront the possibility that what she represents as a loving nest is not sufficient for Bea's needs. The belief that Bea is anorexic brings along a judgment about herself as a failing mother. To confront this possibility would seriously undermine Amy's own emotional stability. Perhaps Amy lives in a very traditional household where women feel guilty for having a career, even when they do take care of their family, etc. When we take into account the broad deliberative context where self-deceptive beliefs belong, their irrationality is less apparent. Given the full story, it is rational for Amy to discount the belief that Bea is anorexic, because this belief threatens the image she has of herself, and it would undermine her emotional stability. For Amy's own sake, it is preferable to suspend the normative force of the evidence that leads to that belief. It is a rational strategy of defense. What it is threatened is not some particular interest or value that are dear to the agent, but her own understanding of her self. Unlike the liar, the self-deceptive agent does not try to pursue a specific interest or promote an interest. She seems engaged in a much broader and worrisome enterprise. Still, we can recognize some continuity, which can be captured in terms of instrumental or strategic rationality. Self-deception is instrumental to the stability of her self, and to this extent it is a rational strategy. This means that self-deception is not a totally pathological phenomenon. Indeed, its basic processes are continuous with other rational epistemic strategies that underlie correct processes of belief formation. Self-deception is a phenomenon typical and distinctive of animals that hold ideals and representations of themselves. It is because we are self-reflective animals capable of designing representations of ourselves that we are liable to self-deception.[10] The self-deceptive agent is concerned with the

[10] On this aspect see Darwall 1988. To the extent that self-deception requires self-reflection, I agree with Brown (2004) when she holds that the activity of self-emplotment is partially constitutive of self-knowledge and self-deception. But I strongly disagree with Brown's claim that self-deception is a positive epistemic state, for reasons I offer in the text. I have defended a narrative conception of practical identity in Bagnoli 2007. Cf. Holton for a completely different account that takes mistakes about the self to be a necessary condition for self-deception.

coherence and stability of her emotional and epistemic system as any rational agent would be. This is what she is trying to protect. Is she successful?

Debates about this latter question admit only of positive and negative answers. Mele (2001, p. 50) and Nelkin (2002, p. 394) think she is successful in coming to believe as she desires; Funkhouser (2003) thinks she is not. But if I am right to say that self-deception is continuous with normal rational epistemic strategies, the answer should be addressed in a broader context and admit of qualifications.

Baljinder and Thagard (2003) propose that self-deception results from the emotional coherence of beliefs with subjective goals. I think Baljinder & Taghard are right that the self-deceived agent is concerned with improving the emotional coherence of her overall epistemic and deliberative set (beliefs, emotions, and subjective goals). However, this pragmatic strategy is successful only to the extent that it is limited and circumscribed. As a pragmatic strategy to maintain emotional stability and coherence, self-deception is rather limited, and it is important to notice how. First, its success crucially depends on its selective nature. It works only if it is a circumscribed phenomenon. Secondly, because it needs to be so circumscribed, its advantage cannot but be temporary. Typically, as a pragmatic strategy, self-deception comes to an end. Third, when it comes to an end, the self-deceptive agent realizes that stability based on purely pragmatic reasons is not enough, because it fails to afford agential authority, which is necessary to self-knowledge and autonomous agency.[11]

Self-deception is a failure of authorship, which is a dimension of self-knowledge as well as of autonomous agency. The notion of authorship as the capacity to endorse a thought as one's own and justify it on the basis of reasons. Reasons are considerations that make an act intelligible and justifiable. More importantly in this context, reasons convey the relation of authorship. They express a relation between the agent and the action, such that the action can be imputable to the agent as hers. Justifying an action or a belief on the basis of a reason is thus authorizing it and also claiming authorship on it. Hence, actions and beliefs are expressive of one's agency insofar as they are supported by reasons. It is crucial for us that we act and think on the basis of reasons,

[11] That the success of this strategy crucially depends on the limitation of scope is an interesting aspect that makes self-deception similar to deception. The systematic liar is self-defeating as much as the global self-deceiver. After all, Kant was right.

because this is the way we exercise our agency on the world. The threat to our authorship can be more or less tragic and disruptive, depending on the nature of the claims at stake and their relation to our selves. The significance of self-deception varies correspondingly.

6. The Moral Problem of Self-Deception

Self-reflective agents exert a special kind of authority on their mental life. This kind of authority is fundamentally first-personal and, under this reading, it is conceptually linked to (or even identified with) autonomy. Self-deception, I argued, is a pragmatic or defensive strategy for maintaining the stability of the self. Its success is limited, and its costs are high: it protects the agent's self by undermining the authority she has on her mental life. To this extent, self-deception is more akin to alienation and estrangement, and in this final section I propose that we dwell on this similarity in order to appreciate it morally problematic dimension.

According to constitutivist views of agency, the moral person assumes responsibility for herself by regulating her life by her own best judgment. Moral integrity thus amounts to a form of self-government. Rational agents are responsible for the constitution of such self-government. But there are rather different views about how to conceive of self-government. Contemporary accounts of self-government tend to be rather minimalist in terms of the requirement for full authorship and rational self-government.[12] For instance, in his early work, Harry Frankfurt suggested that the upshot of practical reflection aiming at self-government is a «radical separation of the competing desires, one of which is not merely assigned a relatively less favored position, but extruded entirely as an outlaw» (Frankfurt, 1988, p. 170).[13] The aim of this strategy is not so much to resolve the conflict by annulling one desire as to produce a "well-ordered self" by removing the internal obstacle. Interestingly, this aim is achieved by altering the nature of the conflict: once one of the conflicting desires is disavowed, there would be no internal division. Disavowal is a way of disowning some mental state as external, and thus distancing and dissociating oneself from it. Hence, disavowal is not simply a disclaimer; it's an act of choice determining withdrawal of ownership and authorship. The aim of

[12] I borrow this characterization from O'Neill 2004, pp. 13-26.
[13] See also Frankfurt 1988 (pp. 63, 66-67), 2001 (p. 11), 1988 (p. 172), 1999 (p. 136).

self-deception is analogous to the operations of the self that Frankfurt describes as distinctive of autonomous agency. This means that self-deception is more akin to other normal rational activities of the self. But it also indicates that to make sense of its aberration we need to provide stricter constraints than those Frankfurt adopts.

My (Kantian) proposal is to adopt universal criteria of rational scrutiny of reasons.[14] In forming beliefs, adopting attitudes, and take responsibility for ourselves as agent, we should rely on considerations that could be shared by all other rational agents. It is possible to construct such reasons and take them as authoritative if we take ourselves as members of a community of agents with equal standing, governed by norms of mutual respect and recognition. Recognizably, this is a Kantian requirement of practical rationality.[15] It requires that our judgments and actions be intelligible and justified to all relevant others. Insofar as they have equal standing, others are entitled to ask for reasons and accept the burden of offering reasons to us. This is to say that they stand in a relation of mutual recognition with us. While self-knowledge and self-constitution are fundamentally first-personal, they always implicate a broader context of shared norms. Of course, I will not be able to argue directly for this claim. The purpose of these final remarks is merely to point out that in order to distinguish self-deception from other rational epistemic strategies, we need some basic *moral* criteria. Appeal to universal norms of shared rationality explains what is morally wrong with self-deception. The self-deceptive agent does not critically review the considerations that count in favor of beliefs on the basis of shared norms. In order to protect her stability she relies on less demanding constraints.

The morally disturbing feature of self-deception is its partiality.[16] First, it undermines agential autonomy. Out of fear and concern for herself, Amy settles on standards of justification that are lower than any rational agent would adopt, and thus loses grip on her agency. She thereby trades off her autonomy for a limited security and comfort. But she puts herself in no safer place. As we

[14] I develop this view in Bagnoli 2007a, 2007b. See also O'Neill 1985, 2004.

[15] «The concept of every rational being as one who must regard himself as giving universal law through all the maxims of his will, so as to appraise himself and his actions from this point of view, leads to a very fruitful concept dependent upon it, namely that of the kingdom of ends» (Kant, 1785/1996, p. 83.

[16] For a different characterization of what it is wrong with self-deception, see Darwall 1988, Baron 1988.

saw, her pragmatic strategy requires that she be insulated from the world, and this is not possible in the long run. The self-deceptive agent routinely fails herself. Secondly, she fails others. Her partial concern with her safety makes victims. As a result of Amy's self-deception, Bea's desperate call for help is not heard. It may be objected that this happens only in some special harmful cases of self-deception, and it cannot be generalized. But the point is that the self-deceptive agent is inclined to discount reasons that concern others, when such reasons are threatening for herself. It is against this possibility that moral criteria are put forward. Our ordinary epistemic life abounds with small-scale cases of self-deception, and it may seem excessive moral zealousness to treat them as signs of moral failures. The Kantian requirement is not there for the moral fanatic to express her harsh disapproval, but for the reflective agent to prevent that such apparently innocent cases make casualties.

REFERENCES

Bagnoli, C. (2007a). The Authority of Reflection. *Theoria: An International Journal for Theory, History and Foundations of Science*, *22*,(1), 43–52.

Bagnoli, C. (2007b). *L'autorità della morale*. Milano: Feltrinelli.

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

Baron, M. (1988). What is Wrong with Self-Deception. In B. McLaughlin and A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 431–449.

Bermúdez, J. (1997). Defending Intentionalist Accounts of Self-Deception. *Behavioral and Brain Sciences*, *20*(1), 107–108.

Bermúdez, J. (2000). Self-Deception, Intentions, and Contradictory Beliefs. *Analysis*, *60*(4), 309–319.

Brown, R. (2004). The Emplotted Self: Self-Deception and Self-Knowledge. *Philosophical Papers*, *32*(3), 279–300.

Darwall, S. (1988). Self-Deception, Autonomy, and Moral Constitution. In B.P. McLaughlin & A.Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 407–430.

Davidson, D. (1986). Deception and Division. In J. Elster (Ed.), *The Multiple Self*. Cambridge: Cambridge University Press, 79–92.

Frankfurt, H. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Frankfurt, H. (1999). *Necessity, Volition, and Love*. Cambridge: Cambridge University Press

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

Holton, R. (2001). What is the Role of the Self in Self-Deception?. *Proceedings of the Aristotelian Society*, *101*(1), 53–69.

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

Kant, I. (1996/1785). Groundwork of the Metaphysic of Morals. In I. Kant, *Practical Philosophy*. (tr. and ed. by M.J. Gregor). The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press, 37–108.

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

Moran, R. (1988). Making Up Your Mind: Self-Interpretation and Self-constitution. *Ratio (new series)*, *1*, 135–151.

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge. Princeton*. New Hersey: Princeton University Press.

Nelkin, D. (2002). Self-Deception, Motivation, and the Desire to Believe. *Pacific Philosophical Quarterly*, *83*(4), 384–406.

O'Neill, O. (1985). Consistency in Action. In N. Potter, & M. Timmons (Eds.), *Morality and Universality*. Dordrecht: Reidel, 159–186.

O'Neill, O. (2004). Self-Legislation, Autonomy, and the Form of Law. In H. Nagl-Docekal, & R. Langthaler (Eds.), *Recht, Geschichte, Religion: Die Bedeutung Kants für die Gegenwart*. Sonderband der Deutschen Zeitschrift für Philosophie Berlin: Akademie Verlag, 13–26

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

Responsibility and Self-Deception: A Framework

Dana Kay Nelkin *
dnelkin@ucsd.edu

ABSTRACT

This paper focuses on the question of whether and, if so, when people can be responsible for their self-deception and its consequences. On Intentionalist accounts, self-deceivers intentionally deceive themselves, and it is easy to see how they can be responsible. On Motivationist accounts, in contrast, self-deception is a motivated, but not intentional, and possibly unconscious process, making it more difficult to see how self-deceivers could be responsible. I argue that a particular Motivationist account, the Desire to Believe account, together with other resources, best explains how there can be culpable self-deception. In the process, I also show how self-deception is a good test case for deciding important questions about the nature of moral responsibility.

Introduction

Self-deception is a phenomenon that manages to strike us as both very common and yet not easy to characterize. It is easy to see how one person can deceive another; after all, one person knows the truth and, using any of a variety of techniques, can influence another to believe a falsehood. But how can one person deceive herself?

Some have decided that self-deception is impossible, but most theorists writing on the subject have continued to assume that it is a real phenomenon, while acknowledging that the correct model of self-deception must diverge in at least some ways from that of interpersonal deception. For the last several decades, those working in the area have tended to occupy one of two main positions on the question of what self-deception is: Intentionalism and Motivationism. Intentionalists preserve at least two key components from the

* University of California, San Diego, USA.

model of interpersonal deception, arguing that self-deception is intentional and that the self-deceived person holds a true belief while at the same time wrongly believes the contrary. In order to address the question of how one person who knows the truth could possibly convince herself of its contrary, the Intentionalist picture is most often combined with the view that a single person can be divided, or partitioned in some way, so that one part of her believes the truth, and brings about the contrary belief in the other part of her.[1] Such a picture allows us to think of self-deception as very like interpersonal deception while also maintaining a great deal of explanatory power. It helps explain why it is that self-deception is a sophisticated cognitive activity, not possessed by young children, for example. It also helps explain why we often hold self-deceivers responsible for their deception and for the consequences that follow from it. It can account for a wide variety of cases of self-deception, including cases in which self-deceivers believe things about the world that they want to be the case, as well as cases of so-called "twisted" self-deception, in which they believe things about the world that they would rather not be true. (For example, consider the case of a husband who desperately wants his wife not to be having an affair, but, worried about getting caught off guard, convinces himself that she is.) All that is required to account for both cases is that self-deceivers intentionally engage in the formation of a belief they know to be false.

Despite these theoretical virtues, and the fact that Intentionalism may once have been the dominant picture, it has lost ground in recent years to Motivationists. They reject Intentionalism on the grounds that it either leads to paradox or, at a minimum, to the unnecessary and unsupported postulation of strongly autonomous parts of the mind.[2] Opponents of Intentionalism (myself included) claim to be able to capture (most of) its theoretical advantages without the metaphysical and psychological complexity of partitioning.

Motivationists have in common the commitment to the idea that self-deception involves a kind of motivated state, while rejecting the commitment to the deception being an intentional action. Beyond this common commitment, motivationists divide in their answers to several further questions about the nature of self-deception, including these:

[1] See Pears 1984, for example.
[2] See Mele 1987 and Johnston 1988 for examples.

1) What is the guiding motivation? Assuming it is a desire, a desire for what? (the content question)
2) What is the product of self-deception about p? A belief that p? A sincere avowal that p? A pretence that p? A belief that one believes p? (the product question)
3) How does the motivation generate the product of self-deception? (the process question)
4) What accounts for the irrationality in self-deception? (the irrationality question)
5) Is there a belief that not-p? (the contrary proposition question)
6) If the product of self-deception is a belief that p, must that belief be false? (the truth value question)

To see the disagreements starkly, it helps to consider some examples. Some argue that the product of self-deception is a false belief that *p* and that there is no contradictory true belief, while others argue that self-deception requires a true belief that *not-p*, and no contradictory false belief that *p*. It is puzzling how a self-deceiver could have no false belief about which she is self-deceived, but various alternatives are suggested in its place in answer to the product question. For example, there are those who argue that rather than having a false belief, the self-deceiver sincerely avows a false claim (the avowal view)[3]; those who argue that the self-deceiver pretends that the false belief is true (the pretence view)[4]; those who argue that the believer has a false belief about her own states of mind, rather than about the object of self-deception (the failure of self-knowledge view)[5]; those who argue that the believer acts in some ways as if she believes the false belief (the behavior view)[6] and combinations thereof.

Notably, motivationists also divide on the question of the nature of the motivation in question. Must it be a desire or could other emotions, such as anxiety suffice?[7] If self-deception must be driven by a desire, what is the content of that desire?[8]

[3] See, for example, Audi 1997 and Funkhouser 2005.
[4] For example, Gendler 2007.
[5] See, for example, Scott-Kakures 1996, Fernandez 2011, and Funkhouser 2005.
[6] See, for example, Audi 1997 and Funkhouser 2005.
[7] See, for example, Barnes 1997.
[8] See, for example, Mele (1997, 2000, 2001) for an account in which the desire need have no particular content, and Mele's earlier (1987) for an account in which the desire must be the desire that p be true, where the product of self-deception is the belief that p. See Nelkin 2002 for a different account to be explained shortly.

The best account of self-deception will answer all of these questions in a plausible and coherent way, as well as yield explanations of some important and well-recognized features of self-deception. Ideally, the account would explain why self-deception takes considerable cognitive sophistication (and thus, why young children do not seem to be capable of it), why we often attribute responsibility on the part of self-deceivers for their self-deception and its consequences, and why we have the thought that self-deception shares similarities with other-deception.

In this paper, I will focus centrally on one challenge that faces all motivationists, and in doing so bring out the virtues of one particular kind of motivationist account. One of the claimed virtues of intentionalism is that it can explain why we often hold self-deceivers responsible and, indeed, often blame them for their deception and for the consequences that follow. If self-deception is an intentional act, then it is chosen by the agent herself, in full knowledge of what she is doing, and is a paradigm object of blame. But if we leave intentions out of the picture, and we think that self-deception is paradigmatically an *un*intentional (and often unconscious) process, then it becomes less obvious that the self-deceived are responsible for their states and for their consequences.

In an earlier paper, I argued for a particular motivationist account, the Desire to Believe account, that seemed most naturally to explain the fact that we often hold people responsible for their self-deception.[9] I believe that the account is especially well-suited for this task, but I also believe that there are more questions to be raised for motivationist accounts, including the one I defend. And in this paper, I articulate these questions, and develop answers based on the Desire to Believe Account. It is worth noting that some (but not all) of what I say could also be adopted by other motivationist accounts. I begin in section 2 by setting out the motivationist account I favor. In sections 3 and 4, I then elaborate some challenges to accounting for responsibility, homing in on where the important issues lie. While I will not here defend a comprehensive theory of responsibility, I will make a start in identifying the issues that must be resolved in attributing responsibility for self-deception, as well as locating self-deception relative to other sorts of objects of responsibility in which we have confidence in our attributions of responsibility (or non-

[9] See Nelkin 2002.

responsibility). In the process, I also hope to show how self-deception can provide an illuminating test case for certain theories of responsibility.

1. The Desire to Believe Account

It will be helpful to lay out some preliminary methodological assumptions. First, I begin with the assumption that the problems with the intentionalist picture are difficult ones, and that if a motivationist picture can succeed in accounting for cases of self-deception and can explain the phenomena we think need to be explained, then we should adopt it.

Second, an account that offers a set of necessary and sufficient conditions would be useful and neat, but it may be that there are blurry boundaries, and that there isn't a perfectly neat set of necessary and sufficient conditions. In that case, it would be better to adopt an account of conditions that are sufficient for self-deception, and that also characterize all the central cases of self-deception, distinguishing it clearly from phenomena that we believe are distinct (such as certain "cold" or unmotivated kinds of belief formation like simple cognitive error, as well as other sorts of "hot" or motivated belief formation such as wishful thinking.)

Endorsing this approach, I aim to combine insights of both intentionalist and motivationist models in order to arrive at a model for being self-deceived about a proposition, say, *p*. The model offers a sufficient condition for being self-deceived, and, I believe, necessary conditions for all the central cases of self-deception that distinguish it from other well-recognized phenomena. The key and distinctive aspect of the account is its answer to the content question: the guiding motivation in self-deception about *p* is a desire to believe that *p*. Accounts that leave open the content of the desire in question, such as Mele's (1997), are appealing in their flexibility, but they also suffer from failing to capture what distinguishes self-deception from other sorts of "hot" belief formation. For example, consider the case of Otis who is motivated to have an answer to every question. When asked whether the 1991 Braves would have prevailed over the 1999 Yankees, his desire to have an opinion motivates him to focus on a particular set of statistics (while ignoring others that might induce doubts) allowing him to form the view that the Braves would have won. Intuitively, this is not a case of self-deception. Or consider the case of Ben, a small child who comes to believe, against his evidence, that his babysitter is not a nice person. He desires his parents' return, and through a complex defense

mechanism forms this belief as a result.[10] These are cases of "hot" or motivated biased belief formation, but do not seem to be ones of self-deception. The category of motivated and biased belief seems, intuitively, larger than that of self-deception.

Traditionally, accounts of self-deception that are more restrictive in the content of the desire in question have tended to identify the relevant desire as the desire that *p*.[11] But, as Mele (1997) points out, this excludes clear cases of self-deception, namely, those of the "twisted" variety. We can see, however, that the desire to *believe* that *p* is plausibly attributed in both paradigmatic straight and twisted cases, and nicely excludes cases like that of Otis and Ben and the Babysitter and other cases of motivated belief that are not self-deception. This answer to the content question also shows why the intentionalist view is appealing, even though false: intentionalist views take it that the intention to deceive either arises from, or is partly constituted by, a desire to generate a belief that *p*. This aspect of the intentionalist view is thereby preserved.

How does the view answer the other questions? As for the product of self-deception, I believe that the most natural answer is that it is a belief. If it is, then once again self-deception will retain a key feature of deception in general. Further, taking it to be a belief explains the (several) kinds of behaviors that the self-deceived person then engages in.[12] To take an example, the mother who is self-deceived in believing her son will return keeps his bedroom undisturbed, sincerely swears that he will return, and makes sure that the house numbers are always lighted. Her behavior is well explained by her believing that he will return.

Finally, in conjunction with identifying the desire to believe as the content of the guiding motivation, understanding the product of self-deception as a belief gives self-deception a kind of intelligibility that is otherwise lacking. *Were* the agent to become aware of her desire to believe, she would be able to see immediately that the product satisfied her desire. No special knowledge of

[10] I discuss these cases in more detail in Nelkin 2002.

[11] See, for example, Mele 1987.

[12] As mentioned, there have recently been a number of alternative suggestions made as to what the product of self-deception is. See notes 3-6. Some also argue against the claim that belief is the product on the grounds that certain behaviors of the self-deceiver, such as the avoidance of evidence, does not fit well with it. But I believe that this behavior can be well accommodated by understanding it as motivated treatment of the evidence.

defense mechanisms and their function would be required to see that her desire had been satisfied.[13]

As for the process, I believe that in its details this is an empirical question, and there may be a great variety of ways that the motivating desire operates to result in the belief that *p*. But we can say some things very generally about it. It seems that the desire has an influence on the agent's treatment of the evidence available to her – either in the selection of data she focuses on, or in the inferences she draws from it – so that she sees the evidence as supporting the belief that *p*, even though it in fact provides greater support for *not-p*.[14]

This answer to the process question also leads naturally to the question of where the irrationality is to be found. It need not be in starkly contradictory

[13] Mele (2009) has argued recently that the key aspect of my account that underlies this intelligibility claim is incorrect. In particular, he offers a counterexample to the claim that a desire to believe is necessary for self-deception. He asks us to «imagine two jealous husbands with very similar evidence in very similar circumstances. Each acquires the false, unwarranted belief that his wife is having an affair--the belief that a, for short» (2009, p. 268). Both treat the evidence they have in similar biased ways. One husband, Jack, is motivated by the desire to believe that a. The other, John, lacks that particular desire, but «does have desires that contribute to his having acceptance and rejection thresholds for a that are just like Jack's. Suppose for good measure, that John has a desire not to acquire a false belief that his wife is innocent of infidelity» and that this desire is what motivates his biased treatment of the evidence and acquisition of the belief that a (p. 268). Because the two husbands are so similar, «it is very plausible that if Jack is self-deceived, so is John» (p. 269). I do not believe that these cases give us good reason to reject the Desire to Believe account. First, as mentioned, the account is consistent with there being blurry boundaries between self-deception and other sorts of irrational motivated belief formation. But there are boundaries, nonetheless, and it is crucial to evaluating the example to distinguish between John's "having desires that contribute" to the biasing as Jack does and his having the very particular desire that Mele offers us "for good measure." Cases like Otis and Ben and the Babysitter show us that restriction on the content of desires is essential to distinguish self-deception from other sorts of cases, and to my knowledge, Mele does not respond to this concern. Equally importantly, the case of John is one in which the content of the motivating desire is actually very similar to that of Jack's. It is a desire not to have a false belief that p, rather than to acquire a belief that not-p. Thus, I believe that Jack's case falls at best in the blurry boundary area of self-deception. This is precisely because the content is so similar as to retain something approaching the intelligibility provided by the Desire To Believe account. (Both Jack's and John's desires have as their objects beliefs with 'p' embedded in its content as the relevant object of desire.) And yet the content is not so similar as to be as easy to see that a desire is fulfilled by successful self-deception; and the content is different enough that it fails to capture the key similarity to other-deception that the account provides. All other things equal, the closer the content of the desire gets to the belief that not-p, the more clearly it is a case of self-deception, on the Desire to Believe account, and, conversely, the farther the content gets from the belief that not-p, the less clearly it is a case of self-deception. If, as I have argued, we confine ourselves to the most specific version Mele offers of John's situation, these cases do not appear to give us reason to doubt this.

[14] See Kunda 1990 for a classic statement of a theory of motivated belief.

beliefs, as it is on the intentionalist picture, but on this motivationist picture, we find it in the logical tension between the self-deceptive belief and the agent's evidence. This in turn can help explain the seemingly odd behaviors that are found in many paradigmatic cases of self-deception. On the one hand, the product of self-deception is the belief that *p*, but there is also avoidance of evidence that better supports *not-p*, for example. Both of these behaviors can be explained by appeal to this sort of mechanism of motivated selective evidence gathering.

On the Desire to Believe account, it is not necessary to have a belief that *not-p*. It is consistent with the account that some cases of self-deception include such a belief, and therefore that some self-deceivers do have contradictory beliefs, but it is not required, as long as the other conditions are met.[15]

How does the account answer the truth value question? I now believe that the account can be open on this question without loss of explanatory power. Typically, we assume that if someone is deceived, she believes something that is false. But by coincidence, it could turn out that though her evidence fully supported *not-p*, and her desire to believe *p* led her, via a biased treatment of evidence, to believe that *p*, she got lucky (so to speak) and *p* is true. Since the psychological process is the same regardless of the truth value of *p*, I think it is reasonable to treat even such a "lucky" case as one of self-deception. But if one prefers instead to treat such a case as very like self-deception, but not *strictly* self-deception, I see no objection to doing so.

Putting the pieces together then, we have a view according to which a person is self-deceived with respect to *p* when she believes that *p* as a result of the biased treatment of evidence which is in turn motivated by the desire to believe that *p*, and when the evidence available to the person better supports *not-p*. In addition to the advantages already described, the view succeeds in explaining why a certain degree of cognitive sophistication is required for self-deception, since it requires having a desire to believe – a second order desire

[15] Some have argued that "at some level" one must believe the truth, e.g., Funkhauser (2005) and Audi (1997). If this turned out to be true, it could be added as a condition to the account. And the kind of partitioning it would require would still be substantially less robust than that of the intentionalist picture in which the mind is divided into true sub-agents. I think it is psychologically plausible that we do have such partitioning of beliefs in some instances. But I do not think it is necessary to account for all of the features of self-deception. The behavior that it is thought to be essential to explain seems to me to be explainable simply by the strong desire to believe that p, operating through a mechanism of selective evidence gathering.

about one's own mental states. This is a significant advantage of the view over accounts that leave the content of the guiding desire unrestricted. And as we have seen, it preserves several aspects of the intentionalist picture, showing why, even if that picture is incorrect, it has been regarded as attractive.

2. Motivationism, Self-Deception, and Responsibility: First Pass

With this particular motivationist account in view, we can turn to the question of whether, and, if so, when self-deceivers are responsible for their self-deception and its consequences.

One answer to the question of when one might be responsible for self-deception is that no one ever is, on the grounds that no one is responsible for anything. This position on responsible action has been forcefully defended and has a number of adherents.[16] I will here set aside this challenge, however, and concentrate on reasons for thinking that the move to motivationism causes a special reason for skepticism.

I will here adopt a very general approach to moral responsibility that understands responsibility to depend on control.[17] In particular, for an agent to be responsible for her actions, she must be responsive to reasons. There are a number of ways that this idea has been developed, and we will see that the details may matter when it comes to judging the responsibility for self-deception. For example, some have defended mechanism-based approaches to reasons-responsiveness, arguing that the responsible agent must act on a mechanism that is reasons-responsive, while others have defended agent-based approaches, arguing that it is the responsible agent herself that must be responsive to reasons.[18] For now, let us begin with the intuitive idea that to be

[16] See, for example, Pereboom 2001.

[17] I here sidestep the important issue of whether moral responsibility is consistent with determinism, for the reason that I want to concentrate on the particular question of whether motivationism has special difficulties accounting for it that intentionalism does not. It is worth noting, however, that incompatibilists–those who deny that responsibility is consistent with determinism–often accept conditions like those described here as necessary, even if not sufficient. (For example, see O'Connor 2001.)

[18] The most notable defenders of a mechanism-based approach are Fischer and Ravizza, who argue that to be responsible one must act on a moderately reasons-responsive mechanism for which one has taken responsibility (2000). Levy (2004) adopts this account in discussing self-deception. Defending an agent-based approach, Wolf (1991) argues that the responsible agent must be able to act on the reasons there are. Some, like Wallace (1994), also defend an agent-based approach, and combine this with the claim that a responsible agent must have general rational capacities. In contrast, others,

responsible, or *a fortiori*, blameworthy for one's actions, one must be able to respond to the reasons that there are.

Questions about the relationship between specific models of self-deception and responsibility have been raised continuously throughout the contemporary discussion of the phenomenon. For example, in a classic article advocating an intentionalist picture of self-deception, Demos writes: «A man who lies to himself is blameworthy because he acts with knowledge of the facts and thus may be held responsible for his erroneous belief» (1960, p. 589). Writing a few years later, Fingarette seems to draw a quite different conclusion:

> There is thus in self-deception a genuine subversion of personal agency and, for this reason in turn, a subversion of moral capacity. The sensitive and thoughtful observer, when viewing the matter this way, is inclined not to hold the self-deceiver responsible but to view him as a 'victim'. (Fingarette, 1969, p. 140.)

And writing recently on the topic, Neil Levy suggests that the motivationist conception of self-deception in particular «has neither need nor place for attributions of moral responsibility to the self-deceived in paradigmatic cases» (2004, p. 294).[19] Is there reason to think that motivationists, in particular, have difficulty accounting for moral responsibility in paradigmatic cases?

The argument that most cleanly distinguishes cases satisfying intentionalist conditions from those satisfying motivationist ones is based on the claim that we can only be responsible where there is intentional action. Choice, it has been argued, is the locus of responsibility, and it does not make sense to hold people responsible for things that they did not choose.[20]

But this view is undermined by a number of apparent counterexamples, and we do not have to look to the hard cases of self-deception to find them. Cases of recklessness seem to qualify as responsible actions. For example, it has happened that though people do not intend to create a risk to their neighbors,

including Fischer and Ravizza and Wolf, argue that the relevant abilities must in some sense be exercisable in the particular situations in which they are responsible. I develop and defend a view that is agent-based and that requires abilities to exercise rational powers in particular situations in Nelkin 2011.

[19] This strong claim appears to be tempered by other claims in Levy's paper that make his view somewhat difficult to pin down. In the conclusion of his paper, for example, he acknowledges that self-deceivers may "often" be responsible (2004, p. 310). But setting this aside this uncertainty, his paper makes explicit the important question of just how much and when self-deceivers are responsible. See also the interesting treatment of Linehane (1982) that also focuses attention centrally on this issue.

[20] See Alexander & Ferzan (2009) for an articulation of this theory.

they consciously disregard knowledge of the risk in setting off firecrackers in the street, for example.[21] Despite not harming or even creating risks intentionally, they are blameworthy for acting as they do. This can be explained by their having had the ability to respond to reasons – an ability that they failed to exercise.

Still, even if we do not draw a line around intentional action, there have been tempting reasons for drawing it in another place that also excludes culpable self-deception on the motivationist picture, namely, around awareness. That is, it has been argued that people cannot be responsible for acting in harmful ways (or unreasonable-risk-enhancing ways) if they are unaware of the risks in question.[22] For example, if I am completely unaware that my light switch has been hooked up to a stick of dynamite currently located in my neighbor's kitchen and I flip the switch, thereby causing the dynamite to explode, I am not blameworthy for anything I did. Having no access to the relevant reasons, I couldn't have responded to them. Here, too, intentionalism and motivationism seem to fare differently. If the self-deceiver intentionally deceives, then it simply follows that she knows of the risk in question. (In fact, she is trying to maximize it.) But if self-deception is an unintentional, and likely unconscious process, then she need not be aware at all of a risk of generating a self-deceptive belief.

But this requirement on responsible action also appears to be too strong. There are cases in which one is unaware of the risk one creates (or fails to stop), but it remains the case that one *ought* to have been aware of it, and so is blameworthy for failing to acquire awareness and act accordingly. Again, we need not look to self-deception itself to see that this is the case. A person who tells an offensive joke may not be aware of the risk that his audience will take offense, but might very well be blameworthy for having done it and responsible for the effects.[23] Thus, if the criterion for responsibility is not awareness, but rather that one "should have known", *and* if self-deception at least sometimes satisfies this criterion, then it is possible to see self-deception as a case in which people are sometimes responsible. Exactly how often and when will then depend at least in part on whether the self-deceiver satisfies the "should have

[21] See the Model Penal Code on recklessness and Alexander and Ferzan (2009, p. 25).
[22] This is what Sher (2009) calls the "Searchlight View".
[23] For this and related cases, see Sher (2009, p. 28).

known" condition, and, of course, the conditions under which one "should have known."[24]

How we are to understand these conditions is itself a matter of fairly intense controversy. But before entering into that debate, I think it is already possible to see how the Desire to Believe Account is in one way better suited than other motivationist accounts to show that such conditions are at least sometimes satisfied.

On an *unrestricted* desire account, it is admittedly not obvious how a self-deceiver could be responsible. For example, if the product of self-deception is the belief that *p*, and the guiding desire is a desire for *q*, then the operative mechanism might be one that is quite complex and that we could not expect the self-deceiver to be aware of. For example, citing research on jealousy, Mele (2001) suggests that a case of self-deception might have stemmed from a desire to have closer relationships, and through a protective mechanism lead to a belief that one's spouse is having an affair. If the content of the desire is so far removed from the content of the product of self-deception, it seems unreasonable to expect the self-deceiver to have even been on guard against such a process.[25] But cases that satisfy the conditions of the Desire to Believe Account have a kind of immediate intelligibility that these cases do not. Were the agent to be aware of her desire to believe that *p*, she could immediately see that her belief that *p* satisfies her desire. This is not yet to say that she is responsible for her deception; but it does make clear how it could be comparatively easier for her to be on guard not to form this kind of motivated belief against the evidence.[26]

Of course, it is open to the advocate of an unrestricted account to allow that some cases satisfy the narrower conditions of the Desire to Believe account, and given that the account only aims to give sufficient conditions, can allow even that some cases are intentional (though that would admittedly require allowing for the strong partitioning that there is reason to be skeptical of). Still,

[24] It is also possible that in some cases of self-deception, people are aware of the risk in a general way, even if not of the specific process.

[25] For further discussion of this case, see Nelkin 2002.

[26] In this way, the Desire to Believe account already anticipates one line of criticism Levy (2004) levels against motivationist accounts that try to preserve the claim that (some) paradigmatic instances of self-deception are one for which people are responsible. He there criticizes other motivationist views precisely on the basis of the differential content between desire and belief (2004, p. 308), asking how one could possibly see the relationship between them. But this criticism does not apply to the Desire to Believe account.

thinking of cases with the desire to believe as *central* allows for comparatively more attributions of responsibility.

At the same time, it is important to note that the Desire to Believe account does not entail that all cases are ones for which one is responsible either. This flexibility is welcome, I believe. After all, figuring out what we are responsible for – if anything – in the way of actions, omissions, and states of all kinds has itself been the source of enormous controversy, and at least some of this controversy rests on debates about the empirical facts concerning the capacities of human beings. We should be cautious in approaching the question of how often, if at all, people are responsible for their self-deception and its consequences. So far, then, I am making a simple comparative claim: the Desire to Believe account has an advantage over other motivationist accounts in that the surface intelligibility of the relationship between guiding motivation and belief makes it easier, all other things being equal, to either be aware of the non-rational process of belief formation, or at least to be on guard against it.

3. Developing a Framework

If what I have argued so far is correct, then the Desire to Believe account has one advantage over other motivationist accounts in accounting for responsibility. But this leaves open all sorts of questions, including the conditions under which any particular instance of self-deception is something for which the self-deceiver is responsible. In this section, I will spell out some issues whose resolution is needed to make progress in answering these questions, and distinguish two sorts of strategies we can take toward making particular determinations of responsibility. I will conclude by showing how each can work.

First, it is important to get clear about exactly what the self-deceiver is supposed to be responsible for. We should distinguish between the process of self-deception, the immediate product of self-deception, and its more indirect consequences. The point made so far on behalf of the Desire to Believe account addresses the process directly, as the account shows how a person could more easily be on guard to avoid the process itself given its surface intelligibility. But even if one had no knowledge of the process, or even any reason to be on guard against it, one might still be responsible for the product. How can this be? Suppose, as is plausible, that one has a duty to form beliefs

that conform to the available evidence, particularly in cases in which much is at stake. Then even if one is not (and has no reason to be) aware of the self-deceptive process that generates a belief undermined by the evidence, it can still be that one ought to critically examine such beliefs and eliminate them. Alternatively, it can be that one ought to engage in simultaneous processes, which would compete with the self-deceptive one by including seeking out and carefully evaluating relevant evidence. These latter points are consistent with a variety of motivationist accounts of responsibility, not just the Desire to Believe account.

These points also illuminate the path to an insight about the nature of responsibility itself. They do so by supporting one general way of developing the reasons-responsiveness approach to responsibility. To see how, recall that there are different ways that the approach has been developed: we can require that the responsible agent act on a *mechanism* that is itself responsive to reasons, or we can require that the responsible agent *herself* be reasons-responsive in the relevant circumstances.[27] On the former view, if the motivated biasing mechanism of self-deception is not reasons-responsive (as seems plausible), then the self-deceiver will not be responsible. In contrast, on the latter view, it is not exonerating that a non-reasons-responsive mechanism is operating. What matters is whether *the agent* could have either prevented that mechanism from operating, or instead put another into action. Since there is good intuitive support for the idea that self-deceivers can be responsible, the agent-based approach to responsibility and the motivationist picture of self-deception provide each other with mutual support.[28]

Finally, in figuring out what agents are responsible for, we should note that it is not easy to say what the relationship is between our responsibility for our

[27] See note 18.

[28] I agree here with one part of DeWeese Boyd's (2010) discussion of Levy's application of a mechanism-based account. He writes: «However, the question isn't whether the biasing mechanism itself is reasons responsive but whether the mechanism governing its operation is, that is, whether self-deceivers typically could recognize and respond to moral and non-moral reasons to resist the influence of their desires and emotions and instead exercise special scrutiny of the belief in question» (2010, section 5.1.). Where I part company is that I reject the necessary attribution of reasons responsiveness to any mechanism--whether the biasing one or the governing one. (I am actually unsure how one would individuate the biasing mechanism and the mechanism "governing its operation".) Thus, I also reject the equivalence of reasons-responsiveness of governing mechanisms of biasing mechanisms with that of agents. But I agree with the second half of the equivalence claim, namely, that what matters for responsibility is what the agent's abilities are. Further, I claim that the case of self-deception actually helps adjudicate between mechanism-based and agent-based views.

actions and attitudes on the one hand, and their consequences on the other in general. Even where self-deception is not at issue, this is not obvious. Must the consequences be foreseeable? Must they be foreseeable to have a high probability of occurring, given one's actions or attitudes? How high must the probability be? These are difficult questions to answer. Nevertheless, it seems reasonable to conclude that where the consequences are fairly obvious results of one's actions or judgments, we have a better *prima facie* case for one's being responsible for the consequences if one is responsible for the action or attitude.

This point leads naturally into a second fundamental issue that we must address in determining responsibility. Although it might be possible to make some generalizations about sets of cases, instances of self-deception vary on a number of dimensions that can be independently relevant to responsibility (and to the degree of responsibility). For example, as we've just seen, what could reasonably be expected in terms of perceived risk plays a role in determining responsibility for consequences generally. And this would seem to apply to self-deception no less than to other cases of responsible action or omission. Suppose the self-deceived husband couldn't have possibly predicted that his wife would attempt suicide on learning of his belief that she is having an affair. Then the extent of his responsibility for such a consequence and of his blameworthiness for his judgment will depend on the risks as he could reasonably understand them at the time.[29] This case shows that there is a second factor at work in addition to degree of risk and that is severity of the harm risked, as well.

In other words, what is at stake in forming the self-deceptive belief plays a role in determining blameworthiness. A case that brings this out perhaps even more starkly is a case of a parent who is self-deceived in believing that her child is not abusing his own children. The failure to treat the evidence appropriately in this case results in her not reporting her child or protecting her grandchildren from further abuse.[30] The high stakes are not by themselves sufficient for attributing a high degree of blameworthiness, but they are one

[29] I here set aside the very large question of moral luck in consequences of one's actions. I instead assume that one is responsible for the consequences of one's actions, whatever they are, but that one's degree of blameworthiness depends not on the actual consequences, but on what it was reasonable to expect in terms of perceived risk. Thus, if the risk were high that his wife would attempt suicide and could easily be discerned, but she did not in fact attempt it, he would be equally blameworthy as in the situation in which she did.

[30] For a similar kind of case, see Barnes 1997.

factor that could potentially distinguish the case from another, like it in other ways, save for the fact that the stakes are not so high. Interestingly, Levy does not consider cases of this sort in arguing that motivationists ought to abandon the claim that self-deceivers are responsible in (some) paradigmatic cases. But it is precisely in cases like this that we can see the cost of abandoning what is an intuitively powerful thesis. Even if it were a cost ultimately to be borne, cases like this show how significant it is, and how strong the arguments would have to be for paying it.[31]

The flip side of high stakes, understood as harms to others, is the potential cost to oneself in avoiding the self-deception. This, too, is a factor that might vary from case to case. If a person could simply not go on living once having acknowledged the truth about his spouse's fidelity, for example, that seems a different sort of case than a case of someone for whom recognizing the truth would merely cause some feelings of embarrassment. A related but separable factor is the simple level of difficulty required to avoid the self-deception or to eliminate its product. There is likely a strong correlation between high stakes and difficulty, but it is possible that even where the stakes are not so high, it might, for some other reason be difficult to avoid.

Taking these two points together, we see that in making specific attributions of responsibility, we will need to take each case on its own terms and distinguish between process, product, and consequences on the one hand, and specific features of the case relating to the stakes and level of difficulty on the other. This suggests taking a fairly individualized approach to particular cases.

Yet at the same time, there may be ways of grouping certain sorts of cases together once we understand better the conditions for responsible negligence. As mentioned in the last section, this is itself a matter of great debate among both philosophers and legal theorists.

The question of how we should treat negligence – or, as Alexander and Ferzan put it, «inadvertent creation of unreasonable risks» – is not settled (2009, p. 69). And although there are relatively few skeptics about culpable

[31] Linehan (1982) and Jenni (2003) offer powerful examples of reported mental lives of Nazi doctors who collaborated–unknowingly?–in the deaths of hundreds of innocent people. While we would need to know a great deal about the cases to determine whether they were genuinely self-deceived about what they were doing, the fact that self-deception is even a relevant hypothesis in explaining their behavior provides a good example both of how much can be at stake and the intuition that self-deceivers can be responsible.

negligence, there is also much debate about *how* one can be responsible for something of which one is unaware, when faced with skeptical arguments.[32] The question is how we can be responsible for not knowing something, or not recognizing it, when we are unaware precisely of what we are supposed to know. In a recent article on the subject, Moore and Hurd (2011) reject skepticism, sticking with their «strong, bottom line intuition that blame can rightly be attached to many of the examples of negligent conduct» that appear in courts of law. But they also conclude that they are unable to come up with a «new, unifying theory of why negligence is culpable», settling for a disjunction of several sorts of conditions. (Moore & Hurd, 2011, pp. 191-192).

In light of the unsettled nature of the debate about culpable negligence in general, there are two strategies we can take to self-deception in particular. The first is to defend a general theory of negligence, and then apply it to a range of cases of self-deception. The second is to leave open what the full theory of negligence is, and instead take the more modest approach of comparing cases of self-deception to other sorts of negligence about which we have some confidence. Let us take each in turn to see how it might be developed.

One general theory of how people can be responsible for negligence and its consequences is a kind of "tracing" account.[33] According to this sort of view, one might be responsible now for one's self-deceptive belief even though one is completely unaware that it is self-deceptive, as long as one's belief is due to an earlier moment of choice during which one was aware of the risk of biased belief, could have chosen to put obstacles in the way, and did not. At the earlier time, one was reasons-responsive, and one chose badly. One's responsibility for the later deception and further consequences "traces back" to that moment. While I do not think that this covers all cases that intuitively count as culpable negligence, it might very well account for a significant number of cases of culpable self-deception.[34] It may be that there are moments in typical

[32] For two excellent recent treatments, see Moore & Hurd (2011) and Sher (2009).

[33] See Sher (2009) for a discussion of this kind of view.

[34] Levy considers this sort of account, but quickly moves on, on the grounds that it only explains why self-deceivers are "sometimes" responsible (2004, p. 304). He claims that motivationists make the "much stronger" claim that "at least typically self-deceivers are culpable" (2004, p. 304) and that they retain the "presumption" that self-deceivers are responsible. (2004, p. 310). The various claims lead to the question of what it is that motivationists have actually claimed. On my reading, at least some motivationists make nothing so strong as the claim that responsibility is a presumption in cases of self-

cases of self-deception when self-deceivers are aware of a choice to look into a piece of evidence more systematically, for example, and they choose not to, thereby allowing the process of self-deception to continue. Ultimately, it is an empirical question what sort of awareness accompanies any particular instance of self-deception. Far from being a disadvantage for the account, though, I take this flexibility to be an advantage for the reasons spelled out earlier.

But tracing is not necessary for explaining culpable negligence in general, and, as I will argue, it is not necessary for accounting for culpable self-deception. For example, consider that people are subject to general epistemic norms to pay attention to evidence on all sides of a question and even to seek out certain kinds of evidence, at least when the stakes are significant.[35] We have an obligation to take due care in our approach to evidence on important matters. One's failure to do so might then be culpable, even when it does not trace back to an earlier moment of conscious decision not to treat the evidence in a certain way. One might be able to respond to the reasons of taking due care, and yet one fails to do so.

Levy (2004) argues that self-deception is not culpable on the basis of this sort of non-tracing grounds. He begins by suggesting that the relevant conditions under which we could be culpable for such epistemic failures in self-deception cases (such as failures to consider evidence on both sides) are cases in which «(1) the subject matter is important...and (2) that we are in some doubt about its truth» (Levy, 2004, p. 305).[36] Having set out these conditions, he then claims that they are rarely (if ever) met in cases of self-deception. He suggests that «concurrent doubts are ruled out almost by definition: effective self-deception seems to *preclude* the concurrent satisfaction of (2). Successful acts of self-deception leave me in no doubt about the proposition concerning which I am self-deceived» (Levy, 2004, p. 307). And, further, there is no reason to think that they have repressed doubts that were present earlier in the process. It is worth noting that this argument

deception. So in the end, it is not clear what the ultimate disagreement is between Levy and his targets here.

[35] Among the many treatments of this sort of obligation are Fitzpatrick 2008 and Hurd & Moore 2011. For an early version, see Clifford 1877.

[36] Levy argues that condition (1) is not sufficient on the grounds sometimes even when a matter is important, we have no obligation to consider evidence on both sides. As an example, he offers that of the Holocaust scholar who has no obligation to consider the arguments of those arguing that the Holocaust was a hoax. This may be true, but only if the scholar has already considered and responded to the general category of reasoning. Thus, I am not convinced that a second condition is needed.

assumes that the product of self-deception is a belief. I agree that it is, but it is important that not all motivationists do. And one reason for rejecting the claim that belief is a product, for some theorists, is precisely self-deceivers' behavior that is claimed to be characteristic evidence of doubt. More importantly for our purposes here, even if the product of self-deception is a belief with significant consequences for behavior and other attitudes, this is perfectly consistent with doubt. For example, I believe that my spouse and I made the right decision about how much to limit exposure to TV for our children in their early years, but I am not without any doubt about it. In fact, I suspect that for many, any of a large number of parenting decisions is a good candidate for belief with doubt. And the more there is at stake, the more significant the obligation to examine our evidence.

Thus, while there is much work to be done in discovering the correct and complete theory of culpable negligence, there is also no obvious general reason for thinking that self-deception will fail to satisfy its conditions, at least some of the time. To support this claim further, let us turn to the comparative project of examining cases of self-deception alongside other kinds of cases of apparently culpable negligence.

Suppose that instead of being motivated by a desire to believe her son innocent of any possible crime, the grandmother described earlier is simply distracted by loud talk radio whenever her grandchildren and son are in her house. Because she is distracted, she doesn't register well-known signs of abuse, or if she registers them, she doesn't spend the mental energy to investigate further. The details could be filled out in different ways, of course, but as described so far this is easily conceivable as a kind of neglect, and a kind of culpable negligence in not pursuing the evidence of something so important. If asked by a friend whether her son abuses his children, she might sincerely answer that he does not. The question before us is whether there is any difference inherent in the kind of process at work that could make the distraction process and consequences something for which the agent is blameworthy and the self-deception process and consequences something for which she is not. The process in each case might be opaque to the grandmother, the evidence available the same, and her general reasoning abilities identical. Still, one might be tempted to think that the process matters because in the case of self-deception, the motivating desire operates in such a powerful way that it is "irresistible", and so exculpatory. It may be that the desire is strong in many cases, but, again, I think we rarely have evidence that

such desires are irresistible. Further, we lack evidence that it is any stronger than the power of distraction, or of any other potential causes such as intellectual laziness. It is true that the presence of the desire means that there is something at stake for the mother; but there might be something at stake even if the desire does not play a *causal* role. If anything, particularly if she recognizes her own desire, she may have extra reason to be on guard against such a self-deceptive process. (She might also know about herself that she is easily distracted in important situations and so also have extra reason to be on guard here, too, of course.) The upshot thus far is that there is nothing that we know about self-deception that would seem *essentially* excusing or mitigating relative to other kinds of erroneous belief formation against the evidence in cases of significant stakes. In fact, in some cases, self-deception might be more blameworthy than in other such cases.

4. Conclusion

The question of whether and when self-deceivers are responsible for their self-deception and its consequences brings together two independent and important, albeit controversial issues, namely, the nature of self-deception and the conditions under which we are responsible. If the motivationist picture is correct, then it brings the more specific, but no less controversial issue of the conditions for culpable negligence into play, as well. In this paper, I have briefly argued for a particular motivationist account, the Desire to Believe account, and then shown how it forms part of a plausible view of when self-deceivers are responsible and preserves the intuitive idea that in at least some high stakes cases, self-deceivers are responsible for their deception and its consequences. I have also argued that the Desire to Believe account, along with motivationism more generally, offers mutual support to one particular way of developing the powerful idea that we are responsible when we are reasons-responsive agents.

REFERENCES

Alexander, L. and Ferzan, K., with Morse, S. (2009). *Crime and Culpability: A Theory of the Criminal Law*. Cambridge: Cambridge University Press.

Audi, R. (1997). Self-Deception vs. Self-Caused Deception: A comment on Professor Mele. [Open Peer Commentary on Mele 1997] *Behavioral and Brain Sciences*, *20*(1), 104.

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

Bermudez, J. (1997). Defending Intentionalist Accounts of Self-Deception. [Open Peer Commentary on Mele 1997] *Behavioral and Brain Sciences*, *20*(1), 107–108.

Clifford, W. (1877/2008). The *Ethics of Belief*. [reprinted in L. Pojman & M. Rea (Eds.) (2008) *Philosophy of Religion: An Anthology*. Boston: Wadsworth]

Davidson, D. (1986). "Deception and Division." In J. Elster (Ed.), *The Multiple Self*. Cambridge: Cambridge University Press, 79–92.

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

Deweese-Boyd, I. (2010). *Self-Deception*. The Stanford Encyclopedia of Philosophy (Fall 2010 Edition). <http://plato.stanford.edu/archives/fall2010/entries/self-deception/>.

Fernandez, J. (2011). Self-Deception and Self-Knowledge. *Philosophical Studies*.

Fischer, J.M., & Ravizza, M. (2000). Responsibility and Control: A Theory of Moral Responsibility. Cambridge: Cambridge University Press.

Fingarette, H. (1969). *Self-Deception*. Berkeley: University of California Press; reprinted, 2000.

Fitzpatrick, W. (2008). Moral Responsibiilty and Normative Ignorance: Answering a New Skeptical Challenge. *Ethics*, *118*(4), 589–614.

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

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

Jenni, K. (2003). Vices of Inattention. *Journal of Applied* Philosophy, *20*(3), 279–295.

Johnston, M. (1988). Self-Deception and the Nature of 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.

Levy, N. (2004). Self-Deception and Moral Responsibility. *Ratio* (new series), *17*(3), 294–311.

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

Mele, A.R. (1997). "Real Self-Deception" and "Author's Response". *Behavioral and Brain Sciences*, *20*(1), 91–102, 127–136.

Mele, A.R. (1999). Twisted Self-Deception. *Philosophical Psychology* 12: 117–137.

Mele, A.R. (2000). Self-Deception and Emotion. *Consciousness and Emotion, 1*(1), 115-137.

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*. Oxford: Oxford University Press, 260–276.

Moore, M. & Hurd, H. (2011). Punishing the Awkward, the Stupid, the Weak, and the Selfish: The Culpability of Negligence. *Criminal Law and Philosophy*, *5*(2), 147–198.

Nelkin, D. (2002). Self-Deception, Motivation, and the Desire to Believe. *Pacific Philosophical Quarterly*, *83 (4)*, 384–406.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. Oxford: Oxford University Press.

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

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

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

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

Sharpsteen, D. & Kirkpatrick, L. (1997). Romantic Jealousy and Adult Romantic Attachment. *Journal of Personality and Social Psychology*, *72* (3), 627–640.

Sher, G. (2009). Who *Knew? Responsibility Without Awareness*. Oxford: Oxford University Press.

Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

What Does the Self-Deceiver Want? *

Patrizia Pedrini †
patpedrini@gmail.com

ABSTRACT

According to a recent theory of the motivational content of self-deception (Funkhouser, 2005), the self-deceiver wants to be in a state of mind of belief that *p*, upon which her want that *p* be true would be merely contingent. While I agree with Funkhouser that the self-deceiver is considerably moved by an interest in believing that *p*, which makes it possible for her to relate to reality in a highly prejudiced way, I will argue that it is unlikely that the self-deceiver's primary want to believe, or interest in believing that *p* occurs as the result of a merely contingent interest in *p* being true. I will finally assess various consequences of the view I favor, regarding the self-deceiver's avoidance behaviour, "twisted" self-deception, and whether we should provide a unifying account of "straight" and "twisted" self-deception.

* I am most grateful to Eric Funkhouser and Dana K. Nelkin for crucial discussions on an earlier version of this paper, and for their pertinent and generous comments; to Akeel Bilgrami and Achille Varzi for many exceptionally stimulating conversations, during my two years at Columbia University in New York City, about my interest in self-deception, among other topics, and the route I was even then showing an inclination to take; to Alessandro Pagnini for having welcomed and encouraged over the years my interest in self-deception, and to Tito Magri for agreeing to publish my second book in Italian on this topic; to the audience in Milan, at the 2011 European Conference of Analytic Philosophy, where a shorter version of this paper was accepted and presented; to the incredibly good audience at the University of Rijeka Department of Philosophy, where I served two semesters as fixed-term professor and, in November 2011, gave a senior lecture that allowed me to discuss in depth one of the last versions of the paper. I also wish to thank the Italian Mensa Society formally for offering me one-year funding to work on what will hopefully be my first book in English on the topic of self-deceptive motivation.

† Senior Research Fellow in Philosophy funded by the Mensa Society; Fixed-Term Professor, Dept of Philosophy, University of Rijeka; Assistant Fellow, College of Letters and Philosophy, University of Modena & Reggio-Emilia, and Dept. of Philosophy, University of Florence.

What does the self-deceiver want? Does she want to reach a state of mind, that is, the belief that *p*, which she likes, or wants, to believe, or does she also want reality to be exactly as she wants it to be, that is, *p* to be *true*? In other words, what is the operative desire that leads her to self-deception? Does she have a *self-focused desire* to be in the state of mind of belief that *p*, regardless of whether *p* is true or false, or does she have a *world-focused desire* that *p* be true? According to a recent theory of the motivational content of self-deception (Funkhouser, 2005), the self-deceiver wants to be in a state of mind of belief that *p*, which she finds pleasant or anyway "interesting", or "important" for her to be in, while not necessarily also being focused on the task of dispassionately ascertaining how things actually stand in the world. On this account, the self-deceiver is thought to have a self-focused desire (Funkhouser, 2005, p. 296) to be in a certain state of mind of belief that *p*, or a "desire to believe" (Nelkin, 2002) that *p*, and this would be the leading motivation for self-deception. She can contingently have a world-focused desire (Nelkin, 2002, p. 296) that the world be such that *p* be true, but such desire is not intrinsic to the self-focused motivation for self-deception and in fact is sometimes lacking. The contingency of her world-focused desires that *p* be true are demonstrated by her "avoidance behavior", according to Funkhouser: typically, the self-deceiver actively avoids evidence suggesting that *p* is false (2005, pp. 297–298). Were she to be dispassionately interested in the truth-value of *p* and in believing what is true, she would not avoid such evidence. Furthermore, according to Funkhouser, the self-focused desire account of the self-deceptive motivation has the advantage, as we will see, of unifying two kinds of self-deception that the traditional world-focused desire accounts cannot explain: "straight" and "twisted self-deception". It also has the consequence of helping us to single out a phenomenon that Funkhouser dubs "apathetic" or "indifferent" self-deception (2005, p. 298), which is described as a kind of self-deception where there is *no* world-focused desire of a contingent kind either, but just a self-focused "desire to believe" that *p*.

In what follows, I will agree with Funkhouser that the self-deceiver is very much attracted by the qualitative aspects of believing that *p*, and that she may be considerably moved by an "interest" in believing that *p,* which makes it possible for her to relate to reality in a highly prejudiced way, but, contrary to Funkhouser, I will argue that it does not seem likely that the self-deceiver's want to believe, or interest in believing that *p* occurs as the result of a merely

contingent interest in *p* being true. On closer inspection, it transpires that the self-deceiver finds it pleasant, or interesting, or important, as the case may be (as we will see, depending on the kind of self-deception she embarks on), to believe that *p* and is attracted to it *exactly because* she wants *p* to be true, and so she does not embrace the self-deceptive belief that *p* in a way which could in principle be considered as independent of the interest in being *p* true, which I take to be *fundamental* and not merely contingent. Considering the self-deceiver's interest in *p* being true as merely contingent significantly underdescribes the psychological complexity of the motivation that triggers self-deception; furthermore, it leads to counterintuitive conclusions on the very nature of self-deception, which could be easily confused with scenarios of false beliefs "artificially self-induced", where, however, we lose the grip on important specificities of the phenomenon of self-deception and its motivation that are *just* linked to the self-deceiver's world-focused desire, or want, more generally, that *p* be true.

At the same time, there is an alternative account of the self-deceiver's avoidance behaviour that does not force us to conclude that she is merely contingently focused on wanting reality to be such as *p*, but rather mainly focused of wanting to believe that *p*. Typically, the self-deceiver actively looks for evidence suggestive of the truth of *p*, while avoiding the evidence suggesting that *p* may be false. Also, she generally gives some treatment of the sources of evidence that *p* may be false (typically, she does so in a motivationally biased way), as opposed to just avoiding the evidence against *p* altogether. She typically explains to herself why such sources of evidence are not worth attending, and this epistemic work is the symptom that she is interested in the way the world is and in the truth-value of what she believes. All avoidance behavior shows, as we will see, is that the interest in *p* being true is strongly biased by the interest in believing that *p*, but it does not also show that the interest in *p* being true is not the fundamental engine of the very desire, or want, to believe that *p*. For these reasons, as I will explain, one could even doubt that "apathetic" or "indifferent" self-deception exists at all.

Finally, I will critically assess the prospects of the unified account of "straight" and "twisted" self-deception reached by defenders of the self-focused desire account and, in this connection, I will say a few words about why some people think that cases of twisted self-deception offer the strongest support to the "desire-to-believe" account of the self-deceptive motivation. I think the question of whether twisted self-deception really offers this support

is, at best, still unsettled, and I will try to suggest considerations that might be useful for a hopefully more decisive future defence of a world-focused approach to the motivation for self-deception.

1. Can the self-deceiver's world-focused desire, or interest, that *p* be true be *merely contingent* upon her self-focused desire to believe, or interest in believing, that *p*?

It is a widespread intuition that when we believe some proposition *p*, we also, intrinsically, believe that *p* is true. Believing that *p*, on this view, is just taking *p* as true; conversely, if we take *p* to be true, we, by the same token, believe that *p*. That is, believing is said to be truth-oriented and the aim of belief seems to many to be that of representing the world as the believer takes it to be. Thus, one who submits to such a view on what believing is, might easily conclude that if one is motivated to believe that *p*, one must be also intrinsically motivated to take *p* as true, and that the motivation to believe that *p* is intelligible in the light of the motivation to take *p* as true. As far as I can see, however, there are cases where the two kinds of motivation can part company. The most obvious case is the case of the motivational set of someone who wants to acquire an "artificially self-induced" false belief or other mental states. Under the heading of "artificially self-induced" false belief or other mental states, more generally, I do not necessarily refer to Matrix-scenarios such as the voluntary implantation of a belief or a mental state *via* the use of futuristic computerised machines. I also include more ordinary cases of voluntary acquisition of mental states *via* everyday, do-it-yourself "techniques". These techniques may generally include drugs, alcohol, and other addictive substances, for instance. We are all familiar with real or fictional subjects who strongly wish to cut themselves off from reality, at least for a while: they may just want temporarily to forget how reality is and what they believe it is like; or they may want to acquire a joyful mood to replace their beliefs and/or other unpleasant states of mind; or they may just want to experience what it is like to become convinced, by means of the stimulating effects of a substance on their cognition and memory, that a certain belief that *p* is true. It may be an issue how exactly those psychological mechanisms actually work, and what exactly the causal chains, initiated by the substance that lead us to acquire a belief that we ultimately want, must be. Also, it is far from clear that using substances can always lead us to believe exactly what we would like to believe. It strikes many people that often it does not, but

we can agree that sometimes it can, and I suggest, for the sake of argument, focusing on the very case in which this outcome is successfully achieved. If this artificial achievement is possible at all, however causally complex and descriptively unclear it may be, we can say that people who successfully reach their doxastic goals via any of these causal chains achieve *artificially* what they could not achieve *epistemically*: since they cannot become convinced, before they enter the self-induced artificial causal chain, that *p* is true, given the evidence at their disposal, they try to reach that conviction by artificially modifying their perception, or cognition, of the available evidence, forgetting it altogether, altering their reasoning, and so on. Be that as it may, the crucial point of the illustration is that the motivational set of these subjects seems to demonstrate that they have a self-focused desire to acquire, or interest in acquiring, a certain state of mind, including beliefs, while lacking the world-focused desire, or interest, that those states of mind, particularly the beliefs they successfully reach, be representative of the world. They want to enter for a while an inauthentic representation of reality where it seems to them that reality is different from what it actually is, and if the outcome of the artificial modification of their cognition successfully matches their interests, they end up believing that reality is exactly as they want it to be. Note that some may presumably do so because they would like reality to be different, but such world-focused interest does not seem to be intrinsic to the motivation to acquire the artificially self-induced belief. For not only does it seem reasonable to say that substance users, who are motivated to acquire the belief that *p* because they would like *p* to be true in some possible, fictional world, are not also necessarily motivated to *establish* that *p* is true in the real world, as they just seem to want to *believe* that *p* simply for the sake of the pleasure, where appropriate, or, more generally, satisfaction (which, as we will see, may even be accompanied by discomfort) they get in so believing, regardless of the truth-value in the real world of the proposition they come to believe; also, we could easily imagine subjects who experiment with a substance just because they want to acquire a certain belief that *p* which it is important for them to acquire, without having any contingent interest in *p* being true, or any sense that *p* being true really matters to them. That is, one may want to feel what it is like to believe that *p* for a while, because it is important to believe so, without having any particular preference regarding *p* being true at all in the real world. As an example, assuming that there is a technique for coming to believe that one is a brilliant mathematician, it may be pleasant, or important, or interesting to

believe for a while that one is a brilliant mathematician, and thus one could be motivated by the qualitative aspects of consciously believing that one is a brilliant mathematician to acquire that belief, without having any interest at all in actually being a brilliant mathematician in the real world.

Now that we have a slightly clearer description of the motivational set of a subject who has a self-focused desire to believe that *p*, or more generally an interest in believing that *p*, without intrinsically desiring that *p* be true, or having an interest in *p* being true, let us ask whether the motivational set of the self-deceiver is different in any relevant respect from that of such a subject. Let us consider the case of Nicole, described by Funkhouser (2005, p. 302). Nicole is Tony's wife. She sincerely believes that she is convinced that her husband is not having an affair with her best friend, Rachel, despite having excellent evidence of the affair being at least likely. For example, Nicole's friends say that Tony's car is parked in front of Rachel's house at times when he had told Nicole he was going out with his friends; also, Tony has a significantly diminished sexual interest in Nicole; and so on. The way she reaches this conviction is instructive: it is not simply that she misinterprets the evidence that this affair may at least be likely by looking for stories to explain why such evidence should not count as conclusive; she also carefully engages in avoidance behavior, such as keeping away from Rachel's house at times when Tony says he is out with his friends, even at the cost of changing the route she would otherwise have taken. Funkhouser is in part arguing that Nicole's avoidance behavior shows that she "deep down" knows the truth, and does have the belief that her husband *is* having the affair, but I will not discuss the tenability of this here, although I have my doubts that we really need to postulate a "deep down" knowledge of how things stand in self-deception (I address this concern at length in another project and I shall briefly return to it in the next paragraph); nor am I interested in the claim that Nicole falsely believes that she believes that her husband is not having an affair, although she does not actually have the corresponding first-order belief. Rather, what I am interested in here is Funkhouser's claim that what the self-deceiver wants, or is interested in, is primarily a state of mind of belief, and that her wanting that *p* be true can be treated as contingent upon the self-deceiver's "desire-to-believe" want. To establish this conclusion, Funkhouser seems to be drawing here a (largely undeclared) inference to the best explanation about the motivation for avoidance behaviour: since she avoids reality, the best explanation for her avoidance of reality is that she doesn't *want* to know the

truth, that she *wants* to protect herself and her favoured opinion from the impact of reality, and so she primarily wants to believe that *p*. That is, in order to explain her avoidance behaviour, Funkhouser attributes to Nicole the motivation provided by a self-focused desire to believe that *p* as primary, while the motivation provided by the world-focused desire that *p* be true would be, at best, contingent upon such primary motivation. An explanation of the precise nature of such contingency is not completely spelled out in the account offered by Funkhouser. Perhaps, the contingent desire that *p* be true is a consequence of the primary "desire-to-believe", in that the self-deceiver must have to do with the world anyway, in order to believe that *p*; or it may be the case that the desire that *p* be true happens to be contingently present because of other psychological coincidences as yet unspecified. I will not develop an analysis of his view on this issue, as it would lead me too far from the major purposes of this article. What I will do, instead, is to show why I think that self-deceivers cannot be moved by the want that *p* be true in a merely contingent way, even if it is apparent that they very much engage in avoidance behavior. After I will have done that, I will try to make a case for an alternative, positive explanation of avoidance behavior and its motivation, that will be coherent with the view I recommend. Before I begin, I would highlight the fact that, in what follows, I am deliberately going to shift away from the "*desire*-centred" terminology used by Funkhouser and Nelkin to defend their views on the self-deceptive motivation (which terminology I have already tried to expand in passing in the first part of the paper) towards a more neutral "*want (or interest)*-centred" language. The reason why I prefer this terminological expansion will be made clear in the last portion of the paper, where I will be offering a number of considerations on "twisted" self-deception, where one believes that *p* even if one does not desire that *p* be true, and on the alleged support that it gives the self-focused desire account of the self-deceptive motivation. I will say something at that point about how I think the motivation for twisted self-deception should be analyzed and how the analysis I suggest about twisted self-deception fits my general view about the motivational content of all kinds of self-deception.

Let me start with the case for claiming that it seems unlikely that the self-deceiver's want that *p* be true could be merely contingent upon her want to believe that *p*. One of the most distinctive features of self-deception is a more or less demanding *epistemic* work on the evidence self-deceivers have or might find, which I believe is what ultimately distinguishes self-deception from other

forms of motivated irrationality. In other forms of motivated irrationality, such as "precipitate cases" of believing (Scott-Kakures, 2002, p. 587), phenomena of jumping to conclusions under the influence of strong emotions[2], wishful thinking, etc., the subject involved does not normally spend time and energy on elaborating "covering stories" to justify to herself the opinion that *p* that she favours and does not typically struggle epistemically with evidence against *p* to arrive at an explanation of why it should not count as undermining *p*. On the contrary, self-deceivers are the champions of a pressing inner dialogue setting out to assess the strength of the evidence undermining *p* and balance it with the strength of the evidence suggesting that *p* may be true. Often they also need to share their "findings" on the strength of the evidence with others. We are all familiar with friends with documented unhappy relationships who call us on the phone at night or write us suspect letters to tell us and *explain to us why* they are really happy, *why* they are not in the position to believe that their partners are unfaithful, *why* they believe their love affair is not really over, *why* they think they have new evidence that that man or woman does like them, and so on. Most of us may, perhaps, have occasionally made such phone calls or written such letters and have experienced the inner epistemic negotiations that have encouraged us to declare our dubious conclusions to our friends. In all, self-deceivers have quite complex "convincing" stories, elaborated by means of what is generally an intense epistemic work on why they believe what they do and declare – complex stories that are lacking in other forms of motivated irrationalities and fundamentally aimed at explaining why *p* is true.

Now, if this epistemic work is, as I think, one of the most prominent specificities of self-deception, its fascinating and disconcerting hallmark, we may well ask if a purely self-focused want to believe that *p* can satisfactorily provide an explanation for the motivation that triggers such an epistemic endeavour. Certainly, the self-focused want to believe that *p* can sometimes successfully account for cases of artificially self-induced false beliefs or other mental states (and, with qualifications, also for other forms of motivated irrationality), but it seems fair to say that the motivation for the epistemic work that typically underpins the doxastic end-state of self-deception cannot but be a world-focused want that *p* be true. We need not deny that the self-deceiver is attracted to the sense of importance she attaches to believing that *p*, which she

[2] See Lazar, 1999, p. 281.

may find pleasant, or else unpleasant but important in the light of certain other wants and convictions, but her epistemic work to establish that *p* is true seems to show that the importance she places on believing that *p* is due to the importance she places on knowing that reality is such that *p* is true. In other words, the world-focused want that *p* be true is not detachable from, and contingent upon, the self-focused want to believe that *p*; rather, it seems to be intrinsic to it. If it were just contingent, we would be in need of an explanation why self-deceivers try to justify their convictions epistemically. This tells us that the self-deceiver's motivational set cannot, and should not, be confused with the motivational set of someone who may even be indifferent to the way things stand and just want to acquire a state of mind for the sake of the importance to her of being in that state; nor can, or should, it be confused with the motivational set of someone who embraces a conclusion on the heat of the moment. Such confusions make us lose our grip on the nature of self-deception and seem to underdescribe the complexity of the motivation that prompts it.[3]

To complete the argument, we now need to try to explain why, then, self-deceivers engage in avoidance behaviour at all and how we should qualify their relationship with reality, if they are interested (as they seem) in establishing the

[3] One might point out that an epistemic work could be compatible with the "desire-to-believe" account. That is, even if the subject is moved by a desire to believe, she may still need to stay focused on the world to secure her doxastic conclusion. My initial sense about this objection is that being focused on the world for the "instrumental" reason of securing a doxastic conclusion that one wants to secure is significantly different from wanting to secure a doxastic conclusion because one wants the world to be as one would like it to be. The crucial difference that I see lies fundamentally in the relationship one has with one's beliefs. If one by default relates first-personally to one's beliefs as states that are representative of the world, and does not try to manipulate those states independently of their representational goal, having with them a third-personal relationship, then, when it comes to self-deceiving, one would tend to establish the truth value of what one believes, and this is the epistemic work that I see the self-deceivers to be doing, as opposed to someone who artificially self-induces mental states. Furthermore, the whole point of the discussion is not whether someone who has a primary desire to believe can also be world-focused to secure that doxastic result – even the artificially self-inducing believer is focused; perhaps she sets out not to take another substance that can work as an antidote, within a certain time-span, and in that sense she clearly "keeps an eye" on the world. Rather, the whole discussion in hand ultimately rests on the question whether, *given what believing is*, people can really be wanting to believe that *p* while their wanting that *p* be true can be treated as merely contingent. And this is exactly why I designed cases of artificially self-inducing beliefs, as the extreme of a spectrum that helps us see that the self-deceiver more probably occupies a place close to the opposite extreme, given her intense, and presumably not merely instrumental, if she relates first-personally to her beliefs, epistemic work.

truth, or at least the credibility, of what they come to believe, but nonetheless avoid pieces of evidence that could lead them to ascertain how things actually stand.

2. Avoidance behaviour and the self-deceiver's relationship with truth and reality

Avoidance behaviour, surfacing either in self-deception or in other psychological predicaments, is undoubtedly a tricky phenomenon that may be difficult to elucidate completely, both phenomenologically and explanatorily. At the phenomenological level, it may not always be clear what is the intentional object of avoidance, if it need be consciously represented, and so on. At the explanatory level, questions arise as to *why* we engage in it at all, and what motivation there is for it. Just one thing seems to be conceptually intuitive and phenomenologically manifest: there is always something in what we avoid (an aspect of it, a thought or a feeling that it prompts, etc., however represented) that we fear or find upsetting, distressing or unacceptable. If this starting assumption is workable, no doubt those who engage in avoidance behavior ultimately do not want to get in touch with such sources of fear, upset, or distress. It is this very assumption that presumably inspired Funkhouser to conclude that self-deceivers "deep down" know the truth, but are somehow scared by it and so avoid it. On this basis, also, it is appealing to infer that *all* the self-deceivers want is to acquire the belief that *p*. I will argue in the remainder of this paragraph that while the starting assumption is correct, and it is thus likely that self-deceivers are scared by "something" that a dispassionate contact with reality may reveal, avoidance behavior does not show as yet that the leading motivation to self-deception is a "want-to-believe that *p*", upon which the want that *p* be true would be merely contingent; on the contrary, it shows that self-deceivers have *a strong interest in p being true*, as made manifest by their epistemic work, but this *interest in p being true is heavily biased by the very want that p be true*. In order to understand this, it is crucial to unpack important details of avoidance behavior, which may easily remain undisclosed.

Let us consider again the case of Nicole. Nicole, as we have seen, avoids Rachel's house at times when Tony says he is out with his friends. An impartial observer may say that she does that ultimately because she is scared by the idea that going there might disclose to her new evidence in support of the

hypothesis that Tony is having an affair with Rachel. If such a hypothesis were confirmed, her love dreams could not survive – nor, perhaps, her marriage. So, it is manifest that, at the very least, she is scared by the way reality *may* be, that is, by the *possibility* that a certain proposition *p* may be true. I am clearly working here with the hypothesis testing model of self-deception made famous by Alfred Mele (2001), where self-deceivers do not start their self-deception by already believing that not-*p* and trying to convince themselves that *p*; rather, the favoured hypothesis that *p* is raised by the corresponding desire that *p* and then tested in a biased way. On the contrary, Funkhouser assumes that the self-deceiver "deep down" knows that the favoured hypothesis that *p* is false, and this is an additional component of his account that encouraged him to conclude that self-deceivers are not primarily motivated by world-focused desire, or, more generally, want, as appropriate: if they were, they would not try to avoid what they already know. But even if Funkhouser were right to say that they start their self-deception by already knowing how things stand, many (e.g., Bermùdez, 2000) have argued that such knowledge could have been suitably undermined by biased epistemic work on it and brought back to the status of a hypothesis in need of a new test. So, Funkhouser would have to show that in all cases of self-deception the alleged "deep down" knowledge that not-*p* is never turned into the corresponding hypothesis that *p*. In this way only could he substantiate his subsequent claim that the intentional object of avoidance is the known truth, as opposed to just the possibility that a feared hypothesis may be true, but this is far from being proven by the examples he gives as they stand.

However, if the possibility that a *feared hypothesis* may be true is the intentional object of avoidance behavior, new light is shed on Nicole's motivation to engage in such avoidance, and the tenability of the general account Funkhouser promotes about the self-deceptive motivation is deeper in trouble. For on this alternative, positive account of Nicole's avoidance behavior, she would not be trying to avoid the truth she somehow knows as such because she is just interested in acquiring or maintaining the belief that *p*, but precisely *to establish that reality is the way she wants it to be*. In other words, she is so interested in establishing that p is true that she carefully avoids contact with sources of evidence suggesting that the favoured hypothesis that *p* may be false. In all, I acknowledge that a fatal bias affects the self-deceiver's relationship with her interest in truth and the way the world is, but I reject the claim that the self-deceiver is only contingently moved by a concern with how she takes the world to be. The bias affecting her relationship with reality is due

to the want that *p* be true, in turn presumably motivated, as we will see, by other convictions, values, character traits and so on, which would have to be uncovered and described case by case. If one pursues this line of reasoning, one is led to suppose that it is the self-deceiver's want that *p* be true that leads her to forge the belief that *p*, and so, also, possibly her want to believe that p, as opposed to thinking that the want to believe that p is what triggers the whole motivational process of self-deception. On this reading of the motivation for self-deception, avoidance behaviour is acknowledged as one of the major symptoms of the fact that the self-deceiver wants the world to be such that *p* be true, perhaps because she fears the possibility that the world be such that *p* be false, which would be distressing, or discomforting, or upsetting to her. That is, avoidance behavior would be no decisive evidence for the "want-to-believe" account.

If my claim is correct, I am, however, left with the task of explaining how my favoured view could accommodate the alleged cases of "apathetic" or "indifferent" self-deception introduced by Funkhouser, and also how it deals with cases of twisted self-deception. Examples of apathetic or indifferent self-deception, according to Funkhouser, would be cases of beliefs typically acquired upon peer pressure: some people may want to believe what their peers believe without having any preference whatsoever about the truth of those beliefs. Here again, I believe such cases would need to be fully unpacked before issuing claims as to their nature. To begin with, even if we can agree that those who self-deceive upon peer pressure are attracted to the importance they attach to belonging to a group and thus sharing opinions on sensitive matters with their peers, if an epistemic work is performed by the self-deceiver to justify what she comes to believe, then we have a clue that her self-deception is not "indifferent". Secondly, the reasons why she wishes to share those opinions with her peers should be more clearly analysed. For it may be that she delegates to peers the authority to entertain true opinions on sensitive matters. On this hypothesis, she would wish to share her opinions because she takes those opinions to be true, even if she does not embark on the epistemic work to establish that *p* is true. Once again, the self-deception would not be "indifferent" at all. Pending further analysis, cases of apathetic self-deception should not be treated as clear cases in which the self-deceiver is primarily led to the self-deceptive belief that *p* by being moved by a mere want to believe that *p*, and with the manifest absence of any want that *p* be true.

3. On the significance and the prospects of unifying the leading motivation for "straight" and "twisted" self-deception

Finally, let me devote a few words to Funkhouser's attempt at providing a unifying account of the motivation prompting two varieties of self-deception, the so called "straight" and "twisted" self-deception. According to Funkhouser, world-focused desire accounts of self-deception cannot provide a unified explanation of why some self-deceivers end up falsely believing that *p* while they want *p* to be the case ("straight" self-deception, e.g., Nicole self-deceptively believes that her husband is not having an extramarital affair while she wants him not to be having one), and others end up falsely believing that *p* while they do not want *p* to be the case ("twisted" self-deception, e.g., John self-deceptively believes that his wife is having an extramarital affair although he does not want her to be having one). That is, on the world-focused desire accounts, straight and twisted self-deception would be accounted for by two different sorts of motivations: a desire that *p* be the case would motivate straight self-deception, while a hostility toward *p* being the case would, *mysteriously*, motivate twisted self-deception. Funkhouser's conviction is that self-focused desire accounts have the advantage of offering a unified treatment of the motivation prompting both varieties of self-deception, while explaining away the mystery affecting the motivational drive to twisted self-deception: both would be triggered by a desire to acquire a belief that *p*, and not by two different sorts of motivation, namely, a desire that *p* for straight self-deception, and a fear, or dislike, or repugnance that p for twisted self-deception. Also, besides the advantage of achieving explanatory unification across different varieties of self-deception, twisted self-deception is used in this line of reasoning as the crucial case that seems to lend the best support to the "desire-to-believe" account, as it is the kind of self-deception in which one seems to best appreciate how a desire to believe that *p* can move someone to believe that *p*, any desire that *p* be true clearly being absent, as the twisted self-deceiver does not desire that *p* be true at all. In the space available, I will just briefly set out two main clusters of considerations, largely incomplete, to provide a general outline of a line of research on this issue that I would like to develop fully elsewhere.

First, even before assessing whether twisted self-deception can be really moved by a primary want to believe that *p* (upon which any sort of world-focused want that *p* be true would be merely contingent, if not absent), it is

worth asking what explanatory advantage is gained by the explanatory unification of the two varieties of self-deception. Suppose, for the sake of argument, that the self-focused want account is correct and all sorts of self-deceivers primarily want to acquire a certain false belief. At this point, a crucial question regarding the self-deceivers of both sorts needs to be answered: why is it that straight self-deceivers want to acquire a belief that they like, while twisted self-deceivers want to acquire a belief that they dislike? I take this to be a perfectly legitimate question that all accounts of the motivation triggering self-deception should answer. A clue to the answer to this question seems to lie in the self-deceiver's relationship with reality, a relationship shaped by her values, desires, fears, other beliefs, and so on. Perhaps, the straight self-deceiver Nicole may like the belief that Tony is not having an affair because she conceives of a good marriage as based on fidelity and forcibly wants her private world to achieve this ideal; perhaps, the twisted self-deceiver John may dislike the belief that his wife is having an affair because he conceives of a good marriage as based on fidelity, but he also has a paranoid conviction that many marriages do not achieve this ideal, and so wants to test whether his is among them. Perhaps he also wants to prove to himself that his paranoid conviction is right, maybe because having a confirmation of his convictions will help him to reduce anxiety, in ways as yet unspecified. Note at this point that the answers to this question on both forms of self-deception cannot but be disjoint: different values and personalities, a different relationship with reality, as well as different fundamental wants shape the two forms of self-deception. That is, if one presses questions upon why either sort of self-deceivers is moved to self-deceive in the specific way they do, the analysis initially provided by the defender of the desire-to-believe account needs to "go deeper", in search of the deeper wants that move and shape the specific variety of self-deception in question, which a "want-to-believe" account does not seem to trace. In other words, the explanatory unification seems to be achieved at the expense of an in-depth grasp of the individual's specific motivation to achieve a particular kind of self-deception.

The stage is perhaps now better set to turn to the second cluster of thoughts I would like to promote. I would like now to consider briefly the very nature of motivation that drives twisted self-deception as such, independently of the prospects of the explanatory unification I have discussed in brief. My intention was to foreshadow my fundamental instinct towards the issue of what may move someone to self-deceive in believing what she fears, or finds

upsetting, or anyway does not like to be so, earlier in the paper by deliberately shifting, on as many occasions as I could, away from the "desire-centred" analysis of the various self-deceptive motivations, including twisted self-deception, and the related terminology, towards a broader "want-centred" analysis of them. When I think of cases of twisted self-deception, and when I think of cases of straight self-deception as well, two general features of the motivational set of a self-deceiver of any sort strike me as central: 1) it is not clear that a desire for something is invariably the most crucial, or deepest, motivational drive for a subject, in general, both in the practical domain, where we typically analyse motivation for action, decisions and so on, and in the theoretical domain, where we can sometimes, as in the case of self-deception, trace motivation to a certain reasoning, direction of cognition, and so on; 2) even when a desire presents itself as the motivational drive for an action or a train of thought etc., it seems to me that the desire-driven motivation is not necessarily the whole motivation story that one could in any case tell to explain the case in point. The analysis of a motivational set can very often, though perhaps not always, "go deeper", and it should do so, if appropriate. I believe that many, even if presumably not all, cases of both straight and twisted self-deception hold out material for deeper analysis, which may prove to be instructive as to the tenability of any "desire-to-believe" account of the self-deception motivation, and more generally, of any "want-to-believe" account. It seems to me likely that deeper drives, further motivations, typically shaped by ground values and more or less hidden convictions of various kinds that a subject has, forge the "surface-motivation" for self-deception, which can be either a desire, or a fear, or any other motivational state one may find appropriate to attribute to a subject in a specific case. "Deeper down", however, there seems to be a much more complex psychological world to explore. As far as I can see, it is this exploration only, subtle and demanding as it may be, that can allow us to hope for a chance of grasping the specific motivation that ultimately moves someone to engage in a specific sort of self-deception. General human drives may be identified in all sorts of self-deception, of course, but while many think that one of the most ever-present motivational drives attributable to a subject who self-deceives is a desire for something, I have the feeling that a deep fear might instead better explain what the subject wants and seeks by self-deceiving – fear of the psychological pain that certain possible states of affairs may cause. If one believe that there is no renouncing the desire-centred analysis, it is perhaps more tempting to look for

an overarching desire – in the account in question, a "desire to believe" – which is then thought to be capable of unifying both straight and twisted self-deception, which unquestionably seem to be moved, but perhaps only superficially, by a desire and a fear respectively.

So, to sum up the general thoughts that guide my research on the issue:

a) The deeper motivational drive for self-deception may well be a fear, instead of a desire – as I said, fear of the psychological pain that certain possible states of affairs may cause, to be coupled with the psychological specificities of the subject involved, her other beliefs and values, her other wants, which only can explain why her self-deception was triggered and what she ultimately want as a person, more than simply as a self-deceiver. Self-deception is an extraordinary window on the psychological structure of an individual subject, and I believe that this explanatory richness should not be lost for the sake of any unification, still less for a unification in the name of desires[4].

b) Once one takes this route, and does not look for a desire only (overarching or otherwise) to explain self-deception in general, but rather looks for deeper wants, case by case, twisted self-deception, in which desires are not (at least superficially) prominent, takes on a new light and seems to have a chance of being accommodated, in ways as yet unspecified, in a "want-centred" account.

c) It also seems easy to accommodate in a "*world-focused* want" account. Twisted self-deception has been considered as the variety of self-deception that lends best support to the "desire-to-believe" account, given that there is no manifest world-focused desire that *p*, but rather a fear that *p*, so if a desire is thought to be motivationally necessary to move self-deception, and no world-focused desire is present in twisted self-deception, then having a "self-focused desire-to-believe account" at hand may seem helpful. But if twisted self-deception is shown to be

[4] There might be a worry of regress, here, about the ultimate motivational source. It may be said that the fundamental fear itself may be due to a desire to feel pleasure. The issue is intriguing, and to deal with it satisfactorily would take me too far from the present purposes. For the time being, I just remark that the drive towards pleasure, and the desires that spring from it, and the fear of feeling pain, and the more specific fears it causes, might well be the two sides of one and the same coin, and so one depends upon the other; yet, what it is most salient, and causally primary, in one specific case of self-deception as opposed to another, may still be one side only of the coin.

driven by other wants, perhaps driven in turn by the want not to feel pain caused by specific contact with a specific reality, thematically sensitive for an individual, then twisted self-deception begins to appear less problematic for a "world-focused want" account than might initially have been thought.

d) Finally, if straight self-deception is equally deeply driven not by superficial wants, but rather by wants that go deeper than the surface desire that *p* be true, maybe an explanatory unification can still be achieved, although on different grounds.

I hope that I have, in the space available, at least established the general theoretical background for a future project I wish to pursue about the motivation for self-deception, and that I have sufficiently clearly set out my reasons for exploring views towards which my philosophical instinct tends to lead me.

REFERENCES

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

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

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

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

Nelkin, D.K. (2002). Self-Deception, Motivation, and the Desire to Believe. *Pacific Philosophical Quarterly*, *83*(4), 384–406.

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

Narrative and Self-Deception in *La Symphonie Pastorale*

Julie Kirsch *
kirschj@dyc.edu

ABSTRACT

Is it possible to develop a personal narrative that is not fictitious or self-deceptive? In this essay, I will look at the way that personal narratives contribute to self-deception. In so doing, I will consider the narrative that the narrator or pastor of André Gide's Pastoral Symphonie develops while reflecting upon his romantic relationship with his blind adopted "daughter", Gertrude. Although the pastor's narrative is largely self-deceptive, we need not fear that all narratives are equally delusional. When a narrative is not self-deceptive, it can make a positive contribution to self-knowledge and moral understanding.

1. Introduction

To what extent are our personal narratives works of fiction? What processes contribute to personal narratives that are largely fictitious or self-deceptive? And is it possible to develop a narrative that is *not* fictitious or self-deceptive – one that makes a positive contribution to self-knowledge and moral understanding? I will begin this paper by looking at the way that personal narratives contribute to self-deception (Sec. 2). In so doing, I will consider a fictional narrative that provides us with a vivid illustration of how this can happen. The narrative that I will discuss is presented by the narrator, or *pastor*, in André Gide's *La Symphonie Pastorale* (1955) (Sec. 3). I will argue that narrative need not always lead us astray in epistemic and moral matters. As my discussion of *La Symphonie Pastorale* will show, we can distinguish in a principled way between self-deceptive and non-self-deceptive narratives.

* D'Youville College, Buffalo, New York.

When a narrative is not self-deceptive, it can make a significant contribution to self-knowledge and moral understanding (Sec. 4).

2. Narrative And Self-Deception

In recent years, philosophers have shown considerable interest in issues involving narrative. They have examined the nature and function of narrative (Currie 2010). They have asked questions about the role that it does, or should, play in our theorizing about particular moral issues.[1] And they have invoked narrative in debates about rationality, action, and personal identity.[2] But philosophers have said remarkably little about the contribution that narrative might make to self-deception and attempts at self-knowledge. This paper is largely an attempt to fill this gap in our theorizing about narrative.

As Daniel Hutto has pointed out, there is little agreement about what a narrative is. For this reason, it is unlikely that any set of necessary and sufficient conditions for narrative will satisfy all theorists (Hutto, 2007, p. 1). Fortunately, for the purposes of this paper, we can bypass these difficulties and understand narrative broadly. A narrative, as I will understand it here, is an oral or written interpretation of a series of events that is presented in sequential order. Narratives do not just report or list events, they interpret them. As an interpretation, a narrative attempts to provide meaning, purpose, or closure to the events in question. Narratives are constructed from a perspective and are in principle incomplete and selective in what they represent. Although there are weaker and stronger ways of understanding narrative, this account captures the core ideas that are found in most others. Moreover, it highlights the qualities or properties of a narrative that are especially useful when thinking about the etiology of self-deception.

Before we can understand the contribution that narrative makes to self-deception, we need to know what self-deception is. Self-deception in my view, and most views, involves holding false beliefs. As Alfred Mele puts it, this is a "lexical" criterion for self-deception (Mele, 2001, p. 51). By definition, a

[1] For a discussion of the role that narrative should play in our theorizing about ethics, see Misak 2008; hereafter abbreviated ENED. See also Misak 2005; hereafter abbreviated "ICU." In "ICU" Misak presents a narrative involving her experience with ICU psychosis. On the basis of this narrative, she argues that medical paternalism is appropriate in a certain limited range of cases.

[2] For a theory of personal identity understood in terms of narrative, see Schechtman 1996. David Velleman develops an account of reasons and agency based in part upon narrative in Velleman 2006.

person who is self-deceived holds at least one false belief. A second widely accepted condition for self-deception involves motivation. A self-deceiver must be motivated to hold the false belief in question: her motivation to believe that *p* makes a causal contribution to her falsely believing that *p*. A self-deceiver holds the false belief in question because she is motivated to do so. If it were not for this motivation, we would expect her to see the world more clearly.

In most, but not all, cases of self-deception, people are motivated to accept positive or flattering views about themselves and their loved ones. We are all too familiar with such cases: An aspiring young writer may be self-deceived about the profundity of her thoughts. A self-absorbed mother may be self-deceived about how caring and supportive she is of her children. And a small-town chef may be self-deceived about the sensitivity of his palette and the innovativeness of his signature dish. But philosophical disagreements arise when we try to understand the shape that this motivation takes. Some theorists, such as Donald Davidson[3] and David Pears[4] require that self-deceivers intentionally deceive themselves. Other theorists, myself included, deny that this is a necessary condition for self-deception. In my view, a self-deceiver's motivational state plays a causal, but not intentional, role in getting her to believe falsely that *p* (the belief that she is self-deceived in holding). A person's motivation to believe that *p* may cause her to gather and interpret evidence relevant to *p* in a biased way. This, in turn, may make it more likely that she will believe that *p* rather than ~*p*.[5] How might the process of constructing a narrative contribute to self-deception on this account?

As we have already seen, the process of constructing a narrative involves interpretation; narratives do not just report events, they interpret them. Given that narratives are largely interpretive, they are subject to various forms of distortion. Indeed, a recent study conducted by Elizabeth Marsh and Barbara Tversky suggests that the *majority* of stories that we tell are distorted in some way.[10] Marsh and Tversky asked participants, 33 undergraduate students, to record «what, when, and how they told others about events from their lives» (Marsch & Tversky, 2004, p. 491). For each retelling, students filed two forms: one form asked them to describe the original event, and the other asked them to describe the retelling of the event (Marsch & Tversky, 2004, p. 294).

[3] See Davidson 1998.
[4] See Pears 1985.
[5] See Mele (2001, pp. 25–93) for a detailed account of how this can happen.

Along with these forms, students submitted answers to a number of questions about each retelling. Among other things, students evaluated each retelling for accuracy. Marsh and Tversky report that students labeled 42% of their retellings as "inaccurate" (Marsch & Tversky, 2004, p. 496). Curiously, they labeled 61% of the same retellings as "distorted in some way"; this broad category includes retellings that were exaggerated, minimized, selective, or additive (Marsch & Tversky, 2004, p. 496). Students apparently believed that their retellings could be distorted in one of the aforementioned ways without being inaccurate. What these results imply is that, more often than not, people share distorted accounts of their experiences with others. This finding is especially important given that one's retelling of an event can influence one's memory of an event; distorted retellings of events tend to result in distorted memories of events (Marsch & Tversky, 2004, p. 500).

I want to suggest that this practice of telling distorted stories to others can contribute to self-deception. If distorted retellings lead to distorted memories, and memories ground our beliefs, then there is a relatively straightforward way in which distorted retellings lead to distorted beliefs (or self-deception). It is worth noting that this might happen with even greater frequency than the Marsh-Tversky study predicts. After all, the Marsh-Tversky study only provides us with data concerning *self-reported* distorted retellings. It does not provide us with data concerning distorted retellings that are *not reported*. It is reasonable to suppose that we sometimes provide distorted accounts of events to others without realizing that this is what we are doing. The Marsh-Tversky study also (understandably) neglects the number of distorted retellings that we share with ourselves *sotto voce*. If we routinely tell ourselves distorted stories, then we may routinely form distorted or false beliefs.

3. La Symphonie Pastorale

Thus far, I have argued that narrative plays an important role in self-deception. The way that a person retells events can influence a person's beliefs and memories about those events. I should add here that this causal sequence is often reversed: Just as a person's retelling or narrative can influence her beliefs, so also can her beliefs influence her retelling or narrative. Most cases of self-deception probably involve causal sequences that move in both directions. There is generally an intimate and mutually reinforcing relationship between a person's beliefs and narrative. It may, therefore, be impossible to sever one

completely from the other and label the former 'cause' and the latter 'effect'. I now want to take a closer look at the way that narrative can make this happen and contribute to self-deception. In so doing, I would like to consider the narrative that André Gide presents in *La Symphonie Pastorale.*

In *La Symphonie Pastorale*, a pastor recounts the development of his love for a blind girl, Gertrude, whom he has adopted. The pastor finds Gertrude in the home of her aunt who has just died. At this point in the novella, Gertrude can neither see nor speak; she is vulnerable and destitute without any means of support. Against the wishes of his wife, Amélie, the pastor decides to bring her into their home and teach her how to speak and read Braille. The pastor claims that his decision to care for Gertrude is motivated by Christian teachings and considerations of virtue. He describes Gertrude as the "the lost sheep" who is deserving of compassion and privileged treatment. As the novella progresses, one begins to suspect that the pastor's motives are not entirely pure and Christian.

To his great dismay, the pastor soon discovers that his son, Jacques, is in love with Gertrude. Still not acknowledging his own love for Gertrude, the pastor mistakes his jealousy for indignation. He is furious with Jacques and forbids him from pursuing a relationship with Gertrude. Eventually, the pastor makes some progress towards understanding his feelings for Gertrude and the reality of their situation. Unfortunately, this personal revelation does not spare Gertrude and the pastor of a great tragedy. Gertrude, with the support of the pastor, undergoes an operation that enables her to see. Interestingly, the operation allows Gertrude to see new dimensions of the *moral* world as well as the *physical* world. When she is reunited with the pastor, Amélie, and the children, she can see the sadness in the face of Amélie. It is only at this point in the novella that she appreciates her sin and the gravity of her actions. She also realizes that she is in love with the handsome young Jacques, not the pastor. She tells the pastor that she imagined him to have the face of Jacques while she was blind. Realizing that she cannot have Jacques (who has at this point entered the priesthood) – that their marriage is impossible – she takes her own life.

Towards the end of the novella, the pastor acknowledges that his "earlier self" was mistaken about the nature of his relationship with Gertrude. While he does not admit to being self-deceived as such, he is aware that his earlier interpretation was in some way flawed or naïve. But there is additional textual evidence that the pastor's interpretation of his relationship with Gertrude is mistaken. We can sense the pastor's mistake through the words of others

woven into the narrative that he constructs. Perhaps the most compelling evidence of the pastor's mistake is provided to us by the perceptive but stoic Amélie. Amélie presents an alternative interpretation of the pastor's predicament through her cryptic and carefully chosen words. When the pastor confronts Amélie about Jacque's relationship with Gertrude, she shares with the pastor her understanding of his mistake. The pastor is angry with Amélie for not having warned him about Jacque's interest in Gertrude. Consider Amélie's reply in the following exchange:

> "I've seen it coming on for a long while. But that's the kind of thing men never notice."
> It would have been no use to protest, and besides there was perhaps some truth in her rejoinder, so, "In that case," I simply objected, "you might have warned me."
> "She gave me the little crooked smile with which she sometimes accompanies and screens her reticences, and then, with a sideways nod of her head:
> "If I had to warn you," she said, "of everything you can't see for yourself, I should have my work cut out for me!." (Gide, 1955, p. 145)

Amélie is in the background, as it were, observing the simultaneous development of two interwoven relationships: the relationship between Gertrude and Jacques, and the relationship between Gertrude and the pastor. She takes the pastor to have been blind to both. As the conversation continues, Amélie signals in her "enigmatic" and "oracular" way that the pastor may not know what he really wants (Gide, 1955, p. 146). The implication is that the pastor has romantic feelings for Gertrude but misinterprets them to himself and others.

Thus far, I have said that the pastor takes himself to have made a mistake. He acknowledges that his initial interpretation of his relationship with Gertrude was flawed. While engaged in a moment of self-reflection, the pastor explains what he takes to be the nature and source of his error:

> Now that I dare call by its name the feeling that so long lay unacknowledged in my heart, it seems almost incomprehensible that I should have mistaken it until this very day – incomprehensible that those words of Amélie's that I recorded here should have appeared mysterious – that even after Gertrude's naïve declarations I should still have doubted that I loved her. The fact is that I would not then allow myself that any love outside marriage could be permissible, nor at the same time would I allow that there could be anything whatever forbidden in the feeling that drew me so passionately to Gertrude [...]. For I should have

> considered love reprehensible, and my conviction was that everything reprehensible must lie heavy on the soul; therefore, as I felt no weight on my soul, I had no thought of love. (Gide, 1955, pp. 152–153)

Clearly, the pastor believes that he had some evidence that what he felt for Gertrude was not love. In fact, he offers a clever little piece of reasoning to account for his mistake: If he loved Gertrude, then he would have felt the weight of this love on his soul. Given that he felt no weight, there must have been no love. The pastor takes himself to be guilty of a simple, unmotivated mistake.

The problem with the pastor's self-diagnosis here is that it is incomplete. While the pastor is forthcoming about this piece of explicit reasoning, he is silent about the role that desire plays in giving it shape and pushing it along. Among other things, his moment of self-reflection overlooks the convenient interplay that we find between his reading of Christianity and relationship with Gertrude. Throughout the novella, the pastor constructs a liberal reading of Christianity that supports his relationship with Gertrude. He does not feel the weight of a "reprehensible" love precisely because he has interpreted away its reprehensibility. He appeases his conscience with his reading of Christianity and the thought that the Lord has entrusted him with Gertrude's sweet and pious soul (Gide, 1955, p. 109).

What makes the pastor's interpretation of Christianity especially suspicious is the fact that he imposes it upon Gertrude through blatant acts of censorship. While teaching Gertrude about Christianity, he omits passages about sin that might distress her. As a result, he presents her with a selective and incomplete understanding of the moral world. This, if you recall, is what she objects to after her operation. But Gertrude confronts the pastor about this concern even before her operation. She insists that he respect her preference for knowledge, not a delusional happiness:

> No, let me say this – I don't want a happiness of that kind. You must understand that I don't [...]. I don't care about being happy. I would rather know. There are a great many things – sad things assuredly – that I can't see, but you have no right to keep them from me. I have reflected a great deal during these last winter months; I am afraid, you know, that the whole world is not as beautiful as you have made out, pastor – and in fact, that it is very far from it. (Gide, 1955, p. 164)

The pastor is Gertrude's primary source of information about the visual and moral world. The only world that she knows is the world that he presents to

her. When Gertrude reflects upon this world, she senses that something is not right. Gertrude, from her exceedingly limited point-of-view, judges the pastor's presentation of the moral world to be incomplete.

With this textual evidence in view, I want to suggest that the pastor is not just mistaken, he is self-deceived. The pastor is self-deceived because he holds false beliefs about his relationship with Gertrude that are caused, in part, by his motivational state. I now want to take a closer look at the role that narrative plays in contributing to his self-deception. When reading *La Symphonie Pastorale*, our epistemic situation is very much like Gertrude's, albeit more extreme: our sole access to information about the world is the pastor's narrative. Although we approach the novella with certain background beliefs and assumptions, our only source of information about the pastor's predicament is what he presents to us in his narrative. The brilliance of Gide's novella is that it allows us to detect the pastor's self-deception from the inside, as it were. Even what we learn from other characters in the novella is presented to us through the pastor's narrative.

In constructing his narrative, the pastor imposes a particular order and interpretation upon a series of events involving Gertrude and himself. His narrative is not an artifact that he finds readymade. Instead, it responds to a series of events that leave him with the conceptual space for various forms of modification and distortion. The pastor can bring the events together in any number of ways that provide us with a sufficiently intelligible account of what happened. His motivation plays a crucial role in the way that he does this. In unpacking his motivation, we can identify three salient desires (understood broadly): (1) the desire to preserve his commitment to Christianity, (2) the desire to maintain his relationship with Gertrude, and (3) the desire to think of himself as good, according to the teachings of Christianity. This motivational set shapes the way that he interprets his relationship with Gertrude and those around him. As I have already explained, it causes him to accept an interpretation of Christianity that is consistent with his relationship with Gertrude, as well as an interpretation of his relationship with Gertrude that is consistent with his Christianity: the two are mutually reinforcing.

Notice that this explanation does not require that we view the pastor as *intentionally* deceiving himself. Indeed, we have no evidence in the novella that the pastor tries to get himself to believe anything at all. Instead, the pastor accepts the interpretation that he does because it seems plausible to him at the time. We can account for this appearance of plausibility by appealing to the

pastor's motivational state. Given that the pastor wants to think well of himself, he interprets his relationship with Gertrude in a way that reflects positively upon him. It takes less evidence to convince him that his relationship with Gertrude is innocent than it would to convince him that his relationship with Gertrude is romantic or inappropriate.

The pastor's self-deception contributes in an important way to his decisions and actions throughout the novella. We might say that the interpretation that he accepts of himself and his situation enables him to make the choices that he does. If he had felt the weight of an improper love, then he might have acted differently. If he had appreciated the nature of his relationship with Gertrude, then he might not have hidden the truth about the world from her. Instead, he keeps Gertrude and himself in a perpetual darkness. The pastor's self-deception affects others in the novella as well; it extends to both Jacques and Amélie. He is insensitive to Amélie and oblivious to the way that he is hurting her. And he selfishly separates Jacques from the woman he loves while condemning him for such love.

4. Narrative and Self-Knowledge

There are a number of lessons to be learned about narrative and self-knowledge in *La Symphonie Pastorale*. One lesson is that not all narratives are created equal; a narrative can be more or less truthful. We can appreciate this point by contrasting the pastor's narrative with the narrative that Amélie, or an impartial viewer, would likely construct. This would seem to imply that a certain kind of self-knowledge is achievable for us. If we judge some individuals to be self-deceived, then we seem to imply that others are not self-deceived (or are at least *less deceived*, as Philip Larkin might put it[6]). There are good reasons to be sceptical about the possibility of certain forms of self-knowledge. But self-knowledge, understood as that which self-deceivers lack, seems to be achievable for us. The second but related lesson is that we can evaluate the truthfulness of our own narratives. The fact that a given narrative is mine does not render it incorrigible. In *La Symphonie Pastorale*, the pastor comes to see the truth (or the partial truth) about his relationship with Gertrude and revises his narrative accordingly.

[6] See Larkin 1960.

Although the pastor's narrative is a work of fiction, it represents the way that narrative can contribute to self-deception in the real world. We are all very much like the pastor in that we understand the events of our lives in narrative form. We retell and remember events in an incomplete way and from a particular point-of-view. In weaving together the events of our lives, we can do a better or worse job. The fact that a narrative is selective and told from a point-of-view does not entail that it is riddled with distortions and inaccuracies. While a narrative cannot reproduce reality in its every last detail, it can be more truthful than not and make a contribution to our understanding of the world and ourselves.

In writing about autobiographical narratives, Cheryl Misak has made similar observations. Misak argues that narratives ground our theories in experience and allow us to deliberate in an informed way about important moral issues. Although narrative «is rife with exaggeration, omission, and self-deception» we should not abandon it altogether (ENED, p. 627). If we discover that two or more narratives make inconsistent claims, we should take seriously the possibility that one of the two narrators «got things wrong» (ENED, p. 629). We should not simply retreat to the relativist claim that each person is right or blameless «from his perspective» (ENED, p. 629). In evaluating narratives, Misak claims that we should use many of the same strategies that govern ordinary theory choice. We should assess a narrative based upon «internal coherence, consistency with other evidence, simplicity, explanatory power, and so on» (ENED, p. 630). We should also consider the motivation behind the narrative and whether or not the narrative is consistent with the experiences of others (ENED, p. 630).

As Misak observes, we can evaluate narratives in a non-arbitrary or principled way. Your narrative is not just as good as mine because yours reflects your perspective and mine reflects my perspective. It might be objected here that while these principles may be of some use, they cannot help us choose between narratives in difficult cases. Consider, for instance, the disputes that sometimes arise in response to published autobiographical works. The writer Isabel Allende has commented on the fact that her family members often reject the way that she retells events in her memoirs. Indeed, her stepfather called her a mythomaniac (Allende, 2011). In an interview, Allende explains why her family – specifically her stepfather – rejects her autobiographical narratives:

> Yes. He [Allende's stepfather] says that I am liar. When I was writing "Paula" it was the first time that I wrote a memoir. In a memoir one is expected to tell the

> truth. My stepfather and my mother objected to every page because from my perspective the world of my childhood, of my life, is totally different from the way they see it. I see highlights, emotions, and an invisible web – threads that somehow link these things. It is another form of truth.

It is interesting that Allende refers to her memoirs as a *form* of truth, and not as the truth full stop. By claiming that her memoirs present readers with a form of truth, Allende seems to acknowledge that her retellings are not entirely truthful. But this may not be exactly what she intends to say here. It may be the case that what Allende writes in her memoirs is not false, but imbued with interpretation. In weaving together the events of her life, she includes information about their highlights and emotional character. But notice that this is exactly what one does when creating a narrative or writing a memoir. Should we conclude from this that discrepancies about certain narratives are inevitable? Is the narrative that Allende's stepfather would write just as truthful as her narrative? Can we ever make decisions about such cases?

In support of Allende, it might be argued that she, as an artist, is able to recognize qualities of events that her family members would overlook. In a Millean vein, we might suppose that people differ in terms of their natural aesthetic and intellectual capacities. After all, we admire great writers not just for their technical skill, but also for their sensitivity and ability to interpret and express emotion. If this is the case, then we may have grounds for thinking that Allende's narrative is superior to that which her stepfather might construct. While his narrative may be more truthful than not, it may be incomplete and deficient in this respect. What this shows is that we not only want a narrator to be truthful, we also want her to be sensitive, perceptive, and discerning. When evaluating a narrative, we are not just interested in the number of events presented, but in the way that these events are represented; quality matters as well as quantity. A person's history, education, and natural abilities can all play some role in determining what she is and is not sensitive to. We would not expect a great poet to perceive a situation in the way that a five-year-old child would, and vice versa.

Not everyone will be completely satisfied by this explanation. After all, as I have argued elsewhere, the fact that Allende is deep and imaginative may make her especially vulnerable to self-deception and other forms of distortion (Kirsch, 2009). Perhaps, as her stepfather would likely suggest, Allende is more inventive than she is sensitive; she creates more than she observes. The more general concern might be that all narratives involve a certain degree of

invention. Indeed, this may be something that we as a society support and encourage. We applaud those who seek the hidden meaning behind a divorce, a reunion, an injury, a recovery, or any other more or less momentous happening in life. Are we not encouraging people to invent meaning where there is none to be found? Are we not prompting them to engage in self-deception?

Consider one of Jean-Paul Sartre's well-known stories on a similar theme. When Sartre was in prison, he met a "rather remarkable" Jesuit man who shared with him the story of how he joined the order (Sartre, 1975, p. 356). This man had suffered numerous tragedies and failures in life. At the age of eighteen, his sorrows peaked with the demise of a sentimental affair and the failure of his military examination. In response to these sad events, the man could have regarded himself as a complete failure. Instead, as Sartre observes, he "cleverly" interpreted his most recent failings as a sign from God that only religious achievements were possible for him (Sartre, 1975, p. 356). In Sartre's view, the man made a *choice* to view his situation in this way. After all, Sartre points out, he could easily have chosen to become a carpenter or a revolutionary (Sartre, 1975, p. 357). If Sartre is right, there is an element of choice in the way that we tell our individual stories. God's sign was not written in the events themselves; rather, the Jesuit man "invented" the sign or chose to see it there. While I would not describe the Jesuit man as having made a "choice" to interpret his life as he did, Sartre's account is largely correct. In constructing narratives, and in interpreting the events of our lives, we are often selective, partial, and in search of meaning.

When evaluating narratives, it is not the case that anything goes. Your narrative is not beyond criticism in virtue of the fact that you "wrote" it. As I have tried to show, we can and do judge narratives, including our own, to be more or less truthful. However, in certain cases, it may be difficult or impossible for us to distinguish between competing narratives. Narrators should be sensitive, perceptive and discerning, but not deceptive and inventive – unless they are just trying to sell books. It is this conceptual space for interpretation that self-deceivers exploit in deceiving themselves. Although we are probably all guilty of some distortions in telling the stories of our lives, we are not all systematically self-deceived or self-deceived on a grand scale. In real life, as in fiction, we can distinguish between the pastors and the Amélies.

When narratives are truthful, they can help us make sense of our personal and moral lives. The process of constructing a narrative involves bringing

together a series of disparate events into one more or less unified story. When this is done well, constructing a narrative makes it possible for us to reflect upon aspects of our lives that might otherwise go unnoticed. The information that we acquire in the process provides us with moral orientation and allows us to understand our obligations to others. Without it, we are like the blind Gertrude, lost in moral darkness and oblivious to the sorrows of others. On a theoretical level, we can also benefit from the autobiographical or real-life narratives of others. They can provide us with valuable moral insight and, as Misak has shown, ground our abstract moral theories in experience (ENED, p. 626).

At this point, a sceptic might question the explanatory force or usefulness of understanding self-deception (and, with it, self-knowledge) in terms of narrative. It might be objected that narrative only describes the way that people pursue or acquire self-knowledge when narrative is understood broadly. But when our understanding of narrative is sufficiently broad, we deflate it of any conceptual intrigue or significance; it becomes conceptually bankrupt.[7] Why not abandon talk of narrative altogether? First of all, the purpose of this paper has not been to present an account of self-deception in terms of narrative alone. Rather, I have tried to show that narrative can enhance our understanding of self-deception and supplement current theoretical work on the topic. Even a broad account of narrative can help us understand the causal processes that contribute to self-deception. Thinking in terms of narrative highlights the role that selectivity, perspective, and interpretation play in the way that we retell the events of our lives. While I would not object to considering these properties of narrative individually, thinking of them collectively has its advantages: (1) It allows us to see how they interact with each other in a familiar way. (2) It encourages us to look at autobiographical and fictional narratives that can deepen our understanding of how self-deception works. And (3) it reveals how one false or self-deceptive belief can spread and infect others. Theorists often focus upon a single isolated belief, the belief 'that *p*,' in accounting for the nature and possibility of self-deception. Thinking about self-deception in terms of narrative can help us appreciate the global nature of self-deception and its tendency to spread. It is often the case that a person's self-deception is not limited to the belief that *p*;

[7] For a critique of the narrative approach in general, see Strawson 2004 and Lamarque 2004.

rather, it spills over into her other beliefs and is woven into a narrative that she constructs about her life.

It is worth noting here that my discussion of narrative and self-deception (and, with it, self-knowledge) is compatible with most accounts of self-knowledge. Nothing that I have said thus far is contingent upon a conceptually demanding account of self-knowledge. Nor does it depend upon our having immediate, introspective access to our mental states. Indeed, the view defended here is even compatible with interpretational accounts of self-knowledge, such as the one advanced by Peter Carruthers (2010). According to Carruthers, we acquire knowledge about ourselves by observing our external and internal behavior (where this includes both inner speech and imagery, p. 83). It is possible, I would like to suggest, that we develop autobiographical narratives in response to this kind of observation. Carruthers and others have gestured in this direction in accounting for *self-knowledge*.[8] However, in so doing, they imply that all narratives are in the same category and equally fictitious. In their view, our narratives are all alike in being so many stories that we invent in an effort to make sense of our behavior. My account of self-deception provides us with some grounds for resisting this claim. Even if a narrative is based entirely upon behavior, it can be more or less consistent with what really happened (or with the actual behaviors in question). Admittedly, there may be deeper theoretical reasons for being sceptical about the possibility of self-knowledge in general. But it is not within the scope of this essay to address these formidable concerns – concerns with which I am deeply sympathetic.

5. Conclusion

The purpose of this essay has been to show that narrative can make an important, though not unavoidable, contribution to self-deception. Given the avoidability of self-deception, this paper is just as much about the possibility of self-knowledge as it is about the possibility of self-deception. As soon as we divide the world into self-deceivers and non-self-deceivers, we acknowledge that a certain kind of self-knowledge is possible for us. This self-knowledge is the kind that the pastor in *La Symphonie Pastorale* lacks. By examining his narrative, and comparing it with the insights and interpretations of others, we

[8] Daniel Dennett presents a version of this account in Dennett 1991.

can see where he goes wrong. We can imagine a pastor who is not self-deceived, or who is at least *less* self-deceived. If self-knowledge is within the realm of the possibilities for the pastor, then there may be some hope for the rest of us.

REFERENCES

Allende, I. (2011). *Questions and Answers*. http://www.isabelallende.com/curious_frame.htm Accessed on February 5, 2011.

Carruthers, P. (2010). Introspection: Divided and Partly Eliminated, *Philosophy and Phenomenological Research*, *80*, 79–111.

Currie, G. (2010). *Narratives and Narrators.* Oxford: Oxford University Press.

Davidson, D. (1998). Who Is Fooled. In J.-P. Dupuy (Ed.), *Self-Deception and Paradoxes of Rationality*. Stanford: CSLI Publications, 1–18.

Dennett, D. (1991). *Consciousness Explained*. New York: Little, Brown and Company.

Gide, A. (1955). La Symphonie Pastorale. In A. Gide, *Two Symphonies*. (tr. by D. Bussy). London: Cassell and Company Ltd.

Hutto, D. (2007). Narrative and Understanding Persons. In D. Hutto (Ed.), *Narrative and Understanding Persons*. Cambridge: Cambridge University Press, 1–15.

Kirsch, J. (2009). Maladies of Fantasy and Depth. *Social Theory and* Practice, *35*, 15–28.

Lamarque, P. (2004). On Not Expecting Too Much from Narrative. *Mind & Language*, *19*, 393–408.

Larkin, P. (1960). Deceptions. In P. Larkin, *The Less Deceived*. New York: St. Martin's Press .

Marsh, E.J., & Tversky, B. (2004). Spinning the Stories of Our Lives, *Applied Cognitive Psychology*, *18*, 491–503.

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

Misak, C. (2008). Experience, Narrative, and Ethical Deliberation. *Ethics*, *118*(4), 614–632.

Misak, C. (2005). ICU Psychosis and Patient Autonomy: Some Thoughts from the Inside. *Journal of Medicine and Philosophy*, *30*(4), 411–430.

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

Sartre, J.-P. (1975). Existentialism Is a Humanism. In W. Kaufman (Ed.), *Existentialism from Dostoevsky to Sartre*. New York: Plume, 345-369.

Schechtman, M. (1996). *The Constitution of Selves* . Ithaca: Cornell University Press.

Strawson, G. (2004). Against Normativity, *Ratio*, *17*, 428–452.

Tversky, B., & Marsh, E.J. (2000). Biased Retellings of Events Yield Biased Memories. *Cognitive Psychology*, *40*, 1–38.

Velleman, J.D. (2006). The Self as Narrator. In J.D. Velleman, *Self to Self*. Cambridge: Cambridge University Press, 203–223.

The Therapeutic Value of Intellectual Virtue

Mark Young *
mark.young@rdc.ab.ca

ABSTRACT

The focus of this article is to offer an account of how the development of one's intellectual character has therapeutic value in the attempt to overcome self-deception. Even stronger, the development of intellectual character has necessary therapeutic value in regard to self-deception. This account proceeds by first consulting the predominant psychological theory of virtuous character offered by contemporary virtue ethicists and virtue epistemologists. A motivational/dispositional account of self-deception is then offered and connected to the former account of intellectual character. By connecting these two sets of literature the therapeutic value of intellectual virtue is displayed. The problem of self-diagnosis is then presented as well as intellectual character as a necessary therapeutic measure to assure agents that they are not self-deceived.

1. Introduction

To display the therapeutic value of intellectual virtue the first step will be to become familiar with the predominant neo-Aristotelian theory of virtuous character. After this has been presented a motivational/dispositional account of self-deception, which is congruent with virtue psychology, will be offered. These two sets of literature then will be explicitly connected in order to display the therapeutic value of intellectual virtue. Finally, an argument will be offered for the claim that intellectual virtue has necessary therapeutic value in the attempt to overcome self-deception.

* Red Deer College, Alberta, Canada.

2. The Psychology of Virtuous Character

The goal of this section is to become familiar with the predominant psychological theory of virtuous character offered by contemporary virtue theorists, which is based on Aristotle's notion of virtuous character.[1] Such familiarity is indispensable to ultimately understanding the therapeutic value of intellectual character in the attempt to overcome self-deception.

The first point to note about virtuous character is that the virtues entail attempts to correct certain natural shortcomings. To be virtuous entails a conscientious effort to overcome excessive or deficient psychological motivations and dispositions, or to regulate personal desires, that are held to lead to inappropriate behaviour. For example, to be courageous the agent must overcome unwarranted contrary desires for safety, and to be temperate an agent must overcome an excessive desire for pleasure.[2] The types of psychological dispositions identified for correction by virtue theorists are often understood to be selfish desires, but this is not always the case; for to be virtuous can also entail the correction, or altering, of the influence of positive and altruistic desires, since an agent can be altruistic to her own detriment and the detriment of others.[3] In regard to the intellectual virtues it is claimed that agents are susceptible to cognitive excesses and deficiencies, i.e., intellectual vices, which must be replaced and corrected by the appropriate character traits deemed intellectual virtues.[4]

The virtues are also typically held to be motives that contain at least some emotive content, and it is this emotive content that is held to initiate activity toward specific ends.[5] Since the virtues do not simply involve acting through the influence of one's emotions, but instead through the influence of those emotions deemed worthwhile, these emotive states are connected to the fulfillment of specific values. For example, a compassionate individual has certain feelings associated with compassion, such as love and sympathy, which then initiates compassionate activity.[6] The virtues, as motivations for action,

[1] See Aristotle 1993, pp. 19–20; Axtell 1997, pp. 2, 14; Hursthouse 1999, pp. 8–12, 15–16; Merritt 2000, pp. 367, 374–376; Sherman & White 2003, pp. 39, 42.
[2] See Foot 1997, pp. 169–170.
[3] See Irwin 1996, pp. 48–49.
[4] See Montmarquet 1993, p. 23; Zagzebski 1996, pp. 105, 152–153. Axtell 1997, p. 14; Axtell 1998, p. 495; Sherman and White 2003, p. 42; Fairweather 2001, pp. 67–70.
[5] See Zagzebski 1996, p. 131–132; Hursthouse 1999, pp. 99–100, 108.
[6] See Zagzebski 1996, pp. 131–132; Hursthouse 1999, pp. 99–100.

are also held to be persistent. Virtue theorists recognize that some motives are episodic since they only occur at particular times, but it is also proposed that certain motivations are persistent and therefore dispositional. That is, individuals possess certain motivations which are enduring and initiate behavior consistently and are therefore considered to be dispositions.[7] This is an important aspect of virtue psychology that has been part of the tradition at least since Aristotle,[8] which is the idea of the motivational self-sufficiency of virtuous character. What this entails is that the virtues, when they become integrated into agents, become robust character traits that dispose agents to act in certain ways regardless of external conditions. So, for example, if an agent is generous she will remain such even if resources are scarce. Agents may need to rely on social conditions to initiate the development of virtuous character traits, such as educational institutions and the family, but once the virtues are fully integrated into the character of an agent that agent will possess enduring dispositions that acts as impetuses for action.[9] In fact, once a virtue has become fully integrated it is held that counter inclinations, such as fear, do not exercise any influence. Such inclinations do not simply compete with virtuous motives to determine action in the fully virtuous agent, but instead are completely silenced by the relevant virtue, or virtues.[10] The virtues therefore enable agents to act consistently, and to adopt the necessary skills needed to act in accord with various virtues.[11] For example, the agent who possesses the intellectual virtue of intellectual conscientiousness, or, as it is sometimes referred to, the love of truth, will develop those skills that will enable her to better achieve true beliefs. Thus, it is generally held by virtue theorists that the virtues tend to have motivational components, which are emotive, dispositional, robust and consistent.

Besides motivational and dispositional impetuses for action the virtues are also held to influence an agent's perception and reasoning.[12] Focusing first on perception it is proposed that the virtuous agent does not simply want the right things, through being motivated or disposed by the virtues, but is also able to apprehend the "salient aspects of the relevant situation" through the influence

[7] See Zagzebski 1996, p. 132; McDowell 2003, p. 134.
[8] See Aristotle, *NE*, 1100b, 1105a.
[9] See Merritt 2000, pp. 366–368, 374–376; Hursthouse 1999, p. 123. McKinnon 1999, p. 29.
[10] See McDowell 2003, p. 125.
[11] SeeZagzebski 1996, p. 133.
[12] See Louden 1997, p. 206.

of the virtues.[13] This occurs because the virtues constitute the agent in a particular way. We have already seen this in regard to the idea that the virtues act as motivations and dispositions for agents, and therefore influence the agent's choices and actions. Since the virtues can influence the agent in this way they can also influence how agents perceive and think.[14] In order to display how the latter is the case we will focus on one particular moral virtue, i.e., courage, and then generalize these remarks to other virtues.

A common claim among virtue theorists is that a virtue is a psychological disposition which is itself a mean between two extremes.[15] These extremes tend to be inappropriate ways of feeling, desires or motivations, which can then obscure agent perception. In the case of courage the two extremes are cowardice and rashness. The coward is overcome with inappropriate fear, and/or desire to save himself, and this then causes him to perceive the particular situation as more dangerous than it actually is. The rash agent, on the other hand, is overconfident. Such an agent perceives the situation as less dangerous than it actually is, and in this way does not perceive the situation accurately. The courageous agent, though, is held to perceive the situation accurately, and therefore will act appropriately. Such an individual has silenced the influence of irrational fears, and therefore does not give inappropriate weight either to his personal safety or to the dangers involved in a situation. The courageous agent is also aware of his own limitations, and hence what his actual options are in the situation. For example, it is generally held that courageous actions entail facing an immediate danger, but this is not always the case. It could be that in a particular situation the courageous act entails retreating from immediate danger. The agent who possesses the virtue of courage knows whether it is better to retreat or to face the immediate danger since the psychological disposition associated with courage enables such an agent to recognize considerations that either warrant retreat or making a stand.[16] Courage is therefore held to enable the agent to perceive whether there is a genuine threat that cannot be overcome, or whether, through personal effort, the threat can be overcome.

The general psychological theory that underlies this description of the influence of courage on human perception is that the agent's affective and

[13] See Annas 1998, p. 40; Hursthouse 1999, pp. 207–208; Sherman & White 2003, p. 36.
[14] See McKinnon 1999, pp. 29–30.
[15] See Aristotle, *NE*, 1106b; Zagzebski 1996, pp. 96–97.
[16] See Adams 2000, pp. 39–40; Irwin 1996, pp. 45–46; Wallace 1973, pp. 64–66.

motivational states influence her cognitions. That is, the virtue theorist ascribes to a psychological theory which proposes that psychological states such as desires, passions, motivations and dispositions, and not merely the agent's various beliefs, influence human cognition. The vices are those affective states that detrimentally affect agent perception since they disable the ability to perceive accurately. This is exemplified in the perceptions of the rash agent as well as the coward. Such agents perceive the same situation differently from the courageous agent, and this is explained due to their divergent psychological constitutions. That is, the virtues constitute the agent in a particular way, which in turn produces accurate perceptions and choices. The virtues therefore make agents sensitive to particular aspects of situations, as well as specific warranted expectations, by constituting the agent in specific ways. They shape and order the agent's concerns and interests. They cause agents to be concerned with courageous, benevolent, fair, charitable acts, and so on, and in this way influence agent perceptions in particular situations. Thus the virtues do not simply remove vicious obstacles, but they also provide a type of knowledge, or understanding, that guides the agent in her various perceptions.[17]

Since the virtues cause agents to perceive in specific ways they also cause agents to reason in specific ways by influencing their perceptions of facts, situations, principles and so on. For example, for the courageous agent certain aspects of situations will appear salient, and decisions made are based on the agent's perception of those salient aspects. So, the virtues do provide an impetus for action by being motivational and dispositional, but they also fulfill a role in the reasoning process of agents when reasoning does occur. That is, there are instances when little to no reasoning occurs, and the relevant virtue, or virtues, shapes perception and a virtuous action results without deliberation.[18] In other situations, though, deliberation occurs before the virtuous action results, and such deliberation is also guided by the virtuous, or vicious, state of the agent. For example, an agent who possesses the virtue of charity reasons through its influence in various ways to bring about charitable acts. In such situations the impetus for an action is not simply a virtue acting as a motivational or dispositional state, but rather the agent acts because of specific reasons and such reasons appear warranted, or appropriate, due to her

[17] See Irwin 1996, pp. 40, 48–49, 53; McKinnon 1999, pp. 32–33; McDowell 2003, pp. 122–127, 135–137, 140; Hursthouse 1999, pp. 11–12, 111, 129–131, 207–208.
[18] See Hookway 2003, p. 184.

virtuous perceptions. The agent does not have to be aware that she is acting from some general impetus for behaviour, such as charity, but instead may mention more situation specific reasons. For example, the agent does not have to say "I did act X because it was courageous," but rather can cite reasons such as "Someone had to save him," or "I knew that I could save him if I tried." The significant point is simply that it is through the influence of a virtue, or virtues, that the agent recognizes such reasons as warranted and compelling. Once the virtues are fully inculcated into the character of the virtuous agent such an agent does not always have to be cognizant that she is reasoning due to the influence of a virtuous disposition. The virtue in question instead simply constitutes the agent in a particular way to shape her understanding and then this understanding is applied to specific situations. The virtues therefore first facilitate appropriate perception, and then, in turn, facilitate appropriate reasoning based on those perceptions.[19]

It is also generally held that the virtues not only enable agents to reason correctly, but also enable such an agent to act in accord with appropriate reasoning. The idea is that the virtues remove, or replace, inappropriate psychological mechanisms, i.e., the vices, from having a deleterious influence on motivations and perceptions, and this includes the reasons for which the agent acts. This means that the virtuous agent will also act in accord with the outcome of her virtuous deliberations. That is, the virtuous agent first deliberates through the cognitive filter of the virtues to come to specific conclusions, and then she is able to act in accord with the conclusions of virtuous deliberation through the motivational/dispositional capacity provided by the relevant virtue, or virtues.[20] This claim is significant for later attempts to refine the specific contribution of the intellectual virtues. For it will entail not simply that such virtues enable agents to perceive and reason correctly when it comes to assessing whether some belief is true, but also to believe in accord with those perceptions and virtuous deliberations. Thus, the intellectual

[19] See Irwin 1996, pp. 48–50. Watson 2003, p. 234; Crisp 1996, p. 17; Pence 1984, pp. 287, 289; MacIntyre 1981, pp. 161–162; McKinnon 1999, pp. 29–30, 34, 44; McDowell 2003, pp. 133–136; Hursthouse 1999, pp. 108, 111, 123–129, 136, 145.

[20] See MacIntyre 1981, p. 162; Irwin 1996, pp. 46, 49–50; Annas 1998, p. 40; Hursthouse 1999, pp. 11–12, 92, 102–103, 108–109, 123–125, 129–130, 136. McKinnon 1999, pp. 29–31, 34, 44; Annas 2003, p. 289.

virtues will dispose agents to not only reason and perceive in certain ways, but also to believe in certain ways.[21]

Having become familiar with how the virtues influence the psychology of agents we must briefly become familiar with one final claim concerning virtuous character. This is the claim that the virtues are teleological. To say that the virtues are teleological means that the virtues possess a particular *telos*, or end, to which they are directed. Broadly speaking, the particular end of the moral virtues is proposed to be the "good," and that of the intellectual virtues is the "true."[22] The claim that the virtues are teleological deserves mentioning since, in what follows, a description of intellectual character is offered where the specific *telos* of such character is true belief and the relationship to this end is instrumental. That is, it is argued that the intellectual virtues fulfill an instrumental role in enabling agents to obtain and sustain true beliefs, and it is for this reason that such virtues can act as a therapeutic means to overcome self-deception. Before this claim can be made, though, familiarity with a specific theory of self-deception is required. So, in the next section, a motivational/dispositional account of self-deception is summarized and then connected to the theory of virtuous character outlined in this section.

3. The Motivational/Dispositional Account of Self-Deception

The focus of this section is set out a theory of self-deception that coheres with the theory of virtuous character outlined in the previous section. It is a theory that ascribes a causal role to motivations and dispositions in occurrences of self-deception. Two general ways in which motivations and dispositions can fulfill a causal role in self-deception are identified. First, agents can be motivated, or disposed, to favor a particular belief, or set of beliefs, and this then causes the agent to gather evidence in a way that will either confirm, or conform to, that cherished belief, or set of beliefs. Second, motivations, or dispositions, can cause agents to miss disconfirming evidence altogether. With the latter situation evidence is not reinterpreted to either conform to, or confirm, some cherished belief or set of beliefs, but instead disconfirming

[21] See Montmarquet 1987, pp. 486–487; Montmarquet 1993, pp. 43, 65; Axtell 1998, pp. 498–499; Zagzebski 1996, p. 149; Fairweather 2001, pp. 67–69; Hookway 2003, p. 188.
[22] See Watson 2003, pp. 230, 241; Annas 2003, pp. 21–22.

evidence is ignored altogether.[23] Consideration of these two ways in which motivations and dispositions can initiate self-deception will then facilitate appreciation of how intellectual character can act as a therapeutic means to overcome it.

We will begin by focusing on the first way in which motivations and dispositions can initiate self-deception, and specifically consult two explanations of self-deception offered by Alfred Mele and Herbert Fingarette. According to Mele, an agent's desire for some belief, or set of beliefs, can cause that agent to engage in acts of both negative and positive misinterpretation. Negative misinterpretation occurs when the agent's desire leads that agent to misinterpret evidence as not disconfirming a particular belief, or set of beliefs, although, in the absence of such a desire, the evidence would easily disconfirm the agent's belief or beliefs. For example, an agent could have evidence that his partner does not love him and yet his desire for his partner's love could cause him to ignore such evidence in order to maintain a belief that she does. Positive misinterpretation occurs when the agent interprets evidence, through the influence of some desire or motivation, as counting in favor of her belief when in fact it does not.[24] For example, an agent who wants to maintain a view of himself as generous will misinterpret his actions in specific situations as conforming to this virtue. This will occur even if there are significant reasons to believe that the agent is not generous.[25] Hence, what the agent does in such situations is provide an explanation to himself that makes the evidence fit together so as to confirm, and conform to, his desires or motivations.

Fingarette's explanation of self-deception focuses more on the motivation to maintain a complex web of beliefs, which he refers to as a specific "cover-story." According to Fingarette, in cases of self-deception agents possesses a cover story to which facts are bent so as to confirm the cover story. The agent skillfully interprets aspects of his engagement in the world in order to maintain the plausibility of the cover story and make it as natural and internally consistent as possible even when the evidence continues to mount against this story. This is accomplished by engaging in inventive acts of rationalization in order to fill in the gaps of the cover story not confirmed by the evidence to

[23] See Sanford 1988, pp. 161–162, 169; Johnston 1988, p. 75; McLaughlin 1988, pp. 39, 52–53; Audi 1988, pp. 97–99, 101–105, 107–108; Mele 2001, p. 29–30.
[24] See Audi 1988, pp. 97–99, 103–105, 107–108; Mele 2001, pp. 26–27.
[25] See Mele 2001, p. 11.

which the agent is exposed.[26] Fingarette is not alone in advocating such an explanation of self-deception, for psychologists who conduct research on self-deception offer a similar explanation. For example, Shelly Taylor proposes that the belief formation of agents is often influenced by the attempt to maintain a self-schema. A self-schema is an organized sets of beliefs about an agent's personal traits and role in the world. Agents attempt to maintain beliefs associated with their self-schemas, for example that they are witty or kind, and this causes them to form false beliefs in specific situations. The self-schema therefore acts as a filter through which specific information is interpreted. If incoming evidence does not conform with the self-schema, then it is either modified or ignored.[27] Agents desire to see themselves, as well as loved ones and cherished beliefs, in a positive light, and attempt to avoid the anxiety that could arise if they were confronted with a belief they do not want to be true. It is therefore a general desire, in this case the desire to maintain a favoured cover-story, or self-schema that is the impetus for specific acts of self-deception.[28]

The second way in which motivations can initiate instances of self-deception, as mentioned, is by simply causing agents to miss disconfirming evidence in the first place. No positive or negative misinterpretation occurs in such situations, but instead disconfirming evidence is ignored altogether. Through the influence of one's motivations an agent either evades an issue altogether or the agent engages in selective attention and evidence gathering. For example, the agent will be hypersensitive to evidence that confirms what the agent is motivated to believe, so that her attention is constantly focused on confirming evidence and fails to acknowledge evidence that would disconfirm a cherished belief. In situations of evasion and selective attention no misinterpretation occurs, since the evidence is never acknowledged. The agent simply ignores the evidence due to the influence of a desire to maintain some belief, cover story or self-schema. For such an agent only specific aspects of situations, i.e., those aspects which confirm, and conform to, the agent's motivations, are perceived as salient, and this is directly the result of the agent's specific desires or motivations. For example, an agent who wants to

[26] See Fingarette 2000, pp. 34, 37–40, 46, 48–49, 52, 61–63, 69–71.
[27] See Taylor 1989, pp. 13–15, 154–155.
[28] See Audi 1988, pp. 97, 101–102, 105, 107–108; Johnston 1988, p. 66, 73, 86; Taylor 1989, pp. 8–45; Sanford 1988, pp. 157–159; Fingarette 2000, pp. 65–69, 86, 139, 142, 145; Asendorpf & Ostendorf 1998, pp. 961–962; Anderson, Srivastava, Beer, Spataro, & Chatman 2006, p. 1095.

believe that her husband is faithful, and is strongly motivated to maintain this belief, ignores evidence that attests to his infidelity while focusing on the evidence that attests to his devotion even when this evidence is minimal. Through the influence of specific desires or motivations, then, agents either evade disturbing evidence altogether or engage in selective evidence gathering, so that it is only the evidence that confirms the motivationally biased belief that is recognized while evidence that disconfirms such belief is not even acknowledged.[29]

Another aspect of the motivational/dispositional account of self-deception that deserves mentioning is that the motivations which initiate self-deception tend to be self-serving. It is a desire to see oneself, as well as loved ones and cherished beliefs, in a positive light, or to remove the anxiety that could arise if the agent were confronted with a belief that she either did, or did not, want to be true that causes specific instances of self-deception. The agent desires to see herself as a person of a particular type, or to maintain the truth of some favoured explanation or theory, and this then initiates either misinterpretation or selective attention. In such situations the agent is not concerned with the truth of her beliefs about her own character, the character of loved ones, nor about the truth of some cherished belief or set of beliefs. Instead, it is the maintenance of what is favoured, often to remove anxiety and maintain psychological well-being, that motivates the gathering of evidence as well as the explanations provided.[30] Thus, self-serving desires and motivations are often the cause of self-deception and, in turn, false belief. Also, the influence of those desires which initiate self-deception are unconscious. The agent who engages in self-deception is not aware that the process is occurring, and this lack of awareness is indispensable for the success of self-deception. If the agent were to become aware of the fact that she was influenced by specific desires, and therefore was motivated to believe in specific ways, then such desires would no longer be efficacious. This is because the agent would then be aware that her beliefs were the result not of evidence, but rather her own biased psychological states. The agent would thus realize that she was duped by her

[29] See McLaughlin 1988, pp. 42–43; Johnston 1988, pp. 67–68, 75, 87; Audi 1988, p. 105; Taylor 1989, pp. 146, 147; Fingarette 2000, pp. 38–40, 46, 167–169, Mele 2001, pp. 26–27, 51–52; Baier 1996, pp. 53–55; Oksenberg Rorty 1988, pp. 11, 18; Oksenberg Rorty 1996, pp. 77–79.

[30] See Johnston 1988, pp. 66, 73, 86; Audi 1988, p. 97, 101–102, 105, 107–108; Taylor 1989, 8–45; Sanford 1988, pp. 157–159; Oksenberg Rorty 1996, p. 77; Baier 1996, p. 55; Asendorpf & Ostendorf 1998, pp. 961–962; Fingarette 2000, pp. 65–69, 86, 139, 142, 145; Anderson et al. 2006, p. 1095; Deutsch 1996, p. 316; de Sousa 1988, p. 327; van Fraassen 1988, p. 145.

own motivational structure, and would, in turn, no longer be taken in by it. Hence, the agent must be oblivious to the influence of specific desires, or motivations, in order for self-deception to occur.[31]

Finally, the influence of motivations that can cause self-deception are generally not episodic, and hence the dispositional nature of self-deception. That is, agents typically maintain the specific motivations that can initiate occurrences of self-deception so that perceiving, reasoning, and ultimately believing through their influence is dispositional. The agent does not adopt a particular motivation, cover story or self schema, for only a moment, but instead has long-term commitments to them. For example, the agent who desires to see herself as courageous does not do such only momentarily but instead is committed to this belief. This is not to deny that the motivations which initiate self-deception cannot be held only episodically, but rather the point is that typically they are not. Agents can be quite committed to maintaining specific motivations, cover-stories and self-schemas, and will therefore continue to be influenced by them when forming new beliefs. In such situations these patterns of entrenched doxastic behaviour act as an "automatic filtering process" through which evidence and reasons are considered, so that those beliefs that serve the agent's interests, by conforming to what the agent desires, are maintained. Consequently, the psychological mechanisms which cause occurrences of self-deception represent enduring psychological stratagems of the agent, or, more simply, dispositions.[32]

To sum up, then, according to the motivational/dispositional account of self-deception it is the desires of the agent that initiate instances of self-deception. These motivations cause agents to form false beliefs by either

[31] See Johnston 1988, pp. 65–66, 70–76, 78, 87; Audi 1988, p. 94, 102–105, 109; Baier 1996, pp. 54–55; Deutsch 1996, p. 317; Fingarette 2000, pp. 46–49, 60–61, 65–66, 78, 98–99. Another possible impetus for self-deception could be akrasia; i.e., the agent does not believe on the basis of reasons she is aware of. One could easily imagine that motivations/dispositions could also fulfill a role here, as the agent does not believe as she should because she is disposed to maintaining some favourable cover-story. If akrasia can be an impetus for self-deception then there could be instances where the agent is self-deceived and in some way aware that she is. It may be questionable, though, whether self-deception can occur due to akrasia. This is because when an agent is suffering from akrasia she is well aware that some claim is true but does not act on it. Hence, to be self-deceived via akrasia means that the agent holds that some belief is true but then does not believe it. It seems impossible that one could believe and not believe some claim simultaneously, and the account of self-deception offered in this article has avoided this possibility so far.

[32] See McLaughlin 1988, pp. 43–44; Johnston 1988, pp. 66, 87; Oksenberg Rorty 1988, p. 18–19. Taylor 1989, pp. 227–228; Oksenberg Rorty 1996, p. 76–78; Fingarette 2000, pp. 46–47.

causing the agent misinterpret evidence or engage in selective attention and rationalization. The motivations which initiate self-deception are also both self-serving and unconscious. The agent is attempting to maintain some cherished belief, cover-story or self schema, and in order for this process to be effective the agent must be unaware that it is occurring. Finally, the motivation to maintain some cherished belief, or set of beliefs, is not episodic, but instead represent certain habits of the mind and are therefore dispositional in nature.

4. The Mitigating Influence of Intellectual Character

Having achieved a basic understanding of the motivational/dispositional account of self-deception attention can now turn to how the development of one's intellectual character can act as a therapeutic means to overcome self-deception. This account relies significantly on the theory of virtuous character outlined in the first section, as well as some new sources.

The first thing to recall is that the intellectual virtues also involve a motivational component. Specifically, they involve a general desire for true belief as well as a variety of specific motivations, such as a motivation to be open-minded, intellectually humble, intellectually courageous, and so on.[33] Since it is the case that motivations fulfill a role in self-deception it is possible that the motivations associated with intellectual character could act as a means to overcome self-deception. According to the motivational/dispositional account of self-deception when agents form their beliefs they do not simply have to be exposed to the appropriate evidence in order to avoid possessing false beliefs. They also must be motivated in the right way toward that evidence. If agents are motivated to reinterpret evidence in a self-serving manner they will come to believe as they want to believe and not as the evidence suggests. What seems to be required, then, to overcome self-deception is not simply to re-expose agents to the evidence, or to even expose them to further evidence, since such evidence will be filtered through their motivational structure. Instead, in order to enable agents to obtain true beliefs in such situations it appears that it is their motivation structure that must be altered. A possible way to overcome self-deception, then, is to replace the self-serving motivations associated with self-deception with motivations focused on obtaining true

[33] See Johnston 1988, pp. 68–69; Fairweather 2001, pp. 68–69.

beliefs. The virtues of intellectual character provide such a motivational structure, and therefore seem to be what is required to obtain true beliefs.[34]

That it is the motivational structure of intellectual character that is required to overcome self-deception is further confirmed when we revisit other aspects of virtuous character outlined in the first section and then compare this to what was proposed in the previous section concerning self-deception. Recall first the claim that the virtues entail attempts to overcome natural shortcomings and personal desires that can exercise an inappropriate influence on agents.[35] For example, to become temperate the agent must overcome a strong desire for pleasure. This appears similar to what occurs with the motivations and dispositions that lead to self-deception, and therefore lends support to the claim that the intellectual virtues could act as a means to overcome self-deception. For, as stated in the previous section, agents who engage in acts of self-deception are typically motivated by self-serving desires. The agent wants to maintain specific beliefs about herself, and others, or to simply maintain some meaningful belief, in order to avoid the anxiety that could result if their falsity were exposed.[36] From the perspective of the intellectually virtuous agent such desires, or motivations, are inappropriate and must be overcome. They are inappropriate from such a perspective, for what matters to the intellectually virtuous agent is to obtain true beliefs. In such a situation the intellectually virtuous agent attempts to mold her motivational structure so as to not be subject to inappropriate motivations, or dispositions, that could lead to false beliefs. The types of motivations and dispositions to be thwarted include the very general self-serving dispositions outlined above, but also very specific motivations and dispositions. Examples include: a tendency to believe too easily, i.e., credulity; fear of questioning one's beliefs; being dogmatic; being diffident in regard to one's beliefs and intellectual abilities; being overconfident; being concerned with status as opposed to truth, and so on.[37]

The motivational/dispositional account of self-deception therefore corresponds to the explanation of human psychology advocated within the virtue perspective. Agents are influenced by natural but inappropriate shortcomings which can be overcome through the influence of the virtues. In this case the natural shortcomings pertain to the beliefs of the agent, and the

[34] See Fairweather 2001, pp. 69–71, 78; Leon 2002, p. 423.
[35] See Roberts & Wood 2003, pp. 261, 263.
[36] See Code 1984, p. 42; Gouinlock 1993, p. 300.
[37] See Sherman & White 2003, p. 42; Roberts & Wood 2003, p. 263.

attempt to maintain desirable yet unwarranted beliefs. The intellectual virtues therefore become correctives to such dispositions because they are directed toward obtaining true beliefs, but also because they are specific motivational/dispositional components that can answer to the motivational/dispositional components that lead to self-deception. Instead of being disposed to sustain and obtain beliefs that confirm, and conform to, self-serving desires the agent is disposed to have beliefs that are true. The possession of a general disposition towards true beliefs, as well as the other more specific dispositions of intellectual character, then influence how the agent forms beliefs just as the self-serving motivational/dispositional structure influenced belief formation to cause self-deception. In this situation, though, since the agent is focused on truth, or obtaining true beliefs, it will be this disposition that will be fulfilled as opposed to the self-serving disposition.[38]

So far, then, we have a fairly good understanding of why intellectual character is therapeutically relevant for overcoming self-deception. Intellectual character is relevant since obtaining true beliefs is not merely a matter of exposure to the appropriate evidence, but also a matter of the motivational, and/or dispositional, structure of the agent. Agents can be influenced in their belief formation by self-serving motivations and dispositions, and the intellectual virtues can act as correctives to these natural short-comings in order to facilitate true beliefs. The next aspect of intellectual character to be explored to display its therapeutic value for overcoming self-deception is the effect of such character on the perceptual and rational capacities of the agent.

In the first section significant attention was given to the idea that virtuous character can influence an agent's perceptions and rational capacities. Inappropriate motivations and dispositions were said to obscure, or contaminate, agent perception, while the virtues were proposed to mitigate this influence to enable the agent to perceive accurately. This explanation of the role of virtuous character is congruent with the explanation of occurrences of self-deception considered in the previous section. Recall that the perceptual capacities of agents who engage in acts of self-deception are significantly influenced by their motivations and dispositions. The agent perceives situations in a way that either confirms, or conforms to, what is desired which then influences the beliefs formed. Intellectual character can mitigate this perceptual influence by replacing self-serving dispositions with dispositions

[38] Zagzebski 1996, pp. 146–147, 154; Fairweather 2001, p. 72.

for true beliefs.[39] For example, an agent who wants to maintain some cherished belief will overestimate the evidence in its favour, and avoids being cognizant of evidence that disconfirms his belief.[40] If the agent were instead constituted by the motivational/dispositional structure of intellectual character, then the evidence would not be overlooked. The agent would be disposed to maintaining beliefs only if they are true, since he would be guided by a general desire for true beliefs as well as other more specific dispositions. Thus the agent would be open to both confirming and disconfirming evidence for his beliefs, and would perceive this evidence as salient due to the influence of his intellectual character.[41]

This influence of intellectual character on agent perception also means that the virtues influence agent reasoning. In the first section it was proposed that the virtues cause agents to perceive in specific ways and therefore also cause agents to reason in specific ways. This occurs by influencing the agent's perception of evidence in particular situations, and therefore the content of the agent's deliberations. Through causing appropriate perceptions the virtues ensure that the evidence the agent relies on in her deliberations is accurate. The intellectually virtuous agent does not reason based on a self-serving interpretation of the evidence, but instead based on an interpretation of the evidence that is directed at achieving true beliefs. This influence of intellectual character is therefore similar to the role of the self-serving motivations and dispositions that lead to self-deception. Self-serving motivations can initiate a rationalization process so that the beliefs formed conform to the content of these motivations. Intellectual character mitigates the possibility of false beliefs by replacing the latter impetuses for rationalization with a disposition toward true belief. Instead of desiring to maintain some cherished belief, and having her perceptions and deliberations influenced by such a desire, the intellectually virtuous agent is motivated to obtain true beliefs and this, in turn, influences both her perceptions and deliberations and therefore disposes her to obtain and sustain true beliefs.[42]

Another aspect of intellectual character must be dwelt on to strengthen the connection between occurrences of self-deception and the mitigating influence of intellectual character. Recall that it was proposed that not only do

[39] See Sherman & White 2003, p. 36.
[40] See McLaughlin 1988, p. 43.
[41] See Fairweather 2001, p. 71.
[42] See Hookway 2001, pp. 190–192; Reed 2001, p. 517.

the virtues clear away inappropriate motivations so that agents can perceive accurately aspects of various situations, but also that the virtues provide a type of understanding for the agent. This occurs partly through constituting the agent's concerns and interests in specific ways, for example by providing a concern for true believing, but also because the virtues are held to be instructive concerning how the agent should think and act in particular situations. Consider, for example, intellectual virtues such as open-mindedness and intellectual humility. The agent who is intellectually humble realizes that some of her beliefs, if not all, could be false, and that she can always learn from others. The agent who is open-minded is not simply willing to listen to the positions of others, but admits to himself that such positions could actually be true while his own beliefs could be false. Other general influences of intellectual character include causing the agent to carefully scrutinize the evidence, to consider alternative explanations and arguments, and to be thorough in her inquiries.[43] With such virtues, as well as others, the agent who possesses intellectual character therefore possesses a certain type understanding of her current beliefs and how she should interact with others when forming new beliefs. With such an understanding in hand she is then willing to question the beliefs she has and is well aware that they could be false. She is therefore less susceptible to the motivations and dispositions that could lead her to self-deception. For example, instead of being motivated to maintain a belief of oneself that one charitable, which can then lead to instances of self-deception, the agent is willing to admit that such a belief could be wrong; especially if this is pointed out to her by someone else. Thus, the virtues of intellectual character can also help to overcome self-deception by providing a certain type of understanding for the agent.

Another way in which the psychology of self-deception lines up with the psychology of intellectual character, which also displays how the latter can mitigate the possibility of the former, is the fact that neither is considered episodic. In the last section it was pointed out that the desires which lead to self-deception represent enduring psychological stratagems of the agent and are therefore dispositional in nature. As such these habits of the mind act as an "automatic filtering process" through which evidence and reasons are considered, so that those beliefs that serve the agent's interests, by conforming

[43] See Fairweather 2001, p. 73; Hookway 2001, p. 194.

to desires, are maintained.[44] A similar role was also ascribed to virtuous character in the first section. It was proposed that once the agent has fully integrated the virtues he achieves a firm and unchangeable character so that virtuous behavior follows naturally and without effort. His perceptions, deliberations and choices are a consequence of his enduring virtuous character and not situational factors. In fact, this occurs to such an extent that virtuous perceptions, deliberations and choices often occur automatically and unconsciously.[45] To acquire the virtues, and this includes the intellectual virtues, the agent must do such consciously and conscientiously, but once they are fully integrated they also become an automatic filtering process through which beliefs are formed.[46] For example, an agent will at first often have to make an effort to be open-minded, but through diligent effort and attempts to be open-minded this intellectual virtue will become fully integrated into his character. Once the virtue of open-mindedness becomes fully integrated the agent will be open to the claims of others so that the beliefs he does form will be automatically and unconsciously influenced by this intellectual virtue. The psychological mechanisms of intellectual character therefore mirror the psychological mechanisms of self-deception. The intellectual virtues also represent enduring "habits of the mind" that influence the agent in her belief formation typically, although not always, at an unconscious level.[47] The difference between the psychological mechanisms of self-deception and the intellectual virtues is that the former lead to false beliefs while the latter lead to true beliefs.

5. The Necessity of Intellectual Character in the Attempt to Overcome Self-Deception

By combining literature on self-deception with virtue psychology literature an understanding of how intellectual character can act as a therapeutic means to overcome self-deception has emerged. Through the development of one's intellectual character an agent can mitigate the influence of motivations and dispositions that lead to instances of self-deception to obtain and sustain true

[44] See Oksenberg Rorty 1996, p. 78.
[45] See Johnston 1988, p. 88; Sherman & White 2003, p. 36; Foley 2001, p. 224.
[46] See Sherman & White 2003, p. 43; Hookway 2003, p. 184; Hookway 2000, pp. 152–153, 155–156.
[47] See Hookway 2000, pp. 150, 152, 155–156; Audi 2001, p. 83.

beliefs. I want to now argue for a stronger claim: that the development of intellectual character fulfills a necessary therapeutic role in the attempt to overcome self-deception. This necessary therapeutic role is that intellectual character is required to assure agents that they are not self-deceived. If such a claim can be established, then the intellectual virtues would fulfill an indispensable role in any attempt to overcome self-deception.

The reason for the claim that intellectual virtue fulfills a necessary therapeutic role in any attempt to overcome self-deception is due to the nature of self-deception itself. Any agent who suffers from self-deception takes her beliefs to be true just as the agent who does not suffer from self-deception. Both can even cite reasons for their respective beliefs, even though one agent's set of reasons are false, or insufficient, while the other's are true. This is because, as touched on earlier, those agents who suffer from self-deception often rationalize the false beliefs they have. The problem that arises is that through mere introspection the agent can be duped by her own assessments and the reasons offered for her beliefs. There is always the possibility that when an agent says to herself "My belief is true because I can see that it is so, and because I can offer reasons for this claim," that she is in fact self-deceived. This is because introspectively things seem the same to both the self-deceived agent and the non-self-deceived agent.[48] The self-deceived agent is as convinced as the intellectually virtuous agent that her assessments of her beliefs are accurate and, ultimately, her beliefs are true. Consider the example, proposed by Hillary Kornblith, of Jack who is self-deceived in regard to his own mental states and how they influence his beliefs. Jack is paranoid and insecure, which often causes him to react with anger toward others. Upon introspection, though, Jack is unaware of his own anger, and how his insecurity and paranoia influence him to obtain and sustain false beliefs concerning what others think of him. If Jack engaged in introspective assessment of the mechanisms which influence his beliefs, and whether his beliefs concerning others are true or not, he would not be able to discern that his beliefs are false or that they were formed through misleading mechanisms. This is because Jack would continue to be influenced by self-deceptive mechanisms that lead him to believe that his beliefs concerning both others and his own mental states are true while they are not. Jack would continue to believe that he is not paranoid, insecure and angry, and that others speak negatively about him even though they do not. He would be

[48] See van Fraassen 1988, pp. 123–135.

just as convinced, upon introspective assessment, of the truth of his beliefs as the intellectually virtuous agent even though his beliefs are false and the mechanisms that lead to them are misleading.[49]

Since it is the case that from the introspective point of view the phenomenal experience of the self-deceived agent is indistinguishable from the phenomenal experience of the agent with true beliefs intellectual character becomes an indispensable therapeutic measure to overcome self-deception. More specifically, intellectual character is necessary to assure agents that they are not self-deceived. Through mere introspection an agent can be duped by her own assessments and not be able to detect that her beliefs are false and that she is self-deceived. Hence, she cannot rely on introspective assessment in order to determine whether her beliefs are true or not. Rather, she must rely on psychological dispositions that have been identified as truth-conducive. This is especially the case since self-deception occurs unconsciously. That is, not only is self-deception undetectable from an introspective point of view, but the mechanisms which lead to self-deception operate without the agent being aware of them. In fact, as previously pointed out, self-deceptive mechanisms have to be unconscious in order to be effective, for if the agent is aware of them she will ultimately not be duped.[50] It is due to these two reasons, then, that intellectual character is necessary to assure the agent that her beliefs are true. For if it is the case that agents can never distinguish between instances where they are self-deceived and instances where they are not then the only assurance, or guarantee, they can have that they are not self-deceived is that they have attempted to secure true beliefs, and avoid self-deception, through an attempt to be intellectually virtuous. As pointed out in the previous section, how agents can attempt to avoid self-deception is through developing their intellectual character. The motivations and dispositions identified as intellectual virtues not only compel agents to be careful and thorough when forming beliefs, they also replace those motivations and dispositions that lead to instances of self-deception. It is therefore only through developing one's intellectual character that an agent can assure herself that her beliefs are not the result of self-deceptive mechanisms. The virtues of intellectual character therefore offer the best protection against the imperceptible mechanisms that lead to self-deception, which means that intellectual character is necessary to

[49] See Kornblith 1998, pp. 50–52; van Fraassen 1988, pp. 123–135, 140, 144–145.
[50] Johnston 1988, p. 65–66, 70–76, 78, 87; Audi 1988, p. 94, 102–105, 109; Baier 1996, pp. 54–55; Deutsch 1996, p. 317; Fingarette 2000, pp. 46–49, 60–61, 65–66, 78, 98–99.

assure agents that their beliefs are true. Intellectual character is not always causally necessary to obtain and sustain true beliefs because non-intellectual preferences do not always exert their influence. This will become even more apparent with the next paragraph. Nonetheless, intellectual character still fulfills a necessary role in the attempt to acquire true beliefs, since it provides a guarantee for the agent that her beliefs have not been the result self-deceptive mechanisms.[51]

By claiming that the intellectual virtues are necessary to assure agents that their beliefs are not the result of self-deceptive mechanisms it must be made clear that the claim is not that the intellectual virtues ensure, or make certain, that the agent's beliefs are true. The intellectual virtues do not infallibly produce true beliefs. It is always possible that an agent could be completely intellectually virtuous and still not obtain true beliefs. The agent could be immersed in a misleading environment which could then make the acquisition of true beliefs impossible even if the agent is completely intellectually virtuous. Hence, the intellectual virtues cannot ensure, or make certain, that the agent's beliefs are true. What is meant, then, by proposing that the intellectual virtues provide a guarantee for the agent that her beliefs are true is that they guarantee that the agent's beliefs are not the result of self-deceptive mechanisms that could lead to false belief. The guarantee that intellectual character provides is therefore not infallible. Nonetheless, it is a guarantee that intellectual character

[51] Of course, a possible objection at this point is how do we reliably discover what character traits are intellectual virtues if self-deception is always a possible undetectable threat. Could we not also be deceived when identifying the intellectual virtues? If so, then it would seem that the intellectual virtues may provide very little assurance against self-deception. A complete response to such an objection cannot be achieved in the context of this article, but an outline of a response I developed elsewhere can be offered. There are two aspects of this response that are intimately connected. First, to reinforce the claim that the intellectual virtues are merely necessary to assure agents their beliefs are true and second to rely on a doxastic community in the identification of the intellectual virtues. In regard to the latter, the claim is that in order to identify the intellectual virtues one will have to rely on various legitimate epistemological methods established by the community. This is meant to solve problems with identifying the intellectual virtues, since one is not relying merely on introspection to identify the virtues. Hence, one does not have worry about how via introspection self-deception is undetectable. The question that then emerges is why must we rely on the intellectual virtues to assure us our beliefs are true if we ultimately rely on the community when identifying the virtues? This is where the claim that the intellectual virtues are merely necessary to assure us our beliefs are true and not sufficient becomes relevant. They are necessary for the reasons presented in this article; i.e., our beliefs are shaped by motivations/disposition and we therefore require truth-conducive motivations/dispositions to overcome them. But the intellectual virtues are not sufficient, since other epistemological practices also have to be reliable to secure true beliefs.

is necessary for since self-deception is undetectable from the introspective point of view. The agent cannot discern whether she is self-deceived via introspection, and therefore must be intellectually virtuous to assure herself that he beliefs are true.

It may be objected that intellectual character is not always necessary to assure agents that their beliefs are true for two reasons. First, it is likely the case that we can identify situations where self-deceptive mechanisms will not exert any influence and therefore intellectual virtue will not be required to overcome their influence. For example, when an agent forms the belief "There is a cat on the mat" based on immediate perceptions it does not seem that self-deception is a valid concern because misleading motivations and dispositions will likely not exercise their influence. Second, it could be proposed that a guarantee that one's beliefs are not the result of self-deceptive mechanisms could be provided via interaction with others. For example, if I want to discover if I am self-deceived in some particular situation all I may have to do is consult some other agent to aid in the identification of the truth-value of my beliefs. Both of these possible objections do not lessen the therapeutic value of intellectual virtue in many instances where self-deception is possible, but they nonetheless appear to display that intellectual virtue is not necessary to assure agents that their beliefs are true. In order to make this stronger claim, then, both of these possible objections must be addressed.

Beginning with the first objection, it is true that even the perceptions of agents can be shaped by self-deceptive mechanisms, but the above example appears to provide a clear-cut case where such mechanisms likely would not fulfill a role in belief formation. Consequently, the claim that intellectual character is necessary to assure agents that their beliefs are not the result of self-deception must be limited to situations where the latter is a valid concern. Fortunately, given what has been claimed concerning self-deception, such situations are easy to identify. Self-deception is a valid concern whenever it is possible for motivations and dispositions to influence belief formation, since the former are the impetuses of self-deception. When it comes to beliefs such as "A cat is on the mat" it is highly unlikely that any agent could be misled by her own motivations or dispositions, and therefore self-deception is not a valid concern and intellectual character is not required to overcome it. Nonetheless, the misleading influence of motivations and dispositions is a valid concern in many situations, and intellectual character would be necessary in such situations to assure agents that their beliefs are true due to the imperceptible

influence of such mechanisms. No attempt will be made to demarcate the possible situations where motivations and dispositions can influence the belief formation of agents, since such demarcation is not required. Rather, the following simple principle can be offered. Intellectual character is necessary to assure agents that their beliefs are true in all situations where it is possible for motivations and dispositions to mislead agents. By offering such a principle all possible situations where the misleading mechanisms of self-deception can exercise their influence are covered without having to engage in the task of identifying them specifically.

In regard to the objection that intellectual character is not necessary to assure agents that they are self-deceived, since consultation with others could also provide a guarantee, we have to keep in mind that the agent has to respond to the insights of others. That is, when confronted by a claim by some other that one is self-deceived the agent in question will have to accept the claims of others and especially accept them as true over his, or her, own introspective assessments. Now, whether an agent would accept the claims of another over his, or her, own introspective assessments, can really only determined empirically. We would have to investigate agents to see whether they would acquiesce in the judgments of others or not. Nonetheless, it does seem warranted to claim that intellectual character is still required in these situations to overcome self-deception, and this is again due to the nature of self-deception itself. Recall that the self-deceived agent is convinced by his own reasoning processes that certain things are true, and the fact that he is self-deceived is undetectable. When confronted by some other who claims that the agent is actually self-deceived the self-deceived agent will have to trust in the claims of this other over his own assessments. This means that the self-deceived agent will have to be more concerned with getting at the truth than confirmation of his own reasoning processes. The agent will have to be either motivated to get at the truth, or disposed toward the truth. Otherwise the agent will just trust in his own assessments and dismiss the comments of this other. If the agent lacked a concern for the truth, then he, or she, would still be more concerned to maintain the particular cover-story which is the impetus for his, or her, self-deception. For example, if Jack were confronted by one of his coworkers who attempted to tell Jack that he was insecure, or even paranoid, it is doubtful that Jack would be open to such remarks, and this is because Jack would be convinced by his own reasoning processes over the suggestions of others. Consequently, in order to even be open to the insights of others and

agent must be intellectually virtuous. Hence, intellectual character still would be necessary in such situations.

6. Conclusion

The goal of this article was to set out a therapeutic means that agents could employ to overcome self-deception. The therapeutic means advocated was the development of one's intellectual character. The case for intellectual character was made by first setting out the standard psychological theory of virtuous character. This theory was then connected to literature on self-deception and the intellectual virtues. What emerged was a description of how the intellectual virtues could act as a means to overcome self-deception. More specifically, the psychology of intellectual character appears to mirror to psychology of self-deception except that the focus of such character is the maintenance of true beliefs as opposed to a particular self-schema or cover-story. After these claims concerning the therapeutic value of intellectual character were advanced a stronger claim concerning the necessary therapeutic value of intellectual character was proposed. Specifically, it was claimed that intellectual character fulfills a necessary therapeutic role in combating self-deception due to the nature of self-deception itself. Agents who suffer from self-deception cannot detect its occurrence via introspection. Hence, the only assurance agents have that they are not self-deceived is that they are intellectually virtuous. Consequently, it seems that intellectual character fulfills an indispensable therapeutic role in the attempt to overcome self-deception.

REFERENCES

Adams, D. (2000). Virtue Without Morality. *Contemporary Philosophy*, *22*(3–4), 38–44.

Anderson, C., Srivastava, S., Beer, J.S., Spataro, S.E., & Chatman, J.A. (2006). Knowing Your Place: Self-Perceptions of Status in Face-to-Face Groups. *Journal of Personality of Social Psychology*, *91*(6), 1094–1110.

Annas, J. (1998). Virtue and Eudaimonism. *Social Philosophy and Policy. Winter*, *15*(1), 37–55.

Annas, J. (2003). The Structure of Virtue. In M. DePaul, & L. Zagzebski (Eds.), *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: Clarendon Press, 15–33.

Aristotle (1947). Nicomachean Ethics. In R. McKeon (Ed.), *Introduction to Aristotle*, Toronto: Random House of Canada Limited, 308–543.

Asendorpf, J.B., & Ostendorf, F. (1998). Is Self-Enhancement Healthy? Conceptual, Psychometric and Empirical Analysis. *Journal of Personality and Social Psychology*, *74*(4), 955–966.

Audi, R. (1988), Self-Deception, Rationalization, and Reasons for Acting. In B.P. McLaughlin & A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 92–120.

Audi, R. (2001). Epistemic Virtue and Justified Belief. In A. Fairweather, & L. Zagzebski (Eds.), *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. Oxford: Oxford University Press, 82–97.

Axtell, G. (1997). Recent Work on Virtue Epistemology. *American Philosophical Quarterly*, *34*(1), 1–26.

Axtell, G. (1998). The Role of the Intellectual Virtues in the Reunification of Epistemology. *The Monist*, *81*(3), 488–508.

Baier, A.C. (1996). The Vital but Dangerous Art of Ignoring: Attention and Self-Deception. In R.T. Ames, & W. Dissanayake (Eds.), *Self and Deception: a cross-cultural philosophical enquiry*. New York: State University Press, 53–72.

Code, L. (1984). Toward A "Responsibilist" Epistemology. *Journal of Philosophy and Phenomenological Research*, *XLV*(September), 29–50.

de Sousa, R.B. (1988). Emotion and Self-Deception. In B.P. McLaughlin, & A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Los Angeles: University of California Press, 324–341.

Deutsch, E. (1996). Self-Deception: A Comparative Study. In R.T. Ames, & W. Dissanayake (Eds.), *Self and Deception: a cross-cultural philosophical enquiry*. New York: State University Press, 315–326.

Fairweather, A. (2001). Epistemic Motivation. In A. Fairweather, & L. Zagzebski (Eds.), *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. Oxford: Oxford University Press, 63–81.

Fingarette, H. (2000). *Self-Deception: With a New Chapter*. Los Angeles: University of California Press.

Foley, R. (2001). The Foundational Role of Epistemology in a General Theory of Rationality. In A. Fairweather, & L. Zagzebski (Eds.), *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. Oxford: Oxford University Press, 214–230.

Foot, P. (1997). Virtues and Vices. In R. Crisp, & M. Slote (Eds.), *Virtue Ethics*. Oxford: Oxford University Press.

Gouinlock, J. (1996). *Rediscovering the Moral Life: Philosophy and Human Practice*. Buffalo, New York: Prometheus Books.

Hookway, C. (2000). Regulating Inquiry: Virtue Doubt and Sentiment. In G. Axtell (Ed.), *Knowledge, Belief, and Character: Readings in Virtue Epistemology*. New York: Rowman and Littlefield Publishers Inc., 149–160.

Hookway, C. (2001). Epistemic Akrasia and Epistemic Virtue. In A. Fairweather, & L. Zagzebski (Eds.), *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. Oxford: Oxford University Press, 178–199.

Hookway, C. (2003). How to be a Virtue Epistemologist. In M. DePaul, & L. Zagzebski (Eds.), *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: Clarendon Press, 183–202.

Hursthouse, R. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.

Irwin, T.H. (1996). The Virtues: Theory and Common Sense in Greek Philosophy. In R. Crisp (Ed.), *How Should One Live?: Essays on the Virtues*. Oxford: Clarendon Press, 37–56.

Johnston, M. (1988). Self-Deception and the Nature of the Mind. In B.P. McLaughlin, & A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Los Angeles: University of California Press, 63–91.

Kornblith, H. (1998). What is it Like to be Me?. *Australasian Journal of Philosophy*, *76*(1), 48–60.

Leon, M. (2002). Responsible Believers. *The Monist*, *85*(3), 422–436.

Louden, R. (1997). On Some Vices of Virtue Ethics. In R. Crisp, & M. Slote (Eds.), *Virtue Ethics*. Oxford: Oxford University Press, 201–216.

MacIntyre, A. (1981). *After Virtue: a study in moral theory*. Notre Dame, Ind: University of Notre Dame Press.

McDowell, J. (2003). Virtue and Reason. Virtue Ethics. Edited by Stephen Darwell. Malden MA: Blackwell Publishing. 121–143.

McKinnon, C. (1999). *Character, Virtue Theories, and the Vices*. Peterborough, ON: Broadview Press.

McLaughlin, B.P. (1988). Exploring the Possibility of Self-Deception in Belief. In B.P. McLaughlin, & A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Los Angeles: University of California Press, 29–62.

McLaughlin, B.P. (1996). On the Very Possibility of Self-Deception. In R.T. Ames, & W. Dissanayake (Eds.), *Self and Deception: a cross-cultural philosophical enquiry*. New York: State University Press, 31–51.

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

Merritt, M. (2000). Virtue Ethics and Situationist Personality Psychology. *Ethical Theory and Moral Practice: An International Forum*, *3*(4), 365–383.

Montmarquet, J.A. (1993). *Epistemic Virtue and Doxastic Responsibility*. Lanham, Maryland: Rowman and Littlefield Publishers, Inc..

Oksenberg Rorty, A. (1988). The Deceptive Self: Liars, Layers, and Lairs. In B.P. McLaughlin, & A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Los Angeles: University of California Press, 11–28.

Oksenberg Rorty, Amélie. (1996). User-Friendly Self-Deception: A Traveler's Manual. In R.T. Ames, & W. Dissanayake (Eds.), *Self and Deception: a cross-cultural philosophical enquiry*. New York: State University Press, 73–89.

Pence, G.E. (1984). Recent Work on Virtues. *American Philosophical Quarterly*, *21*(4), 281–297.

Reed, B. (2001). Epistemic Agency and the Intellectual Virtues. *The Southern Journal of Philosophy*, *XXXIX*, 507–526.

Roberts, R.C., & Wood, W.J. (2003). Humility and Epistemic Goods. In M. DePaul, & L. Zagzebski (Eds.), *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: Clarendon Press, 257–279.

Sanford, David H. (1988). Self-Deception as Rationalization. In B.P. McLaughlin, & A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Los Angeles: University of California Press, 157–169.

Sherman, N., & White, H. (2003). Intellectual Virtue: Emotions, Luck and the Ancients. In M. DePaul, & L. Zagzebski (Eds.), *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: Clarendon Press, 34–53.

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

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

Zagzebski, L. (1996). *Virtues of the Mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge: Cambridge University Press.

Self-Deception, Delusion and the Boundaries of Folk Psychology *

Lisa Bortolotti †
l.bortolotti@bham.ac.uk

Matteo Mameli ††
matteo.mameli@kcl.ac.uk

ABSTRACT

To what extent do self-deception and delusion overlap? In this paper we argue that both self-deception and delusions can be understood in folk-psychological terms. "Motivated" delusions, just like self-deception, can be described as beliefs driven by personal interests. If self-deception can be understood folk-psychologically because of its motivational component, so can motivated delusions. Non-motivated delusions also fit (to a large extent) the folk-psychological notion of belief, since they can be described as hypotheses one endorses when attempting to make sense of unusual and powerful experiences. We suggest that there is continuity between the epistemic irrationality manifested in self-deception and in delusion.

* Lisa Bortolotti would like to acknowledge the intellectual support of the Health and Happiness Research Cluster at the University of Birmingham and the financial support of a Wellcome Trust Research Expenses Grant ("Rationality and Sanity: Implications of a Diagnosis of Mental Illness for Autonomy as Self Governance" - WT092835MF) in the preparation of this paper.
† University of Birmingham, UK, and Macquarie Centre for Cognitive Science, Australia.
†† King's College London, UK.

1. Introduction

1.1. Self-deception

In a fairly uncontroversial characterisation, self-deception involves beliefs that are acquired and maintained in the face of strong counter-evidence and that are motivated by desires or emotions (Deweese-Boyd, 2010). Self-deception is thought to be a widespread phenomenon in the general (non-clinical) population. Here is an example of self-deception. In spite of having at her disposal evidence to the contrary, Sylvia believes that she failed the driving test because the examiner was prejudiced against female drivers. Her belief responds to the need of preserving a positive image of herself as a competent driver. Here is another example. In spite of having at her disposal evidence that powerfully indicates that her son robbed a bank, Janet still believes that he is innocent. Her belief protects her from the acknowledgement of a truth (that her son is guilty) that is painful for her to accept.

There are two opposed philosophical accounts of self-deception. According to the traditional account, self-deception is due to the *doxastic conflict* between the false belief one acquires ("I failed the test because the examiner was prejudiced against female drivers") and the true belief one denies ("I failed the test because I drove badly").

According to the rival account, self-deception is due to *biased treatment of evidence*: there is a bias against considering or gathering evidence for the true belief. Sylvia never acquires the belief that failing the test was due to her poor driving, because she neglects evidence that points in that direction.

In the doxastic conflict account of self-deception, one has two contradictory beliefs, but is aware of only one of them, because one is motivated to remain unaware of the other (e.g., Davidson, 1982; 1986). On this view, when one deceives oneself, one believes a true proposition ("I failed the driving test because I drove badly") and acts in such a way as to cause oneself to believe the negation of that proposition ("I failed the test not because I drove badly but because the examiner was prejudiced against female drivers").

Doxastic conflict is problematic for two reasons. First, it involves accepting that one can believe a proposition and its negation at the same time, and some philosophers think that this is impossible (leading to the *static* paradox of self-

deception). Second, it suggests that one can intend to believe something that one knows to be false – and thus be the perpetrator and victim of a deceitful strategy all at once (leading to the *dynamic* paradox of self-deception). The solution some traditionalists offer for these puzzles consists in postulating *mental partitioning*: According to Davidson, one can have two mutually contradictory beliefs as long as one does not believe their conjunction. The idea is that each of the two beliefs is in a different compartment or partition of the mind, and this prevents the subject from recognising and eliminating the inconsistency.

If this account of self-deception prevails, the scope for identifying an area of overlap between self-deception and delusion is limited, as many delusions (those that are not "motivated") cannot be plausibly characterised as the simultaneous holding of two contradictory beliefs. That said, compartmentalisation can be observed in many people with delusions, when one's delusional belief is insulated from one's other beliefs that conflict with it.

A more revisionist solution to the puzzles generated by the doxastic conflict view leads to endorsing the competing account of self-deception. This account emphasises the differences between deceiving another and deceiving oneself. In the latter case, when the deceiver and the deceived are the same individual, deception need not be intentional, and the deceiver need not believe the negation of the proposition that she is causing the deceived to believe. If Sylvia wanted to deceive her father about the reason why she failed the driving test, the conditions for her deceiving him would be that she knows that she failed the test because she drove badly, but she intends to make her father believe otherwise. Self-deception works differently. Sylvia deceives herself if she genuinely comes to believe that she is not to blame for failing the test.

Al Mele argues that the conditions for self-deception are as follows. First, one's belief is false. Second, one treats the evidence relevant to the truth of the belief in a motivationally biased way. Third, this biased treatment of the evidence is what causes one to acquire the false belief. And finally, the evidence available to one at the time of acquiring the belief lends better support to the negation of one's belief than to the belief one acquires (Mele, 2001, pp. 50–51).

The ways in which the treatment of evidence can be motivationally biased are varied: one might misinterpret the available evidence, focus selectively on

those aspects of the available evidence that support one's belief, or actively search for evidence that supports one's belief, without also searching for evidence that disconfirms it (Mele, 2009). Motivationally biased treatment of evidence is not just relevant to the acquisition of the false belief, but also to its maintenance. One holds on to the false belief because one keeps neglecting some of the relevant evidence.

This deflationist approach is explanatory, and avoids the so-called paradoxes of self-deception. More importantly for our purposes here, the approach highlights the continuity between the phenomenon of self-deception and other instances of epistemic irrationality in ordinary beliefs and in delusions.

1.2. Delusion

According to the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV, APA, 2000) and to the dominant theory of delusion formation in cognitive psychology (Coltheart, 2005), a delusion is a belief held with conviction and rarely challenged or revised. Delusional beliefs are typically implausible and unsupported by evidence (Bortolotti, 2010). As clinical delusions are symptoms of schizophrenia, dementia, and other psychiatric disorders, it is also important to add that they tend to disrupt day-to-day functioning (McKay, Langdon & Coltheart, 2005).

The content of some delusions (e.g., delusions of jealousy and of persecution) can be mundane and not dissimilar from that of false beliefs that we routinely find in the non-clinical population. Other delusions have more bizarre content. The Cotard delusion, for instance, is the belief that one is dead or disembodied. Delusions of infestation involve believing that insects are crawling under one's skin.

All delusions are currently thought to have an organic cause and are explained in neuropsychological terms, by reference to brain damage, perception failures, reasoning biases and cognitive deficits. But the formation of some delusions is also likely to include motivational factors. In a variety of anosognosia, people may fail to acknowledge the paralysis of a limb. This denial can be seen as a defence mechanism: one comes to believe that one's

arm is not paralysed because it is too hard to acknowledge that one permanently lost the use of one's arm.

There are similarities in the surface features of self-deception and motivated delusions – both phenomena typically involve beliefs that are badly supported by the evidence and that conflict with one's other beliefs or attitudes. Moreover, in both cases the beliefs are strikingly resistant to counterevidence. Further similarities can be found in the function of the beliefs: they serve to either preserve positive emotions, deny unpleasant or disturbing facts, or satisfy some other pressing psychological need.

Given what we know about self-deception and delusion, there are at least two features that distinguish the two: (a) whereas in self-deception beliefs are *always* motivated, not all delusional beliefs are motivated; (b) whereas delusions are symptoms of psychiatric disorders, are accompanied by other symptoms, and typically impair functioning, cases of self-deception are widespread in the non-clinical population.

Let us examine some of the differences and similarities in more detail.

2. The overlap between self-deception and delusion

There is no consensus on whether self-deception and delusion significantly overlap.[1] McKay and colleagues adopt the following approach to the issue:

> [S]ome (perhaps all?) delusional states may arise without self-deception, via processes that are not remotely motivated. [...] Conversely, self-deception may occur in a benign manner such that the resulting doxastic states do not sufficiently disrupt functioning to warrant the label *delusion*. (McKay et al., 2005, p. 315)

In this section, we will consider the notion of motivation as it applies to delusions and assess one interpretation of the view that self-deception is somehow more "benign" than delusion.

[1] See Mele 2009, Levy 2009, Davies 2009.

2.1. Motivation

Some delusions have been described as extreme cases of self-deception. They are considered cases of self-deception because they seem to have a defensive function. [2] They are considered extreme because they tend to disrupt functioning to a greater extent than standard cases of self-deception, and to result in the endorsement of more implausible and more tenacious beliefs.

One delusion that seems to fit this description is that of a delusion named "reverse Othello syndrome" which is the opposite of a delusion of jealousy. It consists in believing (incorrectly) that one's partner is faithful and in obstinately refusing to believe the contrary. The belief can plausibly be regarded as part of a defence mechanism against the suffering that the acknowledgement of the infidelity of one's partner would cause.[3]

Another example is anosognosia, the denial of illness. One well-known case is that of a woman (FD) who suffered from a right hemisphere stroke causing left hemiplegia (Ramachandran, 1996). FD could not move without a wheelchair and could not move her left arm. But when she was asked whether she could walk and engage in activities that require the use of both hands (such as clapping), she claimed that she could.

Vilayanur S. Ramachandran puts forward an explanation for this sort of cases. Behaviours giving rise to confabulations and delusions are an exaggeration of normal defence mechanisms which have an adaptive function. They allow one to preserve a positive self-image in the face of threatening negative events. The mind aims at maintaining a coherent system of beliefs that can guide behaviour. In normal subjects, the left hemisphere produces confabulatory explanations aimed at preserving the *status quo* ("My wife still loves me"; "My arm still moves"), but the right hemisphere detects discrepancies between the hypotheses generated by the left hemisphere and reality as it is perceived and it forces a revision of the belief system.

[2] See Ramachandran & Blakeslee 1998, Hirstein 2005.

[3] For a detailed description of one such case see Butler 2000 and the discussion by McKay et al. 2005, p. 313.

In patients with reverse Othello syndrome and anosognosia, the discrepancy detector in the right hemisphere malfunctions. A man suffering from the reverse Othello syndrome, for instance, claimed that his partner was faithful to him, whereas she had left him some time before (Butler, 2000). Thus, he failed to revise his belief in his partner's fidelity. In anosognosia, patients deny their own impairments even if they cannot help experiencing the effects of such impairments. In a conversation reported by Ramachandran (1996), FD asserted that her left arm was pointing at the doctor's nose, whereas her arm laid motionless.

Cases like these seem to support the claim that some delusions are motivated in much the same way as instances of self-deception are. It is prima facie plausible to regard delusions such as the reverse Othello syndrome and anosognosia as cases of self-deception, although the question can only be settled once we agree on an account of self-deception and find that it does fit the behavioural manifestations and the causal history of the beliefs. There are other delusions that deliver a boost to self-esteem: in erotomania, one believes that a person of higher status loves them (in secret); in delusions of grandeur, one believes to be a genius (unbeknownst to others); and delusions of persecutions often explain away instances of personal failure. For instance, a man can believe that he was fired because his colleagues conspired against him, whereas he was fired for his incompetence. Such delusions can also qualify as cases of self-deception in some circumstances. That said, it is important to stress that, even when motivational factors contribute to the formation of delusions, their presence is not sufficient to give rise to the delusion. Other factors (e.g., perception failures, brain damage, cognitive deficits, reasoning biases) need to be in place.

In addition, it is difficult to find any plausible role for motivational factors in the genesis of delusions such as the Cotard delusion, the belief that one is dead or disembodied. This delusion does not have an obvious adaptive function, and there is no fundamental role for motivational biases in the explanation of how the subject comes to hold or retain the delusional belief. Thus, the overlap

between self-deception and delusion can only be a partial one.[4] Motivational factors contribute to the formation of some but not all delusions, and only some delusions can be plausibly seen as the product of a psychological defensive mechanism. That said, there are still some interesting questions to answer. When delusions are motivated, are they *extreme* cases of self-deception? Are delusions in general more puzzling, less understandable, than standard cases of self-deception?

2.2. The boundaries of folk psychology

Recently, different conceptions of the relationship between delusion and self-deception have emerged. According to Keith Frankish (2011), both delusion and self-deception can be described by using the folk-psychological notion of belief, as long as the existence of different *types* of beliefs (roughly, behavioural dispositions and policies) is acknowledged. On his account, delusions and self-deception are continuous and motivational factors can contribute significantly to the formation of (at least some) delusions.

> Perhaps patients adopt delusions because they answer some emotional or other psychological need, rather than because they are probable. (Frankish, 2011)

Andy Egan (2009) also maintains that delusion and self-deception are alike, but takes the opposite line: *neither* can be accounted for satisfactorily by using the folk-psychological notion of belief. He argues that both delusion and self-deception should be regarded as *in-between* states. They represent how the agent takes things to be, and in this respect they are similar to beliefs. But they also convey how the agent wants things to be, and in this respect they are similar to desires. Egan suggests that they may be "besires", mental states that display at once features typical of beliefs and features typical of desires.[5]

[4] See McKay et al. 2005 and Davies 2009.

[5] Maura Tumulty (2011) and Eric Schwitzgebel (2011) also develop an account of delusions as in-between states.

In contrast to Frankish and Egan, Dominic Murphy (2011) highlights the discontinuity between delusion and self-deception. He maintains that instances of self-deception are understandable from a folk-psychological perspective, whereas delusions are not. We want to concentrate on Murphy's view here.

Murphy uses the following example to argue that self-deception is understandable from a folk-psychological perspective and to argue for the existence of a discontinuity between self-deception and delusions.

> It is easy to imagine parents who refuse to acknowledge that their child is guilty of a heinous crime, despite sufficiently overwhelming evidence to convince everyone else that the guilty verdict is the right one. Let's suppose that the child is guilty, and that everyone else believes this because it is the correct inference to make given the evidence. The mother of the guilty man has no relevant evidence not possessed by others, but the cost to her of admitting her child's guilt is too great. (Murphy, 2011)

In line with the accounts of self-deception we cited earlier, Murphy claims that self-deception involves having beliefs that carry emotional commitment and are fixed by personal interests rather than by a careful consideration of the available evidence. According to Murphy, such personal interests offer an acceptable explanation of both the conflict between belief and evidence and the "rigidity" of the belief. Murphy recognises that the epistemologist would consider desire-driven beliefs as not rational, but he thinks that they are an understandable manifestation of human nature.

Typically, delusions are also poorly supported by evidence and scarcely responsive to counter-evidence, but these features cannot (always or entirely) be explained by the influence of desires. Moreover, the content of delusions is somehow more "unbelievable" than the contents we routinely deceive ourselves about. There is nothing absurd in believing that a man is innocent, even if the belief is clearly false given the evidence at one's disposal, but there is something deeply unsettling about the content of many delusions. We take this to be the point of Murphy's next example.

> Let's consider another case, this time the (real) case of a person I'll call Ed. Ed was sleeping rough, and heard a tree in a park tell him that the park was a good place to stay. So Ed settled down for the night in the park. But a little later, the sprinklers in the park erupted and Ed was drenched. Thereupon Ed heard the

> tree tell him that it was very sorry: trees like to be watered, and the tree had not understood that Ed would not appreciate a good soaking. Ed accepted the tree's apology and went on his way. [...] Ed's traffic with trees is evidence of something mentally abnormal about him. (Murphy, 2011)

Murphy argues that delusion (but not self-deception) remains mysterious from a folk-psychological perspective.

> Ed [...] seems incomprehensible in folk terms; he is a suitable case for treatment. Delusions, I suggest, are attributed [...] when we run out of the explanatory resources provided to us by our folk understandings of how the mind works. (Murphy, 2011)

If motivational factors can contribute to the formation of at least some delusions, then Murphy's view about the discontinuity between self-deception and delusion *in general* is problematic. If the fact that a desire motivates a belief is sufficient for the folk-psychological understandability of such belief, no matter how impervious to evidence the belief turns out to be or how implausible, then only those delusions that are not motivated defy folk-psychological explanation. This view is compatible with the claim that, among delusions, those that are not motivated lack the folk-psychological understandability that both instances of self-deception and motivated delusions have. On this account, delusions occurring in anosognosia and the reverse Othello syndrome are amenable to folk-psychological explanation, while the Cotard delusion and delusions of infestation, as well as Ed's delusion about the talking tree, are not.

One may want to deny that the fact that a desire motivates a belief is sufficient for the folk-psychological understandability of such belief. On this view, even motivated delusions are discontinuous with self-deception because the role of motivational factors in their formation cannot provide an adequate explanation of the bizarre content of the resulting beliefs or of their imperviousness to counterevidence. Thus, a mother's love for her son can explain why she refuses to believe that he committed a crime, but one's desire not to be paralysed cannot explain the denial of the paralysis.

We find this latter view implausible. The denial of a serious physical impairment can surely be explained, at least in part, by reference to the relevant motivational states and is therefore folk-psychologically understandable, just

like the refusal of a mother to acknowledge that her son is guilty of a heinous crime. Both beliefs seem to have a defensive function and respond to a psychological need. There are many relevant similarities between the two cases epistemically, such as neglect and misinterpretation of evidence, implausibility and tenacity of the belief. Even considering the role of cultural norms, there seems to be no important difference: just like the acknowledgement that one's son is guilty of a crime, the acknowledgement of a serious and permanent impairment is something people have a reason to avoid. In both cases, from a folk-psychological perspective, it is not surprising that people sometimes believe what they would like to be true.

Let us now consider the more modest claim that *non-motivated* delusions are not understandable within the framework of folk psychology. We think this claim should be resisted too. In his analysis, Murphy focuses on the agent's reasons for her treatment of evidence. One could say that in self-deception (and in motivated delusions) evidence is neglected or misinterpreted *for a reason* (e.g., personal interests that are culturally recognisable) but in non-motivated delusions evidence is neglected or misinterpreted for no reason. When one considers the question whether one's right leg is paralysed, one might neglect to consider as relevant evidence the fact that one can no longer climb stairs. This evidence is neglected or discounted due to one's desire to believe that one's right leg is not paralysed. When one considers the question whether one is disembodied, one might neglect to consider whether one can move, talk and feel. This evidence is neglected or discounted but it is not clear why, as there seems to be no interest in believing that one is disembodied.

An issue that needs addressing is how demanding we take the folk-psychological notion of belief to be. Murphy claims that in the case of the mother deceiving herself about the innocence of her son the belief is in some respects faulty on epistemic grounds (i.e., not supported by or responsive to evidence) but not necessarily irrational. The belief does not conflict with behavioural generalisations that belong to our folk theory of the mind in an extended sense.

> These resources [provided to us by our folk understanding of the mind] do not just include folk psychology in the narrow sense of theory of mind, but a much richer body of beliefs and expectations about the role of hot cognition and personal interests in fixing belief [...] and the role of culture in shaping

> people's assumptions about what counts as legitimate evidence. (Murphy, 2011)

If the folk-psychological notion of belief were very demanding, and required that all legitimate beliefs be supported by and responsive to evidence, then both self-deception and delusion would fail to count as instances of belief. After all, according to the demanding interpretation of the folk-psychological notion of belief, desires do not interfere directly in the formation of beliefs at the expense of evidence – there are no *besires* in old-school folk psychology. Mental states are beliefs in virtue of their relationship to other beliefs (e.g., inferential relations), their relationship to behaviour (e.g., action-guiding potential), and especially their relationship to evidence. A mental state that is formed on the basis of partial evidence and that is scarcely responsive to new evidence would fall short of being a belief in a rigid, uncompromising framework.

But the folk-psychological notion of belief seems to be compatible both with the idea that in some cases desires play a role – even a direct role – in the formation of beliefs and with the idea that there are irrational beliefs. Folk-psychology can allow for the case of someone who believes that she has become disembodied after her experience of herself in relation to the rest of the world suddenly changed. After all, the relationship between unusual experiences and bizarre delusions is the relationship of evidence supporting a belief. Folk-psychology can also allow for Ed's delusional belief that the tree talked to him. The delusion is not without a reason if (we are elaborating the original example here) Ed heard voices in the park but saw nobody around. There are probably no *good* reasons to suppose that a tree is talking, but we do not need *good* reasons to establish the comparative claim with self-deception. Wanting one's son to be innocent is not a good reason to believe that he is.

In some respects, non-motivated delusions seem to be even more typical cases of belief than the case of the mother refusing to accept that her son is guilty. Not only instances of Cotard delusion, delusions of infestations and Ed's belief in talking trees are likely to interact with other beliefs and to guide action, as standard beliefs do, but such mental states are there to make sense of weird experiences with specific contents, experiences which would otherwise be inexplicable to those who are not acquainted with the form that psychotic symptoms can take. This is not the whole story, of course. Delusions are

irrational beliefs because they are not revised when counterevidence becomes available. But being scarcely responsive to some of the available evidence is one of the features delusions have in common with cases of self-deception, so the continuity between the two phenomena is not compromised.

To sum up, if folk psychology can allow for beliefs formed in order to satisfy a desire, and for beliefs that are poorly supported by and scarcely responsive to evidence, then it can also account for delusions.

3. Epistemic irrationality

Philosophers explain the status of self-deception and delusion differently. As we saw, some suggest that they are types of beliefs and some suggest that they are in-between states, which share some features with beliefs and other features with imaginings or desires. We would like to suggest that both self-deception and delusion are beliefs that violate norms of epistemic rationality. This claim is consistent with accepted definitions of both delusion and self-deception, but in order to make the claim meaningful one needs to formulate a notion of epistemic rationality and to distinguish it from other notions of rationality.

There are (at least) three forms of rationality that apply to belief-like states: *procedural*, *epistemic* and *agential* rationality (Bortolotti, 2009). Procedural rationality concerns the relationship between a belief and one's other beliefs. A clear violation of procedural rationality is inconsistency among one's beliefs. Epistemic rationality concerns the relationship between a belief and the available evidence. A clear violation of epistemic rationality is hanging on to a belief that has been repeatedly challenged by reliable evidence. Agential rationality concerns the relationship between a belief and behaviour. A clear violation of agential rationality is acting in a way that conflicts with one's belief.

Delusion and self-deception may violate more than one set of norms, but they are typically beliefs at odds with the evidence. Norms of epistemic rationality govern the acquisition, maintenance and revision of beliefs. Epistemically irrational beliefs can be badly supported by one's initial evidence or scarcely responsive to evidence that becomes available at a later stage. Evidence in support of the hypothesis that if the sky is red at night, then the weather will be good on the following day ("Red sky at night; shepherds

delight") should be weighed up by a rational subject before she takes the hypothesis to be true. Further, if evidence against the hypothesis becomes available after the hypothesis has been endorsed, and this evidence is sufficiently powerful, robust and so on, then the rational subject should come to doubt the previously formed belief, suspend judgement until new evidence becomes available, or reject the belief altogether.

As we previously discussed, forming a hypothesis ("My son is not guilty", "My left arm can move", "Insects are crawling under my skin") that is not supported by all the available evidence is not necessarily problematic. What seems problematic is to endorse such hypothesis as a belief and to hang onto the belief in the face of evidence that openly conflicts with it. Suppose the son confesses the crime to his mother and she discounts his confession. Suppose the patient continues to believe that he is not paralysed after the doctor explains to him in no vague terms what his situation is. In these circumstances, if the hypothesis is not shaken by such challenges but crystallises into a tenacious belief, then something is amiss.

As you may remember, Murphy agrees that the mother's belief in the son's innocence is epistemically irrational, as it is not supported by the evidence. Murphy also thinks that the mother's behaviour is folk-psychologically understandable and that we would not consider it as irrational *tout court*. One way of making the point is that the mother's belief is epistemically irrational but it is pragmatically rational for her to have that belief, in the sense that her life would be worse (all things considered) if she gave up the false belief and acknowledged that her son is indeed guilty. What is interesting is that *some* delusions also seem to work in the same way. By definition (at least the DSM-IV definition), delusions are epistemically irrational beliefs, but it is *not always* pragmatically irrational to be delusional. Imagine you are in Ed's shoes. The alternative to believing that the tree just talked to you is to concede that you hear voices and something is seriously wrong with you.

Aikaterini Fotopoulou explains that after brain damage or memory loss, personal narratives can be disrupted, undermining people's sense of coherence. This is often associated with increased anxiety and depression. Despite their poor correspondence with reality, delusional and confabulatory beliefs represent attempts to define one's self in time and in relation to the

world. Thus, they are subject to motivational influences and they contribute to preserving one's identity (Fotopoulou, 2008, p. 542).

People with delusions and confabulations construct distorted or false self-conceptions. They may claim that they live in a different place from the one where they live, or that they have a different profession or a different family from the one they do. The personal narrative they construct is not «anchored and constrained by reality» (Fotopoulou, 2008, p. 548). These distortions are exaggerated by brain damage or memory loss and exhibit self-serving biases – people reconstruct and interpret events in a way that is consistent with their *desired* self-image.

For the sake of creating a coherent self-image, people enhance their life-stories. In dementia, amnesia, anosognosia, people revisit their present and their past and attempt to establish continuity between the conception they had of themselves before the accident, the memory loss, the illness, and the conception of themselves afterwards. In this reconstruction, people tend to preserve a positive image whenever possible. Maintaining coherence with the previous self-image and promoting a more positive self-image take priority over preserving accuracy. The preference for internal coherence over correspondence has consequences.

The obvious disadvantage is that losing touch with reality can create a gulf between the person with the delusion and the surrrounding social environment. In the most serious amnesic conditions there is often a lack of "shared reality" between confabulators and the people who were once closest to them, which can be very distressing for patients and their families (Fotopoulou, 2008, p. 560). In general, given that delusions are ill-grounded and often bizarrely false, people with delusions are not likely to be believed and taken seriously by others.

These observations on distorted memory and enhanced self-narratives in the clinical population affected by delusions and confabulations apply also to self-deception. In this respect, the oft-perceived gap between delusions as a clinical, pathological phenomenon and self-deception as a homely form of epistemic irrationality seems to shrink. Non-clinical subjects also tend to present their current selves in a way that is both coherent with their past, and largely favourable (Wilson & Ross, 2003), giving rise to common instances of self-deception. Self-deception can also result in a gulf between one's version of

reality ("The examiner was biased against female drivers", "My son is innocent") and the version of reality other people share and accept. The clinical case helps us realise that the development of self-narratives is always a *reconstructive* exercise, even when memory and reasoning are not seriously compromised.

> A self-conception is not just the set of facts we might learn about ourselves; it is an interpretation of these facts within which values are prioritized, emotions are labeled, and attitudes are endorsed or rejected. Importantly, the process of organizing what we know about ourselves into a self-conception is partly a creative or constructive process. (Tiberius, 2008, p. 116)

The fact that delusion and self-deception involve irrational beliefs does not mean that they bring no benefits at all. As previously suggested, delusion and self-deception may have some pragmatic benefits. They protect the subject from undesirable truths, keep anxiety and depression at bay, and help maintain a coherent sense of self (Bortolotti & Cox, 2009). They allow people to keep constructing self-narratives when personal information is not available, and to construct self-narratives that are more positive than the evidence suggests, preserving self-esteem in the face of serious set-backs.

4. Conclusion

In this paper, we revisited a topic that has engaged philosophers of mind in recent years, the potential overlap between self-deception and delusion. Our purpose was to show that, although the two phenomena are distinct, there is considerable continuity between them. We argued against the claim that delusion does not fit the folk-psychological notion of belief, whereas self-deception does. If instances of self-deception can be understood folk-psychologically, then delusions can too.

By appealing to the notion of epistemic irrationality, we suggested that in self-deception and delusion the relationship between belief and evidence is unhealthy, which causes delusional and self-deceiving people to form inaccurate accounts of themselves and of the events that concern them. As a result, the delusional and the self-deceived may reject the view of themselves or of reality that people around them share in order to preserve a positive and coherent sense of self.

REFERENCES

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision.*

Bayne, T., & Fernandez, J. (Eds.) (2009). *Delusions and Self-Deception: Affective Influences on Belief Formation.* Hove: Psychology Press.

Bortolotti, L. (2009). *Delusions and Other Irrational Beliefs.* Oxford: Oxford University Press.

Bortolotti, L. (2010). Delusion. In E.N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)* <http://plato.stanford.edu/archives/fall2010/entries/delusion/>.

Bortolotti, L., & Cox, R. (2009). Faultless ignorance: strengths and limitations of epistemic definitions of confabulation. *Consciousness & Cognition*, *18*(4), 952– 965.

Butler, P.V. (2000). Reverse Othello syndrome subsequent to traumatic brain injury. *Psychiatry: Interpersonal and Biological Processes*, *63*(1), 85–92.

Coltheart, M. (2005). Delusional belief. *Australian Journal of Psychology*, *57*(2), 72–76.

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

Davidson, D. (1986). Deception and Division. In J. Elster (Ed.), *The Multiple Self.* Cambridge: Cambridge University Press, 79–92.

Davies, M. (2008). Delusion and Motivationally Biased Belief: Self-Deception in the Two-Factor Framework. In T. Bayne, & J. Fernandez (Eds.) *Delusions and Self-Deception: Affective Influences on Belief Formation.* Hove: Psychology Press, 71–86.

Deweese-Boyd, I. (2010). Self-deception. In E.N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*,

<http://plato.stanford.edu/archives/fall2010/entries/self-deception/>.

Egan, A. (2008). Imagination, Delusion, and Self-Deception. In T. Bayne, & J. Fernandez (Eds.), *Delusions and Self-Deception: Affective Influences on Belief Formation*. Hove: Psychology Press, 263–280.

Frankish, K. (2011). Delusions, Levels of Belief, and Non-doxastic Acceptances. *Neuroethics*, DOI: 10.1007/s12152-011-9123-7.

Hirstein, W. (2005). *Brain fiction: self-deception and the riddle of confabulation*. Cambridge, MA: MIT Press.

Levy, N. (2009). Self-deception without thought-experiments. In T. Bayne, & J. Fernandez (Eds.), *Delusions and Self-Deception: Affective Influences on Belief Formation*. Hove: Psychology Press, 227–242.

McKay, R., Langdon, R., & Coltheart, M. (2005). "Sleights of mind": Delusions, defences and self-deception. *Cognitive Neuropsychiatry*, *10*(4),305–326.

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

Mele, A. (2009). Self-deception and delusion. In T. Bayne and J. Fernandez (Eds.), *Delusions and Self-Deception: Affective Influences on Belief Formation*. Hove: Psychology Press, 55–70.

Murphy, D. (2011). The folk epistemology of delusions. *Neuroethics*, DOI: 10.1007/s12152-011-9125-5.

Ramachandran, V.S. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: some clues from anosognosia. *Medical Hypotheses*, *47*(5), 347–362.

Ramachandran, V.S., & Blakeslee, S. (1998). *Phantoms in the Brain: human nature and the architecture of the mind*. London: Fourth Estate.

Schwitzgebel. E. (2011). Mad belief? *Neuroethics*, DOI: 10.1007/s12152-011-9127-3.

Tiberius, V. (2008). *The Reflective Life: Living Wisely With Our Limits*. New York: Oxford University Press.

Tumulty, M. (2011). Delusions and not-quite beliefs. *Neuroethics*, DOI: 10.1007/s12152-011-9126-4.

Wilson, A., & Ross, M. (2003). The identity function of autobiographical memory: Time is on our side. *Memory*, *11*(2), 137–149.

Remnants of Psychoanalysis. Rethinking the Psychodynamic Approach to Self-Deception*

Massimo Marraffa †
marraffa@uniroma3.it

ABSTRACT

This article reflects on the phenomenon of self-deception in the context of the psychodynamic approach to defense mechanisms. Building on Giovanni Jervis' criticism of psychoanalysis, I pursue the project of a full integration of that approach in the neurocognitive sciences. In this framework, the theme of self-deception becomes a vantage point from which to sketch out a philosophical anthropology congruent with the ontology of neurocognitive sciences.

1. Debunking the Unconscious

According to the Cartesian doctrine of the perfect transparency of the mind, the latter is simply *res cogitans*, and thought, its defining attribute, is explicated in terms of awareness (*conscientia*). As he writes in the *Replies to the second set of objections*: «I use the term 'thought' to include everything that is within us in such a way that we are immediately aware [*conscious*] of it» (Descartes, 1641/1988, p. 113). Here there is no margin for the notion of unconscious mentality: «there can be nothing within me of which I am not in some way aware» (1641/1988, p. 77).[1] Many philosophers will follow him.

During the second half of the 19th century, however, the unconscious insistently claims its own rights. Neurologists and psychiatrists had drawn

* This essay is one of a series of papers (see Marraffa, 2011a,b,c, forthcoming) in which I have been trying to reconstruct and develop Giovanni Jervis' work on three themes: the unconscious, consciousness, and identity.

† University Roma Tre, Italy.

[1] Interpretations of Descartes's account of consciousness (like everything else in his philosophy) differ significantly. Here we are following John Cottingham's authoritative interpretation (see, e.g., Cottingham, 1988, p. 153).

attention on phenomena such as convulsive "great" hysteria, dissociative fugue, or multiple personality disorder, which could hardly be reconciled with the consciousness-dependent conception of mind originating from Descartes. After ruling over most of the philosophical views concerning introspective self-knowledge, Cartesian mentalism was shaping the early experimental psychology. It is comprehensible, then, that philosophers, psychologists and neuroscientists were bewildered about phenomena that appeared to be mental but went beyond the sphere of awareness and conscious control.

As Livingstone Smith (1999) has convincingly shown, during the second half of the 19th Century two strategies were adopted to reconcile the existence of supposed unconscious mental phenomena with the consciousness-dependent conception of mind. The first option consisted in denying that such phenomena were genuinely *unconscious*; the evidence for unconscious mental states was reinterpreted as evidence for the possibility of a "dissociation" or "splitting" or "doubling" of consciousness, namely «the total possible consciousness may be split into parts which coexist but mutually ignore each other» (James, 1890/1950, p. 206). The second option consisted in denying that such phenomena were genuinely *mental*; the evidence for the existence of unconscious mental states was reconceptualised as evidence for *neurophysiological dispositions* for genuinely (i.e., conscious) mental states.

The two strategies are still options in current Anglo-American philosophy. John Searle has recast the dispositionalist approach to unconscious mental states, whereas the "partitionist" approach to self-deception has revived the dissociationist option.[2] Let us focus on the latter.

Self-deception is traditionally viewed as a temporary impairment of *normal* belief-forming processes.[3] In addition, it is seen as a phenomenon that gives rise to two paradoxes: the "static" paradox and the "dynamic" one (see Mele, 1997). The partitionist approach to self-deception aims to dispel the static paradox by dividing the agent into two (or more) sub-agents, whose minds include the belief that *p* and the belief that non-*p* respectively. And it tries to

[2] See Livingstone Smith (1999, chapters 14–16). At p. 141 the author interestingly notes that Searle's idea that "the ontology of the unconscious is strictly the ontology of a neurophysiology capable of generating the conscious" coincides with what the physiologist Ewald Hering had claimed in 1870.

[3] «Normal», that is, «from the analytic philosopher's point of view, where the central important epistemic goal seems to be the generation of true beliefs» (Sage, forthcoming).

dissipate the dynamic paradox by postulating that the deceived sub-agent cannot access the deceiving sub-agent's activities.

Donald Davidson is often considered the main "partitioner", but actually his partitionism is very moderate. Davidson thinks that when one runs across such (apparent) absurdities of reason as akrasia or self-deception, the personal psychology framework is not to be given up in favor of the subpersonal one, but rather it must be enlarged or extended so that one can find somewhere else the rationality set out by the principle of charity. On this perspective, the division of the mind is a *metaphoric* device to coherently describe (within the personal-level explanatory framework) a phenomenon (self-deception) that otherwise would be unintelligible. As Davidson puts it, a mental division is nothing but "a metaphorical wall" that keeps two contradictory beliefs separate. Consequently, we do not need to postulate «two minds each somehow able to act like an independent agent»; it is sufficient to imagine «a single mind not wholly integrated; a brain suffering from a perhaps temporary self-inflicted lobotomy» (Davidson, 1998, p. 8).

A stronger version of partitionism – appropriately defined as "homuncularist" by Johnston (1988, p. 63) – was suggested by David Pears. Here the psychological partitioning is no longer Davidson's metaphorical wall; rather it is a conceptual reconstruction of Freud's second topographical model of the mind. The psyche is divided into a "main system" and a "sub-system"; the latter is «built around the nucleus of the wish for the irrational belief» and it is «organized like a person» (Pears, 1984, p. 87). Now, as Jon Elster points out, Pears ascribes to the sub-system an internal rationality («it is an efficient, quasi-altruistic manipulator of the main system» (Elster, 1984, p. 1388). And this implies that the sub-system both has all sorts of propositional attitudes regarding the main system, and it is «able to weigh and choose between alternative ways of satisfying the wishes of the main system» (*ibid.*). But then, Elster very properly concludes, «these requirements almost inexorably imply that the subsystem must have some kind of consciousness» (*ibid.*).[4]

Thus we find again here that same need of reabsorbing the discourse on the unconscious into the discourse on consciousness that led some fin-de-siècle researchers to reinterpret the evidence for unconscious mental states as evidence for the possibility of a *dédoublement* of consciousness. On the basis

[4] In this connection, see the entry "Topique" in Laplanche & Pontalis (1967), where it is rightly pointed out that Freud's second topographical model of the mind has an anthropomorphic character.

of such a conclusion, it might appear strange that Davidson's (1982) and Pears' (1982) theories of self-deception are offered as defenses of Freud's theory. For is it not true that Freud put forward a subpersonal psychology (a "metapsychology") that aimed to go beyond the psychology of consciousness? As a matter of fact, the psychological partitioning approach really captures an aspect of Freud's theory of the unconscious; but unfortunately, it is an aspect that – as we will now see – is the main limit of Freud's theory.

2. Troubles with the Freudian Unconscious

When, in the last decade of the 19th Century, Freud intervenes in the dispute on the unconscious, he takes sides against the predominant "consciousness-centric" mentalism and in favor of the reality of occurrent and intrinsically unconscious mental events. The Freudian theory of the unconscious is, therefore, *programmatically* against the psychological partitioning insofar as this treatment of self-deception remains – as we have argued – within an introspective-intuitive psychology of consciousness. The problem is that, *as a matter of fact*, Freud failed to get himself out of that psychology.

Freud's view of the relationship between conscious and unconscious mind is the ground of the conception of consciousness still dominant in the current non-specialized (and sometimes philosophical) culture. The common culture about the mind is a largely psychodynamic culture. Of course, this culture represents an advance on the Cartesian thesis of the transparency of the mind, which informs the image of human beings typical of 19th Century middle class ethics, against which Freud polemicized. According to Victorian anthropology the essence of the human being in its highest expression, that of "the civilized gentleman", lies in the full control exerted by self-consciousness over mind and behavior. But if this anthropology was dominated by the idea of consciousness (and conscious agency) so that a person could say «If I did it, it is *evidently* because I chose it, because I wanted to do it», in the average culture of the mind one realizes that people are tossed about by instances which they do not always control very well, so that sometimes anyone can legitimately say «I did it but I hardly know why», thus implying that one is at least somewhat at the mercy of one's own psychological world (Jervis, 2011, p. xxi).

The psychodynamic culture of the mind, therefore, makes an important correction to the idea of a psyche consisting in conscious and self-transparent

intentions; but it is only a partial correction. In the average culture of the mind, influenced by psychoanalytic psychodynamics, holds what was the most evident limitation of the Freudian view of the unconscious: the definition of the unconscious is still given "by difference" from – and in some respects also depending on – the definition of consciousness; the latter is taken as a self-evident, primary quality of the mind, although it is then criticized and "downsized" in comparison with the traditional idealistic conception. Accordingly, the Freudian mind «is still dominated by the model of the conscious elaboration of choices, and within it the unconscious plays its tricks here and there, but nothing more» (Jervis 2011, p. xxii). Like all the psychoanalytic ideas, the Freudian unconscious is a sort of enlargement or extension of the everyday commonsense psychological framework, which is a psychology of consciousness.[5]

(One might remark that in recent years a number of philosophers, influenced by Davidson, have argued that the extension of our ordinary psychological conception of mind is a strength of the psychoanalytic theory.[6] This move is the basis of a defence of psychoanalysis against well-known epistemological challenges.[7] But as will become clear in the next section, the metaphilosophy inspiring this essay rejects any form of antinaturalism that deprives science of the domain of the mental construed as a space of reasons rather than causes. In our perspective, the right question to ask is how and to what extent the folk-psychological conceptual framework should be rectified in light of neurocognitive sciences, in which – pace Kandel (2005) – not much of psychoanalytic theory can be integrated.)

[5] See Manson, who rightly notices that in Freudian psychoanalysis the hypothesis that consciousness is not a necessary condition of mentality is applied only to «a few exceptional or anomalous cases (slips, neuroses etc.), and relative to a conception of mind as paradigmatically conscious» (2000, p. 163). And see also O'Brien and Jureidini, who argue that «[j]ust as much as the mental entities that parade across our consciousness, those that inhabit the [psychoanalytic] unconscious are [...] "personal-level" phenomena [...] in terms of their contents at least, unconscious ideas are conjectured to be indistinguishable from their conscious counterparts in all things save the fact that consciousness of them is absent» (2003, p. 143).

[6] These philosophers think that «the grounds for psychoanalysis lie [...] in its offering a unified explanation for phenomena (dreaming, psychopathology, mental conflict, sexuality, and so on) that commonsense psychology is unable, or poorly equipped, to explain» (Gardner, 1999, p. 684).

[7] In this perspective, «[p]sychoanalytic explanations, like ordinary psychological explanations, may be exempt from the epistemological and methodological standards of experimental science» (Manson, 2003, p. 179).

Furthermore, it should be emphasized that if Freud still preserves the primacy of consciousness, this is not because he develops a phenomenology, which has this consciousness as a methodological source of its investigation of reality. In other words, Freud does not develop a theory of subjectivity at all, and not even a theory of knowledge that starts from subjectivity. The very concept of subjectivity, or experientiality, was not part of Freud's toolkit. His way of theorizing more than neglecting the subjective dimension, tends to translate it into objective terms, like a collection of mechanisms and energies. Described with a very original and sometimes informally imaginative idiom, places, forces and events in the Freudian mind (ego, id, super-ego, censorship, libido, cathexis, and so on) never cease to be markedly reified. All Freud's thought is characterized by the influence of positivism: the mind is a world of facts, or even objects. But these objects are more metaphorical than real, more imagined than described. It might be said that the Freudian psyche is a collection of imaginary interfaces of the nervous system; his theory of mind is the psychologization of a very personal speculative-introspective neurology. During the development of his thought after 1900, the way in which the psychological dimension becomes autonomous from the neurological one – from which Freud had started – never becomes detached from an objectivistic (and indeed one could say: subjectively objectivistic) way of conceiving the mind (see Jervis, 2011, pp. xxii-xxiii).

Freud then claims to describe in accordance with a positivistic objectivism neurobiological mechanisms as constitutive of the mind. But although these mechanisms aim to explain many dimensions of the affective and emotional life, they are not supposed to explain consciousness. In spite of the unconscious and its energy-driven instincts, the Freudian adult self-consciousness is once more "assumed" or "given". So we find in his work the persistence of a partial endorsement of the Cartesian model of the subject, which postulates a perturbing corporeal influence on the mind ("les passions de l'âme") but also rigidly safeguards a primary (and in Descartes transcendent) principle of human rational awareness.

Briefly, psychoanalysis is a personal psychology that is masked as subpersonal psychology.[8]

[8] This is the gist of the famous objection that Sartre makes to Freud, when he rejects the idea of a censor mechanism (see Sartre, 1943, pp. 87–88). If Sartre's criticism is translated into the idiom of the explanatory levels, we obtain the claim that psychoanalysis (and, we add, the homuncularist partitionism) moves from the personal level to the sub-personal one, «but it ends up having to re-

3. Consciousness as Seen from the Bottom Up

Today the response of a psychologist to the above-discussed discontents over psychoanalysis would be claiming that cognitive science can count on a genuinely subpersonal level of analysis – the information-processing level, wedged between the personal sphere of phenomenology and the subpersonal one of neural facts – which no longer takes consciousness as an unquestionable assumption, as a non-negotiable given fact. The cognitivist mind is a process of construction and transformation of *representations*; and a mental representation is an explanatory hypothesis in a computational theory of cognition; it is a structure of information (somehow encoded in the brain), which is individuated exclusively in terms of intra-theoretical functional criteria.[9] Cognitive scientists introduce mental representations to explain intelligent behavior not differently from what physicists do when they posit entities like spin, charm and charge.

Cognitive science, therefore, challenges the traditional link between consciousness and intentionality, thus opening a conceptual space to build a consciousness-independent conception of the unconscious. As Dennett (1991) puts it, first the cognitive scientists develop a theory of intentionality that is independent of and more fundamental than consciousness – a theory that treats equally any form of unconscious representational mentality; and then, they proceed to work out a theory of consciousness on that foundation. In this perspective, consciousness is «an advanced or derived mental phenomenon» and not, as Descartes wanted, «the foundation of all mentality» (Dennett, 1993, p. 193).

In viewing consciousness no longer as something that explains, but rather as something that needs to be explained, analyzed, dismantled, cognitive science amends the Freudian thought on the basis of Darwinian naturalism. Differently from Freud's introspective-intuitive description of the unconscious, cognitive science follows Darwin's anti-idealistic methodological lesson and proceeds *bottom-up*, attempting to reconstruct how the complex psychological functions underlying the adult self-conscious mind evolve from the more basic ones. This attempt does not appeal to our introspective self-

import the personal level at the sub-personal, in order to get all the sub-personal bits to do what they are supposed to do» (Gardner, 2000, pp. 100–101).

[9] In this context, the phenomenological aspects are considered to play a role in the mental life only insofar as they can be explicated in representational terms. See Lycan (2008).

knowledge, but instead appeals to those disciplines that investigate the gradual construction of self-consciousness as introspective reflexivity (Jervis, 2007, p. 152).

In this bottom-up perspective, it becomes possible to distinguish different forms of consciousness, which range from the simplest environmental monitoring to sophisticated forms of self-monitoring.

First, studies in cognitive ethology and developmental psychology tell us that neither infants under one year of age, nor most animals have the slightest idea – not even a confused one – of their own existence. They are conscious in the sense that they are able to form a series of representations of objects and operational plans of action, and hence to interact with persons and things in flexible but not self-conscious ways.

Second, some species take a step beyond the basic interactive monitoring of the environment that characterizes the simple consciousness of all animals. Great apes like chimpanzees, and in our species infants from 15–18 months of age, can be said to attain a state in which they are able to make a clear distinction between their own physical bodies and the surrounding environment. (More precisely, they first become capable of physical self-monitoring, i.e., focusing attention on the material agent as the (physical) executor of actions; and then their bodily self-monitoring comes to completion as the objectivation of a *proper body*, and thus as a rudimentary self-consciousness.)

Finally, it is only in human species, and only after the age of 3 or 4, that some unconscious psychological functions come to self-present themselves in accordance with the modes of self-conscious subjectivity. This is human consciousness in the traditional sense: self-consciousness as introspective recognition of the presence of the virtual space of the mind, separated from the other two primary existential spaces, i.e., the corporeal and extracorporeal spaces (see Jervis, 2007, p. 153).

By unearthing the non-primary but derived, constructed and partial character of self-consciousness, the cognitivist bottom-up approach can be regarded an *anti-phenomenology*, i.e., a critique of the subject, of its alleged givenness. The term "anti-phenomenology" was coined by Paul Ricoeur, who used it to define Freud's methodological approach. Ricoeur calls this approach «an *epoché* in reverse» (1970, p. 118). Freud's inquiry into the unconscious is an *epoché* in reverse because «what is initially best known, the conscious, is suspended and becomes the least known» (*ibid.*). Consequently, whereas the

phenomenological tradition pursues a reduction of phenomena *to* consciousness, Freud's methodological approach aims at a reduction *of* consciousness: the latter loses the Cartesian character of first and last certainty, which stops the chain of methodical doubts on the real, and becomes itself an object of doubt. However, as we have seen above, in reality Freud's inquiry into the unconscious really starts from consciousness taken as given; and this makes psychoanalysis a dialectical variant of phenomenology (Jervis, 2011, pp. xxxi-xxxii). In contrast, cognitive science, fortified by a consciousness-independent concept of intentionality, rightly qualifies as an anti-phenomenology.

This allows us to estimate all the distance that separates the new cognitivist mentalism from the "consciousness-centric" mentalism that characterized the early experimental psychology, and from which the Freudian theory of the unconscious failed to disentangle itself. Under the influence of positivism, the introspectionist psychologists reified subjectivity. In most cases the 19th century experimental psychology did not understand consciousness in an experiential or subjective sense, but as an objective field, within which it was supposed to be possible to break down mental contents, viewed as measurable objects. As an antidote against the positivistic attempt to reify phenomenological experience, information-processing psychology provides us with a repertoire of tools to penetrate the nature of self-conscious subjectivity, making it possible to conceive phenomenological data not as tangible and measurable objects, but as the result of the self-presentation of unconscious psychobiological functions.[10]

4. Disunity and Opacity

Against the Cartesian conception of introspective consciousness as transparent awareness of our own mental processes and contents, Freud suggested that it is a construction packed with self-deceptions.[11] This theme

[10] The term "psychobiological function" points to my endorsement of teleofunctionalism, according to which «what makes a given type of mental state the type that it is, is its distinctive job or function within its subject's psychobiology» (Lycan & Neander, 2008).

[11] Although Freud does not offer an account of self-deception as such, his writings reveal very important characteristics of it that are not acknowledged by his "analytic" interpreters. See, e.g., Hållén (2011), who discusses self-deception in the context of Freud's writings and criticizes Davidson's and Gardner's analyses of the phenomenon.

can be considered the "strength" of Freud's conception of the unconscious.[12] A legacy, however, that can be capitalized provided one is willing to replace Freud's personal-level notion of dynamic unconscious with the new unconscious of neurocognitive sciences.

To begin with, Freud describes a *primary* self-deception when he sets up a contrast between the composite, non-monadical character of the mind and its unitary phenomenology. In the "feeling of our own ego" (*Ichgefühl*), Freud writes, the ego (*das Ich*) «appears to us as something autonomous and unitary, marked off distinctly from everything else» (1930/1962, p. 13). But this appearance is *deceptive*: as a matter of fact the ego is heterogeneous, heteronomous and secondary. In fact, it is the organized part of the id, which is totally unconscious and unstructured pulsionality, with which the ego is continuous «without any sharp delimitation» and «for which it serves as a kind of façade» (*ibid.*). Consequently, the ego is *both* the partial structure of the disparate psychological functions, *and* the apparatus that has, inter alia, the function of presenting to consciousness the immediate but illusory certainty of the existence of «a mind that is fully conscious of itself, integrated, unitary, rational and controllable» (Jervis, 2011, p. 43).

Today many behavioral, neuroimaging and computational investigations offer robust evidence for the composite, non-monadical nature of the mind-brain. In particular, since the early 1980s a *modularist* conception of the mind-brain has loomed large in psychology and neuroscience. The concept of modularity is to be placed in the framework of the crisis of the "pyramidal" conception of the mind, historically associated with the hierarchical conception of the cerebral functions dating back to the 19th Century. Against this view of mental life as a homogeneous and hierarchically-ordered field, ruled by consciousness and rationality, Noam Chomsky and David Marr have envisioned – in the wake of R. Mountcastle, D. Hubel and T. Wiesel's studies on the specializations of neurons – a less unitary, homogeneous, and hierarchically-ordered mind: its structure is *modular*, consisting of a bunch of distinct subsystems, that perform highly specific functions independently of each other (see Carruthers 2006).

Thus the neurocomputational architecture of our minds is composite and de-centralized, not monadic; and its appearing to consciousness as unitary is –

[12] This point is made by Jervis (2007, pp. 149–50). On Jervis' reconstruction of Freud's theory of the unconscious, see Marraffa (2011a,b).

as Freud suggested – a primary self-deception. To take just one famous example, in Dennett's narrative theory of personal identity the unitary consciousness of "self" is a short-lived "virtual captain" that occurs when a coalition of semi-independent, often domain-specific information processing mechanisms implemented in far-flung regions of the brain, has temporarily prevailed over other coalitions in the contest for the control of such activities as self-monitoring and self-reporting. Each of these short-lived phenomena is the 'me' of the moment, and they are connected to earlier fugacious selves by the autobiographical memory.[13] But then, "[i]f the temporary coalition of conscious states that is winning at the moment is what I am, is the self, each temporal chunk of 'self' is likely to be found in different parts of the brain from other such chunks and there will be many [neural correlates of consciousness] of unified consciousness in many different places" (Brook & Raymont, 2009, §7).

Freud's hypothesis that the presentation of the unconscious to introspective consciousness gives rise to deceptive beliefs about ourselves has found a rich source of evidence in the experimental social psychology literature on cognitive dissonance and self-attribution. Famously, in the experiments reviewed by Nisbett and Wilson (1977) the causes of the participants' behavior and attitudes (judgements, preferences and choices) were inaccessible motivating factors (e.g., subliminal cognitive inputs). However, when explicitly asked about the motivations (causes) for their behavior or attitudes, the subjects did not hesitate to sincerely affirm their plausible motives. The two psychologists explained this pattern of results by arguing that the subjects did not provide reports of real mental states and processes due to a direct introspective awareness; rather, they drew on repertoires of *rationalizations* seen as acceptable by mutual consent, and from time to time applied them, more or less stereotypically, to what needed to be justified.

Nisbett and Wilson's article was published in 1977. In the following thirty years the experimental literature on self-knowledge has increased substantially. Research in social and group psychology (e.g., Wilson, 2002; Wegner, 2002), in cognitive neuroscience (e.g., Hirstein, 2006) and cognitive neuropsychiatry

[13] Here I am following Brook & Raymont's (2009, §7) account of Dennett's view of the neural architecture of unified consciousness. The authors make clear that not any kind of autobiographical memory will be appropriate here; it must be «memory of the having, feeling, or doing of earlier experiences, emotions, actions, and so on» (Brook & Raymont, 2009, §5.2).

(e.g., Carruthers 2011) makes a very strong case for some version of a «symmetrical or self/other parity account of self-knowledge» (see Schwitzgebel, 2010, §2.1). According to the theory-theory version of this account, the attribution of psychological states to oneself (first-person mindreading) is an interpretative activity that depends on mechanisms that exploit the same folk theory of mind used to attribute mental states to other people. Such mechanisms are triggered by information about mind-independent states of affairs, essentially the target's behavior and/or the situation in which it occurs. The claim is, then, that there is a functional symmetry between first-person and third-person mentalistic attribution.

On this perspective, self-knowledge is not introspection insofar as this is construed as a direct access to the *causes* of our attitudes and behavior. In most cases of everyday life the explanation of the motives (being able to say "why") plays a *justificatory* role rather than a *descriptive* one. "Introspection" is then a misnomer for the capacity to explain one's behavior and attitudes *ex post* as the products of a rational and autonomous agent.

Moreover, Carruthers (2011) has extended this reappraisal of introspection beyond the causes of attitudes, to the attitudes themselves. According to Carruthers, we do not access propositional attitude events like judging and deciding via introspection; our only form of access to them is via self-interpretation, turning our mindreading faculty upon ourselves and engaging in unconscious interpretation of our own behavior, circumstances and sensory events like visual imagery and inner speech. Carruthers, therefore, develops a version of the symmetrical account of self-knowledge in which the theory-driven mechanisms underlying first- and third-person mindreading can count not only on observations and recollections of one's own behavior and the circumstances in which it occurs/occurred, but also on the recognition of a multitude of perceptual and quasi-perceptual events. Thus introspective consciousness comes out still more drastically downsized. True, agents have a sort of "perceptual" introspection. But this information is nothing but the raw material for an interpretative activity in which the access to the inner life is the access to an imaginary dimension generated by the folk-psychological theories driving the mindreading system.

Finally, Carruthers (2008) has put forward the hypothesis that Descartes' belief in the self-transparency of the mind reflects an innate feature of the human mind. According to this hypothesis, the mindreading system operates with a model of its own access to the rest of the mind that is essentially

Cartesian, assuming that subjects know, immediately and without self-interpretation, what they are experiencing, judging and intending. This assumption may have great heuristic value, greatly simplifying the mindreading system's computations. Moreover, as Wilson (2002) suggests, it may make it easier for subjects to engage in various kinds of adaptive self-deception, helping them build and maintain a positive self-image (a suggestion that anticipates the topic of the next section).

5. A Baconian Approach to Defense Mechanisms

Self-consciousness as introspective reflexivity is largely a theory-driven activity of re-appropriating the outputs of unconscious cognitive processing – this is the main point of the preceding section. Now what I want to emphasize is that such an activity is characterized by self-apologetic defensiveness: the description-narration of one's own inner life gets organized on the basis of the fundamental need «to construct and defend a self-image endowed with at least a minimal solidity» (Jervis, 1997, p. 33).

So we finally come to grips with the theme of defense mechanisms. But in view of neurocognitive sciences, the way in which Freud and his successors in the psychodynamic tradition have dealt with the study of psychological defenses must undergo a radical revision.[14]

We have already said that Freud's conception of the unconscious suffers due to an insufficient emancipation from the Cartesian model of the mind and the relationship between reason and passions. Descartes traced the errors of judgment and conduct back to the emotional, visceral, impulsive-instinctual, "animal" sphere of the body – this allowed him to safeguard the assumption of a primary (and for him transcendent) principle of human rational awareness. This ideology persists in non-specialist culture in the present day. The Cartesian faith in reason as producer of truth, the idea that what is clear and

[14] The notion of psychological defense is a psychoanalytic notion par excellence, whereas self-deception is a classical philosophical topic. Nevertheless, as McKay, Langdon and Coltheart (2009) rightly point out, defense mechanisms typically involve self-deception. Rationalization is a good example. The classic fable of the fox and the grapes, which nicely illustrates the "rationalization of disengagement", is a defensive maneuver through self-deception (see Elster, 1983). A variation of the sour grapes paradigm consists in rationalizing certain situations of intrapsychic conflict such as the cognitive dissonance investigated by Festinger in the 1960s, which illustrates the rationalization of "engagement".

distinct cannot be false, and that errors are essentially a sort of derailment due to drive-visceral interferences, is implicit also in Freud's system of thought.

But the Cartesian conception of error pays heavy tribute to philosophical predecessors of the modern era. It had already found an implicit refutation in Francis Bacon's work, which traces the errors of judgment and conduct back to the forms of doing and knowing that are peculiar to the psychological essence of human beings. In Bacon, contrary to Descartes, the conscious and rational mind naturally produces errors: the human understanding, he writes, «is like an uneven mirror receiving rays from things and merging its own nature with the nature of things, which thus distorts and corrupts it» (1620/2000, p. 41). We could say, in current terms, that Bacon sees the mind's errors, illusions, and self-deceptions as intrinsic to the ordinary cognitive-affective processes. It is on these grounds that he claims the necessity of a system of tests through which our spontaneous tendency to make errors is "dug out" and rectified by the method of research, on the base of a rigorously empiristic methodological principle.[15]

It is this Baconian perspective that has been taken by research traditions such as psychology of thought and social psychology. Thus, for example, social psychology tells us that stereotypes, the dynamics of prejudice, the structurally unreliable or diverting nature of many programmatic and principled avowals, are structures of bad faith which originate from cognitive mechanisms underlying the etiology of social attitudes. In such a perspective, then, self-deception can no longer be conceived as a *pathology* of belief-formation, the temporary crisis of a fundamentally rational agent, which can be explained only in terms of a non-rational psychological sphere, consisting of passions, instincts, emotions, and which can be clearly demarcable from the workings of our self-conscious rationality.[16] Now self-deception is a natural inclination of the human mind, a property inherent to belief-formation mechanisms (see, e.g., Bayne and Fernández, 2009, pp. 5–6).

This gives rise to a *reinforcing overturning* of the psychodynamic questioning about defenses. Now «the aspects of ambiguity, self-deception, and [...] sufferance of human life» can no longer be conceived as «interferences

[15] See Jervis (1993, pp. 122–123), who refers to Paolo Rossi's works on this topic (see, e.g., Rossi 1968).

[16] On the other hand, the folk concept of emotion is not a natural kind, i.e., a category that groups together a collection of objects whose properties are correlated by virtue of a causal mechanism that makes it possible projection and induction. On this point, see Griffiths (1997).

that are restrictively connected to affective and emotional factors (and hence negatively affecting a self-conscious rationality safeguarded as primary)»; they are to be seen as aspects «globally constitutive of the mind and behavior» (Jervis, 1993, p. 302). What needs to be explained, then, is not «how and why some defense mechanisms exist, but rather how all the structures of knowledge and action are by themselves, integrally, a matter of defenses» (Jervis, 1993, p. 301). In short, defense mechanisms are mechanisms that permit us to think and act. Although their most manifest function is that of protecting from anxiety, defense mechanisms are the primary instruments for setting up order in the mind. Consequently, we are now able to capture something that is already in Freud but which the Cartesian model prevented him from thoroughly articulating: the defensive processes are something more than bulwarks against anxieties and insecurities that perturb the order of our inner life; actually, defense mechanisms are the very structure of the mind – the Freudian ego itself is a defense. Here are the roots of the clinical theme of the fragility of the ego, namely that intimate personal insecurity that seems to originate from insufficiencies in the primary relationship between mother and child (what Michael Balint termed "basic fault"). But the theme is much wider, and it has to do with a philosophical anthropology that is congruent with the ontology of neurocognitive sciences.

Then let us ask ourselves: who is the subject of a dynamic psychology based on the cognitive-science ontology of unconscious psychobiological functions? After undergoing the above-mentioned "reinforcing overturning", the ideas of the unconscious and defense mechanisms have no longer the function of downsizing the traditional image of a subject with a primary identity and force; on the contrary, they certify the non-existence of a human subject of that kind. What, more than anything else, defines the real human subject is its original lack of ontological consistency. Unlike Descartes' soul-like ontologically guaranteed consciousness-substance, the image of the subject that cognitive sciences deliver us is that of a multiplicity of functions that in presenting themselves to consciousness exhibit a "façade" made of representations of the self. But it is a façade that is inextricably marked by *bad faith*; that is, «it is something inauthentic and bi-dimensional, i.e., "shallow", which tends to pass itself off – in accordance with our insuppressible tendencies to self-deception – as the 'solid', or 'deep', structure of the person» (Jervis, 2011, p. 45).

These dynamics of the representations of the self are the dynamics of the *subjective identity*, namely the consciousness of the self as description of the

self. I know that I exist insofar as I know that I exist "in a certain way", as describable identity, constant through changes. This theme is well captured by William James: every day, at each awakening, I find again my own body and my own mind, namely myself as known identity – «Each of us when he awakens says, Here's the same old self again, just he says, Here's the same old bed, the same old room, the same old world» (James, 1890/1950, p. 334).

However, self-consciousness as finding oneself again as known identity, as feeling of being-here as being-here in a certain way, is a *precarious* acquisition, continuously constructed by the subject and constantly exposed to the risk of not being here. If the subject's self-description becomes uncertain, she soon feels that the feeling of existing vanishes. This can occur for various reasons: because of a sudden breakdown of self-esteem; on the occasion of unexpected emotional upheavals; in some cases of psychoses or loss of memory; when the continuity of the tissue of our sociality is broken, as it can happen when one is suddenly thrown in some dehumanizing total institution (Jervis, 2011, pp. 131–132).

It is therefore the precariousness of this description of identity that makes intelligible the primary defensiveness of the self-constructing subject. The human subject constitutes itself as a repertoire of defensive maneuvers that must cope with its ontological insubstantiality. It could be said that the mind achieves its appearance of unity in the act of mobilizing tricks against the threat of its breaking down. And it is worth noting that such an activity – aimed to defend one's own self-describability and, indissolubly, the cohesiveness of one's own self-conscious consistency – is not restricted to an individual, psychodynamic dimension, i.e., to the intrapsychic defenses and the interpersonal maneuvers to which we appeal in the relationship with other people and our social environment. For it also has a collective, anthropological dimension, where the defenses consist in the construction of a system of references, in part symbolic and ritual, which give meaning to one's own being in the world.[17]

[17] See Jervis (2011, p. 92), who is building on Ernesto de Martino's seminal work on the "crisis of presence". This is a breakdown in the sense of self that occurs in the confrontation with death, in cases of psychological dissociation, alienation, and «loss of subjectivity, i.e., of one's ability to act on the world rather than simply to be a passive object of action» (Saunders, 1993, p. 882). According to de Martino, overcoming the crisis of presence is the fundamental task of culture.

6. Conclusion

Self-deception can be seen as a paradox of rationality only within the framework of the Cartesian conception of a self-transparent, unified and integrated self. Once we abandon the Cartesian theory of the subject, and invoke the subpersonal framework of neurocognitive sciences, self-deception is, in its primary form, a way of alluding to a mismatch: the description of the self as a description of identity is irreducibly out of phase, i.e., heterogeneous, with respect to the much more composite reality of the neurocognitive unconscious. Our mind is not self-transparent, i.e., essentially it eludes us, and also "deceives" us; and it deceives us just starting from its pseudo-transparency and consciential pseudo-unity. The mind contains non-truth-tropic cognitive mechanisms that generate the reassuring effect of a unitary egoic subjectivity that is master of the contents of consciousness. This effect is a "façade" whose deceptive character will be denied if human beings must feel their own autonomy, and thus experience themselves as persons. Or equivalently, the activity of narrative re-appropriation of the products of the unconscious cognitive processing is ruled by the fundamental need «to construct and protect a self-image endowed with at least a minimal solidity, and that is, in practice, solid enough to confirm to ourselves that we exist without dissolving ourselves» (Jervis, 1997, p. 33). This is the framework within which we can understand the construct of defense mechanisms, and with it all variety of self-deception.

REFERENCES

Bacon, F. (1620/2000). *The New Organon*. Edited by L. Jardine, & M. Silverthorne. Cambridge: Cambridge University Press.

Bayne, T., & Fernandez, J. (2009). Delusion and Self-Deception: Mapping the Terrain. In T. Bayne, & J. Fernandez (Eds.), *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*. New York: Psychology Press, 1–20.

Brook, A., & Raymont, P. (2010). The Unity of Consciousness. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/fall2010/entries/consciousness-unity/>.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Carruthers, P. (2011). *The Opacity of the Mind*. Oxford: Oxford University Press.

Cottingham, J. (1988). *The Rationalists*. Oxford: Oxford University Press.

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

Davidson, D. (1998). Who is fooled? In J. Dupuy (Ed.), *Perspectives on Self-Deception*. Cambridge: Cambridge University Press, 1–18.

Descartes, R. (1641/1988). Author's replies to the second set of objections. In J. Cottingham, D. Murdoch, & R. Stootho (Eds.), *The Philosophical Writings Of Descartes*. Cambridge: Cambridge University Press, vol. II, 93–120.

Dennett, D. C. (1991). *Consciousness Explained*. New York: Little, Brown & Co.

Dennett, D.C. (1993). Review of J. Searle, The Rediscovery of the Mind. *The Journal of Philosophy*, *60*, 193–205.

Elster, J. (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Elster, J. (1984). Managing to deceive ourselves. *The Times Literary Supplement*, *4261* (November 30), 1388.

Freud, S. (1930/1962). *Civilization and its Discontents*. Translated and edited by J. Strachey. New York: Norton.

Gardner, S. (1999). Psychoanalysis, contemporary views. In R.A. Wilson, & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge (MA): MIT Press, 683–685.

Gardner, S. (2000). Psychoanalysis and the personal/sub-personal distinction. *Philosophical Explorations*, *3*(1), 96–119.

Griffiths P. E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: Chicago University Press.

Hållén, E. (2011). *A Different Kind of Ignorance: Self-Deception as Flight from Self-Knowledge*. Uppsala University, PhD dissertation.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation*. Cambridge (MA): MIT Press.

James, W. (1890). *The Principles of Psychology*. New York: Dover, 1950.

Jervis, G. (1993). *Fondamenti di psicologia dinamica*. Milan: Feltrinelli.

Jervis, G. (1997). *La conquista dell'identità*. Milan: Feltrinelli.

Jervis, G. (2007). The unconscious. In M. Marraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the Mind*. Berlin: Springer, 147–158.

Jervis, G. (2011). *Il mito dell'interiorità*. Tra psicologia e filosofia. Edited by G. Corbellini, & M. Marraffa. Turin: Bollati Boringhieri.

Johnston, M. (1988). Self-deception and the nature of mind. In B.B. Mclaughlin, & A. Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 63–91.

Kandel, E. (2005). *Psychiatry, Psychoanalysis, and the New Biology of Mind*. Arlington (VA): American Psychiatric Publishing.

Laplanche, J., & Pontalis, J.-B. (1967). *Vocabulaire de la psychanalyse*. Paris: Presses Universitaires de France.

Livingstone Smith, D. (1999). *Freud's Philosophy of the Unconscious*. Dordrecht: Kluwer.

Lycan, W. (2008). Representational Theories of Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/fall2008/entries/consciousness-representational/>.

Lycan, W.G., & Neander, K. (2008). Teleofunctionalism. *Scholarpedia*, *3*(7), 5358.

Manson, N. (2000). A tumbling-ground for whimsies? The history and contemporary role of the conscious/unconscious contrast. In T. Crane, & S. Patterson (Eds.), *The History of the Mind-Body Problem*. London: Routledge, 148–168.

Manson, N. (2003). Freud's own blend: functional analysis, idiographic explanation, and the extension of ordinary psychology. *Proceedings of the Aristotelian Society*, *2*, 179–195.

Marraffa, M. (2011a). Precariousness and bad faith. Giovanni Jervis on the illusions of self-conscious subjectivity. *Iris*, *3*(2), 171–187.

Marraffa, M. (2011b). Jervis e la genealogia nascosta della coscienza umana. In G. Jervis, *Il mito dell'interiorità. Tra psicologia e filosofia*. Edited by G. Corbellini, & M. Marraffa. Turin: Bollati Boringhieri, XI–LVI.

Marraffa, M. (2011c). Jervis, De Martino e il mito dell'interiorità. *Rivista di Filosofia*, *2*, 241–259.

Marraffa, M. (forthcoming). Troubles with self-consciousness. Jervis on introspection and defense mechanisms. *Medicina nei secoli*, *23*(1), 2012.

McKay, R., Langdon, R., & Coltheart, M. (2009). "Sleights of mind": Delusions and self-deception. In T. Bayne, & J. Fernandez (Eds.), *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Hove: Psychology Press, 165–185.

McWilliams, N. (1994). *Psychoanalytic Diagnosis*. New York: Guilford Press.

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

Mele, A.R. (2009). Delusional Confabulations and Self-Deception. In W. Hirstein (Ed.), *Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy*. Oxford: Oxford University Press, 139–157.

Nisbett, R.E., & Wilson, T.D. (1997). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

O'Brien, G., & Jureidini, J. (2002). Dispensing with the dynamic unconscious. *Philosophy, Psychiatry and Psychology*, *9*(2), 141–153.

Pears, D. (1982). Motivated irrationality, Freudian theory and cognitive dissonance. In R. Wollheim, & J. Hopkins (Eds.), *Philosophical Essays on Freud*. Cambridge: Cambridge University Press, 279–288.

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

Ricoeur, P. (1970). *Freud and Philosophy. An Essay on Interpretation*. New Haven and London: Yale University Press.

Rossi, P. (1968). *Francis Bacon: From Magic to Science*. London: Routledge.

Sage, J. (forthcoming). The Evolutionary Basis of Self-Deception. <http://www4.uwsp.edu/philosophy/jSage/Sage%20Evolutionary%20Basis%20of%20Self-Deception.pdf>.

Sartre, J.-P. (1943). *L'être et le néant*. Paris: Gallimard.

Saunders, G. R. (1993). "Critical Ethnocentrism" and the ethnology of Ernesto De Martino. *American Anthropologist*, *95*(4), 875–893.

Schwitzgebel, E. (2011). Introspection. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/fall2010/entries/introspection/

Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge (MA): MIT Press.

Wilson, T.D. (2002). *Strangers to Ourselves*. Cambridge (MA): Harvard University Press.

Commentary

Much Ado About Truth: On Seduction, Deception, and Self-deception

Clancy Martin *
martincw@umkc.edu

Alan Strudler †
strudler@wharton.upenn.edu

ABSTRACT

Seduction, while not an unqualified good, is something most people enjoy and desire, especially when it's done in the right way. However seduction almost always involves techniques of deception and self-deception, and risks trust and other moral goods we associate with truthfulness. We examine various accounts of seduction, and focus in particular on two texts: Kierkegaard's *Diary of the Seducer* and Shakespeare's *Much Ado About Nothing*. We do not draw any strong conclusions about the moral status of seduction; rather, we use the phenomenon to explore the complicated philosophical and psychological terrain of how truth, trust, deception and self-deception may interact in a process with which we are all intimately familiar.

1. Introduction

Think about who you would most want to seduce you. Now ask: by seducing you, need this person wrong you? Only the puritanical will say yes. Seduction done by the right person at the right time and in the right way can be a dream come true. Romantic love – one of the highest goods – rarely occurs without seduction. Seducing rightly is tricky, however. It requires, we will maintain,

* University of Missouri, Kansas City, USA.
† University of Pennsylvania, USA.

the complicity of the seduced. And once such complicity occurs, it becomes puzzling why one might say that seduction occurs. We hope to make this puzzle more strongly felt. In doing so, we hope to shed some light on self-deception, deception of others, and the relationship between the two.

Seduction can be crude, and much worse. Sexual predators seduce, scar, often ruin children. Charismatic ministers seduce their parishioners, cajoling them to donate their wealth to projects that only bolster ministerial egos. Demagogues seduce the public with empty promises of tax-free increases in public benefits and blood-free victories in war. All these seducers wrong their victims not only by the harm that they cause but also by the way that they cause it. Seducers have at best an unstable relationship with truth, and at worst, a hostile relationship. A seducer has plans for the prospective seduced, but does not reveal them, and more likely hides or lies about them. Seduction, it will be plain, often involves deception. Nevertheless seduction can be morally praiseworthy or at a minimum quite fine, even if it is not always so.

At the most general level it may seem easy to say how seduction involves wrong. It may violate a person, playing with his will in ways that disrespects him, by duping him into attaching his affection to an illusion. But the wrong in seduction cannot be reduced to deception. Seduction often also involves preying on a person's vulnerabilities by stimulating his desires in destructive ways that he has a hard time controlling, something as close to offering alcohol to a struggling alcoholic as deceiving someone. A seducer can understand a person's weaknesses and wrongfully exploit those weaknesses without ever deceiving that person. Much seduction, however, seems deceptive at its core. The most successful seductions may involve both the exploitation of the weakness of the seduced and deception. Interestingly, both these strategies involve an attack on a person's autonomy.

2.

2.1. Seduction and falsehood

Take a simple and common case of seduction. When Smith lies to Jones about his love for her, he seduces her through falsehood. Why not just say that what makes seduction wrong in these cases, is deception, and that what makes deception wrong, when it is wrong, is the violation of autonomy that deception involves? Contemporary Kantians, including Barbara Herman, Christine

Korsgaard, Thomas Hill, and Onora Oneill take some variant of this autonomy line against deception, and the line seems easily extended to cover seduction (Herman, 1993; Korsgaard, 1986; Hill, 1984). Sarah Buss rejects the appeal to autonomy that the neo-Kantians make in their discussions of deception. She says that deception's apparent clash with autonomy provides no "key" to explaining why deception is sometimes wrong (Buss, 2005). Indeed, she says that claiming an essential clash between autonomy and deception involves metaphysical error (Buss, 2005, p. 213). We disagree with Buss on the explanatory value of autonomy and on her diagnosis of metaphysical error, and will explain why. On other matters, we agree with her. We agree that seduction and hence deception are sometimes morally fine. Indeed, we think that this agreement may transcend the three of us, that at least some of the Kantians Buss selects as her foil are more open to welcoming deception and seduction than Buss acknowledges. So in this paper we investigate when deception is wrong and why, using seduction as our base case of deception. In the end we hope to shed light on the complexity in the idea of respecting autonomy that forms the heart of so many analyses of the wrong in deception.

Before ascending to theory, it will be good to get something more about which to theorize. Here we join Buss in focusing on Johannes' seduction of Cordelia in Kierkegaard's *Diary of the Seducer* (1843/1987). Johannes' seduction is unusual both in its motivation and technique. These details will soon matter for our argument, but not yet. Focus now on the big picture. To cause Cordelia to fall in love with him, Johannes deceives her about his love for her and his intentions for their future. When she does fall for him, he abandons her. Can this have anything to do with Cordelia's autonomy? We think so. We think that he flouts Cordelia's autonomy. It is surprisingly hard to say just how he does this.

Cordelia chooses to allow herself to fall in love with Johannes. She could have resisted, let us suppose, but did not. In fact it is crucial for Cordelia that she understands herself as actively engaged in the process of Johannes's courtship of her: the depth of passion she comes to have for him is the result, she thinks, of her having freely decided to commit herself to him. Cordelia views her choice as an expression of her deepest values, an expression of her autonomy. But it was not. She had been deceived – conned – in ways that undermine her prospects for making a choice that expresses her deepest values. Buss never comes to terms with the con, we will argue. Cordelia's love was aimed at a man who, it turns out, did not meet the description of Johannes

as she understood it. She trusted him to be truthful about himself, but he was not. Her choice to allow herself to fall in love was not autonomous because it flowed from a con, departing from her deepest values, which could only have been realized by a man satisfying a description very different than a true description of Johannes. Here we believe that we echo Barbara Herman, who says that deception is wrong because it causes a person to act on desires that are not hers «all the way down» (1993, p. 228). But Herman's words are vague, and so far ours are, too. The best analysis of "all the way down," we will maintain, requires reflection on the structure of conning, which we soon endeavor to present.

But first back to Buss, and her limits. She offers a causal interpretation of Herman's idea of "all the way down" that is consistent with, but we think not required by Herman's text, and that we find uncharitable. This interpretation provides that a person acts on desires that are hers all the way down when these desires are not caused by anything external to her. Because Cordelia's desire was caused by Johannes, it was not hers all the way down. This interpretation relies on a notion of causation as mere influence; it is hopeless, we think. Typically a person's desires are influenced by something external to the person, as Buss observes. Every person who falls in love is influenced by the object of her affection. So on this causal interpretation of "all the way down," acts of love are never autonomous, nor are virtually any other acts. Although Buss rightly mocks this causal view, she lacks clear textual evidence that it is Herman's view she mocks, and it does not strike us as a view that Herman (or anyone) would be likely to defend. Unfortunately, Herman herself does little to say what she means by her vexing but provocative phrase, "all the way down."

2.2. Buss's argument and the problem of autonomy

We think that there is a credible interpretation of the idea that a person's act is not autonomous unless it expresses a desire that is hers all the way down. It is the idea that, at least for important choices, an act expresses autonomy only if it expresses one's deepest desires relative to the object of one's choice. On our view, wrongful deception always wrongly impinges on autonomy. Our argument for this view will be indirect. We aim to show how our argument is needed to explain why Buss's argument fails. Then we will develop an alternative.

Here is Buss's main argument that wrongful deception lacks an essential link to impinging on autonomy:

(1) A person acts autonomously if he makes the choice that he sees as justified in the circumstances.
(2) Deception does not prevent a person from making a choice that he sees as justified in the circumstances.
(3) Deception does not undermine autonomy. Whatever is wrong with deception must be something else.[1]

The lure in this argument comes from its hard-nosed stance regarding autonomy. Autonomous choice is resilient. It occurs even when based on false belief, and this seemingly has implications for the relevance of deception to autonomy: if false belief generally does not undermine autonomy, then why should false belief caused by deception undermine autonomy?

Against Buss, we will argue that in the right circumstances, though not in all circumstances, false belief caused by deception undermines autonomy: sometimes enough false belief undermines autonomy. And we will argue that deception may undermine autonomy in ways that cannot be understood simply in terms of the false belief that it causes: sometimes factors other than quantity matter. (Think of false beliefs that are tailored to one's weaknesses.) Still, we recognize that Buss is on to something when she suggests that autonomy may survive false belief. As Columbus first sailed across the Atlantic, he falsely believed, because he trusted his day's science, that he might run into India. He made the decision to head to India based on the best evidence available, and saw himself as justified in making it. His decision was autonomous, or at least not deficient in autonomy, on Buss's account, because it satisfies (1).

We doubt that satisfying (1) carries the weight that Buss suggests. If the reasons for which one sees one's choice as justified in the circumstances are sufficiently defective reasons, it may undermine the autonomy of one's choice. Some reasons for doubting (1) can be derived from skepticism about the work of Harry Frankfurt, who identifies a free action as one that issues from desires that mesh together in the right way.[2] While Frankfurt's focus was a connection between properly meshing desires and free action, Buss's focus is on properly meshing beliefs and autonomous action. Doubts can be raised about the

[1] For Frankfurt's defense of this view against critics, see Buss & Overton 2002.
[2] See for example, Stump 2002.

importance of meshing. Variants of cases that have been offered as counterexamples to Frankfurt's account of freedom can be used against (1), we believe. Perhaps these cases would not move Buss, as they do not move Frankfurt. They move us. We will argue, moreover, that the cases prove stronger against Buss than against Frankfurt. And we will later offer a very different argument against (1).

2.3. Frankfurt's dog

Consider "dog". Imagine that you love dogs but that your spouse, an ingenious neuroscientist, hates them. So she secretly implants in you the minimal constellation of desires needed to get you to wholeheartedly donate your pet schnauzer to the pound. Frankfurt would say that you have freely chosen to take your dog to the pound, though not all metaphysicians would agree that such alien desires could be a source of free choice. We can imagine Buss similarly saying that your choice was autonomous, because you see it as justified, even though you only see it as justified because of your spouse's sneaky move. We believe that this case is perplexing in ways that Buss's account does not allow her to acknowledge. Thus, if one were to discover that an outsider had implanted these anti-dog desires, it is simply unclear how one should respond. From an internal point of view, these desires seem impeccable. Because they mesh well with one's other desires, one is badly positioned to disown them or complain about them. For that reason, one may feel constrained to see the choice as autonomous. But matters are not so simple. Knowing the history of the desires should create a creepy feeling, a sense of alienation from the desire. That this particular history includes someone else's desire to manipulate your choice toward the direction you have in fact chosen heightens the sense of alienation. So we think that a lucid person who discovers that his desires have been implanted should feel confounded about his choice to take the dog to the pound. He should feel perplexed about which course of action, or choice, is in fact free, authentic, or autonomous. Any theory that gives an easy answer misses the complexity of the phenomena. We raise the topic of the controversy regarding Frankfurt's analysis of freedom not because we hope to make a new contribution to resolving the controversy, but because we think that the sources of skepticism about Frankfurt's view, whatever problem they create for him, create worse problems for Buss. Frankfurt faces a problem of alien desires. If some of a person's desires are

alien, then the fact they mesh with other desires he has does little to improve the autonomy of a choice rooted in those desires. The problem of alien desires has a cognitive counterpart. If one's beliefs have an alien genesis, then the fact that they give you a reason to feel justified may leave you with impaired autonomy. This seems clearly true when the beliefs are implanted artificially. But similar impairment occurs when the genesis of beliefs is ordinary deception.

2.4. Guilt-free pancakes

Consider a purely cognitive case, "pancake". You are a brain in a vat. It didn't start out that way, but as you watch the Super Bowl one Sunday afternoon scientists pluck you from your armchair and drop you into their vat. Now you see only what they want you to see, and they have been wholly successful in getting you to think that life had proceeded normally since the Super Bowl. You think that you are choosing and then eating pancakes for breakfast each morning, choosing to jog and then jogging each afternoon, and so on. We think it plain that you do not autonomously choose to eat your pancakes (though the success of our argument does not hinge on this). Despite the fact that your choice was wholehearted, the choice has a suspect history that destroys its authenticity. What made these putative pancakes seem attractive to you was wholesale illusion. If you had known even a fraction of the truth, you would have felt repulsed by this fake food. Perhaps Frankfurt would nonetheless find your choice suitably free; perhaps Buss would follow in finding the choice autonomous. But there is a difference between "dog" and "pancake" that makes it harder to find autonomous choice in the later case than in the former. The difference concerns plausible answers to a telling counterfactual question. In "dog", which involves instilled desires, if one asks: now that you know about the instilled desires, would you choose otherwise, it is hard to say. You have no alternative value set available to you that can serve as the basis of a choice. But in "pancake" you can say: these are not even pancakes! I do not even have a mouth! In an important respect, one cannot make the same choice to eat pancakes once one knows the history one's beliefs. In contrast, once one knows the history of one desires in "dog", one can still choose to take the dog in. Indeed, apparently Frankfurt thinks that one might reasonably do so (although if we know a little bit about husbands and

wives and how they respond to one another's attempted manipulations, it seems highly unlikely).

The conclusion that we draw from "pancake" is that no matter how much internal meshing attaches to the beliefs that undergird a choice, the falsity of those beliefs may well matter in an assessment of the autonomy of that choice. In "pancake", because of the falsity of one's beliefs, one does not make an autonomous choice. We have described "pancake" in a way that involves an illicit path toward the beliefs it involves, but that was only for expository ease. We can describe a variant, "pancake*" , relevantly the same except for the absence of illicit etiology. Suppose that you were not kidnapped and made into a brain in a vat, but that you instead accidentally fell into the vat that had been created as a test. Nonetheless, you were automatically anesthetized, your body stripped away, and the relevant electronics were set to work creating pancake beliefs. In this case, in which no foul play but only nasty accidents occur, it nonetheless seems that your choice for pancakes is less autonomous than one might like. "Pancake*" suggests the following principle:

> P1: The deeper your error regarding the factual grounds for a choice, the less the choice expresses your preferences (or is yours "all the way down") and hence the less it expresses your autonomy.

We think that P1 is roughly true, but requires some qualification. No doubt Buss would simply reject P1. We think that her reasons for rejecting P1 can be accommodated in a suitably qualified principle.

2.5. More ado about autonomy, error and trust

Remember Christopher Columbus. Suppose that going to India was his principal aim in crossing the Atlantic. Columbus would then have made his choice on the basis of false belief. That would not show that his choice was deficient in autonomy, we think. Columbus knew he was taking a gamble. He understood that he might be making a mistake, was aware of the risk that he was mistaken. At a minimum, P1 should be modified to reflect the possibility of an autonomously chosen gamble. (The notion of an autonomously chosen gamble will be crucial to an understanding of the processes of seduction and being seduced, as one would expect.) Suppose, however, that Columbus was not reasonably undertaking a gamble. Instead, he believed, while consciously rejecting the best evidence available, that he would encounter India at the end

of his trip, and he believed this because an astrologist advised him to do so. In that case, his decision was deeply mistaken, not because he took a reasonable gamble, but because he was unreasonable. He was at fault. Still, his faulty reasoning does not excuse him of responsibility for his choice. He chose autonomously, if unreasonably, to head across the Atlantic seeking India. This suggests that not all factual error undercuts autonomous choice: autonomy is not undercut by error that is one's own fault, or error that occurs as part of a reasonable gamble. It also suggests a modification of P1:

> P2: The deeper your factual error regarding a choice, to the extent that it is not attributable to your fault, or simply a reasonable gamble, the less the choice expresses your preferences and hence the less it expresses your autonomy.

Of course, if P2 is correct, then much garden-variety deception, including Johannes' deception against Cordelia, violates autonomy and is therefore wrong. Cordelia's factual error about Johannes's intentions are not attributable to her fault. And we would not say that Cordelia's love for Johannes is predicated on "a reasonable gamble" – as much as love is always a kind of reasonable gamble – because Johannes is playing a very different game than Cordelia supposes he is playing. Cordelia is gambling for love (and Johannes pretends these are also his stakes); Johannes is gambling for a night in the sack. To make Cordelia's innocence that much more clear, we should not forget that what "a night in the sack" means for Johannes: it is the symbol of her relinquishing her autonomy to him.

Buss might resist P2. She seems wedded to coherentist justificatory principles. If your beliefs mesh together in the right way, you are justified in acting on them, no matter what their history, no matter how unreasonable you were in acquiring them. But P2 seems to take care of the cases that motivate Buss to say false belief, and hence, deception, do not undercut autonomy. Her (1) and our P2 are at odds, but perhaps, based on the cases so far presented, she'll take (1). Although we think that P2 can be used to explain away Buss's intuition, she might stick to them. We think, however, that an argument can be made that goes beyond this simple appeal to intuition. This argument appeals to the idea that deception in crucial cases involves breach of trust. We will propose that the involved breach of trust compromises autonomy.

Plainly Cordelia trusted Johannes. He courted that trust. And he breached it. Breaching trust, particularly when trust forms the basis of belief,

compromises autonomy. When Cordelia trusts Johannes about what he says, it follows that she accepts what he says as true, without skepticism. This process is a gradual one – she does not trust him instantly, as no reasonable person does, and especially not in the game of love – but with his persistent courtship, his many devices and ploys, his astonishingly complex and artistic techniques of winning her trust, she comes to believe him wholeheartedly. Because she does so, she transfers the effective locus of her decision-making on the truth of these beliefs to Johannes. Her autonomy with respect to these epistemically significant matters is in his hands. Thus when he deceives her by betraying her trust, he compromises her autonomy and wrongs her. He achieves his goal: he takes her freedom. But he is able to take her freedom precisely because she entrusts it to him (Studler, 2005).

Our harsh remarks about Johannes may seem too easily generalizable, or at least inconsistent with our earlier embrace of seduction, even when it involves deception. Our position is that some seduction involving deception is morally fine. Yet such deception, on our account, may conflict with respecting autonomy, and so seems wrong. How do we reconcile these strands in our position?

2.6. A happy surprise at the airport

We think that it is a puzzling fact of moral life that sometimes one may deceive an innocent person, in ways that surprise him and hence seem to breach his trust, but not wrong him. Consider "airport". Suppose that your friend's spouse is returning from her tour of military duty in Iraq. She asks you to keep her secret, but to get her husband to the airport for her arrival. So you make up a story about how you need his help at the airport, and get him there, where he is delighted to find his spouse arriving. How does this case differ from Cordelia's? In deceiving the husband, you act for his sake and out of respect for him. You do not deceive him to "gain an advantage over him" (in Ingmar Bergman's witty definition of a lie)[3], as we think that Johannes does to Cordelia. Johannes might claim otherwise, saying that he acts to helping her out in the best way available. We think that he deludes himself. Suppose that we are wrong. There would still be a morally important difference between Johannes's deception and the airport deception. The former but not the latter

[3] Ingmar Bergman, *Fanny and Alexander*, 1982.

is paternalistic. Regrettably paternalism may be acceptable when one deals with a person suffering some defect. There need be no defect in the airport husband. Instead there is a good – the surprised delight of finding the spouse at the airport – that can be obtained only through deception. You get the good for his sake, out of respect for him, and not because you see something wrong with him. More generally, we propose this principle:

> P3: If you deceive a person while reasonably seeing yourself as acting for his sake and not seeing yourself as correcting for his defect, and you do so to obtain a good in which he shares and whose existence is essentially tied to deception, then you do not thereby violate his autonomy.

P3 makes sense if there is a class of goods whose acquisition ineluctably involves deception. (We think certain kinds of seduction are among that class of goods.) It varies with the purported beneficiary whether P3 warrants deceiving him. A reasonable person raised on a steady diet of Kantian fervor might resent being deceived into taking the airport trip, and P3 could hardly be used to justify deceiving him. For most reasonable people, as we have said, we think the deception would be morally acceptable, perhaps even morally praiseworthy.

P3 becomes more plausible if one reflects on the experience of falling in love in everyday life. Even Kant admits that in forming friendships – and how much more so in falling in love – we are naturally led to «cover up our weaknesses, so as not to be ill thought of» and that this is necessary for us to «impart our feelings to the other.» (1997, 187–188). And Kant was no expert on love. Every one of us has known the careful, playful, coy and deceptive game that involves luring and withdrawing, approaching and coercing, mixing truth and lie, and knowing that the other person is doing the same, because we both understand that this is the only way to achieve the goal we are mutually seeking: love. The would-be lover who throws himself on his knees and simply declares the truth, the whole truth and nothing but the truth about himself goes home alone at the end of the night. To pretend otherwise is to be even more puritanical about deception than Kant himself was, and to be more puritanical about deception than Kant is not going to help us better understand anything about how and why one deceives.

Take "the good" of P3 as "seduction" or "cultivation of romantic love." If one thinks about seduction, the paternalism we worried over in P3 might be

seen as a kind of conceit: "seeing yourself as acting for his sake" in the context of attempting to seduce someone must be understood as seeing yourself as a good worth having, and a good worth having for the agent you are seeking to seduce. On this ground Johannes' seduction of Cordelia clearly fails the test of P3 even before we get to the important criterion of helping her "obtain a good in which (s)he shares," because he cannot reasonably see himself as acting for her sake. He may consider himself a good worth having, but he is not acting in such a way so as to provide her with that good: he plans to deny her the good as soon as he has culminated his own wish to seduce her. He could only be acting for her sake if he wanted to disillusion her about romantic love – he is himself disillusioned about it, and that is part of the greater lesson Kierkegaard is trying to teach through the novella, that Johannes is himself profoundly confused about the psychological condition he thinks he has mastered – but few reasonable people could sincerely consider such disillusionment a good. Most of us happily go to our graves with the belief, illusion or no, that romantic love and the right kind of seduction are among the finest things in life.

Johannes puts Cordelia into a kind of experience machine, and while the extreme case is good for testing intuitions about why deception is morally blameworthy, it is not representative of seduction in general, and certainly not of the kind of seduction in which people are typically involved. A more representative case of seduction, we think, is the reluctant and mutual seduction that takes place between Benedick and Beatrice in Shakespeare's *Much Ado About Nothing*.[4]

3.

3.1. Seduction, deception and self-deception in *Much Ado About Nothing*

One revealing feature of the mutual seduction of Benedick and Beatrice is Shakespeare's emphasis – the same holds true for the seductions in virtually all of his plays – on the complicity that exists between the seducer and the seduced. Even in the extreme case of Johannes and Cordelia, the complicity of the seduced is present: as Johannes's seduction proceeds, there is a gathering

[4] All references are to William Shakespeare, *The Oxford Shakespeare: The Complete Works 2nd Edition* (New York: Oxford University Press, 2005), by Act, Scene, and lines. For ease of reading, act, scene and line references are internal.

atmosphere of deception, a feeling of "everything I want to believe about him turns out to be true." A telltale sign of self-deception is that one winds up believing precisely what one wanted to believe in the first place – this is not to say that such cases always involve false belief, but that they should certainly raise our epistemological antennae – and Cordelia never calls out Johannes, never fully accepts her own responsibility as an epistemological agent. (Even in love, there is due diligence). As young as she is, one cannot reasonably blame Cordelia for naivete and a little self-deception. But this, again, is why the tale of Benedick and Beatrice offers a richer and more attractive example of seduction and deception than does the tale of Johannes and Cordelia.

On their own account, neither Benedick nor Beatrice believes in romantic love, at least for himself or herself; moreover, each professes a distinct dislike for the other. Beatrice's first words in the play are a jab against Benedick – though we notice she is also asking if he has "returned from the wars?" – which she quickly follows up with a long complaint against him, ending with the remark that he has only one wit, the sole "difference between himself and his horse"(Act I, i, 15-94). For his part, Benedick first greets Beatrice with: "What, my dear Lady Disdain! Are you yet living?" (implying she is not just unkind, but old) to which she replies "Is it possible disdain should die while she has such meet food to feed it, as Signior Benedick?"(I,i, 95-136). They quickly go on to reassure one another that:

> Benedick: [...] it is certain I am loved of all ladies, only you excepted: and I would I could find in my heart that I had not a hard heart; for, truly, I love none.
> Beatrice: [...] I thank God and my cold blood, I am of your humour for that: I had rather hear my dog bark at a crow than a man swear he loves me. (I,i,95-136).

They then proceed to exchange numerous insults.

Notice that both Benedick and Beatrice are already plying their deceptions and in doing so initiating the process of seduction. Benedick's boast that all ladies love him (a timeworn if silly and no doubt ineffective male technique for attracting a woman's attention) is obviously false, and not really a lie: he says it so as to contrast all other women with Beatrice, and to suggest that he could have any woman he pleases except for her. The real deception that Benedick is practicing – the deception, repeated by Beatrice, that sets up both the seduction and the comedy of the play – is the claim that his heart is so hard that it cannot love. Beatrice and Benedick open the play already in sexual tension, which both are pretending does not exist between the two of them and which,

furthermore, on their account, is not the sort of thing either of them is interested in anyway. Benedick deceives Beatrice by insisting that he is not interested in love (he repeats the same claim to anyone who will listen to him throughout the first act of the play). By saying that he loves none, however, Benedick is also revealing to Beatrice that there is no woman he is presently attached to or even interested in. Should he take an interest in a woman, it follows, what a rare and fine thing that would be – and this is intended to pique her curiosity and vanity. Beatrice responds with the same deception, but is more direct and to the point (in a funny way, more honest about her deception): I don't even want to hear promises of love from a man, she says, much less the real thing. Of course we know she has already been asking specifically about Benedick, and hers is also a familiar technique for interesting a lover: he is a warrior, and she is raising a challenge. The conversations Benedick and Beatrice both have with friends shortly after this scene confirm, in indirect but no less certain ways, their attraction for one another. All this is so transparent – such a clear and delightful example of schoolyard flirtation – that the audience knows, only a few minutes into the play, that these two will fall in love before it ends.

But the point of their deception is not only to begin the process of seduction, it is also to protect themselves, because neither is sure of the other's interest. They don't trust one another. Benedick puts it plainly: "Because I will not do them [women] the wrong to mistrust any, I will do myself the right to trust none"(I, i, 220-225). Furthermore, they shouldn't trust one another: if either Benedick or Beatrice were to be too overt about their interest in one another, the other's pride and sense of him or herself as superior to love (to which they both at least pretend, and may partially believe) would end the seduction before it could begin. Beatrice and Benedick mutually seduce one another because they regard one another as equals, and should that equality shift too much in one direction or the other – if one, in other words, came to feel that he or she were losing control or being controlled, if he or she were being diminished in terms of *autonomy* – the seduction would be frustrated. Beatrice is as clear about her autonomy as Benedick is about his trust: "Would it not grieve a woman to be overmastered with a piece of valiant dust? To make an account of her life to a clod of wayward marl?"(II, i, 34-78). Soon we learn that Benedick and Beatrice had been involved before, and something went wrong: Beatrice claims she had lent Benedick her heart a while, but that he had

won it "with false dice"(II, i, 243-284). So for Beatrice there is a particular and we may suppose justified distrust of Benedick.

Here our earlier notion of romantic love as "an autonomously chosen gamble" comes to the fore, because Beatrice and Benedick had previously gambled at love, and Beatrice – at least, on her account – had lost. (Though as cagey as each is with the other, the feeling one has is that both suffered in the failed game.) The problem now is that, because of shared mistrust, both are reluctant to take a chance, to gamble a second time. Beatrice and Benedick seem to view the very idea of gambling on love as a violation of their autonomy: and it takes several deceptions before either of them is willing to admit that "the die is cast," and they are willing actively to try to allow romantic love to take hold.

Nevertheless, the seduction continues. It is through another deceit – one of Shakespeare's classic devices, the masked ball – that the seductive tension between Benedick and Beatrice mounts. They are dancing with one another, each clearly knowing who the other is, but with the comfortable position of enjoying plausible deniability about their epistemic situation. Benedick asks the masked Beatrice what she thinks of Benedick, looking for the least encouragement – "Did he never make you laugh?"(II, I, 114-152) – only to find Beatrice using the mask against him to say even crueler thing about him than she might say to his face, and the words are that much sharper because, he is forced to suppose, she is willing to say them to someone whose identity (he is forced to pretend) she doesn't know.

The leitmotif of the play comes from the song that opens the famous orchard scene, and is a kind of playful leitmotif of our paper:

> Sigh no more, ladies, sigh no more / men were deceivers ever, /
> One foot in sea and one on shore/ To one thing constant never:/
> Then sigh not so, but let them go, / And be you blithe and bonny, /
> Converting all your sounds of woe / into Hey nonny, nonny
> (II, iii, 44-88).

Naturally the ladies can no more let the men go then the men can the ladies – "can't live with 'em, can't live without 'em" – so the advice is ironical: meant truly, in a sense, on its face; but in another sense meant in just the opposite way, that though we recognize and complain about one another's weaknesses and bemoan them, but they are part and parcel of a good we cannot do without.

While Benedick and Beatrice are slow and reluctant to understand this ironic truth about love, their friends are not. So, growing impatient with the

spectacle of Beatrice and Benedick trying to seduce one another but tripping over their pride, freedom and mistrust in the process, three of Benedick's friends deceive him – while he thinks he is deceiving them, by hiding behind the bushes – and have a "secret conversation" in order to convince him that Beatrice is passionately, desperately in love with him, and all but dying from her fear to disclose it to him. In the very next scene, at the opening of Act III, Beatrice's friends, also part of the plan, have the same secret conversation designed for her eavesdropping ears, persuading her that Benedick is in just the same impassioned, prostrate position he supposes she is in for him.

By this point in the play we have Benedick practicing P3 for Beatrice, Beatrice practicing P3 for Benedick, and both Benedick's and Beatrice's friends practicing P3 for each of them. It's comical and charming; seduction is taking place; no one's autonomy is being violated; and while trust is in some sense being betrayed (that is, by Beatrice and Benedick's friends, who are willfully exploiting their eavesdropping – though we should ask, as Shakespeare wants us to ask, whether you can betray the trust of someone who is already betraying your trust be eavesdropping on you), the betrayal of trust does not look morally blameworthy: on the contrary, it's a happy, well-intentioned, even praiseworthy act. Only the worst kind of moral sourpuss could frown down on this playfulness and friendship.

The drama is not yet over: Beatrice will demand a proof of Benedick's love after he professes it, and the proof is terrible enough that it tests their love. The great moment of suspense is captured by Beatrice when she summarizes their position, add how much depends on whether or not she can trust Benedick. Benedick tells her: "I do love nothing in the world so well as you: is not that strange?" And Beatrice replies: "As strange as the thing I know not. It were as possible for me to say I loved nothing so well as you: but believe me not; and yet I lie not; I confess nothing, nor I deny nothing"(IV, i, 265-271). Sounding a bit like Pyrrho or Sextus Empiricus, Beatrice is about to ask Benedick to prove his love by killing his friend Claudio in recompense for the betrayal of her cousin. Shakespeare is subtle as ever: this proof of love is demanded as the enactment of justice for a betrayal of trust.

Happily, after several more demoniacally clever Shakespearean twists and turns, Benedick succeeds in proving his love, and at the close of the play the two are married. But right until the last few minutes of the play they continue to deceive one another, denying their love, because they find themselves in the classic lover's paradox: "who will say the L-word first?" This paradox is a

paradox of trust, and when at last they are confronted with their own professions of love in writing (produced, naturally, by others), the Gordian knot of their distrust is cut, and – to everyone's relief – they are at last free to bind themselves to one another. One of Benedick's friends is about to tease him about marrying, after all he has said against it, and he summarizes his position with one of the most plangent observations about the nature of love in all the vast literature on the subject: "In brief, since I do purpose to marry, I will think nothing to any purpose that the world can say against it; and therefore never flout at me for what I have said against it; for man is a giddy thing, and this is my conclusion"(V, iv, 85-126). He has gained the good he desired, and however giddy and deceptive and full of false belief the process was that got him there, now it doesn't matter.

3.2. The giddiness of self-deception

Benedick's statement, here at the very end of the play – "his conclusion," as it is Shakespeare's – emphasizes what we referred to at the outset as the complicity of the seduced: the willing self-deception that we have thus far sought to illustrate, but not made explicit. While even Cordelia shared some responsibility for her seduction by Johannes, because she never took a step back to examine the constant pressure and manipulation she was experiencing from the man pursuing her, how much more so are both Beatrice and Benedick complicit in their own seduction. They hide behind masks, they lie to themselves about their own feelings and reaffirm their self-deceptions by repeating them to others, they test one another's interest through insults and jabs, they eavesdrop in the hope of learning that their hopes of shared love might be fulfilled. Before long the audience realizes that both Beatrice or Benedick would be willing to twist the truth in any direction she or he pleased in order to gain the good each of them seeks: the seduction of the other. Both are so complicit in one another's seduction and each in their own seduction – think of Benedick's giddy joy as he interprets and reinterprets Beatrice's innocent and casual invitation to come in to the house after hearing his friends' speak of her love for him – that it no longer makes sense to divide seducer from seduced. Each not only seduces the other, both recognize that a kind of mutual self-seduction, an allowing oneself to be seduced, is also necessary. Thus theirs is genuinely an autonomous gamble, because they are involved in the risks of the game from both the perspective of the seducer and the seduced.

Some may worry that explaining complicity in terms of self-deception is explaining one mystery in terms of another. But the reason we appeal to the case of Beatrice and Benedick is that we think Shakespeare's depiction of seduction shows how the two lovers deceive themselves while deceiving one another and being deceived by their friends. On our analysis this makes them complicit, without placing the burden on us of explaining how the self-deception does its work (that is the subject for another paper, and of a vast philosophical literature).

As giddy a thing as Benedick undoubtedly is, we don't want to go too far in endorsing giddiness (or deception, or false belief). But the back-and-forth nature of the romance between Benedick and Beatrice, the alternation of true and false, of frankness and deception, and the very tentative small steps forward into trust: these, we think, are the elements of how the more usual kind of seduction occurs. In the case of Beatrice and Benedick, seduction "was essentially tied to deception," and was practiced to obtain a good in which they both shared. There were elements of conceit, paternalism, and manipulation throughout the case, but neither Benedick nor Beatrice was wronged, and it would be silly to argue that either of their autonomy was compromised. In fact, for both of them it was their proud insistence upon their autonomy – proud almost to the point of irrationality – that made so many deceptions necessary in order for them to accomplish the mutual seduction they both desired. And though the case is exaggerated for comic effect, we think anyone who has been involved in this kind of seduction with the result of romantic love – whether or not that love endured – will agree that Benedick and Beatrice seem familiar.

3.3. Seduction and self-deception

Now that we've had a little foray into grown up seduction, let us bring the case of Benedick and Beatrice back around to our critique of Buss and the case of Johannes and Cordelia. We have said that Buss is wrong in arguing that Cordelia's autonomy is not violated by Johannes, because her autonomy was reasonably informed by her trust in him, and he violated that trust. Her trust was a consequence, in part, of her being an innocent, in part from the sheer quantity of false beliefs Johannes instilled in her, and in part from the cunning with which he tailored those false beliefs to her weaknesses. Our conclusion was that what makes Johannes's deception wrong, at the end of the day, and contra Buss, is a violation of Cordelia's autonomy, when our understanding of

her autonomy is properly robust. Buss misses the connection between autonomy and trust.

But that attack does not undermine the more interesting argument Buss makes. We agree with Buss's intuition that the case of seduction may illustrate why deception need not undermine autonomy, and have employed Beatrice and Benedick to that end. Along the way we have buttressed, if qualified, Buss's argument that an account of the wrongfulness of deception that relies on the wrongfulness of violating a simplistic notion of autonomy is insufficient.

A particularly surprising and interesting byproduct of the Beatrice-Benedick tale is that deception, both of oneself and of others – at least in some seductions – may foster trust rather than betray or destroy it. In scenarios where mutually interested parties begin a seduction with mistrust (and doesn't it usually begin this way?), some deception may be necessary in order for the process of trusting to get off the ground. If trust is importantly linked to autonomy in seduction, as we have argued, then it may be that some deceptions and self-deceptions actually enhance autonomy. Autonomy may not merely survive false belief, but flourish in it.

REFERENCES

Bergman, I. (1982). *Fanny and Alexander*.

Buss, S. (2005). Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints. *Ethics*, *115*(2), 195–235.

Buss, S., & Overton, L. (2002). *Contours of Agency*. Cambridge: MIT Press.

Kant, I. (1997). Lectures on Ethics. In P. Heath, & J.B. Schneewind (Eds.), *The Cambridge Edition of the Works of Immanuel Kant*. Cambridge: Cambridge University Press.

Herman, B. (1993). Leaving Deontology Behind. In B. Herman, *The Practice of Moral Judgment*. Cambridge: Harvard University Press, 208–241.

Hill, T. (1984). Autonomy and Benevolent Lies. *Journal of Value Inquiry*, *18*(4), 251–267.

Kierkegaard, S. (1843/1987). The Seducer's Diary. In *Either/Or* (tr. and eds. By H.V. Hong, & E.H.Hong). Princeton: Princeton University Press, 301–446.

Korsgaard, C. (1986). The Right to Lie: Kant On Dealing With Evil. *Philosophy & Public Affairs*, *15*(4), 325–349.

Shakespeare, W. (2005). *The Oxford Shakespeare: The Complete Works 2nd Edition.* New York: Oxford University Press.

Strudler, A. (2005). Deception Unraveled. *Journal of Philosophy*, *102*(9), 458–473.

Stump, E. (2002). Control and Causal Determinism. In S. Buss & L. Overton (Eds.), *Contours of Agency: Essays on Themes From Harry Frankfurt*. Cambridge, MA: MIT Press.

Commentary

Ambiguity, Opacity and Sartrean Bad Faith *

Mark A. Wrathall †
mark.wrathall@ucr.edu

ABSTRACT

I argue that Sartrean "bad faith" amounts to a motivated failure to apprehend the state of dis-integration that exists between one's facticity and transcendence. This "failure to see" is explained by drawing on Merleau-Ponty's account of perceptual ambiguity and existential opacity.

1. Introduction

I propose that we treat Sartrean bad faith as an instance of what I'm calling the "motivated failure to see." With this formulation, I mean to pick out a class of situations in which:

(1) one is in a position to apprehend something and thus have it figure in one's comportment in the world.

That means, I'm not using "see" narrowly to refer to visual perception, but quite broadly to refer to any way in which something is taken up so that it can

* Previous versions of this paper were presented at the *History of Philosophy Workshop,* sponsored by the Humanities Research Center at Rice University; the 2nd Annual *Southwest Seminar in Continental Philosophy,* held in Denver, Colorado, and the 2011 meeting of the *American Society for Existential Phenomenology*, held at Goodenough College, London. I'm indebted to the participants at all those events for their insights, encouragement, and valuable objections and criticisms. While I can't single out everybody who helped me work through the issues in this paper, I am particularly grateful to Charles Siewert, Iain Thomson, Steven Crowell, Beatrice Han-Pile, Wayne Martin, Christian Emden, Don Morrison, Steven DeLay, J. S. Blumenthal-Barby, Martin Blumenthal-Barby, and Samantha Matherne for their helpful suggestions

† University of California, Riverside, California, USA.

play a role in orienting and guiding our engagement with the world. Next, in a motivated failure to see,

(2) one fails in some way to apprehend the thing that one is in a position to apprehend.

This failure means that one does not apprehend it in such a way that it figures as the thing that it is in one's comportment in the world. That is, it does not figure at all, or it figures via a mistaken apprehension of it. Finally,

(3) one's failure to apprehend is a *motivated*[1] failure.

That is, it is not a simple oversight when one does not apprehend it. Rather, not apprehending it plays a role in maintaining our existential grasp on the situation, even if this existential grasp is not optimal.

Motivated failures to see are quite common, indeed, I would suggest, pervasive. In saying they are *failures*, I do not mean to suggest that a practical or moral failing is involved. That is, these failures do not necessarily make me less able to cope with a situation, or render me unable to make a morally correct decision. Indeed, it is likely that we can only cope at all because of all the things we don't apprehend – things that, in principle, we are in a position to apprehend, but the apprehension of which would distract us from our tasks or obscure our aims. It is possible that motivated failures to see are precisely what allow us to function without constantly being overwhelmed by all that the world has to offer. For instance, I fail to notice what time it is because I'm absorbed in reading my book, even though the clock is right there in front of me. In failing to see the clock, I am not failing in a moral or practical sense since the primary aim of my activity was comprehending the contents of the book. In that sense the failure to see served me well; by not glancing up periodically, I failed to see the time but I also maintained myself in a state of absorption, undistracted by things and events around me that would interfere with my ability to read and detract from my enjoyment of the book. Now there might be cases where I *should* have noticed the time. In those cases, we might want to address the motivated failure. But in countless others, the motivated failure contributes to my ability to maintain the overall grip I have on the situation.

[1] For more on the existential-phenomenological notion of motivation, see Wrathall 2005.

Moreover, in calling it a *motivated* failure, I'm not suggesting that there is necessarily anything *insidious* about the failure. The failure isn't always something we need to hide from ourselves. Rather, the adjective "motivated" alerts us to the fact that we are not concerned with a mere oversight, that I am *moved* to overlook something by the background aim of maintaining my grip on the world. Consider another example. I miss my turn while driving because I am talking to my friend. The failure to see that my turn was coming up is motivated by my primary response to the world being guided by the conversation rather than by what driving requires. I see some things, and not others, as a function of my maintaining myself in the conversation. But I don't need to hide from myself that this is what was going on, and if it is pointed out to me that this was the reason for my missing the turn, I will readily acknowledge it.

There is, however, a class of motivated failures to see which *are* insidious. This class includes Heideggerian "inauthenticity," Sartrean "bad faith," and various other kinds of self-deception. Such attitudes have a reflexive structure – it is not just the case that we are motivated to not see something, we are also motivated to not see *that* we are motivated to not see something. In these insidious versions, the motivated failure will only perform its function if it either goes unnoticed or, should it happen to be noticed, if it can pass for a mere oversight. This is because coming to recognize the motivation for what it was would itself undermine our grip on the world – perhaps even more so than seeing the thing to which the motivation was blinding us. Moreover, members of this class are thought to be self-undermining in a practical or moral sense, inasmuch as they deprive us of something that a responsible or autonomous or moral agent needs to have.

Now, my hope is that by approaching bad faith in terms of a more general account of a motivated failure to see, we can gain some insight into a number of issues surrounding it. For one thing, this approach helps to explain how Sartrean bad faith is possible at all, given that it seems to involve a self-contradiction – indeed, both a first and a second order self-contradiction. As Sartre notes, on the first order contradiction, bad faith is «a certain art of forming contradictory concepts which unite in themselves both the idea and the negation of the idea» (Sartre, 1956, p. 98). And on the second order, he argues that «the first act of bad faith is [...] to flee what it is» (Sartre, 1956, p. 115). That is, in order to act in bad faith, I need to hide from myself that I am acting in bad faith. It thus looks like I have to hide something from myself while

knowing it all the while, and knowing that I'm hiding it from myself. I would suggest, however, that the appearance of an intolerable paradox stems from thinking of such states on the model of epistemic states rather than perceptual states. It is not at all clear how I could not know what I know. But we are well familiar with the way that the perceptual field is characterized by gradients of apparentness, focus, lighting, and so on. We understand that in perception we are responsive to features that are not clearly and focally seen – indeed, that our ability to cope with a situation depends on our selectively not perceiving everything. There is thus less paradox involved in the proposal that we see a situation in the mode of not seeing certain features of the environment.

By illuminating the motivation which supports bad faith and inauthenticity, this approach also helps explain why such states are so pervasive and persistent. Bad faith, Sartre notes, «is an immediate, permanent threat to every project of the human being» (Sartre, 1956, p. 116). Its persistence is a puzzle because, if such states really do deprive us of something worthwhile, it is not clear why we would lapse back into them once we recognize them for what they are. For instance if, as Heidegger maintains, authenticity really brings us an equanimity and unshakeable joy that we are deprived of in inauthenticity (Heidegger, 1927, p. 310), then it is not clear why there is a constant temptation to inauthenticity, as he also maintains. This becomes more intelligible, however, when we understand the function such attitudes play in sustaining an existential grip on the world.

Finally, this approach gives us the structure we need to specify the content of the different varieties of authenticity and inauthenticity. Only by carefully specifying the content are we in a position clearly to assess the contention that such states deprive us of something worthwhile, because we can then determine the benefits and costs of maintaining ourselves in such states.

Moreover, we will eventually be able to distinguish more precisely the Sartrean and Heideggerian variants of authenticity in terms of the kind of failure-to-see that their respective versions of authenticity are correcting. But that is a task for another paper. In this paper, I focus rather narrowly on Sartrean bad faith. And I will discuss only in passing the problems of self-contradiction and persistence – the problems of understanding how bad faith can succeed in hiding itself from itself and persisting even once it is uncovered as such. Instead, I want to concentrate on the third issue – to see what light this approach sheds on the content of bad faith itself. Sartre's account of bad faith is notoriously difficult to understand, and there are a number of competing

accounts of it in the secondary literature.[2] By treating it as a type of a motivated failure to see, I take as my point of departure «the ambiguity necessary for bad faith» (Sartre, 1956, p. 99). Sartre notes: «the condition of the possibility for bad faith is that human reality, in its most immediate being [...] must be what it is not and not be what it is» (Sartre, 1956, p. 112). The constitutive ambiguity makes it possible to take up one of the attitudes or projects of bad faith (I express it in the plural because Sartre describes several distinct forms of bad faith). All of them involve a "flight", that is, a way of comporting oneself that does not take into account some vital aspect or feature of our being in the world.

2. "The Ambiguity Necessary for Bad Faith"

Before going any further, let's get an example out on the table to illustrate the centrality of ambiguity in Sartre's account. Sartre's most illuminating example (I've modified and elaborated it slightly) is the homosexual in denial and his friend, the "champion of sincerity." We are to imagine a man who, by all indications, is gay. Perhaps he seeks out the company of good looking men, and avoids the company of women. He feels sexually attracted to the men. He develops recognizably gay mannerisms of speech and posture. He often finds himself – quite by accident, he insists – in gay venues: internet chat rooms, bars, parks, and so on. But, Sartre explains, he «refuses with all his strength to consider himself» gay:

> his case is always "different," peculiar; there enters into it something of a game, of chance, of bad luck; the mistakes are all in the past; they are explained by a certain conception of the beautiful which women can not satisfy; we should see in them the results of a restless search, rather than the manifestations of a deeply rooted tendency, *etc., etc.* (Sartre, 1956, p. 107)

«Here,» Sartre concludes, «is a man in bad faith who borders on the comic since, acknowledging all the facts which are imputed to him, he refuses to draw from them the conclusion which they impose. (1956, p. 107).

[2] In this paper, I largely ignore the sizeable secondary literature on bad faith. At a future point, I intend to contrast and compare my approach with other interpretations of this important phenomenon.

On the surface, the gay man's bad faith seems obvious and straightforward. He is in bad faith because he is in "flight" from the significance of his own actions. He doesn't want to acknowledge what he is doing, because that acknowledgment would render him unable to maintain his way of being in the world (which is built around understanding himself as straight). It is a general atmosphere of ambiguity – ambiguity over the meaning of his actions, ambiguity over the meaning of the situations he finds himself in, ambiguity over the meaning of his very intentions and motivations – that supports his flight. This obvious, straightforward account of the gay man's bad faith and its dependence on a far-reaching ambiguity captures something important about Sartre's analysis of bad faith. But the straightforward account is complicated by the fact that Sartre insists that his friend, the "champion of sincerity" is in bad faith just as much as the gay man in denial. The friend, as Sartre explains, «becomes irritated with this duplicity. The critic asks only one thing – and perhaps then he will show himself indulgent: that the guilty one recognize himself as guilty, that the homosexual declare frankly – whether humbly or boastfully matters little» that he is gay (Sartre, 1956, p. 107). This complicates the straightforward account to the extent that it left us with the impression that the way to correct the gay man's bad faith is to cut through the atmosphere of ambiguity and own up to the facts as they really are. That the "champion of sincerity" is equally in bad faith shows, however, that it is no simple matter to dispel ambiguity.

Let's look more carefully at the "ambiguity necessary for bad faith." We call an entity or event or situation or motivation "ambiguous" when it can be understood in two or more incompatible ways. For example, at the moment it leaves the pitcher's hand, a baseball pitch is for the batter ambiguous – is it a fastball or a curve ball? At that moment, all the information available to the batter might be insufficient to determine which one it is.

Ambiguity is a product of what, following Merleau-Ponty, I call "opacity." We ordinarily think of opacity in terms of features of the perceptual field that keep us from seeing an entity or event clearly. For example, fog in the baseball stadium or poor lighting would render the atmosphere opaque so as to interfere with the batter's ability to see the pitch. But I want to use "opacity" more broadly to refer to any feature of a situation that renders something ambiguous. Sticking with our baseball example, a pitch is ambiguous for the batter because the world is temporally opaque – that is, the batter can't see

forward in time with precision to tell how the pitch will break as it approaches the plate.

Opacity in this broader sense might render something ambiguous, even if "in itself" or in other respects, it is quite determinate. The pitch is already either a fastball or a change-up. The ambiguity from the batter's perspective is a merely epistemic ambiguity. The world's temporal opacity prevents him or her from knowing it for what it is, but there already is an answer to what it is. When we are dealing with occurrent entities – the "present-at-hand" features of the world – we tend to think of opacity and ambiguity in epistemic terms. The facts are in themselves determinate, but they are rendered ambiguous to us by limitations on our knowledge.

Returning now to our gay man, we run into immediate problems if we understand his ambiguity as an epistemic ambiguity. Take the ambiguity that surrounds his spending the evening in a gay bar. What is the meaning of this action? It *could* be that he went to a gay bar because he wanted to meet other gay men. But it could have been a coincidence – he wanted a drink, he just happened to be passing by, he didn't realize it was a gay bar, once there, he became engaged in a lively political discussion and, in any event, he is an open-minded and liberal man secure in his sexuality for whom it would be silly to leave just because it is a gay bar, and so on. Now, if the meaning of his action is merely epistemically ambiguous, it is ambiguous only for someone who doesn't have epistemic access to the gay man's intentions. And that's where the paradoxical nature of bad faith becomes apparent. For surely the gay man himself is in a position to resolve the ambiguity, even if we are not. He ought to know the facts about why he went into the bar, and thus it would seem that he lacks the necessary epistemic opacity to pull off the flight from his homosexuality. More importantly, if it is an epistemic ambiguity at issue, then the "champion of sincerity" is right – what the gay man needs to escape from bad faith is to simply acknowledge the facts about his actions, intentions, desires, and so on. That is precisely what the champion of sincerity wants his friend to do. He wants him to look squarely at the facts about his actions and motivations, and acknowledge them as facts. This, he thinks, will resolve the ambiguity surrounding his friend's actions. But were the gay man to take his friend's advice, Sartre argues, he will merely trade one kind of blindness for another – he will blind himself to his own role in constituting the meanings of the "facts." He will end up treating his actions and intentions as if they are objective facts that exist as they are independently of his way of being in the

world. Champions of sincerity, Sartre argues, are insidiously motivated by a desire to avoid responsibility for who we are. If we can reduce the gay man's sexuality to an unambiguous objective fact, it «removes a disturbing freedom from a trait,» and it «aims at henceforth constituting all the acts of the Other as consequences following strictly from his essence» (Sartre, 1956, p. 108). Now, to make sense of Sartre's criticism of the champion of sincerity, we need a different notion of ambiguity than merely epistemic ambiguity. If Sartre is right that the champion of sincerity is in bad faith, it must be because there is, in fact, no objective fact of the matter about the meaning of the gay man's actions and the content of his intentions. They must be subject to an ambiguity that no amount of access to his inner states and desires will resolve.

In what follows, I want to draw on Merleau-Ponty to explain what I think Sartre has in mind here. Merleau-Ponty has particularly trenchant things to say about the connection between ambiguity and opacity and the way they support and enable a motivated failure to see. Indeed, I am optimistic that one can find in Merleau-Ponty the tools for a more nuanced and plausible account of this phenomenon than one finds in Sartre. Unfortunately, Merleau-Ponty doesn't offer a concise and systematic account of opacity. In the following section, then, I try to distill from his work a typology of opacities.

3. Merleau-Ponty on Perceptual Ambiguity and Existential Opacity

Through his phenomenology of perception, Merleau-Ponty came to recognize «the indeterminate as a positive phenomenon» (Merleau-Ponty, 1962, p. 5). In calling it a *positive* phenomenon, he means to suggest that indeterminacy is not the result merely of a lack of clarity or attentiveness on our part. Rather the perceptible things that populate the world around us are what they are only because they are "in themselves" indeterminate – what they *are* depends on how they are related to other things around them, and there is no uniquely correct way to relate things to each other.[3] That makes everything in the perceptible world profoundly ambiguous. But ambiguity isn't limited to the things around us. It infects us as well – our thoughts, desires, intentions, and so on. «Ambiguity,» Merleau-Ponty notes, «is of the essence of human existence, and everything we live or think has always several meanings»

[3] For more on phenomenological accounts of perceptual indeterminacy, see Wrathall 2009.

(Merleau-Ponty, 1962, p. 168).[4] In fact, the ambiguity of the world and the ambiguity of consciousness are closely interrelated: «I know myself only in my inherence in time and in the world, that is, I know myself only in ambiguity» (Merleau-Ponty, 1962, p. 344). This interdependence of my self-understanding and my perception of the world is also a key component of Sartre's account of bad faith. But before developing this idea (in the next section), we need to lay the groundwork by developing Merleau-Ponty's insights into the ontological nature of indeterminacy.

According to Merleau-Ponty, what we might call "existential opacity" is a significant source of perceptual ambiguity (whether the ambiguity is located in our perception of the things around us, or in our perception of our own intentional states). The very being of the world is opaque, so that there is no clear, definitive, and unique answer to what and how things "really" are. «Existence,» Merleau-Ponty explains,

> is not a set of facts (like 'psychic facts') capable of being reduced to others or to which they can reduce themselves, but the ambiguous setting of their inter-communication, the point at which their boundaries run into each other, or again their woven fabric. (Merleau-Ponty, 1962, p. 166)

There is a plurality of orders of facts. These orders come into contact but cannot be seamlessly united. The world is existentially opaque because it is never absolutely clear which set of facts is the right one for making sense of any particular entity or event, thus leading to an important kind of perceptual ambiguity: «the perceived, by its nature, admits of the ambiguous, the shifting, and is shaped by its context» (Merleau-Ponty, 1962, p. 10). Such a perceptual ambiguity is not merely an ambiguity for knowledge. It is not, in other words, a matter of our lacking sufficient information about what the facts of the situation are. No amount of additional information about the present state of affairs will eliminate all the ambiguity since the information will always be relative to a particular order or set of facts. But it would be a mistake to think that existential opacity is something we should want to overcome. As this quote from Merleau-Ponty suggests, an existential situation is a setting for human agency only in virtue of the fact that it presents an intersection between

[4] See also Merleau-Ponty 1962: «Consciousness, which is taken to be the seat of clear thinking, is on the contrary the very abode of ambiguity» (1962, p. 331).

different orders or sets of facts. I will come back to this point shortly. But first let's develop this account of perceptual ambiguity a little further.

It helps considerably if we can take a concrete example and see how it manifests the kind of ambiguity Merleau-Ponty describes. Take Sartre's example of the «woman who has consented to go out with a particular man for the first time» (Sartre, 1956, p.96). She knows that this man wants to have a sexual relationship with her (of course, there might well be more to his motivation than this, but for the sake of the example, that is at least an important part of his overall motivation). The woman, Sartre tells us, would be «humiliated and horrified» to agree to go out with him on this basis. Of course, she would also be insulted if he was not sexually attracted to him: «she would find no charm in a respect which would be only respect» (Sartre, 1956, p. 97). She would not feel the same delight at his attention to her if, for example, it were clear that he is not attracted to her, if he could honestly declare that he didn't find her attractive in the least. And so she is in a paradoxical position from the outset: «in order to satisfy her, there must be a feeling which is addressed wholly to her *personality* [...] but at the same time this feeling must be wholly desire; that is, it must address itself to her body as object» (Sartre, 1956, p.97). To maintain herself in the paradox, she needs to fool herself about his intentions while at the same time being aware of and responsive to his sexual attraction to her. How does she pull this off?

Part of the answer is that she exploits the opacity that surrounds each of his acts to convince herself that they express feelings that are driven by higher ideals – love, respect, admiration. He rests his hand lightly on her thigh as they talk. The meaning of this simple act is ambiguous. Is he touching her to punctuate a point he has made in the conversation? To get her attention? To test her receptivity to physical contact? Because he is curious about the texture of her skirt? In the ordinary course of affairs, we reduce such ambiguity by situating the act in a number of ways:

- *Temporally* – meaning unfolds over time. What it means depends (at least in part) on what it develops into.
- *Socially* – meaning is (in part) publicly determined; something means what a relevant community of others understand it to mean.
- *Motivationally* – the meaning of acts is (in part) determined by the motivations and intentions that give rise to the act.

- *Contextually* – meaning depends on which relationships are the *definitive* relationships. Any existing thing (or act or event) could plausibly be taken in relation to any number of different things (or acts or events). In which direction it refers us decides what kind of a thing it is in the first place.

Of course, any or all of these dimensions of meaning can be in play at any given time, thus complicating our effort to disambiguate a situation.

Corresponding to each of the ways of situating something is a type of existential opacity. There is a *temporal opacity*, because we often can't tell what something is without seeing how it develops, and we can't always see how an action will play out. In our example, we can wonder, will the man immediately remove his hand once he has her attention? Will he stroke her thigh with his fingertips?

There is also, Sartre argues, an ineliminable *social opacity* when it comes to the meaning of actions because we are not in control of our meaning for others, and we are never in a position to see clearly how we are interpreted by others. Sartre notes:

> When Pierre looks at me, I know of course that he is looking at me [...] The meaning of this look is not a fact in the world, and this is what makes me uncomfortable. Although I make smiles, promises, threats, nothing can get hold of the approbation, the free judgment which I seek; I know that it is always beyond [...] My reactions, to the extent that I project myself toward the Other, are no longer for myself but are rather mere *presentations*; they await being constituted as graceful or uncouth, sincere or insincere, *etc.*, by an apprehension which is always beyond my efforts to provoke, an apprehension which will be provoked by my efforts only if of itself it lends them force (that is, only in so far as it causes itself to be provoked from the outside) (Sartre, 1956, p. 105).

In the example of the young woman, we can see social constitution and social opacity at work in the prevailing social norms that govern what counts as appropriate contact between a man and a woman. But we see it also in the fact that the man cannot all by himself decide the significance of his hand resting on the woman's thigh. His action will depend for its meaning on how she responds to it.

The opacity that governs this situation also invades even the "inner recesses" of our consciousness, obscuring our own beliefs, desires, intentions, motivations, and so on, from us. Because so much of the significance of the

man's actions depends on factors beyond him, there is a sense in which he himself discovers only "after the fact" what it is he intended or was motivated to do. Thus, there is a *motivational opacity*. Our actions are ambiguous because our conscious life is an opaque setting for states and acts, and that setting derives its opacity in part from the intimate dependence of conscious states on the surrounding world. Merleau-Ponty says of visual experience, for instance, that vision is

> an operation which fulfils more than it promises, which constantly outruns its premises and is inwardly prepared only by my primordial opening upon a field of transcendence [...] Sight is achieved and fulfils itself in the thing seen. It is of its essence to take a hold upon itself, and indeed if it did not do so it would not be the sight of anything, but it is none the less of its essence to take a hold upon itself in a kind of ambiguous and obscure way, since it is not in possession of itself and indeed escapes from itself into the thing seen. What I discover and recognize through the *cogito* is not psychological immanence, the inherence of all phenomena in 'private states of consciousness', the blind contact of sensation with itself. It is not even transcendental immanence, the belonging of all phenomena to a constituting consciousness, the possession of clear thought by itself. It is the deep-seated momentum of transcendence which is my very being, the simultaneous contact with my own being and with the world's being (Merleau-Ponty, 1962, p. 377).

So in sitting at my desk as I write this, there is an obvious sense in which I see the desk. But the "seeing" is not a private, inner experience, clearly and distinctly set before my mind. What it is for me to see the desk, what it is like, what the experience contains – that is something I have only a tenuous grasp of independently of being able to refer directly to the properties of the desk itself. Sartre makes the same point when he argues that «so far as my being is concerned, there is no difference between being and non-being if I am cut off from my project» (Sartre, 1956, p.111). That is, what I am is only fixed when I carry out my intentions. Until that point, there is no difference between saying that I am X and saying that I am not-X. But Merleau-Ponty helps us see that our intentions not only lack content without their projects. In addition, the fulfilment of the intention in an act isn't enough to banish all the ambiguity. In carrying through my projects, my action «fulfils more than it promises and constantly outruns its premises» (Merleau-Ponty, 1962, p. 377). The desk as it is given to me exceeds what I project in intending the desk, and it constantly

exceeds what I have a perceptual hold on at any moment. This means that I am capable of misinterpreting my experience of the desk, that I can mistake something else for a desk, and that I can lose hold of what was present in my experience of the desk even when I have had a relatively clear and distinct perception of the desk. The same holds, quite generally, for other conscious states, including beliefs about our intentions and motivations. I'm often not able to tell clearly what it is that moved me to perform some action, nor what I intended (retrospectively) or intend (prospectively) to do, and I can be mistaken about my motivations and intention – and I'm uncertain or unclear about all this because the motivations and intentions find their meanings in actions which exceed in their "external" consequences what I knew of my intentions.

Much more could be said about temporal, social, and motivational opacity, but I'd like to focus in the remainder of this section on *contextual opacity*. There is contextual opacity when there are a number of different contexts into which an action could be inserted and it is not clear to which context an action or an event belongs. To put it slightly differently, contextual opacity is a situation in which it is not clear what *kind* of situation it is. Ian McEwan's novel *Atonement* (2001) provides a nice illustration of how such opacity can produce a profound ambiguity in worldly actions and events. One of the pivotal moments of the narrative occurs (to describe it with a minimum of contextual information) when a woman beside a fountain prepares to lower a porcelain vase into the basin of the fountain. A man standing beside her reaches over and grabs the lip of the vase. The woman tries to pull the vase away from him; it snaps and a piece of porcelain falls into the fountain basin. The man and woman stare at each other. The woman strips down to her bra and knickers, climbs into the fountain and retrieves the broken piece. She gets out, dresses, and carries the broken vase into the house.

Obviously, with such a bare account of the event, it is hard to make much sense of its meaning. One could imagine an endless series of ways to account for what happened. The ambiguity of this event and of the actions of the characters in it is reduced considerably as we begin to add context to our description of it. The event occurs on a country estate in England in 1935. The woman, Cecilia Tallis, is the daughter of the owner of the estate. The man, Robbie Turner, is the son of a charlady who works on the estate, and he himself is working as the gardener. Cecilia and Robbie are both young, unmarried, and on holiday from their University studies at Cambridge. They grew up together,

but «had fallen out of touch at Cambridge» (McEwan, 2001, p. 26), and more recently their relationship to each other had become further strained. Already these few details point to a couple of broad contexts within which the events can be situated.

For instance, one important context for the event is the social and political context of class relations in early 20th century England. Robbie and Cecilia are very mindful of this context, both in the lead-up to the event in question, and as they themselves try to interpret its meaning afterwards. Just moments before, Cecilia had reminded Robbie that her father is paying for Robbie's education at Cambridge. And she had alluded to Robbie's past affiliation with the Communist party (she begins the encounter by asking «would you roll me one of your Bolshevik cigarettes?»)(McEwan, 2001, p. 25). Cecilia is convinced that the increasing distance between her and Robbie is a result of his cognisance of and resentment at the class difference. She believes he is playing up his status as the cleaning-lady's son to mock and punish her «for being in a different circle at Cambridge, for not having a charlady for a mother» (McEwan, 2001, p. 27). She takes offense at this, thinking that

> she hadn't changed, but there was no question that he had. He was putting distance between himself and the family that had been completely open to him and given him everything (McEwan, 2001, p. 28).

It is in this context that she refuses his offer to fill the vase for her, struggling with him when he reaches out to take it from her. She is refusing to let him to continue to play the part of a social inferior and employee of the estate. When the vase – a precious family heirloom – breaks, she sees in his eyes «not shock, or guilt, but a form of challenge, or even triumph» (McEwan, 2001, p. 29). Cecilia, in other words, interprets Robbie's actions as a rejection of and challenge to the contemporary social norms that govern class relations.[5] She resents this – not because she wishes to uphold these norms, but because she thinks their interaction should be governed by another context: the amicitial context, that is, the context of interactions between friends. Within that context, his behavior is intolerably rude. Friends don't highlight class or wealth distinctions between each other. When Robbie makes a move to undress and retrieve the broken shard, she sees him as again playing up the class

[5] Of course, a repudiation of norms is ultimately meaningful in terms of the very norms it repudiates.

differences. Her reaction is to "show him" that she was not above doing things for herself. And thus she undresses in front of him to "punish" him for invoking the wrong context to govern their interactions, thereby "banishing" him (McEwan, 2001, p. 30).

Robbie, for his part, is also aware that the conflict between the social context and the amicitial context complicates their interactions. He interprets the increasing distance between them – the way «his childhood friend [was] now in danger of becoming unreachable» (McEwan, 2001, p. 80) – as a product of this conflict. At Cambridge, «she always seemed to find it awkward» to encounter him:

> That's our cleaning lady's son, she might have been whispering to her friends as she walked on. He liked people to know he didn't care – there goes my mother's employer's daughter, he once said to a friend. He had his politics to protect him, and his scientifically based theories of class, and his own rather forced self-certainty (McEwan, 2001, p. 79).[6]

But Robbie is also very much aware that the event is further complicated and rendered ambiguous by yet another context – the erotic context. As he reflects later on what had happened, he entertains the possibility that in undressing, «she had wanted to show him just how beautiful she was and bind him to her» (McEwan, 2001, p. 81). He even momentarily weighs the Freudian possibility that «she hid the unconscious desire to expose herself to him behind a show of temper» (McEwan, 2001, p. 81). For present purposes, it matters less whether this is a probable explanation of her motivations than that it is a coherent explanation, for its coherence points to the fact that her actions have significance within an erotic context whether she acknowledges it or not.

Here we see a point of intersection between contextual opacity, social opacity, and motivational opacity. When Cecilia climbs out of the fountain and stares at Robbie, her act is erotically charged regardless of what intentions she thinks she might have. She might be completely oblivious to the sexual dimension of the situation, entirely focused on the act of retrieving the vase shard. And yet, the act has a sexual significance for Robbie. What is salient for him, in that moment, is

> a drop of water on her upper arm. Wet. An embroidered flower, a simple daisy,

[6] Again, *not caring* that a situation is structured by a particular context (even if it is a sincere form of not caring) is still to see the situation as given meaning by that context.

> sewn between the cups of her bra. Her breasts wide apart and small. On her back, a mole half covered by a strap. When she climbed out of the pond, a glimpse of the triangular darkness her knickers were supposed to conceal [...] The way her pelvic bones stretched the material clear of her skin, the deep curve of her waist, her startling whiteness (McEwan, 2001, p. 79).

Bodily actions (and I include in this category speech acts) are ambiguous because, as Merleau-Ponty puts it, the body is not "the transparent integument of spirit" – that is to say, a bodily act does not contain a clear and determinate meaning that could be separated from it and inserted without change into other vehicles, like taking a peanut from its shell and putting it into an M&M. But at the same time, what the bodily act *is* is inseparable from its meaning.

So it's both right to say that a bodily action only is what it means, and that it only means in virtue of what it is. In addition, the meaning depends heavily on how others respond to it, and on which context(s) are activated in their response. The meaning of the event (and the actions that constitute it) is obviously very different if we situate it in a social/political as opposed to erotic context. If one is unable to see through this contextual opacity, then one is unable to disambiguate the meaning of the event, and say definitively what it *was*, what *happened. Atonement* illustrates this by describing an awkward series of exchanges which follow as Robbie, Cecilia, and others struggle to get a clear grasp on the meaning of the utterances and the actions each is performing. They initially fail to see the meaning because these events are profoundly ambiguous, and they are ambiguous because they simultaneously participate in several different meaningful contexts.

Merleau-Ponty insists that ambiguity and opacity of this sort are essential, not accidental features of existence: «ambiguity is of the essence of human existence, and everything we live or think has always several meanings» (Merleau-Ponty, 1962, p. 169). Moreover, it is impossible to separate completely the different contexts which determine the meaning of a thing. They become so interfused that they can't be teased apart:

> this existence is the act of taking up and making explicit a sexual situation, and that in this way it has always at least a double sense. There is interfusion between sexuality and existence, which means that existence permeates sexuality and vice versa, so that it is impossible to determine, in a given decision or action, the proportion of sexual to other motivations, impossible to label a decision or act 'sexual' or 'nonsexual'. Thus there is in human existence a principle of indeterminacy, and this indeterminacy is not only for

> us, it does not stem from some imperfection of our knowledge, and we must not imagine that any God could sound our hearts and minds and determine what we owe to nature and what to freedom. Existence is indeterminate in itself, by reason of its fundamental structure, and in so far as it is the very process whereby the hitherto meaningless takes on meaning, whereby what had merely a sexual significance assumes a more general one, chance is transformed into reason (Merleau-Ponty, 1962, p. 169).

Thus the contexts present themselves, Merleau-Ponty says, not as an explicit set of references, but as a hazy or «ambiguous atmosphere» which is «at all times present» but not necessarily at all times invoked: «there are here blurred outlines, distinctive relationships which are in no way "unconscious" and which, we are well aware, are ambiguous, having reference to sexuality without specifically calling it to mind» (Merleau-Ponty, 1962, p. 168). By bearing on us in this way, sexuality doesn't always preoccupy us, but it is ready to be made an active determinant of meaning at any moment (as the juvenile joke "that's what she said" demonstrates). Other contexts are similarly more or less ready to be invoked.

With this account of ambiguity and the various forms of opacity, I believe that Merleau-Ponty has provided us with the tools we need to make sense of the kind of 'failures to see' that make up Sartrean bad faith. Recall Sartre's example of the young woman (considered above). On their date, the young man expresses a sexual interest in her, which she succeeds in ignoring by «disarm[ing]» his conduct of its «sexual background; she attaches to the conversation and to the behavior of the speaker, the immediate meanings, which she imagines as objective qualities» (Sartre, 1956, p. 96). Sartre explains:

> she refuses to apprehend the desire for what it is; she does not even give it a name; she recognizes it only to the extent that it transcends itself toward admiration, esteem, respect and that it is wholly absorbed in the more refined forms which it produces, to the extent of no longer figuring anymore as a sort of warmth and density.

In other words, she exploits the contextual opacity to disregard the meaning that the actions would have within a sexual context. «But then,» Sartre continues,

> suppose he takes her hand. This act of her companion risks changing the

> situation by calling for an immediate decision. To leave the hand there is to consent in herself to flirt, to engage herself. To withdraw it is to break the troubled and unstable harmony which gives the hour its charm. The aim is to postpone the moment of decision as long as possible. We know what happens next: the young woman leaves her hand there, but she *does not notice* that she is leaving it. She does not notice because it happens by chance that she is at this moment all intellect (Sartre, 1956, p. 97).

Here, of course, we have a clear instance of a motivated failure to see. How is it that the young woman fails to see that she is leaving her hand there? More importantly, how is it that she does not recognize the sexual context which, in some sense, is contributing to the "charm" of the moment? The answer has to do with the way that we "activate" particular contexts in particular situations.

Merleau-Ponty provides important insight into this phenomenon through his explanation of certain pathological cases of forgetfulness (these are cases where, for instance, one forgets how to speak, or forgets the existence of important artifacts). Merleau-Ponty writes:

> our memories and our body, instead of presenting themselves to us in singular and determinate conscious acts, are enveloped in generality. Through this generality we still 'have them', but just enough to hold them at a distance from us. We discover in this way that sensory messages or memories are expressly grasped and recognized by us only in so far as they adhere generally to that area of our body and our life to which they are relevant.

This "general adherence to a relevant area" I take it, amounts to what I've described as "activating a context."

> Such adherence or rejection places the subject in a definite situation and sets bounds, as far as he is concerned, to the immediately available mental field, as the acquisition or loss of a sense organ presents to or removes from his direct grasp an object in the physical field. It cannot be said that the factual situation thus created is the mere consciousness of a situation, for that would amount to saying that the 'forgotten' memory, arm or leg are arrayed before my consciousness, present and near to me in the same sense as are the 'preserved' regions of my past or of my body. No more can it be said that the loss of voice is voluntary. Will presupposes a field of possibilities among which I choose: here is Peter, I can speak to him or not. But if I lose my power of speech, Peter no longer exists for me as an interlocutor, sought after or rejected; what collapses is *the whole field of possibilities*. I cut myself off even from that mode of communication and significance which silence provides (Merleau-Ponty, 1962, p. 162, emphasis supplied).

Merleau-Ponty's perceptual paradigm provides us with an important piece of the puzzle, then. In a motivated failure to see, I don't necessarily have to conceal from myself the very thing which I don't want to apprehend. Instead, by allowing a whole context or field of possibilities to lapse back into the overall atmosphere, the meanings that this context contains also withdraw. The man's sexual actions are most at home in a sexual context. But the young woman blinds herself to this by letting other contexts take the lead in orienting her to the situation – something she can do because the event is inherently ambiguous. Having done that, all manner of possibilities withdraw from view, including the possibility of removing her hand from his, because she now overlooks the context in terms of which such possibilities would make sense. Of course, the sexual significance of the act retains an ambiguous presence to the degree that the sexual context remains hazily in the overall atmosphere.

I suggested earlier that an interfusion of contexts is actually an enabling condition of human freedom. We are always open to a number of different dimensions in which we can act, and thus we are not locked into any of them. An important part of our agency is the ability to switch contextual horizons. But at the same time, we could not move confidently and transparently in any of them if they were all equally salient. That means that the contexts work both by making some relationships salient, but also by withdrawing relationships when the context is not activated. So it would be a mistake, according to Merleau-Ponty, to think that we can or should want to disambiguate completely the situations we encounter in everyday life. We cope with a situation, on Merleau-Ponty's view, not by reducing it to a univocal meaning, but by recognizing which of the meanings, given our current intentions and desires, are salient and operative. Where we run into trouble is in trying to move as if one context were the only relevant one, when in fact it is another one that actually affords the optimal way to cope with the situation.

The ability to act on the basis of an unclear foundation is what Sartre (and Merleau-Ponty) mean by "faith". Sartre describes faith as «adherence of being to its object when the object is not given or is given indistinctly» (Merleau-Ponty, 1962, p. 112). Thus, it's not necessarily a cognitive state. An 'adherence of being' is a reliance in my way of acting in the world. So to have faith is for my actions to rely on something which is either not given at all, or not given clearly and unambiguously. To the extent that ambiguity and opacity as we have described them are pervasive, indeed, are conditions of agency, faith is inescapable. For instance, I have faith that the floor will support my weight –

in my being (standing here), I rely on the floor's supports, even though I have no apprehension of them at all. Or, I have faith that Pierre is my friend, insofar as I trust him to look out for my best interests, even though some of his actions (talking behind my back, arguing with me over petty points) are ambiguous at best. Thus "faith" is not something we should want to do without. What we do want, however, is to make sure that our faith is not bad – is not insidiously blinding us to dimensions of our existence that we lose sight of only to our detriment.

We cannot say, as one might at first be tempted to say, that bad faith is merely faith in something for which we lack adequate justification. This would, as Merleau-Ponty points out, make all faith into bad faith, since

> faith – in the sense of an unreserved commitment which is never completely justified – enters the picture as soon as we leave the realm of pure geometrical ideas and have to deal with the existing world. Each of our perceptions is an act of faith in that it affirms more than we strictly know, since objects are inexhaustible and our information limited (Merleau-Ponty, 1964, p. 179).

One could even, in good faith, base one's existence on something which was false or illusory. The young woman in Sartre's example, for instance, is a different case than another similarly situated woman who in all innocence lacks any understanding of the man's ultimate intentions. She might agree to meet him for lunch, for instance, believing in all good faith that he wanted to discuss with her a work project. There is a difference, then, between on the one hand avoiding the recognition that one's faith is false (that is bad faith), and on the other hand maintaining in all sincerity an orientation to the world that is guided by a faith in something that turns out to be illusory. Bad faith is not the same as a good faith belief in something illusory.

Thus, the motivational element in the notion of bad faith as a "motivated failure to see" is decisive. A helpful illuminating example is the distinction between losing a true love, versus discovering that one was misguided in thinking that one was in love. Merleau-Ponty explains that a true love

> summons all the subject's resources and concerns him in his entire being, whereas mistaken love touches on only one persona [...] True love ends when I change, or when the object of affection changes; misguided love is revealed as such when I return to my own self. The difference is intrinsic. But as it concerns the place of feeling in my total being-in-the-world, and as mistaken love is bound up with the person I believe I am at the time I feel it, and also as, in order to discern its mistaken nature I require a knowledge of myself which I

> can gain only through disillusionment, ambiguity remains, which is why illusion is possible (Merleau-Ponty, 1962, p. 379).

The same could be said for good faith in something false. It would penetrate to every corner of our way of being in the world. A bad faith, by contrast, can only be maintained to the degree that we can avoid carrying it into certain situations or contexts where it will clearly not work. Bad faith is motivated to maintain itself by avoiding those situations or contexts.

So, to summarize the argument to this point, a consideration of Merleau-Ponty's phenomenology of perception has helped us develop and articulate the "ambiguity necessary to bad faith." What an act or event means is dependent on a number of different factors or dimensions, including the context of relationships to which it belongs, how it unfolds temporally, what it means to a relevant community of observers, and what intentions or motivations were responsible for it. Each of these factors is existentially opaque, meaning that the act or event *is* what it is precisely because it doesn't belong uniquely to any particular temporal sequence, social community, meaningful context of relations, intentions, or motivations. Any particular conduct or event can stand within a plurality of different definitive or constitutive relationships. This renders it profoundly ambiguous. It is not clear what things are, what actions mean, what is relevant or important or salient in a particular situation. This is not necessarily a problem – opacity enables action in general by giving us the leeway to *not* attend to things that might interfere with our ability to act, and it enables agency in general by freeing us to switch horizons. But we must necessarily act on faith – that is, launch ourselves into a course of action that follows up particular meanings, without certainty that this is the course that will allow us to best navigate the situation we are in. Put otherwise, our ability to act in the world at all depends on an ordinary "motivated failure to see" – we attend to certain significations by not attending to others. This approach also shows how ambiguity need not be a result of an epistemic failing on the part of the agent – there is no amount of knowledge about the facts or about his or her intentions that can prevent an action or event from being ambiguous when existential opacity prevails.

An ordinary motivated failure to see passes over into bad faith, however, when (a) we are motivated to preserve ourselves in an orientation to the situation that is somehow less than optimal, and (b) we do this by avoiding recognition of those features of the situation that would force us to confront the fact that we are coping in a less-than-optimal way. Let's return now to

Sartre's account of bad faith to see if we can come up with a good general way of specifying what it is that we are motivated not to see in bad faith.

4. Facticity and Transcendence

Bad faith, Sartre explains, «utilizes the double property of the human being, who is at once a *facticity* and a *transcendence*» (Sartre, 1956, p. 98).

By "facticity," Sartre means the brute, concretely existing features of the universe *insofar as* we bestow meaning on them. For Sartre, occurrent entities have no meaning – they just are «a particular "this"» (Sartre, 1956, p. 132), a contingent fact (see Sartre, 1956, p. 29). As we try to make sense of the brute, concrete, particular features of ourselves and the entities we encounter, we uncover their manifold relations to each other, and we experience them as having sensory qualities[7]: «it is impossible to grasp facticity in its brute nudity, since all that we will find of it is already recovered and freely constructed» (Sartre, 1956, p. 132). (Our discussion of contextuality and contextual opacity has already drawn on the idea that the meaningful world as we encounter it is relationally structured.) I as an embodied being,[8] the people and things around me, the history that gave rise to a person like me – all of this is factical, a «recovery and freely constructed» appropriation of brute existence into meaningfulness. Because facticity always involves a meaningful appropriation, I am not absolutely bound by it – I can reconstruct it by making other meanings salient. But because facticity is ultimately grounded in a brute existence, there are limits to how I can appropriate it. We are not at liberty to construct it in any way we wish. Sartre argues, for example, that a café waiter is not inherently, in his brute existence, a café waiter: he «must *play at being* a café waiter in order to be one» (Sartre, 1956, p. 131). Yet given his body, his history, the concrete situation in which he finds himself, «it would be in vain [...] to play at being a diplomat or sailor» (Sartre, 1956, p. 131).

By "transcendence," Sartre means «the pro-ject of self beyond» (Sartre, 1956, p. 52), that is, any going beyond brute contingent existence. All the acts

[7] These, of course, are themselves relational properties – ways contingent entities give themselves to us.

[8] In virtue of my body, I am «an ensemble of structures [...] whose totality is an absolute concrete» (Sartre, 1956, p. 675).

of consciousness – intending, representing, desiring, wishing, imagining, and so on – are acts of transcendence. The factical itself only *is* in virtue of our capacity for transcendence, since meaning always involves a projecting beyond. But the human capacity for transcendence also makes it possible for us to be free, to «remake the *Self*» (Sartre, 1956, p. 72), thereby going beyond and altering our current facticity.

«These two aspects of human reality» – facticity and transcendence – «are and ought to be capable of a valid coordination» (Sartre, 1956, p. 98). Indeed, Sartre argues that a fundamental aim of human existence is to achieve a valid coordination – to be a free manifestation of my facticity, or to bring my facticity into conformity with my free self-projecting. We want to be at one with ourselves, to incorporate the concrete facts of our embodied insertion into the world into our aspirations, so that who we are and what we do is in harmony with our highest ideals (see Sartre, 1956, p. 472). But there are numerous obstacles to achieving a valid coordination. One is that we are beset by the ambiguity and opacity that we outlined in the last section, thus interfering with our ability to even know what our facticity is, or to be clear about what it is that we aspire to be. For Sartre, this locks us into a perpetual cycle of conflict with others. We need them to help us determine our facticity, but we are also constantly being objectified by others, and thus alienated from ourselves by their interpretations of us (see Sartre, 1956, Part Three: "Being-for-Others").

Sartre posits other psychological obstacles to achieving a valid coordination of facticity and transcendence – obstacles which give rise to bad faith. «Bad faith,» Sartre explains, «does not wish either to coordinate them or to surmount them in a synthesis» (Sartre, 1956, p. 98). It is troubled by its facticity, or by its transcendent desires. Under such conditions, we suffer from a "dis-integration" of facticity and transcendence. What we do or what we are stands in contradiction to our aspirations and ideals, and the person in bad faith wishes to avoid responsibility for this state and the painful work of self-transformation that coordinating our facticity and transcendence would entail.

It is to this state of disintegration, and our responsibility for it, that bad faith wishes to blind us. Put in the most general terms, then, bad faith is the motivated failure to see that we are responsible for the dis-integration of our facticity and transcendence. It is motivated by our desire not to take responsibility for this dis-integration. «The very project of flight,» that is, the strategies for not seeing that characterize bad faith, «reveals to bad faith an

inner disintegration in the heart of being [...] But it denies this very disintegration as it denies that it is itself bad faith» (Sartre, 1956, pp. 115-6).

By way of illustration, recall the puzzle we faced over the fact that both the gay man and the champion of sincerity were in bad faith, according to Sartre. We can now see that they actually present complementary forms of a motivated failure to see one's responsibility for the dis-integration in his or her being. For instance, Sartre explains that when the gay man denies that he is gay, he

> would be right actually if he understood the phrase "I am not [gay]" in the sense of "I am not what I am." That is, if he declared to himself, "To the extent that a pattern of conduct is defined as the conduct of a [gay man] and to the extent that I have adopted this conduct, I am a [gay man]. But to the extent that human reality cannot be finally defined by patterns of conduct, I am not one" (Sartre, 1956, p. 108).

So the authentic response to his actions would be to acknowledge that his action belongs to several contexts – several "patterns of conduct" – and is not reducible to any of them. One of these is the erotic context, and thus he ought to acknowledge that he is gay relative to that context. If his conduct were integrated with a self-understanding of his motivations and projects as those of a gay man, then he would not be in bad faith. He is in bad faith to the extent that he tries to deny that his actions receive their sense from that context. But the champion of sincerity is equally in bad faith to the extent that he is trying to restrict his friend's actions to just one context, thereby denying him the capacity for transcendent self- re-creation.

The dis-integration of facticity and transcendence is also obscured by temporal opacity: «Let us note» Sartre reminds us, «the confusing syntheses which play on the nihilating ambiguity of these temporal ekstases»[9] (Sartre, 1956, p. 100). I can try to reduce my transcendence to my facticity if I see only my past, what I have been, blinding myself to my future and my capacity for later changes. Or I can deny my facticity if I overlook what I have been, insisting only on my freedom to recreate oneself. I exploit temporal opacity, in other words, to not see the lack of fit between what I have been and what I aspire to be. The young woman exploits both temporal and contextual opacity to avoid confronting the dis-integration that prevails between her facticity and transcendence. She takes the meaning of the man's interest in her, his glance,

[9] The temporal ekstases are the dimensions of past, present, and future.

his gestures, his touch, as fixed manifestations of the higher – she treats them as if there is no ambiguity to them, they simply and fully exhaust the significance of these acts. And she exploits the temporal opacity «to postpone the moment of decision as long as possible» (Sartre, 1956, p. 97). That is, his acts mean what they will develop into, but they haven't developed into that yet. That gives her leeway to take them as something else.

In summary, then, I am proposing that the content of bad faith as an attitude needs to be understood as a particular type of a motivated failure to see. We characterize the content of bad faith by specifying what is not being "seen," that is, what is not being allowed to figure in our comportment in the way that it is. According to Sartre, what all forms of bad faith fail to see is the way that one's actions and intentions are integrated (or, more precisely, the way in which they lack integration). In virtue of our transcendence, our actions are implicated in a plurality of overlapping contexts, none of which is capable of uniquely determining the meaning of the action. It is only in virtue of our facticity, or rather, in virtue of the integration of our transcendence with our facticity, that our intentions have any specific content at all. But because of the opacity of human existence, it is always ambiguous which context is governing any particular action, which actions are expressing our intentions, and, indeed, whether any particular event is in fact an action expressing an intention or a mere accident. Bad faith exploits ambiguity to obscure the dis-integration of our actions and intentions, thereby reducing our facticity to objective facts, and elevating our transcendence to a radically ungrounded freedom.

Let me conclude by sketching out a few theses about the conditions of the possibility of inauthentic modes of existence. I would propose that these theses hold quite generally for existentialist accounts of inauthenticity, despite other distinctions between them. In a future paper, I will try to apply these theses to the case of Heidegger.

My claim is that the possibility of inauthenticity requires:

1) a domain where it is not clear what the meaning of our acts are (not even to ourselves), a domain where significance escapes intention. Such a domain is one where the meaning of our actions is determined in a significant part by factors "external" to the actor (social norms, worldly contexts, and temporal developments, to name a few). The world and others in the world must sustain a meaning that is independent of my intentions.

2) a domain where, nevertheless, an intention can have a meaning which is independent of the meaning of the action. The type of self-deception involved in inauthenticity requires a mismatch between what my intention means according to the "external" determinants of its meaning, and what it means to me.

The meaning of the intention, in other words, is in part dependent on what it produces, but in part it is identifiable as the intention it is independently of what it produces. As Sartre explains:

> Upon any one of my conducts it is always possible to converge two looks, mine and that of the Other. The conduct will not present exactly the same structure in each case. But [...] there is between these two aspects of my being, no difference between appearance and being – as if I were to myself the truth of myself and as if the Other possessed only a deformed image of me. The equal dignity of being, possessed by my being-for-others and by my being-for-myself, permits a perpetually disintegrating synthesis and a perpetual game of escape from the for-itself to the for-others and from the for-others to the for-itself (Sartre, 1956, p. 100).

Finally, inauthenticity requires:

3) that there is a kind of nothingness to the self. The self admits of multiple equally valid interpretations, over which I exercise some but not exclusive authority.

«Bad faith,» Sartre explains, is «intended to fill up the nothingness which I *am* in my relation to myself,» and in this way «precisely implies the nothingness which it supresses» (Sartre, 1956, p. 83).

REFERENCES

Heidegger, M. (1927). *Sein und Zeit.* Tübingen: Max Niemeyer.

Merleau-Ponty, M. (1962). *Phenomenology of Perception.* (tr. by C. Smith). London: Routledge.

Merleau-Ponty, M. (1964). Faith and Good Faith. In M. Merleau-Ponty, *Sense and Non-Sense* (tr. by H. Dreyfus & P. Dreyfus). Evanston, Ill.: Northwestern University Press, 172–181.

McEwan, I. (2001). *Atonement.* London: Random House.

Sartre, J-P. (1956). *Being and Nothingness.* Trans. Hazel Barnes. New York: Washington Square.

Wrathall, M. (2005). Motives, Reasons, and Causes. In T. Carman & M. Hansen (Eds.), *The Cambridge Companion to Merleau-Ponty.* Cambridge: Cambridge University Press, 111–128.

Wrathall, M. (2009). On the "Existential Positivity of our Ability to Be Deceived". In C. Martin (Ed.) *The Philosophy of Deception.* Oxford: Oxford University Press, 67–81.

Book Review

Delusions and Other Irrational Beliefs

Lisa Bortolotti
OUP, Oxford, 2010

Elisabetta Sirgiovanni[*]
elisabetta.sirgiovanni@isgi.cnr.it

Delusional people are people saying very bizarre things like they are dead, their spouse is a robot, the TV star is talking to them, they are possessed by the devil, aliens are following them, and so on. Even though we know that they are not identical, terms like "delusion" and "mental illness" are often used as synonyms in ordinary language. This comes from what psychopathology tradition handed down: delusion is the key psychopathological phenomenon, although essentially *un-understandable* (Jaspers, 1959). In her book Delusion and Other Irrational Beliefs, Lisa Bortolotti explores the topic of delusion from the epistemological perspective of analytical philosophy.

Do delusional people really believe what they say? This question is as interesting as it is pressing for clinics. From the very beginning however this work is engaged in defending two core ideas. First, understanding belief, regardless of whether it is a "real pattern" or not (Dennett, 1991), is relevant to understanding what delusions are. Second, delusions can be beliefs like others. This is only a small part of what makes this book a fascinating and indispensable work.

The aim of the book is arguing against accounts which deny the doxastic nature of delusion. In philosophy of mind, the claim that delusions are not beliefs is taken as a *modus tollens* argument deriving from the general premise that all beliefs presuppose a background rationality, as assumed by belief attribution theory in the Davidson-Dennett tradition. In other words, since delusions do not meet the rationality constraint (since they are irrational phenomena), they are not beliefs at all.

Chapter 1 is an opening background section devoted both to the rationality constraint in belief attribution theory and to conceptions and taxonomy of

[*] Istituto di Studi Giuridici Internazionali (ISGI-CNR, Rome, Italy).

delusions. The problem of the aetiology of delusion is explored here, and comparisons are made to other similar phenomena like self-deception, obsessive thoughts, confabulation and hypnotically induced belief. This section of the book is accurate and rich. From the first pages, the book impresses us for its scholarliness and for author's deep knowledge of the topic in all its relevant aspects. Dominic Murphy is right in affirming that this book is «a tour de force» (2011, p. 1).

The book structure reflects the main counterarguments to which the author aims to reply. Each chapter is dedicated to common accounts of belief to their relation to the theory of rationality. Beliefs are shown not to be procedurally rational (Chapter 2), epistemically rational (Chapter 3), and agentially rational (Chapter 4). Moreover, as suggested by the book's title, delusions and ordinary beliefs are shown to share the same features of irrationality without compromising either their doxastic nature, or their contribution to the construction and preservation of the conception of the self (Chapter 5).

In this way, the background rationality constraint is shown to be no more than a philosophical myth, and can thus be rejected. That is exactly what experimental psychology has told us for a while (Stein, 1996). The minimal belief account Bortolotti suggests is constructed in terms of possibility. Beliefs must be integrated in a system that has some (not any) inferential relations with other intentional states; they are sensitive (not responsive) to evidence or argument; they can be manifested in behavior; they can be self ascribed and defended with reasons. It is less clear why delusions are pathological whereas other beliefs are merely irrational.

So a question should be raised: can we establish whether delusional people are really believing what they say on the base of belief attribution theory? What is referred to as belief ascription is a heuristic strategy from the observer's perspective, where the interpreter assumes mental states in others on the basis of behavior to explain and predict their actions. Rationality constraint is a heuristic constraint too, which is presupposed in order to make interpretation work. The theoretical background of this story goes back to the problem of the radical translation in Quine (1960): if a native speaker of an unknown language says something illogical, I must conclude I have not understood him properly. According to the principle of charity, a bad translation is more improbable than the explicit violation of logical principles. This is likely to be a conventional rule. Can we characterize delusions as beliefs from the intentional stance? Maybe we cannot. Belief characterization as offered in the

book could account for why we are conventionally justified anyhow (even without the rationality constraint) to expect real beliefs from irrational patients, but we may be wrong about their having real beliefs (maybe some delusions are, some are not). The reason is that we do not know if the folk-psychology interpretative strategy is a sufficient tool for establishing the presence of beliefs. Probably it is not. Maybe holding firmly a belief is not a fact that can be established from the intentional stance, but it could be established by neuroscience, if correlated brain patterns are discovered in future. The alternative view is quite old-fashioned in cognitive science. We may expect neuroscience to empirically find brain patterns of what believing something means. Besides we are prepared to possible cases in which there might be also no clear self-transparence of our beliefs at the first-person narrative level.

Main concerns about the book include problems like natural kinds (are beliefs natural kinds?), tools to denote them (should we use philosophical or empirical tools?) and the relationship between the disciplines involved (folk-epistemology, scientific psychology or neuroscience). Accordingly, we cannot ignore the fact that many contemporary philosophers (the sort called eliminativists) claim that beliefs might not exist at all. Bortolotti intentionally avoids the problem of scientific reduction to some fundamental physical level. She is aware that there is an urgency of causal explanations in psychiatry coming from the medical model (especially, from cognitive neuroscience) and that present psychiatry taxonomy (the Diagnostic and Statistical Manual of Mental Disorders, or DSM) is in the middle of a big crisis. DSM a-causal descriptive approach gives no definitive solution to the problem of delusion and other mental symptoms, so we are looking forward to the neuroscientific reply. Nevertheless she does not commit herself to any hypothesis of underlying causal mechanisms of delusion (although stating to be more congenial to some version of the two-factor theory, p. 35) and of the existence of belief itself. Nevertheless «questions about belief ascription» she writes «are no less important in the age of neuroscience» (p. 1). She is right. Whether delusions are beliefs is a different question from what causes delusions and what are delusions at the level of neurocomputational mechanisms (a certain breakdown of a given neurocomputational mechanism). But a problem is: what remains of this discourse about the belief status of delusion if the notion of belief comes to be replaced by a mature neuroscience?

According to Murphy, this approach «may not serve as a foundation for a developed science of abnormal intentional stance» (2011, p. 4). In a more

recent article, Bortolotti clarifies that even if she uses beliefs as fictions, she wants to give a contribute to the development of such a science by «gradually revising our existing conceptual framework» (2011, p. 13). As for methodology, in the book the author identifies four aims for philosophy: working out the implications of empirical results; suggesting new avenues; drawing some conclusions; assessing the relationship between data and interpretation. The guiding role of philosophy for the scientific domain might be considered to be a little pretentious. Murphy states (2011) that the book approach is that of a *folk epistemology of delusions*. But what must be said is that the book approach is not that of a mere folk epistemology, even more modest than a strong *naturalized epistemology* (Quine, 1969). Quine theorized the view of naturalized epistemology in terms of *replacement naturalism* (Feldman, 2001), according to which traditional epistemology should be abandoned in favor of psychology. And this is not Bortolotti's approach. However, there is also a naturalized epistemology in terms of *cooperative naturalism* (Feldman, 2001) according to which empirical results from scientific psychology allow to make progress in epistemological questions. This seems to be more her approach. Bortolotti in fact claims that philosophical inquiry should not conflict with empirical findings (p. 7). Moreover she uses a lot of data and results from experimental psychology as examples that intervene to solve epistemological concerns. What is unclear is which is supposed to have the last word on conceptual issues, whether the philosophical or the scientific-psychological domain.

Admittedly these remarks should not make one approach the book with suspicion. This book is an important contribution to the recent delusion debate. The book can also usefully work as a cognitive science textbook on delusion. The author introduces the topic in depth, covering all the right issues in a way that no one has done before. The bibliography is also an extremely rich guide for those interested in further exploring the subject, and also for finding sources relevant to disputes in the philosophy of mind.

REFERENCES

Bortolotti, L. (2011). In Defense of Modest Doxasticism about Delusions. *Neuroethics*. Published online, doi: 10.1007/s12152-011-9122-8

Dennett, D.C. (1991). Real Patterns. *Journal of Philosophy*, *88*(1), 27–51.

Feldman, R. (2001). Naturalized Epistemology. In Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/entries/epistemology-naturalized/ Accessed on: 2011, December 21st.

Jaspers, K. (1959). *Allgemeine Psychopathologie*. Berlin: Sprinter.

Murphy, D. (2011). The Folk Epistemology of Delusions. *Neuroethics*. Published online, doi: 10.1007/s12152-011-9125-5

Quine, W.V.O. (1960). *Word and Object*. Cambridge (Ma): MIT Press.

Quine, W.V.O. (1969). Epistemology Naturalized. In W.V.O. Quine, *Ontological Relativity and Other Essays*. New York: Columbia University Press, 69-90.

Sirgiovanni, E. (2007). Verso una Spiegazione Cognitiva del Delirio. *Psicopatologia Cognitiva*, *3*(1–2), 179–195. Reprinted in G. Cardamone & R. Dalle Luche (Eds.) (2009), *Paranoia, Psichiatria e Antropologia*. Pisa: Edizioni ETS, 39–74.

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford: Clarendon Press.

Book Review

The Philosophy of Deception

Clancy Martin
Oxford University Press, 2009

Brad Bolman *
bolman@college.harvard.edu

The release of WikiLeaks documents triggered a debate between two perspectives on truth and transparency. One side wanted to know what was happening. More importantly, they wanted the truth. This view was opposed by another insisting that secrets were necessary – transparency and truth would be destructive in the wrong hands – and that we did not need to know *everything*. After all, they reasoned, a little deception can be a good thing! And in this, there was a kernel of truth: we often find it comforting to forgo truth in order to find safety in our ignorance or in defense of threats to our worldview. Much of modern politics seems to rely on deception, intentional or not, intended to avoid confronting the supposed truths that sustain our world. The distinction between withholding and putting forward false information is blurred. It is commonplace to assume that we do not want to be deceived, but the WikiLeaks debate demonstrated plainly that the desire to end deception and find truth is hardly as clear – and desired – as it might seem. «We cannot imagine social intercourse without opacity» writes Robert C. Solomon in his chapter «Self, Deception, and Self-Deception in Philosophy,» which serves to introduce many of the areas of contention throughout the book (p. 21). Here *The Philosophy of Deception,* a highly diverse and strong collection of many of the leading thinkers on the philosophy of lying and deceit, intervenes.

As Clancy Martin explains in the introduction, «Lies and self-deceptions seem to exist along a continuum,» from the direct lie that is not self-deceived – We did not do *this* (even though we know we did) – to the other extreme, where one is entirely self-deceived – That is not why we're doing *this* (even though it is) – «and in the middle the many cases where the lies we tell others are inseparably mixed up with the lies we tell ourselves» (p. 3). *Philosophy of Deception* engages this idea thoroughly, and from varied perspectives: Mark A.

* Harvard University, Cambridge, MA, USA.

Wrathall's phenomenological investigation into "perceptual deception" contrasts with Kelly Oliver's psychoanalytic examination of the possibility of an inherent self-deception in our existence. Even though the approaches differ, the narrative of the book is clear and gratifyingly cohesive for an edited collection.

The book aims to unite two fields: the study of lying and the study of self-deception. The book's thesis, as Martin argues, is that these two fields «which had been undertaken almost entirely independently, could both benefit from a sustained examination of the many traits they have in common, of the ways they work together, of similarities and differences in their structure, their practice, their ethics» (p. 4). Its overarching purpose, then, is to explore bonds between these two forms of inquiry and ask what syntheses might come out of this reading. The strongest point that the work makes is the importance of the analogy of deception and self-deception. For Martin, « [Mele's] understanding of self-deception can provide us with a more helpful analogy with deception» (p. 11). Why? Precisely because the most fruitful cases to investigate involve beliefs that are not as simple as believing "p and not-p." The analogy demonstrates that self-deception is much less about the attempt to trick oneself, than about a person being affected or motivated in a certain way that falls in line with his/her interests. It becomes a question of the confusing, tricky, vague ways that deception and self-deception manifest themselves because in these difficult situations, in «the way the mind actually works, [that] we are human» (p .11).

Philosophy of Deception is divided into two halves. The first, which deals with "the *how* of deception," focuses on the role of deception in our lives; the second half, which takes a more theoretical direction and presentation, analyzes concepts like lying and self-deception. What makes *The Philosophy of Deception* work is the subtle way in which all of the pieces stand in dialogue with each other. Following Solomon's first chapter on how lying is in many ways a necessary part of social existence, Harry Frankfurt's chapter "On Truth, Lies, and Bullshit," the only previously published material in *Philosophy of Deception*, is enlightening particularly because of how he probes deeper into the way deception changes and alters interpersonal situations. To quote Frankfurt, a lie is damaging precisely because «It reveals that *our own nature* [...] is unreliable, having led us to count on someone we should not have trusted» (p. 40). In effect, then, lies make you feel "a little crazy" by rejecting a personal assumption of the ability to guide oneself through social situations

accurately. Frankfurt's claim that «Lies are designed to damage our grasp on reality, » contrasts with the assertion beginning Kelly Oliver's chapter "Duplicity Makes the Man, Or, Can Animals Lie?" that, «Insofar as unconscious forces drive us beyond our control and even beyond our knowledge, then we are all and always a bunch of liars» (p. 104). If this is the case, as Oliver goes on to investigate, the Lacanian understanding of "lying" can problematize the assumption that lying and deception are predominantly *human* behavior. From that examination of the unconscious, we can jump to a materialist investigation in Paul Ekman's wonderful chapter on "catching" lies through microexpressions. Ekman asks, if learning how to notice and catch lies is possible, why is it that we do not all do it? His conclusion is sobering: «Anyone who says there is an absolutely reliable sign of lying that is always present when someone lies and never present when someone is truthful is either misguided or a charlatan» (p. 133). There are a multitude of approaches here, from William Ian Miller's look at "who we root for" in the classical tales of tricksters – concluding that «It is not always clear» (p. 65) – to David Sherman's call to "remake the social world" through a new understanding of deception in relation to social being. The collection ends with Alfred Mele's "Have I Unmasked Self-Deception or Am I Self-Deceived" which introduces his notion of self-deception as motivationally biased belief acquisition and rebuffs some of his critics. This chapter makes a good end to the book particularly because it immerses the reader in a broad swath of the literature on self-deception while simultaneously leaving the question of deception open to further investigation.

To quote Amelie Rorty's chapter "User-Friendly Self-Deception: A Traveler's Manual," what *The Philosophy of Deception* does well is to «engage ourselves in the Stoic task of understanding the minute details of [self-deception's] operations» (p. 259). The central lesson in the book is a reminder of the risk for any philosopher of believing that any single theory can provide the absolute explanation of the nature of truth and lying. That, this book tells us, is just another form of self-deception. Rather than a feeling of theoretical schizophrenia which is always a risk of an edited volume – of course, our friends Hegel, Kant, and Plato show their faces frequently throughout adding a clear theoretical undercurrent – this collection succeeds in bringing a sustained investigation from multiple angles, cleverly self-referential, questioning, and continually searching. The obvious joke about a book about lying – that the truth about self-deception appears to be an oxymoron – seems

relevant here. Such an investigation into the meaning of deception, self-deception, and truth asks us to consider in our own lives both the power and the risk in investigating the truths – and, of course, the lies – large and small that we think and tell. An answer is not absolutely clear and it is doubtful that it will ever be, but *The Philosophy of Deception* should serve as a rallying point for scholars to continue in the quest to deepen our understanding of the intricate connections between deception and self-deception.

Interview

Amélie Oksenberg Rorty *

Edited by Patrizia Pedrini

AMÉLIE OKSENBERG RORTY is the Findlay Professor of Philosophy at Boston University and a Lecturer in the Department of Global Health and Social Medicine at Harvard University. Her *Mind in Action* (1988) consists of essays in ethics and philosophical psychology; she has also published a number of anthologies on Aristotle's ethics, his poetics and his rhetoric, as well as collection of papers on Descartes' Meditations. Continuing her interest in the philosophy of education (*Philosophers on Education*, 1998), she is now working on a book defending ambivalence: *On the Other Hand: The Ethics of Ambivalence*.

With Brian McLaughlin, Rorty edited and contributed to a seminal collection on self-deception (*Perspectives on Self-Deception*, 1988) which is still a classic on the subject. Rorty and McLaughlin acknowledge that «explaining, or explaining away, the phenomena of self-deception raises many of the central problems in the philosophy of mind» and rightly declare that they use «self-deception as a microcosmic case study that bears on a range of issues dividing contemporary philosophical psychology», because

> [...] disagreements about the existence and analysis of self-deception expresses disagreements about the unity of consciousness, homuncularism in psychological explanations, the criteria for the attribution of belief, the conditions of intentionality and rationality, the primacy of cognition in psychological processes, the relation between motivational and epistemic attitudes, the social formation and malformation of belief and self-deception, and moral constraints on responsible belief. (McLaughlin & Rorty, 1988, p. 1)

Rorty and McLaughlin were aware of the importance of these topics for epistemology and ethics, as well as the philosophy of mind. They therefore divided the collection into sections covering "The Analysis of Self-Deception" (part I), "The Epistemic Dimension of Self-Deception" (part II), "The

* I am grateful to Patrizia Pedrini for her searching questions and to M.R. Amiran, Aaron Garrett, Steven Gerrard, Aryeh Kosman, William Ruddick, and Richard Schmitt for helpful comments

Psychology of Self-Deception" (Part III), "The Social Dimension of Self-Deception" (Part IV), "The Moral Dimension of Self-Deception" (Part V), and finally also "Self-Deception and Literature" (Part VI).

This interview tries to focus both on Professor Rorty's explanation of self-deception and on her views on some ongoing open questions and recent controversies. I asked Professor Rorty to answer six questions, to which she offered extensive, challenging responses. We are all most grateful to Professor Rorty for having generously undertaken this task.

1. In your seminal work on self-deception, you defended the idea that self-deception becomes less mysterious once we accept a conception of the self as a «loosely organized system of relatively autonomous subsystems» (Rorty, 1988, p. 12). The view you held in the paper quoted was brilliantly capable of accommodating a phenomenon that Donald Davidson's view was perhaps making unnecessarily paradoxical. In this sense, you anticipated the spirit of Al Mele's "deflationary view" of self-deception (2001). Would you still subscribe to this view of the self and to how it applies to the explanation of self-deception, or have subsequent reflections changed your mind on this point, or refined your position?

I think that the familiar philosophical puzzles about the apparent incoherence of self-deception rest on views about the 'the self' as a unified and temporally continuous entity capable of acting from rationally monitored reflective self-awareness. So construed, the idea of the self is a theoretical construction, designed to accommodate cultural notions of individual agency and responsibility. Largely for the sake of rationalizing our practices of assigning responsibility, we treat the self as a psychologically and cognitively unified entity, capable of effective self-knowledge. The range of actions for which we hold individuals responsible varies with what we take to be within a normal agent's knowledge and reflective capacities. On the one hand, we hold individuals morally and legally responsible for a wide scope of voluntary agency, including their intentions as well as their actions; on the other hand, we accept a wide and generous latitude of excusing conditions to explain and exonerate failures of responsibility.

I believe that the idea of the self as a unified, conscious and presumptively self-aware entity is an ideal superimposed on a loosely organized system of

relatively independent but mutually supportive and interactive modular psycho-physical subsystems, only some of which are capable of 'internal' scanning. As Carruthers puts it,

> [such] modules might be isolable function-specific processing systems, all or almost all of which are domain specific, and whose operations aren't subject to the will. [These modules] are associated with specific neural structures (albeit sometimes spatially dispersed ones). [Although these modules are typically interactive] [...] their internal operations may [sometimes become] [...] inaccessible to the remainder of cognition. (Carruthers, 2006)

Van Leeuwen goes further: «The capacity for self-deception [...] is a spandrel [...] of other mental traits, i.e., a structural byproduct. The irony is that the mental traits of which self-deception is a spandrel/byproduct are themselves rational» (van Leeuwen, 2007).[1] Although individual persons are presumed to be normally conscious, capable of basic reflective introspection, the scope of their capacities for accurate self-awareness varies considerably. For instance, some people have acute self-knowledge in epistemic matters, but very little understanding of their motivational patterns: they are good at reflecting on what they believe, but are often mistaken about what they desire. Others are sensitive to their sensory and proprioceptive functioning but relatively unreflective and often mistaken about what they value.

Some modular sub-systems of the self function as internal scanners, dispositionally geared to monitor cognitive and psycho-physical operations as the need arises. Individuals vary 1) in their ability to coordinate scanning information with other cognitive and connative functions and 2) in the extent to which they can voluntarily control and direct their scanning operations. Some areas of psychological and cognitive functioning – for instance, high order cognition engaged in theoretical reasoning – tend to be more transparent than those engaged in preferences that were developed in infancy. For some people, conflicts of beliefs and desires are relatively transparent, easy to diagnose. Although they may find such conflicts troublesome, such people may be less subject to self-deception than are those who resist or deflect reflective scrutiny of conflicting beliefs and desires. Patterns of accessible scanning and accurate reporting can be affected by trauma; self-knowledge can become more or less acute with experience and with motivational changes. The more integrated and voluntary are a person scanning functions, the less is she

[1] See also Fodor 2000.

likely to be subject to self-deception. On the other hand, those with a low level of epistemic integration – those who tend not to monitor the consistency of their sub-systems – may simply be inconsistent or be mistaken about their beliefs. Because they never claimed self-knowledge, they may not be self-deceived. A great deal of apparent self-deception involves a contrast between the content of a conscious occurrent belief and that of an unacknowledged – and sometimes vague – dispositional belief. Because the criteria for the attribution of various types of belief vary, and because its ascription can be a matter of degrees, there may sometimes be more (and sometimes less) self-deception than meets the eye. In any case, self-deception is notoriously difficult to ascribe with any confidence because it typically occurs in opaque contexts.

2. The idea of the adaptive fitness of self-deception had been first and importantly defended in your writings on the topic. However, not all the scholars agree that all forms of self-deception are invariantly adaptive for the species, let alone that it will always make us flourish individually (according to criteria for the "flourishing" in question that a scholar might want to specify) or make us happy (e.g., Van Leewen, 2009)[2]. What's your thought about the new arguments produced by those who are sceptic about the adaptive value of self-deception?

The structural capacities for self-deception – the relative independence and compartmentalization of psychological and cognitive sub-systems – are adaptive for survival and for high level functioning. The functional independence of such subsystems promotes specialized and highly developed cognitive and psychological activities; it enables intensive focused attention; it protects sub-systems from doing infectious collateral damage to one another; it enlarges the diversified scope of psychological and cognitive functioning. By bracketing agents' awareness of risk, it enables them to act with confidence and conviction in situations of uncertainty and risk, to be devoted to personal and social commitments when closer scrutiny might distance them, to maintain an even tempo and temperament in the face of the erratic fluctuating circumstances.

[2] See also discussions in Martin 2009.

To be sure, not every instance – or even every type – of self-deception is beneficial, either for the individual or for the species. The psycho-physical structures that are adaptive for effective psychological functioning nevertheless also bring marked disadvantages and vulnerabilities. Functionally and structurally independent sub-systems increase the possibility of the failure of psychological integration; they can conduce to the kind of active disintegration that self-deception and akrasia sometimes represent. The benefits of compartmentalized functionally independent sub-systems are matched by the need for their integration, for accurate transparency and accessibility among them. Because the effective strategies of psychological and cognitive adaptivity are integrally connected to their vulnerabilities (and vice versa), their integration requires constant adjustment in ways that are rarely under voluntary or even conscious control. Ironically, such adjustment obviously presupposes the very integration it is meant to maintain.

Given the advantages of the structural capacities for self deception and the benefits of a great deal of self-deception, why does it have such a bad press? Why do we blame ourselves and disdain others for what is in many ways an adaptive and useful strategy, one that sustains many of our central activities? At least one of the draw backs of self-deception is that it is a powerful instrument of moral indifference and even cruelty. Consider how a self-deceiver might deflect criticism of his behavior by describing a shady negotiation as resourceful rather than as aggressive or by describing a fawning and flowery compliment as tactful rather than hypocritical. The brilliantly inventive and self-deceptive ability to find or to concoct a covering but deflecting description for a morally suspect action can provide the basis for a tangential moralizing justification that masks and disguise great wrongs. It enables us to blind ourselves to our motives and to the effects of our actions on others; even more dramatically, it enables us to ignore or misdescribe what we are actually doing. Self-deception allows us to abstract ourselves from our actions, remaining selectively ignorant of their presuppositions and consequences. Kant's severity describes the matter well: «[The] inner advocate expounds the law to [his] advantage [...] he grows deceitful, making use of the law for his own purposes, [as] a means of self-deception whereby he persuades himself that he has been acting rightly, on principle» (Kant, 1963, p. 137). Taking advantage of Kant's emphasis on the freedom of self-legislation, the self-deceiving Mafioso within adds: "You want moral principles? I can get them for you wholesale."

How does the apparently innocent self-deceiver manage to bring off his own deception? Self-deception is sometimes a free rider on referential opacity (Kaplan, 1986). Even the smallest, most precise actions or character traits are open to multiple descriptions whose tonal connotations, etymologies and classifications implicitly tend to direct its evaluation and justifiability. Although such descriptions are not substitutable *salva moralitate*, the louche self-deceiver treats them as fungible: she substitutes a morally permissible description of an action or trait for one that might be morally suspect. By treating a referentially opaque expression as if it were transparently substitutable *salva moralitate*, she gains ground for justifying the action to which it refers. Referential opacity allows the ingenious self-deceiver to find a resonant principle to justify whatever interests she favors by focusing on an astutely self-serving description of what she does. All she has to do is emphasize some features of her traits or actions as salient, others as recessive. Without actually lying to herself, the self-deceiver can present herself to herself as a morally decent if not actually estimable figure.

Hannah Arendt (2006) argued that the failure to think, the failure to notice or attend to the full description of what we do is often the first step in finding a convenient, apparently reasonable justification for great wrongs. Self-deception can take the form of astutely substituting a thin and morally innocent description of an action for one that would reveal its morally relevant thick description. Consider Eichmann defending himself by saying "I was just following orders to coordinate train schedules." That thin generic description of his action carries relatively neutral implications and expectations about its generic standard aims, settings, and outcomes. It carries an implicit standard justificatory explanation that tends to deflect the kind of attentive questioning that might press for a fuller, thicker description. A more robustly detailed thick description – "I consulted train schedules to plan a timetable for transporting gypsies to Auschwitz" – might have unmasked Eichmann's self-deceptive justification of what he did. But neither the thin nor the thick description of Eichmann's scheduling trains to Auschwitz necessarily reveals his motivational structure: he might have been an ordinary standard issue bureaucrat, primarily focused on doing whatever would undermine his rival in the SS Schutzstaffel. Or he might have been an obsessive compulsive, a man with a tidy, obedient mind whose attention was always focused on the minutiae of whatever he did. Quite independently of his motives or habits, Eichmann can be self-deceived about (the thick description of) his action in constructing a schedule for

transports to Auschwitz. The brilliantly inventive and self-deceptive ability to find or to concoct a covering but deflecting, tangential moralizing justification can mask and disguise moral failures. Eichmann might – or might not – have been self-deceived about his motives as well as about his action. His being self-deceived about his motives might – or might not – have explained his being self-deceived about his action. In any case, the evaluation of his motives is independent of his being self-deceived about what he did.

3. A very new question raised by Eric Funkhouser (2005) is what the self-deceiver wants and whether it ultimately gets what he wants. The controversy is still live and attracts much interest, and I would like to ask you about your current view of the motivational state of the self-deceiver.

Sometimes self-deception just happens: a self-deceiver need not always be motivationally prompted to deceive himself about his beliefs or about anything else, for that matter. A pattern of self-deception can become habitual as a result of a person's psychological history or his social milieu, without any particular motivation on his part. (Ruddick, 1988). Just as a painter can deceive a biographer or art historian, so too she can deceive herself about the merits of her work. Because her parents and friends successfully deceived her about her talent, she came to collude in the deceptive estimation of her talent. To be sure, sometimes such a painter may simply be chronically mistaken, but she might sometimes actively collude in keeping herself from realizing the truth of the matter. She can consistently be inventively obtuse, ignoring or denying the evidence given by critics, collectors and museum curators whom she normally admires and whose judgment she trusts. Her self-deceptive self-esteem can be habitual, without being specifically motivated.

In any case, not all self-deception is deception about the self, or about its beliefs and desires. Very roughly, X is self-deceived about p (where p can be any state of affairs) when 1) X has evidence that p, and 2) X directly or indirectly denies that she has evidence that p (or believes q, where X has evidence that q entails not-p); and 3) there is evidence that X is aware that she both believes and denies that p; and 4) X directly or indirectly denies that she has such evidence. In the second place, although these affirmations and denials can sometimes be motivated, they need not be prompted by a specific concurrent desire. To be sure, beliefs are, in the very nature of the case, truth-directed and truth-claiming, presumptively integrated in a truth-oriented

system of beliefs. In that sense, belief-claims carry second-order implications about the believer's commitment to truth-orientation. Those commitments need not, however, indicate anything about his wants or desires. It is not unusual for someone to want to be free of his commitments: he might sometimes not altogether unreasonably wish he were less committed to telling – or even to discovering – the truth. 'Beliefs' that are fully constituted or determined by non-truth-tracking second order motivations are nevertheless suspect as instances of bona fide beliefs, independently of whether they are self-deceived. Expressions of wishes rather than of beliefs, they may prompt self-deceptive claims without themselves being instances of self-deception.

4. Self-deception seems to involve a failure of self-knowledge (e.g., Scott-Kakures, 2002). Do you think this is correct and how would you characterize this failure?

Self-deception does not involve more failure in self-knowledge than we ordinarily have under 'normal' circumstances. We have very little self-knowledge to begin with: we are rarely able to articulate the scope and details of our values and commitments; we are often mistaken about our basic character traits; we are often at sea about whether we are prepared to affirm the logical entailments or presuppositions of propositions we take ourselves to believe. The limitations of self-knowledge do not necessarily involve self-deception: they typically indicate ignorance, diagnostic errors and sometimes simple disinterest. On the other hand, since self-deception is not necessarily deception about the self, not all self-deception involves a failure of self-knowledge. Sometimes it involves denials in the face of overwhelming evidence of the chicanery of friends or the corruption of colleagues.

Just as deception does not necessarily involve lying, so self-deception does not necessarily involve holding a false belief. It is possible to mislead or deceive someone by distracting them, by redirecting their attention to some inane or trivial truth. So too one can deceive oneself by paying careful and accurate attention to some distracting or tangential feature of one's experience, and so mislead or deceive oneself to ignore what might be most germane in the circumstances. As I suggested in my response to Question 2, referential opacity is the self-deceiver's friend: sometimes the canny self-deceiver need only substitute an alternative description of an action a description that captures his focused attention – to deceive himself about what he is doing.

5. What's your current view about the relationship between confabulation and self-deception? Hirstein (2004) argues for the view that there are some overlaps between the two phenomena but the debate is still open.

Some – but by no means all- self-deception is accompanied by a covering confabulation designed to explain away awareness of counter-vailing evidence to cherished or entrenched beliefs. It seems to me that "the overlap view" is overly and nervously intellectualistic: in practice, in the ordinary course of things, neither believers nor deceivers feel the need to explain –or explain away – the grounds for their attitudes.[3] Just as we do not confabulate to explain errors of judgment unless we are pressed to do so, so too we do not typically need to explain consistently deflected attention by confabulating. Indeed confabulation tends to highlight the self-deception, to make it suspect. Qui s'excuse, s'accuse. Self-deception typically remains unacknowledged and unexplained: the entrenched self-deceiver standardly overlooks the pattern of his denials. Of course someone charged with self-deception – given solid evidence of its occurrence – sometimes confabulates to explain or exculpate himself. In such cases, confabulation accompanies self-deception without being integral to its strategies.

6. Finally, do you think there is any urgent question scholars should address in order to make the current research on self-deception progress further in the light of the new results in philosophy of mind?

- Our understanding of self-deception would benefit greatly from research into the structures of localized, modular sub-systematic patterns of brain functioning and from studies of the integration of cognitive centers with endocrine functioning. Under what conditions does such integration succeed and when does it fail?
- Inter-disciplinary studies in the philosophy of language and the psychology of speech acts – analyses of the relation between the psychology of propositional attitudes and the pragmatics of speech acts – would also be illuminating. What kinds of speech acts qualify as self-deceptive? Can merely expressive non-propositional utterances be self-deceptive? Can wishes and fantasies be self-deceptive? What is the

[3] See the classic studies reported by Richard Nisbett and Lee Ross (1980) and by Daniel Kahneman, Paul Slovic and Amos Tversky (1982).

structure of self-deceptive promising? Can performative or constative speech acts be self-deceptive? Can externalist and internalist standards of the attribution of self-deception be reconciled? [4]

- Anthropological and sociological studies of self-deception would enlarge and correct our present rather provincial understanding of the dynamics – and the norms – of self-deception. Do cultures differ in the domains in which self-deception is prevalent? What sorts of social pressures support or conduce to self-deception? Does the prevalence of forms of high politeness in social inter-action conduce to self-deception? What are the cultural differences in the incidence and areas of common self-deception? Does successful self-deception typically involve social reinforcement? Can religious or social rituals like absolution, forgiveness, penitential prayers be self-deceptive?
- Victorian novels (George Eliot, Trollope, D'Israeli) and political autobiographies (Koestler, de Beauvoir) provide wonderful insight into the subtle processes of self-deception and their occasional unmasking.[5] We have, for instance, much to learn from tracing Eliot's descriptions of Dorothea's self-deceptive admiration for Casaubon and her gradual, reluctant disillusionment. Lydgate's blindness to Rosamond's manipulations highlights the way that naïve self-deceivers sometimes collude in the deceptions that others initiate. Autobiographies of fervent communists who became anti-communists after the Stalin Trials also provide rich examples of the reflections of self-declared self-deceivers, of the strategies they employed in their self-deceptions, of their techniques in resisting contrary evidence, of the occasions of their "breakthrough" self-corrections.
- The current industry of philosophical work on self-knowledge – -initially prompted by Anscombe (1981) and recently developed by Holton (2009), Bermùdez (1998), Cassam (1994), Gertler (2003), Moran (2001), and Hatzimoysis (2011) – would benefit from a closer study of the various domains and strategies of self-deception. It would also be illuminating to locate the varieties of strategies of self-deception within a

[4] See Grice 1989 and Recanati 2004.

[5] For novels, see George Eliot, Middlemarch, 1874; Trollope, He Knew He was Right, 1869; D'Israeli, Sybil, 1845; Henry James, The Wings of the Dove(1902); for autobiographies, see Arthur Koestler, The Invisible Writing and The God that Failed (1949), Simone de Beauvoir, The Force of Circumstance(1963) and All Said and Done(1972).

more general taxonomic frame of the varieties and domains of belief and of self-knowledge.

- Is the idea of collective or interactive self-deception coherent? Can analyses of collective intention and action be applied to self-deception? (Gilbert, 1989; Bratman, 2007). If so, how does it work, what are its 'mechanisms?' What are its implications for the philosophy of mind and the philosophy of language?
- We need a catalogue and taxonomy of the varieties of self-deception, with an account of how their domains and strategies differ from the varieties of self-knowledge. Is there a significant difference between motivated and non-motivated self-deception, between its occurrent and habitual forms? between self-deceptive belief and self-deceptive action or emotion? between direct or active and indirect, passive or collusive self-deception? between self-deception that issues in false belief and that which issues in a true but pragmatically defective belief or action?

  I am especially interested in indirect, passive or collusive self-deception, cases where we collude in being deceived by others. Consider the ways that we knowingly allow ourselves to be conned, "taken in" by political rhetoric and manipulative advertising. We typically know perfectly well that such claims and promises are inflated if not actually false, and yet we find ourselves believing and acting as if they were reliable and trustworthy. What makes us susceptible to internalizing claims that we would typically hold suspect? When and why do we abandon our normal epistemic caution and extend epistemic trust beyond its normal limits?
- I suspect that self-deception is now a fashionable topic in the philosophy of mind because a great deal of post-Wittgensteinian philosophical psychology has focused on perceptual and cognitive transparency.[6] The prevalence of philosophical concern about self-deception is also of concern to consequentialists and neo-Kantians who place heavy emphasis on the underlying unity and effectiveness of the capacities for rational choice or self-construction.[7] Chronic and structural vulnerability to self-deception – endemic and apparently functional patterns of irrationality – appear to threaten effective norms of the rational basis and directives of

[6] See e.g., Siegel 2010.
[7] See e.g., Railton 2003 and Korsgaard 2009.

morality. Integration, self-knowledge and integrity are in high demand precisely because they seem elusive. We are concerned to eradicate self-deception because it seems to threaten our claims to epistemic responsibility, moral integrity and social reliability. The independence of modular sub-systems engaged in high level cognitive thinking from those engaged in sensation and perception — the focused reflexive and transparent awareness of abstract thoughts abstracted from perceptual content — is ironically the very condition that makes us capable of — and vulnerable to — self-deception and akrasia as well as other common and prevalent forms of irrationality.

REFERENCES

Anscombe, E. (1981). The First Person. In *The Collected Philosophical Papers of G.E.M. Anscombe*, vol. II (*Metaphysics and the Philosophy of Mind*). Minneapolis: University of Minnesota Press.

Arendt, H. (2006). *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Penguin.

Bermúdez, J.L. (1998). *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Bratman, M. (2007). *Structures of Agency*. Oxford: Oxford University Press.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Cassam, Q. (Ed.) (1994). *Self-Knowledge*. New York: Oxford University Press.

Fodor, J.A. (2006). The *Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

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

Gertler, B. (Ed.) (2003). *Privileged Access: Philosophical Accounts of Self-Knowledger*. Aldershot: Ashgate Publishing.

Gilbert, M. (1989). *On Social Facts*. London: Routledge.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hatzimoysis, A. (Ed.) (2011). *Self-Knowledge*. Oxford: Oxford University Press.

Hirstein, W. (2004). *Brain Fiction. Self-Deception and the Riddle of Confabulation.* Boston, MA: MIT Press.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kant, I. (1963). Self-Love. In I. Kant, *Lectures on Ethics*. (tr. by L. Infield). New York: Harper & Row.

Kaplan, D. (1986). Opacity. In L. Hahn & P. Schilpp (Eds.). *The Philosophy of W.V. Quine*. La Salle: Open Court, 229–289.

Korsgaard, C.M. (2009). *Self-Constitution: Agency, Identity and Integrity*. Oxford: Oxford University Press.

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

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

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

Moran, R. (2001). *Authority and Estrangement*. Princeton: Princeton University Press.

Nisbett, R. & Ross, L. (1980). *Human Inference*. Englewood Cliffs, NJ: Prentice-Hall.

Railton, P. (2003). *Facts, Values and Norms: Essays toward a Morality of Consequence*. Cambridge: Cambridge University Press.

Recanati, F. (2004). *Literal Meaning*. Cambridge: Cambridge University Press.

Rorty, A.O. (1988). *Mind in Action*. Boston: Beacon.

Rorty, A.O. (1998). *Philosophers on Education*. London: Routledge.

Ruddick, W. (1988). Social Self-Deceptions. In B. McLaughlin & A.O. Rorty (Eds.), *Perspective on Self-Deception*. Berkeley: University of California, 380–390.

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

Siegel, S. (2010). *The Contents of Visual Experience*. New York: Oxford University Press.

Van Leeuwen, N.D.S. (2007). The Spandrels of Self-deception. *Philosophical Psychology*, *20*, 329–348.

Van Leeuwen, D.S.N. (2009). Self-Deception Won't Make You Happy. *Social Theory and Practice*, 35(1), 107–132.

www.humanamente.eu

Humana.Mente – Journal of Philosophical Studies was founded in Florence in 2007. It is a peer-reviewed international journal that publishes 4 issues a year. Each issue focuses on a specific theme, selected from among critical topics in the contemporary philosophical debate, and is edited by a specialist on the subject, usually an emerging researcher with both philosophical and scientific competence.

Humana.Mente wants to be a place for exploring the most recent trends in the international philosophical discussion and wants to give the opportunity to the international community of young researchers to confront each other, and to discuss, control and verify their theories. An analytic perspective is favored, and particular attention is given to the relationship between philosophy and science, without however neglecting the historical aspects of the philosophical topics.

In this issue works by

Alfred R. Mele, Dion Scott-Kakures, Anna Elisabetta Galeotti, José Eduardo Porcher, Eric Funkhouser, Carla Bagnoli, Dana Kay Nelkin, Patrizia Pedrini, Julie Kirsch, Mark Young, Lisa Bortolotti and Matteo Mameli, Massimo Marraffa

Humana.Mente
Journal of Philosophical Studies

Issn: 1972 -1293