teorema
Vol. XXVI/3, 2007, pp. 73-95
ISSN: 0210-1602

Self-Deception, Rationality, and the Self

Thomas Sturm

RESUMEN

Este ensayo constituye un alegato en favor de la opinión de que los filósofos deberían analizar el concepto de autoengaño con el claro objetivo de que su trabajo tenga aplicaciones útiles para la investigación empírica. Es de desear que esto sea así especialmente porque los psicólogos usan a menudo conceptos diferentes e incluso incompatibles de autoengaño cuando investigan las verdaderas condiciones y consecuencias, así como la existencia misma, de este fenómeno. Al mismo tiempo, los filósofos que recurren a la investigación psicológica sobre la cognición y el razonamiento humano para entender mejor el autoengaño no se percatan de que estas teorías y datos están cargados de suposiciones problemáticas. Más expresamente, examino qué conceptos de racionalidad se dan por supuestos cuando describimos los casos de autoengaño como irracionales o adaptativamente racionales, y cómo surgen modelos ontológicos encontrados del yo en los distintos enfoques sobre el autoengaño. Sostengo, en primer lugar, que aunque el yo es casi siempre el objeto de tal engaño, no lo es siempre. En segundo lugar, mientras es el sujeto de engaño, lo es tan sólo de un modo moderado: no necesitamos asumir personalidades múltiples, ni tampoco el autoengaño viene causado o sostenido típicamente de modo intencional. Sin embargo, el evitar el autoengaño está, al menos a veces, bajo el control racional del sujeto. Este enfoque no da por sentada la existencia del fenómeno de autoengaño. Valorar si el autoengaño realmente ocurre es una tarea seria de investigación empírica. Este problema depende a su vez de la cuestión, hasta ahora ignorada, de qué concepto normativo de racionalidad es el que se asume cuando uno considera ciertas creencias como autoengañosas.

ABSTRACT

This essay is a plea for the view that philosophers should analyze the concept of self-deception more with the aim of having useful applications for empirical research. This is especially desirable because psychologists often use different, even incompatible conceptions of self-deception when investigating the factual conditions and consequences, as well as the very existence, of the phenomenon. At the same time, philosophers who exploit psychological research on human cognition and reasoning in order to better understand self-deception fail to realize that these theories and data are loaded with problematic assumptions. More specifically, I discuss what conceptions of rationality are assumed when we describe cases of self-deception as either irrational or as adaptively rational, and how competing ontological models of the self appear in different accounts of self-deception. I argue, first, that although the self typically is an object of such deception, it is not always so. Secondly, while it is the subject of deception, it is so only in a moderate way: We need neither assume multiple selves, nor is self-deception typically brought about or sustained intentionally.

However, the avoidance of self-deception is at least sometimes under the subject's rational control. This account does not take for granted the existence of the phenomenon of self-deception. It is a serious task of empirical research to figure out whether self-deception really occurs. This issue also depends on the question ignored until now of what normative conception of rationality is assumed when one views certain beliefs as self-deceptive.

INTRODUCTION

While philosophers and psychologists are interested in understanding roughly the same phenomena when they study self-deception, they are frequently talking past one another. Much current analyses in philosophy, while being shrewd and careful, tend to be too distanced from empirical research about factual conditions and consequences of self-deception. It would be profitable if philosophers would engage more in thinking about the relevance of their concept-chopping to questions that can be pursued empirically [Brandtstädter & Sturm (2004)]. At the same time, psychologists often make claims about self-deception without having invested a sufficient amount of conceptual analysis. This has problematic consequences for their empirical research. For instance, they sometimes confuse different conceptions of self-deception, and they consequently misunderstand what their studies about the existence, causes, and effects of such deception actually show.

These problems are obviously too complex to be dealt within a single essay. I here confine myself to the following tasks in order to clear the ground for a better interdisciplinary research on the topic. In part I, I shall make clear how philosophical and psychological studies on self-deception differ in not only their methods, but also in their guiding questions. I also present what is currently the most promising attempt to combine philosophical analyses with psychological research, namely Alfred Mele's non-intentionalist or deflationary account of self-deception, according to which self-deception should be understood as a form of biased belief, namely that in which the biasing is motivated. In part II, I give three objections to Mele's non-intentionalism, of which the most important concerns his reliance upon the "heuristics and biases" approach in the current psychology of human rationality. There are strong criticisms against this approach, which might easily affect the attempt to analyze self-deception as a form of motivated irrationality. In part III, I turn to the issue of the role of the self in self-deception. I argue that the self in self-deception is not to be understood as the object, but as the subject of deception. It is so in a sense that cannot be easily subsumed under the heading of intentionalism or non-intentionalism, but which requires that we are rational thinkers and agents. The philosophical analysis of self-deception cannot by itself show what norms of rationality such thinkers ought to follow. However, when it comes to providing convincing instances of self-deception,

these must ultimately be provided by empirical research on the topic. In such research, scientists will have to lay their cards on the table: They will have to explicate and justify their account of rationality, otherwise their empirical demonstrations of cases of self-deception will remain highly problematic.

I. A Philosophical Paradox and a Topic of Psychological Research

Here are two garden-variety examples of self-deception as one finds by the dozen in the current philosophical literature. By listening only to students and colleagues who depend on him, Tom continues to believe that he is a great university teacher while having often received strong evidence to the contrary. Even after many occasions of being laughed at, Mary remains convinced that she is a great opera singer; she listens instead to her rich husband, who is a media tycoon and lets his newspapers publish enthusiastic reviews about her ridiculous performances (as in Orson Welles' movie about William Randolph Hearst, *Citizen Kane*).

Self-deception was not discovered just in the last decades. Plato used the concept, speaking of how discomforting it is if the deceiver is not even a step away from us, and that self-deception must be taken to be the greatest evil [*Cratylos* 428d]. Bishop Butler, Adam Smith, Kant, and others as well wrote about it. However, the philosophical and psychological literatures on self-deception have developed rapidly since the latter half of the twentieth century. There are two new developments.

First, in earlier centuries, philosophers treated self-deception mostly as an ethical problem. In recent decades, often in the mood of Wittgensteinian puzzle-solving and ordinary language philosophy, philosophers have concentrated upon a theoretical problem, the "paradox of self-deception". This arises when one construes self-deception along the model of interpersonal deception. For instance, Rumsfeld might intend to make Powell believe the opposite of what Rumsfeld takes to be the case, and attempts to bring this about by various intentional actions. In self-deception, then, one must deceive oneself intentionally into believing something one does not believe at the very same time. How is such inconsistency possible? How can someone deceive her- or himself about a proposition *p*? Kant [1900ff., vol. VI, p. 430] noted the puzzle, but a closer discussion of it began only in the twentieth century. This discussion currently concerns itself mostly with the questions of whether self-deception has to be understood along the lines of interpersonal deception, whether it involves believing contradictory propositions, and, perhaps most importantly, whether it involves intention. Secondly, psychologists have of course used the concept of self-deception earlier on, but what seems new during the last few decades is the interest in questions that can be answered by empirical means. For instance, can one demonstrate that there really are cases

of self-deception? What experiments could show that? Also, what are the mechanisms and functions of self-deception? How does the phenomenon fit into empirical theories of human reasoning, its potentials and limits?

To begin, how is it possible to solve the paradox of self-deception? There are many different proposals. Some think that self-deception is possible because the conflicting beliefs are held "on different levels of awareness" [Demos (1960); Sahdra & Thagard (2003)]: Whereas the belief a person is more motivated to accept and avow is transparent to the person, and while she denies the contradictory belief, certain indices reveal that "deep" inside she believes otherwise. Fingarette, in turn, has claimed that talk of belief and of unconscious states should be given up. We should rather speak of different "engagements" we have in the world, and which we are, in cases of self-deception, simply unable to spell out [Fingarette (1969)]. More moderate is a proposal such as Audi's. He claims that one of the conflicting propositions is not really held as a belief, but merely "avowed sincerely": the paradox is here avoided by giving up the condition that both propositional attitudes must be beliefs [Audi (1982); similarly, Cohen (1992)]. Davidson, again, claims that self-deception is made possible by a division of our minds into independent sets of states and processes — independent in the sense that the usual logical and epistemic relations between them are broken down, though the states remain causally connected, such that the assumption of several selves within one agent can be avoided (such an assumption has been ascribed wrongly to Davidson by Bird, 1994; it is, however, held by, e.g., Rorty, 1983). Because the usual logical and epistemological relations do not hold, self-deception is said to be irrational [Davidson (1986); see also Greve (2000), p. 17].

These and other attempts are not empirical hypotheses, but proposals for a correct conceptualization of the phenomenon. All proposals have problems I shall not discuss here. Rather, I wish to note first what these proposals have in common, namely the criterion for their correctness: The conceptualization ought to solve the puzzle of self-deception. But considering the variety of options — and I have only mentioned a very limited number of them —, it seems that this criterion is not sufficient to capture the concept of self-deception correctly. One might think that an obvious way out here would be to look at how self-deception is studied by psychological research and theorizing — after all, isn't psychology's task to describe and explain the real phenomenon, instead of merely solving a conceptual puzzle about it?

To see what psychologists do and how easily they get into trouble with self-deception, let us begin with a perhaps surprising question: Does self-deception really ever occur? It has been claimed, for instance, that invalid self-report personality inventories are due more to self-deception than to other-deception or lying [Meehl & Hathaway (1946)]. Also, self-deception has been invoked in order to explain why subjects seem to maintain hypotheses in the face of disconfirmation [Wason & Johnson-Laird (1972)]. But in

none of these studies has it been explicated what self-deception is, and whether it ever occurs. But, contrary to what many traditions assume, and given the paradoxical nature of some folk psychological conceptions of it, is by no means beyond doubt that people really ever deceive themselves. Perhaps self-deception is merely attributed to subjects by outside observers, e.g., in order to blame others for behaving irrationally or even immorally? Psychologists have therefore attempted to provide more serious experimental demonstrations of the phenomenon. In perhaps the first example of such an investigation, Gur & Sackeim have employed Demos' idea that the inconsistent beliefs are held "on different levels of awareness" [Gur & Sackeim (1979); Sackeim & Gur (1979)]. Part of the reason for this preference is that they find self-deception to be similar to perceptual defense, which also implies that people can often be unaware of their representations. People sometimes tend to avoid certain perceptions; but in order for a perceiver to avoid perceiving a stimulus, the stimulus must first be perceived. The solution is found by saying that it is erroneous to assume that perception must be subject to awareness. The concept of perceiving a stimulus equivocates on 'being presented to one's sensory apparatus' and 'being cognized with awareness'. Now, Gur & Sackeim observe that the same idea is implied by Demos' concept of self-deception. Accordingly, they state the following criteria to be necessary and sufficient for ascribing self-deception to any given phenomenon:

1. The individual holds two contradictory beliefs (that p and not p).

2. These two contradictory beliefs are held simultaneously.

3. The individual is not aware of holding one of the beliefs.

4. The act that determines which belief is and which belief is not subject to awareness is a motivated act [Gur & Sackeim (1979), p. 149].

They then use voice-recognition experiments. In a typical experiment, subjects are asked to recognize whether a taped voice is their own or that of another person; at the same time, while subjects report, behavioral indices — galvanic skin responses — are used to find out whether a contradictory belief is also held [experiments using such indices, if for different questions, have been used by, e.g., Tomaka, Blascovich & Kelsey (1992)]. People with negative attitudes about themselves, or with discrepant beliefs about what they believe themselves to be and what they should be, seem to find confrontation with themselves — e.g., with their own voice — aversive. On the other hand, people who score low in such discrepancy have been said to not find self-confrontation aversive. On the contrary, they seek it. Gur and Sackeim indeed claim that self-deception in the sense outlined occurs.

Whether skin responses are really indicative of belief is problematic, and in later responses Sackeim has granted this [Mele (1987); Sackeim (1988)]. A similar point can be made against Quattrone and Tversky, who have adopted Gur's and Sackeim's notion of self-deception in order to explain why people favor actions that are only "diagnostic" of their consequences instead of causing them [Quattrone & Tversky (1986); for criticism, see Mele (1997), p. 96f.]. It has also been pointed out that the task of recognizing one's own voice is not a good task for the ascription of self-deception, since similar results were achieved for subject's recognition of voices of their acquaintances [Douglas & Gibbins (1983)]. Moreover, it is doubtful that Sackeim & Gur have not really employed Demos' [e.g., (1960), p. 588] concept of self-deception: Their examples do not involve an intention on the side of the self-deceiver. Having a motive is weaker than that, since it need not involve any serious practical deliberation about ends and means. Remember the parallel with intentional other-deception: We would not speak of an intentional deception if person A (who believes that p) accidentally caused person B to believe that $\neg p$, even if A also desired that B believes that $\neg p$ but, resisting the temptation, nevertheless tried to bring it about that B believes that p. In other words, it may be doubted that the whole approach of Gur & Sackeim is satisfactory because it does not start from an adequate conception of self-deception. Similar objections may be raised, of course, concerning other characteristics thought to be essential for self-deception on one or another account. One might also claim, for instance, that self-deception demands a deeper division of the self than a difference in levels of awareness, or that it does not require contradictory beliefs, or that it does not require that these beliefs be held simultaneously.

While self-deception is often taken to be a pervasive phenomenon of human life, being present in, say, the denial of illness, the careless behavior of professional drivers, and the overconfident optimism of self-employed people or of soldiers in battle [e.g., Sahdra & Thagard (2003)], it is now clear that the demonstration of its occurrence is no simple matter. To be true, it would be premature to infer from the abovementioned objections that such demonstrations are impossible at all, or that self-deception is a mere social construction, as some have claimed [e.g., Gergen (1985); Lewis (1996)]. The issue is simply wide open. It will remain this way as long as psychologists are less careful about conceptual analysis than philosophers are, while the latter only care about solving the paradox of self-deception instead of contributing their training in conceptual analysis towards empirical studies of the occurrence, mechanisms and functions of self-deception.

Alfred Mele, whose work deals perhaps best with current psychological investigations, claims that the whole background model of intentional interpersonal deception is unnecessary. We should rather think of self-deception as a species of *biased belief*, as Mele calls it, building upon empirical re-

search on the so-called "cognitive illusions", or on the "heuristics and biases" program (HB for short) in judgment and decision making [Mele (1997), (2000); for similar approaches see Lazar (1999); Patten (2003); for a computational model of that approach, see Sahdra & Thagard (2003)]. Biased beliefs are brought about by factors such as preferring certain readily available information, misinterpretating data positively or negatively, looking more often for confirmation instead of falsification, or by selective evidence-gathering [e.g., Wason (1966); Tversky & Kahneman (1974), (1983); Nisbett & Ross (1980)]. Mele furthermore maintains that not all biased beliefs are beliefs one is self-deceived about. In self-deception, the biasing — the selective focusing upon certain kinds of evidence only, or the misinterpretation of certain data — is *motivated* [cf. also Kunda (1990)]. There need not be any intention, any self-caused activity involved in this, and also no simultaneous holding of contradictory beliefs [Mele (1997), (2000); cf. already Siegler (1962) vs. Demos (1960)]. In this sense, Mele's account of self-deception is deflationary. He does not deny that intentional self-deception is possible. He describes his own account as stating sufficient conditions of self-deception, not as necessary ones; hence, other cases are imaginable. However, he thinks that the garden-variety cases can be explained without reference to intentions, inconsistent believing and the like.

On this model, self-deception is still an irrational phenomenon, but not because of the violation of the law of contradiction. Rather, self-deception is now viewed as irrational because it violates certain norms of good reasoning coming from various domains of logic, probability theory, or statistics. Consider the empirical finding that 94% of university professors think they are better-than-average at their job [Gilovich (1991), p. 77]. This, of course, cannot be the case for all of them, so they are somehow deceived. But they are not, or at least need not be, self-deceived or deceived by anyone else. Nor do they need to have inconsistent beliefs. They may have inferred the erroneous belief from the feedback of students who think they are just the best teachers. They may also have ignored certain facts about statistical distributions, or have questionable views about scientific standards. While deception results of such cognitive processes, no one needs to have intended it.

II. Some Illusions of Non-Intentionalism, and the Relevance of Rationality

But is it convincing that self-deception is the motivated species of biased belief or reasoning? Let me note three objections. First, it could be that some cases of self-deception are unmotivated; so Mele's central condition would be superfluous. I might deceive another person without any substantive motivation, just out of sheer curiosity to see if I can do so. Is this not

possible with deceiving oneself as well? The question has been raised, because otherwise the condition that self-deception is always motivated is merely stipulated, or perhaps results from an overgeneralization of certain cases viewed as typical [see Patten (2003)].

More convincing is a second objection. May we not be motivated in many of our biased beliefs without these being cases of self-deception [Kunda (1990)]? If that is so, Mele's claim that his conditions are sufficient for self-deception would be refuted. For instance, it may be pointed to cases where a person possesses biased beliefs that are also motivated, such as racist ones, and where he or she sticks to them in the face of overwhelming counter-evidence [Holton (2000), p. 59]. I do not think this case is convincing (see below, part III.1). There are other, more plausible examples. Already Demos [(1960), p. 588] denied that the scientist who overestimates his competence is best viewed as self-deceived. Although that scientist can have a motive, there is no intention to deceive himself involved, and hence no similarity to other-deception. Moreover, intentionalist conceptions of self-deception according to which the two inconsistent beliefs must not be held simultaneously might strengthen the view that motivation is not enough. I might plan to deceive myself about my school math notes by changing the current record, hoping that my poor memory will help me forget my action. A girl who is told maliciously by her brother that her pet rabbit has died puts her hand over her ears, yells out loud, and runs out of the room. She continues to do so whenever he seems to be about to tell her again, thus maintaining intentionally her belief that the rabbit is still alive [Perring (1997)]. I was once told a story about a philosopher who, having left a talk in which the speaker had argued vigorously for views that cannot seriously be taken to be true, commented that "this was probably a very, very subtle form of self-deception" (I think the talk was about Hegel, the Myth of the Given, or the like). Perhaps that comment was meant such that the speaker must have somehow forgotten his intention, or managed, through reading only books or articles coming from universities he takes to be the centers of good philosophy, to achieve belief in the incredulous things he claimed to be true. I do not think that such cases are typical for what we mean by 'self-deception', and the comment that this was a "very subtle form" probably reflects this. However, there are authors who think otherwise [e.g., Bermudez (2000)].

Of course, Mele would reply that such cases reflect the misconstrual of the concept of self-deception on the side of intentionalists. Cases where the result of the intention to deceive oneself comes about only after delay hardly show that self-deception always involves intentions. But that leads to another impasse: Nonintentionalists like Mele and intentionalists like Demos or Davidson are simply talking about different phenomena. Their conceptualizations are primarily designed to solve the paradox of self-deception, and both accounts may be said to be successful in that regard. But which of them cap-

tures the "real" phenomenon? As long as no other constraints are added, this now seems a matter of mere decision. Alternatively, we have not just one, but different kinds of self-deception.

A third objection has not yet been raised, but is most destructive. It concerns the psychological theories of reasoning or rationality used here. Even philosophers like Mele or Lazar, who try to get in close contact with psychology by referring to research about heuristics and biases, do hardly if ever refer to actual cases or empirical demonstrations of self-deception as they conceive of it. Mele, for instance, argues that the studies by Gur & Sackeim, or by Quattrone & Tversky can be interpreted without assuming that intentions played a causal role in bringing about the self-deception, and he criticizes their pretensions of demonstrating the occurrence of self-deception on other grounds as well. In favor of his own account, however, he presents hypothetical cases, appealing to our intuitions ("Imagine Freddy…"). While certain kinds of self-deception might satisfy Mele's conditions, he owes us real empirical proof. We must take seriously the doubt that self-deception does not really occur, but is merely ascribed by outside observers in order to, say, blame persons for being irrational or pathological.

Now, such a proof is no easy matter, given Mele's account. He does not claim that standard cases of self-deception are irrational because they imply contradictory beliefs, but because they violate certain norms of good reasoning — his reliance upon the HB program commits him to this. According to the former, traditional view, an empirical demonstration of self-deception is difficult, as the problems of studies by Sackeim & Gur or similar ones have shown. On the latter, Melean conception, it becomes much more complex, and it is doubtful whether any clear results are achievable. Why?

Simplifying somewhat, the HB program requires that for every reasoning test a certain norm must be chosen with which to compare the subject's behavior, for instance the material implication from logic (the "Wason selection task"; [Wason (1966)]), the conjunction rule from probability theory (the "Linda problem"; [Tversky & Kahneman, (1983)]), or the Bayesian rule ("base rate neglect"; [Casscells, Schoenberger & Grayboys (1978)]). People are then said to be irrational if they do not use that rule when solving a certain concrete reasoning task but use biases and heuristics instead. Oddly for the present discussion, Wason & Johnson-Laird (1972) have explained the apparently irrational avoidance of items that could disconfirm a certain material conditional in the Wason selection task by invoking that these people are somehow self-deceived. Mele is not to be blamed for this, and his account does not thereby become circular. Yet, it is clear that we should not uncritically take up all results from the HB program in order to explain self-deception. More important than this is another problem. The HB approach has been attacked vigorously in recent years. It is often possible to reinterpret the empirical results allegedly revealing fallacies or biases in human reason-

ing such that the subject's responses are quite rational after all. It has actually been shown that many favorite studies in the HB program are based on experimental artifacts. For instance, it is possible that the alleged result that subjects massively fail to apply the conjunction rule when expected to do so is due to linguistic ambiguities in crucial terms of the test materials. For instance, in ordinary language, "and" and "more probable than" may possess legitimate meanings that deviate from those they have in logic or probability theory. When such ambiguities are removed, the number of false responses decreases drastically. Also, subjects may have understood the test as requiring the application of a different rule of reasoning, such as a conversational rule [e.g. Fiedler (1988); Hertwig & Gigerenzer (1999); Oaksford & Chater (1994)]. Similar points have been made concerning many other alleged cases of biased reasoning [see, e.g., Lopes (1991); Gigerenzer (1991); Gigerenzer & Hoffrage (1995)]. To use another idea of Davidson's here, we should always try to apply the principle of charity, that is, we should avoid as far as possible viewing human behavior as irrational. As these studies show, we not only *should*, but often *can* apply that principle.

I cannot enter these debates more closely here, since they have become complex and are by no means resolved in favor of one or the other approach [Sturm (2007b)]. It seems plausible that there can be non-epistemic reasons that may render some or even many cases of self-deception rational. Self-deception may be viewed as practically rational, as based on subjective goals combined with adequate practical deliberation [e.g., Rorty (1972); Davidson (1986)]. This, however, is a strong variety of intentionalism, adequate at best for a few untypical cases. Less demanding in this respect is the psychological account of rationality which competes with the HB program. It rejects the assumption made by that program that there are universally valid norms of reasoning such as those coming from logic, probability theory, or statistics. Instead, we are invited to favor a conception of "bounded" rationality, according to which even epistemic reasoning (both the concrete instances and the rules governing them) is normatively valid only relative to contents and contexts in which we reason. Typically, judgments of validity are then based upon whether the items or norms of reasoning show an (evolutionary) fitness or adaptivity. Now, this direction of research on reasoning seems to happily converge with a certain tendency in many empirical studies on self-deception. Instead of emphasizing the irrationality of self-deception, they focus on the question of what advantages self-deception might have. Typical advantages cited are the reduction of self-reports of stress, the maintenance of self-esteem and well-being [Jamner & Schwartz (1986); Welles (1986); Lockard & Paulus (1988); Tomaka, Blascovich & Kelsey (1992); Sahdra & Thagard (2003)], or the more effective hiding of one's real beliefs from other human beings, an idea also based on evolutionary considerations [Trivers (1985), (2000)].

However, we should resist the temptation to think that all cases self-deception can be rationalized in one of these ways. Why? Not so much because we should downplay the importance of non-epistemic reasons or of less-than-universal reasoning norms. Rather, the problem is that at least sometimes psychologists use such unclear notions of self-deception that it is unclear whether or not the phenomena rationalized in their studies are instances of self-deception. For example, one study considers the relation between deception, self-deception and social dominance in tennis [Whittaker-Bleuler, (1988)]. By deceiving herself about her level of ability, the state of the match, and so on, a tennis player might be able to hide her insecurity better from her opponent. She might be able to keep her head up in a more natural fashion, and to avoid acts like shaking her head horizontally or going through a stroke motion without a ball. One assumption here is that the degree of self-deception must be high when the player has lost the majority of previous points, and still behaves dominantly by showing, for example, coolness. Pete Sampras almost never showed strong emotions, no matter how the match was going for him. However, no inconsistent believing needs to be involved here. Sampras might not have taken the evidence of previous and current points to be as important as the belief in his fitness, his excellent technique, or his ability to concentrate upon the next point only. It may be doubted that Sampras' behavior is a case of self-deception, given that he won more Grand Slam titles than Rod Laver. Thus, simply picking another theory of rationality cannot without much further ado show that self-deception is rational, because very different conceptions of self-deception are available. This is similar to many other examples of allegedly adaptive or "rational" self-deception: They might only be cases of justified confidence in oneself.

Also, many issues with which we have to deal are risky and uncertain. In cases like the self-deceived husband, we are typically told that his friends tell him that his wife is unfaithful, that she goes out more often than usual, etc. Ultimately, the husband might follow her, observe the unfaithful act, and thereby reduce uncertainty to a minimum. But in countless other cases, risk and uncertainty will remain, such as when we decide whom to marry, what job offers to accept, whether our businesses will be successful, or which experts to trust. The attitudes we adopt here are often viewed as examples of self-deception [Sahdra & Thagard (2003)]. But are they, given that we have to make assumptions that are from being supported by the evidence and may therefore not count as beliefs at all? We need more caution here.

What can be learned from the debate of human rationality in psychology for our present topic concerns the requirements for further psychological research on self-deception. Two points may be noted. First, empirical demonstrations and theories of self-deception require both a clear conception of the phenomenon and an adequate theory of rationality as well. Since we do not possess either, there is nothing to do but to work on both tasks at the same

time. Secondly, over and above showing that the conditions of self-deception hold (depending on what account we should ultimately accept), appropriate demonstrations of the phenomenon would require that (i) a specific norm be chosen; (ii) the norm would have to be a valid one; (iii) the norm would have to be the one to be applied by subjects in the test situation; and (iv) the possibility that alleged cases of self-deception might charitably be reinterpreted as being rational according to some other rule would have to be excluded. I hope it is clear how demanding such conditions are. I do not claim that they cannot be satisfied. But, unless we wish to deceive ourselves, we should reject philosophical accounts that take over theoretical models and alleged research results from psychology in an uncritical fashion.

III.1 The Self in Self-Deception: Not Necessarily the Object of Deception

Now let us leave behind the issues of intentionality and the rationality or irrationality of self-deception and turn to the role of the self in self-deception. It is trivial to say that some conception of the self must be part of the meaning of 'self-deception', but it is not trivial to spell out which one is most adequate. So, what is the role of the self here? The basic options are the following: A person may be deceived *by* herself, or she may be deceived *about* herself. Of course, many cases — like those of Tom or Mary mentioned above — invite the view that both may be true at the same time, but let us ignore this here.

Can it be that the role of the self in self-deception is merely that of an object, such that the idea that one brings about that deception oneself is left out? The question can also be raised in a more psychological way: Where do conceptualizations of self-deception locate the controlling (independent) variables of self-deception? Can explanations of self-deception leave out what goes on "inside" the person? Some instances of self-deception point in this direction, e.g., when we say that people are deceived about their own talents or characters [for a vivid analysis of a literary example, see Sahdra & Thagard (2003)]. It has been claimed that the concept of self-deception sometimes necessarily means this, and that while it does not exclude deception by the self, the latter is unnecessary [Holton (2000)]. A behavioristic approach that locates the controlling variables of self-deception outside of the deceived person comes close to such a view as well. Self-deception is then construed as the absence of self-knowledge, as a lack of knowledge of what one is doing, established perhaps through negatively reinforcing consequences [Skinner (1953), chap. 18; Day (1977)]. That avoids the issues of whether intention is involved, whether self-deception has to be modelled after interpersonal deception, or whether self-deception involves multiple selves. The price to be paid for such advantages is, however, too high. Such accounts ignore, first,

the difference between self-deception and mere ignorance or error about oneself. Secondly, although many cases like those of Tom and Mary are cases where one is deceived about oneself, not all must be: I may be self-deceived about my spouse's actions, or about my children's talents and character, or that there will be no further war in the Middle East this year. The case of the racist mentioned above (part II) is structurally similar. Holton [(2000), p. 59], who claims that this is not a case of self-deception, begs the question.

These considerations also bring out an interesting difference between self-deception and other reflexive types of experience, thought, and action. In self-knowledge, the object of reference is always oneself; it is part of the content of the known proposition. Similarly in self-evaluation or self-control, where reference to oneself has to be a part of the content of the evaluative or prescriptive propositions in question. In self-deception, by contrast, reference to oneself need not be part of the content of the relevant belief. One might object that there is a hidden relation to oneself in at the examples just mentioned. Does not self-deception often relate to one's friends or spouses, one's personal relationship to them and, thereby, one's own self-esteem and well-being? However, the relevant propositions need not contain reference to oneself as the object of deception. Even when they are often derived from self-regarding motives, they need not always be. I might deceive myself about the prospects for another war because of selfless motives. Such cases are not covered by the view that self-deception is merely deception about oneself; hence, we must reject that view.

III.2. But How Can the Self Be the Author of Self-Deception?

What, then, about the idea that the self is the subject or author of the deception? Here we should bring to mind some basic options in the ontology of the self. There are three main tendencies in the ontology of the self, deriving mostly from traditions of early modern philosophy and reactions to it. First, the self is often viewed as a particular, irreducible mental entity. Secondly, there are eliminativist positions that argue that no such self really exists. Thirdly, the self is viewed as reducible to or identical with some set of bodily or mental processes or states.

The ontological disputes connected with these positions may appear to be far removed from psychological research on self-deception. But that is not correct. Interest in the self within psychology has grown during the latter half of the twentieth century [e.g., Mischel (1977); Baumeister (1987), (1999); Markus & Wurf (1987); Greve (2000)], and there has even been an "inadvertent rediscovery of Self in social psychology" [Hales (1985)]. Most importantly, there is a constant ambivalence between viewing the self as *explanans* or as *explanandum* of research. One of the reasons for the renewed interest in

the self in psychology is the recognition that, contrary to behavioristic orientations, we do not merely notice or remember our own behavior, but each of us "instead mediates and regulates this behavior. In this sense, the self-concept has been viewed as dynamic — as active, forceful, and capable of change." [Markus & Wurf (1987), p. 299; cf. Greve (2005)]. Our social and natural environments influence our actions to different degrees, and our actions, our very self-understandings or our personality, are also constantly reshaped by how we perceive environments [Brandtstädter (1998)]. The self-understandings of persons even influence how they act in psychological experiments: Subjects constantly try to protect or enhance their self-esteem in ways that make many areas of psychological research quite difficult [Hales (1985); Morawski (2007)]. Because the self appears to be both an *explanans* and an *explanandum* of psychological research outside of the field of self-deception as well, it makes good sense to critically reflect on the question of whether the self can be the author of self-deception (and hence part of the explanans) in terms of the basic ontological options.

A. *The Central-Headquarters-View.* The first view has nicely been termed the "homunculus" — or the "central headquarters" view of the self [Dennett (1991)]. One thinks of the self as a particular subject of thoughts, experiences, and actions, an internal mental agent. This is certainly due to practical needs: To view ourselves as responsible for our actions, we assume that the self is a substance with powers of thinking and deciding. Another consideration in favor of the view of the self as a particular mental subject is that the self seems to be an object of reference, of quantification and other procedures that lead to a reification of self-related thought and talk. It seems possible to count selves — only one self to a customer is the rule, as Dennett says [(1991), pp. 419f.]. When this rule is violated by human beings who seem to possess different selves, these selves might still be viewed as countable objects of reference; only some human beings have several of them.

It is widely accepted among psychologists that the "central headquarters" view leads to insurmountable problems [cf. W. Mischel (1976); Toulmin (1985); Greve (2000)]. For instance, there lurk regresses if one takes that concept as explanatory. If a homunculus is supposed to explain, e.g., how we autonomously initiate actions, then there must be another little man inside the homunculus making his decisions or beginning his actions, and so on. Secondly, there is no empirical support for the idea that there must be a central instance where our different experiences meet and where our actions begin [Dennett (1989); Dennett & Kinsbourne (1992); Churchland (1995)]. I do not see that such objections are overcome by recent attempts to defend the notion of the self as a simple mental substance [McGinn (1997); Strawson (1997)].

These objections also hold when one thinks that self-deception is due to a division of the human being into several selves. Self-deception then be-

comes a case of interpersonal deception. One might claim that we need not worry about claims that self-deception is due to a split self, since the self is anyhow a "decentered, distributed, and multiplex" phenomenon that is "the sum total of its narratives, and includes within itself all the equivocations, contradictions, struggles and hidden messages that find expression in personal life" [Gallagher (2000), p. 20; cf. Sahdra & Thagard (2003), p. 227]. It is true that talk of "the self" possesses different meanings, and it is also true that the role of the self in such diverse phenomena as the self-concept, self-identity, self-knowledge, self-deception, self-control, self-esteem is far from being always the same. But these are issues that merely demand careful conceptual and terminological analyses; they do not imply that the self is a "decentered, distributed, and multiplex" phenomenon. Moreover, such a viewpoint makes it impossible to distinguish between self-deception and cases of deeper mental pathologies. At least, many if not most garden-variety cases do not involve multiple selves [I agree here with Mele, (1997)]. Self-deceivers are not persons we describe as hearing voices or commands from within that they cannot properly self-ascribe, or as viewing some of their mental states as belonging to another subject.

B. *Eliminativism.* Eliminativists argue that no self exists, or that what we call the 'self' is an illusion due to bad philosophy, outdated folk psychology, or both at once. Such a view is rarely held among psychologists. Rather, certain philosophers and neuroscientists apply their eliminativism to self-related talk as well [e.g., Churchland (1995)]. They claim that the idea that I might be the origin of my decisions and actions is as good an explanation as that the presence of a witch explains why certain cows give less milk than they normally do. There are many arguments against eliminativism, some of which also apply with regard to self-related talk, but it would lead too far afield to discuss them here [Greve (1996); Pauen (2001), pp. 97-106]. Self-related thought and talk seem to be crucial to an appropriate understanding of human experience, thought and action. The assumption that we can therefore dispense with it are at least premature.

With respect to the self in self-deception, eliminativism may be ascribed to those who think that the self is merely the object, not the subject or agent of deception. I have shown why this is unsatisfactory (in part III.1). More interestingly, non-intentionalist accounts of self-deception might also be viewed as eliminativistic. Their denial of the claim that the agent or self — by means of an intention — plays an active role in the deception is, after all, meant as a rejection of the assumption that the self plays any explanatory role in the emergence of the deception. However, the eliminativist thinks (i) that phenomena such as self-deception constitute anomalies of folk psychology or the schema of intentional action-explanation, and (ii) that this reveals that the whole folk psychological idiom should be given up or treated as mere (and

bad) social construction. In contrast, the reductionist safeguards the folk psychological idiom — among other things, by cleansing it of errors it occasionally contains. That might include arguing that some mental phenomena cannot or need not be explained intentionally. Attempts such as Mele's or Lazar's to understand self-deception might thus better be viewed as reductionistic. Yet, in the next section I will clarify one sense in which they are not, and what problem this leads to.

C. *Reductionism*. With respect to the nature of the self, many prefer reductionist views, which identify the self with some set of bodily or mental processes or states and thereby try to safeguard its existence [e.g., Newen, (2000)]. Reductionism goes back to David Hume's famous "bundle" theory, according to which a self is the sum of our mental states held together by certain causal relations.

Such an approach suffers not only from limitations of current scientific knowledge, or from general problems of reductionism in the sciences. What is important in the present context is this: Since reductionists reject the view that terms such as 'I' or 'myself' refer to a particular mental subject, what should be reduced are *properties* (or classes of particulars) shared by those to whom we ascribe a self. Also, just like it makes no sense to ask for a reduction of other indexical terms such as "here" or "now", we cannot wonder what 'I' may be reduced to. It is thus not the self but, for instance, the property of *self-representation* or *self-consciousness* [e.g., Newen (2000)] or the "human sense of the self" [Strawson (1997)] which is taken as explanandum. But shifting explanatory interests in this way leaves open important questions. First, mental representations require a bearer or subject. Since that is so, we can distinguish two qualitatively identical representations — two identical belief-tokens, say — only by referring to their numerically identical bearer. But can there be such identity without an entity, even if that entity exists only briefly [Tugendhat (1979), p. 73]? Secondly, to be conscious of one's own feelings, thoughts, or perceptions, presupposes that one ascribes theses states to *oneself*. But how is it possible for a given set of mental representations to be *my* and not your representations? It is a serious issue whether we can explain this mineness of mental states without falling back to the idea that there must be an irreducible, "objective self" [Nagel (1986), chap. 4].

Non-intentionalism about self-deception may be said to be reductionistic about self-deception as a whole, and eliminativistic about the role of the self as author. But is there anything wrong with the latter? Not exactly the problems just outlined. We should, of course, wonder how to explain that a certain self-deceptive belief is mine and not yours. However, we are interested in the role of the self not so much as *owner* but as *author* of self-deception. Since non-intentionalists try to avoid viewing self-deception as a result of practical deliberation, and instead describe it as a species of biased

belief, driven by emotions or motivations, they thereby deny a role of the self as the author of self-deception. But that leaves an important point unexplained. Mele himself notes:

> […] a detailed understanding of the etiology of self-deception would help reduce the frequency of harmful self-deception. [...] A lively debate in social psychology about the extent to which sources of biased belief are subject to personal control has generated evidence that some prominent sources of bias are to some degree controllable. This provides grounds for hope that a better understanding of self-deception would enhance our ability to do something about it. [Mele (1997), p. 91f.]

In other words, Mele does not reject the idea that we can regulate our own behavior, and that we can improve our self-understanding and self-control. However, then the self is not only part of the *explanandum*, but also of the *explanans* of self-deception. Some things need to be said to support this position. At least two questions are likely to be raised: (1) Do we not thereby reintroduce a homunculus-explanation? (2) In what sense could such a self help to explain self-deception?

To answer these questions, let me briefly explain a little further the distinction between the self as subject/author and as object of self-deception. This distinction has certain standard contexts in which it is used. One standard context is represented by Kant's distinction between the "I as subject" and the "I as object of thought" [Kant (1781/1787), B407-409; (1900ff.), Vol. VII, p. 134n.]. He makes it clear that sometimes we use talk of the self not in order to *describe* ourselves in some way or other, as when we speak of our height or hair color, our beliefs or desires, or our personality. Rather, sometimes such talk is used in order to express that we *do* certain things. The self-as-subject is not an independently existing entity, but a built-in part of certain mental acts. Kant's standard examples of such acts are epistemic ones, as when we make knowledge-claims or think critically about them. Here, the role of the self-as-subject is made possible by the possession of certain capacities, especially the cognitive faculty of understanding ["Verstand" in German; Kant (1900ff.), vol. VII, p. 127]. In the special cases that we call 'self-knowledge', the self is involved in both roles: For instance, Jimmy knows that he believes that sheep don't grow on trees, and expresses that knowledge to himself in the form of "I know that I believe that sheep don't grow on trees". The first occurrence of 'I' here expresses the subject-role, whereas the second occurrence expresses the object-role of the self. (Of course, if Jimmy thinks, "I believe that sheep don't grow on trees", then this occurrence of 'I' expresses a subject-role.) The distinction between the two roles is not one between two different objects, or between two numerically distinct selves; it is the human being that is the common reference point of the different notions [Kant (1900ff.), Vol. VII, p. 134n.].

There are other examples of such a first-person point of view in thought and action, e.g., certain cases of verbal action, as Austin has pointed out in his analysis of explicit performative utterances: "I will" (uttered by the groom), "I shall be there" (used to express a promise), or "I promise to hold on to the principles of constitution" [Austin (1962), p. 60f.]. Here, other capacities and dispositions are necessary, such as an understanding of social rules, or knowledge of how to follow as well as violate them, or sufficient memory (the bride will remind the groom). Equally, such cases can also involve the self in both object- and subject-roles — e.g. "I promise I won't mislay the car keys anymore" — without that leading to a split self.

Now, when a person is self-deceived, the role of the self-as-subject seems different. First, during self-deception, the person cannot express her being deceived in first-person statements. The fact is intransparent to her. Only afterwards can she say seriously things such as "I deceived myself into believing that Hegel was right about the Myth of the Given". Secondly, as already noted, the deception need not involve the self-as-object. Even when it does, self-deception need not involve any split self or a number of homunculi, just like these are unnecessary in the case of self-knowledge or self-control. Thirdly, the role of the self-as-subject in self-deception cannot be that of an ordinary intentional agent, at least in those cases where the condition of simultaneity of inconsistent believing holds. The role of the subject must be weaker than that: The person herself is able to *overcome* the deception, e.g., by intentionally focusing attention upon relevant evidence in the right way, or by critically reflecting her motives.

Yet, the role of the self as critical thinker and author of actions is not so different after all. Remember that the recognition that certain descriptions apply to oneself can be highly essential for action-explanation [Perry (1994)]. When I hear on the news that TS is wanted by the police, I might for a moment not be clear that I am TS; but when I realize that I am TS, that will cause me to perform certain actions, such as quickly donning sunglasses, getting on the next train by whatever means, and hiding there in uncomfortable places, as did Cary Grant in *North by Northwest*. Similarly, when I hear other people saying that TS is self-deceived about his prospects of winning the next tennis match, I may at first not realize that people are talking about me. But when I do, I immediately do something about it. I check whether they are really talking about me or some other TS, and if they are indeed talking about me, I start thinking about whether I might have ignored certain facts about my next opponent on the tennis court. I might of course also attempt to safe the self-deceptive belief, but that only shows that I am involved actively as well. It all depends on the recognition that *my* beliefs are at stake, that I can think critically about them, and that I am responsible for them. The immediacy in which all of this happens would not be possible if the self played no explanatory role at all in self-deception. That involves some assumptions

about human beings being rational thinkers, but at first only moderate ones. Since the boundary between conscious and subconscious processes is permeable [Brandtstädter (2007)], people can become aware of what they have been previously unaware. They can, moreover, reflect their beliefs and desires critically due to our ability to develop second-order beliefs and desires [Frankfurt (1988)], and to develop principles of good reasoning. In short, no homunculi are reintroduced, and the self as subject helps to explain the acquisition and holding of self-deceptive beliefs because only thereby can we understand the immediate efforts of correction, control, or defense of one's own beliefs.

Let me conclude with another remark concerning rationality. I just spoke of rationality in the sense of the human *capacity* of reason, the capacity which allows us to argue, draw inferences, check the justification of our beliefs, and so on. I did not say which *norms* ought to guide the exercise of this capacity. While a conceptual analysis of self-deception cannot provide these norms, it should once again be emphasized that further empirical work requires closer connection to the ongoing debate about the foundation of norms of rationality. That is a task both intentionalists and non-intentionalist accounts have seriously ignored. Because many if not all alleged cases of self-deception might on closer inspection turn out to be something else — mere ignorance or error about oneself or others, wishful thinking, or some stronger mental pathology —, any account should ultimately face the test of reality. Hypothetical examples are not that test. The examples in the literature typically invoke that the self-deceived person is one who "overlooks certain available pieces of evidence", "misinterprets the data", or has "overwhelming evidence" which she or he then somehow ignores or distorts in order to either acquire or retain a cherished belief. Such vague descriptions are either left unexplained, as if it was clear that they are justified by common sense or by some universally accepted epistemology. At the same time, it is taken for granted that self-deception occurs. Or they are connected to psychological theories of heuristics and biases in reasoning, which in turn presuppose an ideal theory of rationality as normatively valid — the norms of logic, probability theory and statistics. I have argued why this is problematic, and I have stated conditions that would have to be satisfied for studies to be acceptable. We need more serious attempts to empirically demonstrate the occurrence of self-deception. These are not to be had without a substantive theory of norms of rationality and of the conditions of their proper application.*

Max Planck Institute for the History of Science
Boltzmannstr. 22,
D-14195 Berlin, Germany
E-mail: tsturm@mpiwg-berlin.mpg.de

NOTES

[*] This paper is in part based on Sturm (2007a).

REFERENCES

AUDI, R. (1982), "Self-Deception, Action, and Will", *Erkenntnis*, 18, pp. 133-158.
BAUMEISTER, R. (1987), "How the Self Became a Problem: A Psychological Review of Historical Research", *Journal of Personality and Social Psychology*, 52, pp. 163-176.
BAUMEISTER, R. (ed.) (1999), *The Self in Social Psychology*, Ann Arbor, Taylor & Francis.
BERMÚDEZ, J.L. (2000), "Self-Deception, Intentions and Contradictory Beliefs", *Analysis*, 60, pp. 309-319.
BIRD, A. (1994), "Rationality and the Structure of Self-Deception", *European Review of Philosophy*, 1, pp. 19-38.
BRANDTSTÄDTER, J. (1998), "Action Perspectives on Human Development", in R.M. Lerner (Ed.), *Theoretical Models of Human Development*, New York, pp. 807-863.
— (2007), "Causality, Intentionality, and the Causation of Intentions: The Problematic Boundary", in M.G. Ash & T. Sturm (eds.), *Psychology's Territories: Historical and Contemporary Perspectives from Different Disciplines*, Mahwah, NJ, Erlbaum, pp. 51-66.
BRANDTSTÄDTER, J. & STURM, T. (2004), "Apriorität, Erfahrung und das Projekt der Psychologie" ["Apriority, Experience, and the Project of Psychology"]. *Zeitschrift für Sozialpsychologie*, 35, pp. 15-32.
CASSCELLS, W., SCHOENBERGER, A. & GRABOYS, T.B. (1978), "Interpretation by Physicians of Clinical Laboratory Results", *New England Journal of Medicine*, 299, pp. 999-1001.
CHURCHLAND, P. (1995), *The Engine of Reason, the Seat of the Soul*, Cambridge, MA, MIT Press.
COHEN, L. J. (1992), *An Essay on Belief and Acceptance*. Oxford, Oxford UP.
DAVIDSON, D. (1986), "Deception and Division", in J. Elster (ed.), *The Multiple Self*, Cambridge, Cambridge UP, pp. 79-82.
DAY, W. (1977), "On the Behavioral Analysis of Self-Deception and Self-Development", in T. Mischel (ed.), *The Self: Philosophical and Psychological Issues*, Oxford, Blackwell, pp. 224-249.
DEMOS, R. (1960), "Lying to Oneself", *Journal of Philosophy*, 57, pp. 588-595.
— (1989), "The Origin of Selves", *Cogito*, 3, pp. 163-173.
DENNETT, D. (1991), *Consciousness Explained*, Boston, Little, Brown & Co.
DENNETT, D. & KINSBOURNE, M. (1992), "Time and the Observer: The Where and When of Consciousness in the Brain", *Behavioral and Brain Sciences*, 15, pp. 183-247.
DOUGLAS, W. & GIBBINS, K. (1983), "Inadequacy of Voice Recognition as a Demonstration of Self-Deception", *Journal of Personality and Social Psychology*, 44, pp. 589-92.

FIEDLER, K. (1988), "The Dependence of the Conjunction Fallacy on Subtle Linguistic Factors", *Psychological Research*, 50, pp. 123-129.

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

FRANKFURT, H. (1988), "Freedom of the Will and the Concept of a Person", in H. Frankfurt, *The Importance of What We Care About*, Cambridge, Cambridge UP, pp. 11-25

GALLAGHER, S. (2000), "Philosophical Conceptions of the Self: Implications for Cognitive Science", *Trends in Cognitive Sciences*, 4, pp. 14-21.

GERGEN, K. J. (1985), "The Ethnopsychology of Self-Deception", in M.W. Martin (ed.), *Self-Deception and Self-Understanding*, Lawrence, Kansas, Kansas UP, pp. 228-243.

GIGERENZER, G. (1991), "How to Make Cognitive Illusions Disappear: Beyond Heuristics and Biases", *European Review of Social Psychology*, 2, pp. 83-115.

GIGERENZER, G. & HOFFRAGE, U. (1995), "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats", *Psychological Review*, 102, pp. 684-704.

GILOVICH, T. (1991), *How We Know What Isn't So*, New York, Free Press.

GREVE, W. (1996), "Erkenne Dich Selbst? Argumente zur Bedeutung der 'Perspektive der ersten Person'" ["Know Thyself? Arguments Concerning the 'First-Person-Perspective'"], *Sprache und Kognition*, 15, pp. 104-119.

— (ed.) (2000), *Psychologie des Selbst* [*Psychology of the Self*], Weinheim, Beltz.

— (2005), "Die Entwicklung von Selbst und Persönlichkeit im Erwachsenenalter" ["The Development of Self and Personality in Adulthood"], in S.-H. Filipp & U. Staudinger (eds.), *Entwicklungspsychologie im Erwachsenenalter* (*Enzyklopädie der Psychologie*, C/V/6,), Göttingen, Hogrefe, pp. 343-376.

GUR, R.C. & SACKEIM, H.A. (1979), "Self-Deception: A Concept in Search of a Phenomenon", *Journal of Personality and Social Psychology*, 37, pp. 147-169.

HALES, S. (1985). "The Inadvertent Rediscovery of the Self in Social Psychology", *Journal for the Theory of Social Behaviour*, 15, pp. 237-282.

HERTWIG, R. & GIGERENZER, G. (1999), "The 'Conjunction fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors", *Journal of Behavioral Decision Making*, 12, pp. 275-305.

HOLTON, R. (2000), "What is the Role of the Self in Self-Deception?", *Proceedings of the Aristotelian Society*, 101 (1), pp. 53-69.

KANT, I. (1781/1787), *Kritik der reinen Vernunft* [*Critique of Pure Reason*], ed. by Jens Timmermann, Hamburg, Meiner, 1996.

— (1900ff.), *Gesammelte Schriften* [*Collected writings*], Academy edition, Berlin, De Gruyter.

KUNDA, Z. (1990), "The Case for Motivated Reasoning", *Psychological Bulletin*, 108, pp. 480-498.

LAZAR, A. (1999), "Deceiving Oneself or Self-Deceived?", *Mind*, 108, pp. 263-290.

LEWIS, B.L. (1996), "Self-Deception: A Postmodern Reflection", *Journal of Theoretical and Philosophical Psychology*, 16, pp. 49-66.

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

LOPES, L.L. (1991), "The Rhetoric of Irrationality", *Theory & Psychology*, 1, pp. 65-82.

MARKUS, H. & WURF, E. (1987), "The Dynamic Self-Concept: A Social Psychological Perspective", *Annual Review of Social Psychology*, 38, pp. 299-337.

McGinn, C. (1997), *The Character of Mind*, Oxford, Oxford UP.
Meehl, P. & Hathaway, S.R. (1946), "The K Factor As a Suppressor Variable in the Minnesota Multiphasic Personality Inventory", *Journal of Applied Psychology*, 30, pp. 525-564.
Mele, A. (1987), "Recent Work on Self-Deception", *American Philosophical Quarterly*, 24, pp. 1-17.
— (1997), "Real Self-Deception", *Behavioral and Brain Sciences*, 20, pp. 91-102.
— (2000), *Self-Deception Unmasked,* Princeton, Princeton UP.
Mischel, T. (ed.) (1977), *The Self: Philosophical and Psychological Issues*, Oxford, Blackwell.
Mischel, W. (1976), "The Self As the Person: A cognitive Social Learning View", in A. Wandersman, P.J. Poppen & D.F. Ricks (eds.), *Humanism and Behaviorism: Dialogue and Growth*, Oxford & New York, Pergamon Press, pp. 145-156.
Morawski, J. (2007), "Scientific Selves: Discerning the Subject and Experimenter in Experimental Psychology in the U.S., 1900-1935", in M.G. Ash & T. Sturm (eds.), *Psychology's Territories: Historical and Contemporary Perspectives From Different Disciplines*, Mahwah, NJ, Erlbaum, pp. 129-148.
Nagel, T. (1986), *The View from Nowhere*. Oxford, Oxford UP.
Newen, A. (2000), "Selbst und Selbstbewusstsein aus philosophischer und kognitionswissenschaftlicher Perspektive", in A. Newen & Vogeley (eds.), *Selbst und Gehirn* [*Self and brain*], Paderborn, Mentis, pp. 19-55.
Newen, A. & Vogeley, K. (eds.) (2000), *Selbst und Gehirn* [*Self and brain*], Paderborn, Mentis.
Nisbett, R.E. & Ross, L. (1980), *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ, Prentice Hall.
Patten, D. (2003), "How Do We Deceive Ourselves?", *Philosophical Psychology*, 16, pp. 229-246.
Pauen, M. (2001), *Grundprobleme der Philosophie des Geistes* [*Basic Problems of the Philosophy of Mind*], Frankfurt, Fischer Taschenbuchverlag.
Perring, C. (1997), "Direct, Fully Intentional Self-Deception is Also Real", (Comment on Mele, 1997), *Behavioral and Brain Sciences*, 20, pp. 123f.
Perry, J. (1994), "The Problem of the Essential Indexical", in Q. Cassam (ed.), *Self-Knowledge*, Oxford, Oxford UP, pp. 167-183.
Oaksford, M. & Chater, N. (1994), "A Rational Analysis of the Selection Task as Optimal Data Selection", *Psychological Review*, 101, pp. 608-631.
Quattrone, G.A. & Tversky, A. (1986), "Self-Deception and the Voter's Illusion", in J. Elster (ed.), *The Multiple Self*, Cambridge, Cambridge UP, pp. 35-58.
Rorty, A.O. (1972), "Belief and Self-Deception", *Inquiry*, 15, pp. 387-410.
— (1983), "Akratic Believers", *American Philosophical Quarterly*, 20, pp. 175-183.
Sackeim, H.A. & Gur, R.C. (1979), "Self-Deception, Other-Deception and Self-Reported Psychopathology", *Journal of Consulting and Clinical Psychology*, 47, pp. 213-215.
Sahdra, B. & Thagard, P. (2003), "Self-Deception and Emotional Coherence", *Minds and Machines*, 13, pp. 213-231.
Siegler, F.A. (1962), "Demos On Lying to Oneself", *Journal of Philosophy*, 59, pp. 469-475.
Skinner, B.F. (1953), *Science and Human Behavior*, New York, Macmillan.

STRAWSON, G. (1997), "The Self", *Journal of Consciousness Studies*, 5/6, pp. 405-428.

STURM, T. (2007a), "The Self between Philosophy and Psychology: The Case of Self-Deception", in M.G. Ash & T. Sturm (eds.), *Psychology's Territories: Historical and Contemporary Perspectives from Different Disciplines*, Mahwah, NJ, Erlbaum, pp. 169-192.

— (2007b), "The Just Cause of the 'Rationality Wars' in Psychology (and Philosophy)", in A. Beckermann & S. Walter (eds.), *Philosophy: Foundations and Applications*, Paderborn, Mentis, pp. 212-229.

TOMAKA J., BLASCOVICH, J. & KELSEY, R.M. (1992), "Effects of Self-Deception, Social Desirability, and Repressive Coping on Psychophysiological Reactivity to Stress", *Personality and Social Psychology Bulletin*, 18, pp. 616-624

TOULMIN, S. (1986), "The Ambiguities of Self-Understanding", *Journal for the Theory of Social Behaviour*, 16, pp. 41-55.

TRIVERS, R. (1985), *Social Evolution*, Benjamin/Cummings.

— (2000), "The Elements of a Scientific Theory of Self-Deception", *Annals of the New York Academy of Sciences*, 907, pp. 114-131.

TUGENDHAT, E. (1979), *Selbstbewusstsein und Selbstbestimmung*, Frankfurt, Suhrkamp (engl. *Self-Consciousness and Self-Determination*, transl. by P. Stern. Cambridge 1986, MIT Press.)

TVERSKY, A. & KAHNEMAN, D. (1974), "Judgment Under Uncertainty: Heuristics and Biases", *Science*, 185, pp. 1124-1131.

— (1983), "Extensional Versus Intuitive Reasoning: Conjunction Fallacy in Probability Judgment", *Psychological Review*, 90, pp. 293-315.

WASON, P.C. (1966), "Reasoning About a Rule", *Quarterly Journal of Experimental Psychology*, 20, pp. 273-281.

WASON, P.C. & JOHNSON-LAIRD, P. N. (1972), *Psychology of Reasoning: Structure and Content*, Cambridge, MA, Harvard University Press.

WELLES, J.F. (1986). "Self-Deception as a Positive Feedback Mechanism", *American Psychologist*, 41, pp. 325-326.

WHITTAKER-BLEULER, S. (1988), "Deception and Self-Deception: A Dominance Strategy in Competitive Sport", in J. S. Lockard & D. L. Paulhus (eds.), *Self-Deception: An Adaptive Mechanism?*, Englewood Cliffs, NJ, Prentice-Hall, pp. 212-228.