SELF-DECEPTION WITHIN THE PREDICTIVE CODING FRAMEWORK

Iuliia Pliushch

MAINZ 2016

Table of contents

I would very much like to thank both my supervisors for their impeccable guidance and the Barbara-Wengeler foundation for more than generous financial support, without which this work might not have seen the light of day.

List of figures

List of tables

Introduction

Have you ever experienced the feeling that in a split of a second your world view has radically changed to incorporate a disturbing insight: you have been deceiving yourself all along and you are not as smart as you thought, or not as healthy as you thought, or your friends are not those who you thought they were and, crucially, you feel to have known it all along? Or maybe you have experienced long and enduring conversations with somebody whom you think to be denying facts and stubbornly defending a positions that must be obviously wrong in his own eyes? Then you may have been self-deceived and may have ascribed self-deception to others. Self-deception (SD), or at least its ascriptions, is a pervasive phenomenon: delusion, self-enhancement, unrealistic optimism and many other phenomena have be associated with it, yet no consensus on how to distinguish self-deception from the mentioned phenomena is in sight.

Self-deception – an enigma or just a superfluous umbrella term for different kinds of cognitive biases? The epistemic aims of this thesis are first, to set philosophical constraints on a satisfactory explanation of self-deception; second, to extend the behavioral and phenomenological profile of the self-deceiver; third, to criticize empirical measures of testing self-deception and to propose improvements; fourth, to enrich the functional profile of the self-deceiver; fifth, to sketch a predictive coding account of self-deception.

Given these epistemic aims my argumentative aims and the structure of the thesis are as follows. The 1st chapter will provide a review of the philosophical literature on self-deception, as well as of the psychological paradigms and questionnaires for testing self-deception. The 2nd chapter will be concerned with my own take on what the explanandum of self-deception looks like, what the behavioral and phenomenological profile of the self-deceiver looks like, as well as how the motivation, process and properties of self-deceptive attitudes might be described. I use the notion 'profile,' instead of the 'necessary and sufficient conditions', because 'self-deception' is a fuzzy notion that is better to be characterized by a set of behavioral and phenomenological constraints – a profile. The 3rd chapter will then characterize different non-evolutionary and evolutionary accounts of self-deception that suggest different kinds of functions that it serves. Last, in the 4th chapter I will introduce predictive coding, emphasize and describe explanatory tools and apply them to construct a tentative predictive coding model of self-deception.

Argumentative aims of the first chapter: Four constraints are to be set on satisfactory explanations of self-deception: parsimony, demarcation, disunity in unity and phenomenological congruency. An explanation should be parsimonious not to postulate unnecessary internal states and should allow for the demarcation of self-deception from other phenomena. Further, each of the explanations of self-deception can be described as introducing a certain kind of disunity into another kind of unity. This is the paradoxical aspect of self-deception: how a certain kind of unity can be disunified? The personal level governed by rules of logic and rationality is usually taken as the unifying element and the kinds of disunities vary greatly (see the product debate on self-deception in 1.2). The last constraint requires that an explanation does justice to how self-deception is experienced, because, despite the fact that per definition self-deception cannot be experienced as self-deception (Borge, 2003), it does not mean that there is no phenomenology whatsoever. In the last part of the first chapter I will argue that questionnaires testing for self-deception suffer from the openness to discussion of the intentional objects of measurements (1.3.1). As for the paradigms for testing self-deception (skin conductance and pain endurance paradigm), I will argue that those confuse the personal and subpersonal levels of descriptions insofar as they make invalid inferences from the subpersonal to the personal level.

Argumentative aims of the second chapter: Intuitions have had a lion's share on which kinds of internal states have been postulated to underlie self-deception (2.1.1) Instead, I will propose to set the behavioral and phenomenological profile of self-deceivers as an explanandum (2.1.2). On the behavioral level, self-deceivers exhibit inconsistency that they nevertheless justify. On the phenomenal level, they may experience tension, but upon relinquishing self-deception insight kicks in. One possible characterization of tension, on the basis of Proust's (2013) theory of metacognitive feelings, is as metacognitive feelings with an indicator function (2.1.3). I will then argue that *subpersonal* goal representations provide the motivation of self-deceivers, that an alternative description of tension might be as a counterfactual goal-directed pull (see Irving, 2015 for the introduction of the term) such that the anxiety and discomfort experienced probably stems from a degree of both (2.2.1) and that self-deception stems from two kinds of selection: selection of a certain world/self-model and of a certain epistemic agent model such that which selection took place can be determined by the degree of transparency (unavailability of earlier processing stages) that self-deceptive attitudes might exhibit (2.2.3). As the personal level description of the epistemic agent model of self-deceivers, I will argue that the dolphin model of cognition best accomplishes this task (2.2.2.3).

Argumentative aims of the third chapter: In 3.1 I will discuss four non-evolutionary ways of defining a function of self-deception where each of the subsequent functions narrows down the previous one: (i) Self-deception serves to reduce cognitive dissonance (3.1.1); (ii) it serves to preserve the stability of the self-concept (3.1.2); (iii) it is the perseverance of self-esteem as a central feature of the self-concept that stands in the foreground (3.1.3); (iv) defense of self-concept and self-esteem are proximal (intermediate) means of reducing the anxiety of death as the ultimate aim (3.1.4). The last function – that of terror-management theory – builds a bridge to evolutionary functions in virtue of possible evolutionary implications: reduction of death anxiety has been argued to have an evolutionary benefit. In 3.2 I will then explore three evolutionary functions offered for self-deception: that it serves the deception of other individuals (3.2.2), that it is a by-product of rational capacities of the human mind (3.2.3.1) and that is an exaptation – a by-product that acquired the function to uphold positive self-perception (3.2.3.2). I will propose a combined function hypothesis, namely that self-deception may have evolved to relieve the anxiety of death, but in virtue of the changes of social relationships in the current environment may have acquired an additional function of other deception. The reason for this proposal is that it would combine the defensive and offensive function and explain at the same time why observers often recognize the self-deception in others. If self-deception served the evolutionary aim of deceiving others, then it should remain hidden from observers, which is often not the case. So, either one explains this as some kind of a race (observers getting better vs. self-deceivers learning new tricks), or one assumes that there has to be something else apart from deceiving others. Anxiety of death is one such handy alternative: it is observer-independent in that the recognition of an observer of somebody else's self-deception is not a condition for the success of that self-deception.

Argumentative aims of the fourth chapter: First, I will review Sahdra & Thagard's connectionist model of self-deception that emphasizes the role of emotions in changing the acquisition of attitudes (4.1). I will then present two personal level error minimization accounts on the basis of which Mele has built his account of self-deception (4.2). This will serve as a contrast for predictive coding as a *subpersonal* error minimization account. Thus, I will then distinguish four different kinds of Bayesian explanation with respect to their explanatory scope – including phenomenology or not (4.3). This is because since predictive coding uses terms that can have a personal, as well as subpersonal reading, e.g. 'inference,' 'prediction error,' 'uncertainty,' care should be taken to avoid confusion. Last, I will review explanatory tools of predictive coding such as free energy, set of attractors, different kinds of ways of changing the generative model of causes of our sensory input and different kinds of inferences, precision and counterfactuals (4.4). The last section will, then, combine them to analyze self-deception (4.5).

Before I start, I want to give the reader a feeling of the debate on the example of the difficulty to satisfy the demarcation constraint on self-deception. Self-deception is a goal- and emotion-directed kind of subpersonal hypothesis testing that leads to belief-like attitudes which may acquire the property of realness. Goal representations are responsive to context in selective ways and may also lead to the creations of new contexts (2.2.1). The intuition that self-deception is intentional, when it is subpersonally selective, arises in virtue of the fact that for the observer the self-deceiver seems to *control* the selectivity in question. The problems recapitulated at the beginning of this section point into the direction of the possession by self-deception researchers of the intuition that one can self-deceive about *every kind of content* one wants, e.g. see Van Leeuwen's criticism of Mele that one can self-deceive also about subjectively costly outcomes that violate the principles of error-minimization (1.1.2.3; 3.2.3.1). It is this control and potential generality of selectivity to any kinds of content that is to be explained. Self-deceptive belief-forming process is seldom a case of explicit deliberation (2.2.2), such that personal level accounts that describe it as such still have to use psychologically plausible subpersonal mechanisms, e.g. different kinds of biases, in order to explain motivated changes in the information that is taken into account or the amount of processing. Given the assumption of Mele's influential theory that every kind of bias may be motivated (1.1.2.3), self-deception cannot be restricted by the kinds of biases that may or may not be self-deceptive, which makes personal level accounts vulnerable to Van Leeuwen's criticism that the notion of 'self-deception' loses its scientific value, if applied to every kind of bias. I agree that, at least in theory, every kind of bias can be broadly described as self-deception, if self-deception is to be defined as motivated distortion of truth. To show this I chose a couple of biases (from the Wikipedia list of cognitive biases at http://en.wikipedia.org/wiki/List_of_cognitive_biases) that have not so far been taken to be self-deceptive and applied the strategy of personal level accounts to them (see table 1).

Illusion/Bias	Hypothetical example
Clustering illusion	Ted is a neuroscientist against whose theory recently an experiment has been published. Ted wants to prove his theory right, given the amount of time and effort that he invested into its developments, as well as the consequences giving up would have for his career and the grants he will be able to get. He conducts a new experiment and falls victim of the *clustering illusion* – he overestimates the importance of small trends in the data and sees an effect where there is none.
Bias blind spot	Vicky's view on whatever topic that is relevant to her self-concept have been challenged by counterarguments. Vicky thinks that those arguments are biased in virtue of the *bias blind spot* – tendency to see herself less biased than others.

Table 1. Hypothetical examples of self-deception

The lesson from this is that the psychological biasing procedures are not a good demarcation criterion for self-deception. In chapter two I will argue what the behavioral and phenomenological profile of the self-deceiver is.

1 Review: Desiderata for the concept of self-deception

> Belief systems may be like works in progress – a potpourri of compatible, semi-contradictory, and totally contradictory ideas. Unless we are writing a dissertation or engaging in psychoanalytic self-examination, we may rarely dedicate the effort required to pull all our beliefs together and come to a logical conclusion on most of what we think.
> (Brown & Kenrick, 1997, pp. 109–110)

In this chapter I will list and specify the requirements on the concepts of self-deception that are mentioned in the philosophical and the psychological literature, as well as modify or discard if necessary. The negative thesis I will come to is that regarding self-deception as a special case of (faulty) belief-formation will not aid in distinguishing it from other phenomena, as it has been argued to be in the case of delusions (Bortolotti, 2010). In the second chapter I will then present my own positive thesis: self-deception (SD) is a motivated kind of subpersonal hypothesis-testing with a specific phenomenological and behavioral profile.

In recent years, the main debate about the definition of self-deception has been dominated by two opposing camps – the intentionalists[1] (Davidson, 1986, 1998; Pears, 1991; Rorty, 1988, 1986, 1994, 2009; Fingarette, 2000[1969]; Bermúdez, 1997, 2000b; Talbott, 1995, 1997) and the deflationarists (Bach, 1981, 1997, 1998; Barnes, 1997; Johnston, 1988; Mele, 2001, 2010, 2012). Recently though, the scale between intentionalists, which are those claiming SD to be a result of intentional action, and proponents of a deflationarist approach, which are those claiming it to have another "motivational or affective basis" (Scott-Kakures, 2012, p. 18) has tipped in favor of the latter, not least because of the tighter conceptual similarity between motivated cognition and deflationary definitions of self-deception, e.g. self-enhancement, optimism, that allow for an easier empirical testing (see Helzer & Dunning, 2012, p. 380 for the view that motivated cognition is a "paradigmatic case of self-deception").

In the following chapter, I will start with the philosophical literature on self-deception that is to be followed by the review of psychological questionnaires designed to test self-deception on the basis of the philosophical definitions of the phenomenon. Typically, three building blocks are emphasized in the self-deception literature: the nature of the motivation (1), the process (2), and the final effect or "product"[2] (3 and 4) (Van Leeuwen, 2007a; Nelkin, 2012; Funkhouser, 2009). It is guided by three main questions: What is the driving force behind SD? Which *kind* of reasoning process led to its occurrence and how was it influenced by the initial motivation? What exactly marks out the resulting cognitive attitude? Although all authors agree that the process of self-deception is to be explained somehow, some authors lay more explanatory weight on motivation, rather than the kind of misrepresentation and vice versa, so that two debates have emerged: the motivational and the "product" one – that will be considered in the following two sections.

[1] Deweese-Boyd (2012) in the Stanford Encyclopedia, as well as Scott-Kakures (2012), Galeotti (2012), Porcher (2012) and other authors of the articles in the issue on self-deception in Humana.Mente (2012) give overviews over the intentionalist-deflationasirts debate.

[2] Van Leeuwen (2007a) proposes calling "the attitude that results from self-deception the *product of self-deception*" (p. 421). In this article, he also gives a brief overview over what he calls the "strategies" for solving the paradox of self-deception (i.e., believing in obviously contradictory propositions).

1.1 Motivation debate: intentionalists vs. deflationary positions[3]

The structure of this section will be as follows: I will first present different intentionalist (1.1.1) and deflationary (1.1.2) positions independent of each other, in order to then summarize the problems for explanations of self-deception and set out constraints in the last part (1.1.3).

1.1.1 Intentionalist positions

In this section I will explore six intentionalist positions, each with a little different kind of flavor, for the reader to get an idea of the benefits and shortcomings of an intentionalist solving strategy. An intentionalist positions, as the name already suggests, holds that intentions are the motivational elements of self-deception: intentions *cause* and *sustain* self-deception. A more general characteristic of intentionalist positions, deduced also from the use of intentions as explanatory tools, is that these are (mostly) *personal* level explanations. This is because 'intentions' are folk-psychological concepts that belong to the personal level description – the description of a person as a whole.

The argumentative results of the section will be that a *personal* level explanation of self-deception is insufficient for two reasons: first, that of *parsimony* and second, that an explanation of self-deception requires an explanation of *disunity in unity*. Personal level is the one that provides the unity, but the subpersonal one brings in the disunifying element. Which positions will I present, in which order and how do they contribute to the argumentation? First, two divisionist positions will be summarized – Davidson's and Pears'. Both explain self-deception as a kind of *compartmentalization*. This means that since personal level beliefs are governed by the rules of logic which enforce consistency, inconsistency in beliefs is explained by postulation of their storage in different compartments which are consistent in themselves. The difference between the two accounts lies in the subsystemic *agency* assumption which Pears defends. I deny subsystemic agency for parsimony reasons and I think that a more abstract description of such a compartmentalization strategy indicates a general characteristic of self-deception, namely that self-deception is *disunity in unity*. Personal level is the unifying element in virtue of the rules of logic that govern it. The subpersonal level is the one doing the explanatory work, namely the one that is disunified and has a causal influence on the kind of unity we find at the personal level. In other words, which set of beliefs we adopt as persons is dependent on the disunified subpersonal workings. Simplified example: I would like to eat a cake, but still believe that I did everything I could not to gain weight. My intention to eat a cake for reasons of pure joy is compartmentalized from my another belief set, containing beliefs such as "Chocolate eating is very important for mental work such as thinking and creative writing, so me eating a cake is just helping me in achieving the best results in my thesis". Those two belief sets are present at the subpersonal level, but only *one* of them can be adopted at the personal level, because the personal level is that of unity. Summarizing, the first two accounts make the disunity in unity problem vivid.

The next two accounts – Fingarette's and Rorty's – present two different kinds of explanation of disunity in unity in self-deception. Fingarette narrows down the unity element from the personal level to personal identity (here I assume a part-whole relationship between personal identity and the personal level). Basically, the answer is that a person can choose which characteristic to embrace as belonging to personal identity. Self-

[3] My choice for the accounts to consider has been influenced by Deweese-Boyd (2012).

deception is, thus, argued to involve personal agency and *narrative construction*, instead of belief formation. Rorty, from a different point of view, explains self-deception by abandoning unity of the self (thus, here the unity element is the self-concept) as an *ideal* requirement that is often violated in virtue of subpersonal disunity. Rorty names different mechanisms for such disunity. There is another crucial difference between these two accounts – Fingarette's account is *constructionist*, while Rorty's – *dispositional* (I will speak at length about this difference in 1.2.7). For Rorty, Davidson and Pears, as well as most other self-deception experts, self-deceptive beliefs are stored and retrieved at the time that they influence action. For Fingarette and Michel (1.2.7), those are constructed at the time they are needed for action, e.g. avowal. Which type of account better explains self-deception? Dispositional accounts face static and dynamic paradoxes. These paradoxes are those that *every* theory of self-deception has to avoid, but particularly those accounts that model self-deception on the basis of interpersonal deception – intentionally bringing about a certain belief oneself (Bermúdez, 2000b, p. 309) – are susceptible to them:

- *static*: how to conceive of the possibility to believe and not-believe *p* at the same time
- *dynamic*: the presence of an impossible deceptive strategy in self-deception - a deceptive strategy aimed to deceive oneself (Mele, 2001, pp. 7–8).

The straightforward way to solve the static paradox would be to abandon the contradictory belief requirement (e.g., Talbott, 1995) and to solve the dynamic paradox – to posit *diachronic* or *intentionally self-induced self-deception* where the deceiver and the deceived do not overlap in time. Yet, diachronic self-deception has been argued not to be of an "ordinary" kind (Talbott, 1995, pp. 31-32, footnote 6) while contradictory belief requirement is often preserved in a modified form (see the "product" debate in 1.2).

Constructionist accounts do not necessarily face the static paradox, yet, are still susceptible to the modified version of the dynamic paradox: How can I choose what to accept as belonging to personal identity or self-concept in the face of contradictory evidence? And this, again, shows the need to resort to something else, apart from the unifying element that explains the disunity, demonstrated in diachronic or synchronic oscillations in the unifying element (oscillations in the accepted set of beliefs for example). The subpersonal level is often used for these explanatory purposes. The dynamic paradox makes every explanatory account embracing personal agency as a characteristic of self-deception problematic.

Bermúdez and Talbott offer yet another explanation for disunity in unity in self-deception. For Bermúdez, the *content* of the intention, e.g. to deceive oneself or to avoid an unpleasant state of the mind, can solve the dynamic paradox and conceal self-deception from oneself. In terms of the introduced explanatory tool – disunity in unity – the unified element is one's adopted belief set and if the content of the intention is congruent with the belief set, then there is no need for the disunified element. Reformulated cake example: If the content of my intention is not to acquire a *false* belief that I eat the cake for the sake of enhanced mental functioning, but the intention is just to acquire a belief that me eating the cake is justified, then there is no logic inconsistency or paradox. The problem with omitting the disunifying element is that, actually, it is not omitted, but another kind of disunity is introduced: disunity in the causal relationship between beliefs. This is because, in virtue of us as supposedly rational creatures, we are assumed to accept something as a belief only if it is justified. If we were aware of that the content of our intention that lead to a certain belief is not justified, we would be forced to relinquish this belief, because our rational capacities would indicated that the acquisition of a belief only in virtue of its pleasantness is not justified. Nelkin (1.1.2.5) will speak of the potential causal connection between the self-deceptive belief and its motivating element as a reason why self-deceivers are responsible for their self-deception. At the same time, *absence* of such a causal connection at the time of being self-deceived, whatever the content of the intention is or whatever other kind of motivational element one postulates, is a disunity introduced into self-deception. Last, Talbott offers a Bayesian account of self-deception according to which self-deception occurs in the case a certain preference in belief acquisition is present. Where this preference and, thus, the causal connection between it and the self-deceptive acquired belief made clear to the self-deceiver, this would undermine the justification for her acquisition and maintenance of the self-deceptive belief.

Again in short, lessons from the six accounts to be presented is that explanations of self-deception differ in the kind of unity they postulate to underlie self-deception and the kind of disunity they postulate to explain how self-deception works. There are at least two kinds of account – constructionist and dispositionalist. Yet, adoption of *personal* agency as a requirement is problematic, because the dynamic paradox is present and *subpersonal* agency is implausible for parsimony reasons. Typically, the unity domain is a certain personal level construct, e.g. 'self' or 'identity'. The explanatory weight then lies on the subpersonal level that offers an explanation for how and why changes in the unity element are made. The question of what the motivation for self-deception is, if not intention, is not clarified in this section and is to be discussed further in the following section (1.1.2) on deflationary accounts of self-deception that argue that certain kinds of *desires* and/or *emotions* provide the motivation for self-deception. The result of that section will be that desires, as folk-psychological constructs akin to intentions, also have to be supplemented by a subpersonal level of the explanation of self-deception. Apart from *parsimony* and *disunity in unity*, two more constraints on an explanation of self-deception will be imposed: first, *congruency between the (subpersonal) explanation and the phenomenological level of description*, an example of which is the tension problem (= contradictory beliefs evoke feelings of anxiety and uneasiness) and, second, *clarification of demarcation criteria*. The last constraint can be subdivided into (at least) the demarcation of the motivation (Nelkin's content dilemma), demarcation of the scope of self-deception (Van Leeuwen's usefulness problem) and demarcation of the mechanism (Bermúdez's selectivity problem). I will write in detail on this in section 1.1.2.

1.1.1.1 Davidson: weakness of the warrant

> Self-deception is thus a form of self-induced *weakness of the warrant*, where the motive for inducing a belief is a contradictory belief (or what is deemed to be sufficient evidence in favour of the contradictory belief).
> (Davidson, 1986, p. 89; my emphasis)

On the example of Donald Davidson's often cited view, I want to demonstrate the central problem for accounts of self-deception resulting from the contradictory belief requirement and still encourages discussion (e.g., see Michel and Newen's, 2010, critique of Mele's account and 1.1.2.3): Even if self-deception arises as a result of the operation of subpersonal mechanisms, a certain kind of *personal level recognition* has to occur for the phenomenon to deserve the label self-deception. The mystery is about the relationship between *mechanisms (subpersonal level causation)* and *reasons (personal level deliberation and justification)* in producing self-deception (e.g., see for more on this distinction Bortolotti, 2010, pp. 183-187 and 1.2.6). Beliefs are often not formed via *explicit deliberation* or search for reasons (p. 183), sometimes they just occur and might be justified by the subject. Now, can subpersonal causes for beliefs be available at the personal level as reasons (Bortolotti, 2010, pp. 183-184 discusses this possibility in the context of delusions, for more on delusions and self-deception see 1.2.6 and 1.2.7)? This seems to me prima facie implausible (for more on the personal/subpersonal distinction see 1.3.4). If self-deception has been caused subpersonally, it might not merit the label 'self-deception' anymore, if it has been caused personally, paradoxes of self-deception would arise. Davidson's account is of a personal kind. He tries to explain how acquisition of contradictory beliefs can be achieved via reason-searching and gives up the unitary picture on the human mind in order to avoid the static paradox.

Davidson (1986) has proposed that self-deception possesses the following features:

- *intentionality* (p. 87);
- *irrationality or a genuine inconsistency* that is a deviation from the person's own norms (p. 79).
- having two *contradictory beliefs that stand in a causal connection*[4] (p. 79).

The second feature imposes a constrain on an explanation of self-deception, namely that it should avoid the *paradox of irrationality* that consists in the fact that an explanation for an instance of irrationality in the subject leads to a certain form of rationalization such that if the rationalization succeeds, irrationality vanishes, as it has been explained away (p. 79). The irrationality involved in self-deception is that of weakness of the warrant, in analogy to the weakness of the will. Weakness of the warrant violates the normative principle of continence while weakness of the will – the normative principle of the requirement of total evidence for inductive reasoning (see table 2). [5]

	Weakness of the warrant (= self-deception)	**Weakness of the will**
Kind of attitude	Cognitive	Evaluative
Analogy	Judging a hypothesis on the basis of less pieces of available evidence than is available.	Acting on the basis of less relevant reasons than one recognizes.
Product	Belief	Intention
Violated normative principle	**Principle of continence** enforces a fundamental kind of consistency in thought, intention and evaluation of action.	**Requirement of total evidence for inductive reasoning** enforces the choice of the hypothesis, most supported by the evidence, from the available set of hypotheses.

Table 2. Davidson: weakness of the warrant vs. weakness of the will. Distinctions from Davidson (1986, pp. 80-82).

Davidson's explanation of the acquisition of self-deception looks as follows: The undesirable thought that *p* motivates the agent to come to believe (through avoidance of attention, evidence search etc.) and sustain (through making boundaries within the mind) the belief that non-*p*, although the belief that *p* does not cease to exist in the mind. Making (non-permanent) boundaries results in the development of in itself rational parts of the mind. For the agent to recognize the inconsistency, one has to erase the boundary first. This

[4] The contradiction cannot be believed, so those contradictory beliefs have to be distinct from each other, even if causally connected: "The distinction we need here is between believing contrary propositions and believing a contradiction, between believing that p and believing that not-p on the one hand, and believing that [p and not-p] on the other" (Davidson, 1998, p. 5).

[5] Noordhof (2003) describes Davidson's approach to the mental as an *interpretist* approach, insofar as *rationality* is according to this sort of approach seen as a necessary condition for an agent to be interpretable (p. 75). Further, for Davidson, the *possession* of the requirement of total evidence is to be distinguished from the *awareness* of the latter: "If we grant, then, as I think we must, that for a person to 'accept' or have a principle like the requirement of total evidence mainly consists in that person`s pattern of thoughts being in accordance with the principle, it makes sense to imagine that a person has the principle without being aware of it or able to articulate it" (Davidson, 1986, p. 83).

way rationality can be preserved, but a unified model of the self has to be sacrificed (Davidson, 1986, pp. 88-89). Consequently, the two critical elements in the explanation of self-deception are for Davidson (1986, 1998):

1. Mental *causes* of mental states are at work that are *not reasons* (Davidson, 1998, p. 7): The desire to change a belief is not a reason for the truth value of the changed belief (p. 8).
2. *Impermanent boundaries* between parts of the minds account for the possibility of the inconsistent beliefs in the self-deceived (1986, pp. 91-92; 1998, p. 8-9).

In the following, I discuss three implications of Davidson's view. First, different aspects of the self-deceptive process might be held to be crucial for it to be successful:

- success conditions for getting self-deceived are to be focused on, e.g. *concealment* of the way the self-deceptive belief has been acquired, instead of the inconsistency and irrationality (Davidson, 1986, p. 91, argues this to be the case for Pears and Bach, see sections 1.1.1.2 and 1.2.3 for their views);
- *continuing* motivation is necessary to uphold self-deception (*threat* as motivation for the boundary) and, thus, contradictory beliefs are necessary to uphold the boundary (this is Davidson's preference).

Second, with respect to the relationship between self-deception (as a resulting attitude) and other attitudes Davidson (1998) argues that it is impossible to draw a firm distinction between self-deception and these attitudes:

> The moral I draw from these examples is brief. Self-deception comes in many grades, from ordinary dreams through half-directed daydreams to outright hallucination, from normal imagining of consequences of pondered actions to psychotic delusions, from harmless wishful thinking to elaborately self-induced error. It would be a mistake to try to draw firm lines within these continua. (Davidson, 1998, p. 18)

Bortolotti's (2010) recent claim that delusion cannot be differentiated from other irrational beliefs by their irrationality (1.2.6) echoes this claim. Nonetheless, Davidson (1986) does make a distinction between self-deception and wishful thinking that consists in the latter being an overly positive phenomenon and often an ingredient in self-deception (see table 3 and 1.2.7 for distinguishing self-deception from other phenomena).

Self-deception	Wishful thinking
- often benign	
- there are reasons for believing the proposition	
- may be positive or negative	- positive
- irrational	- not (necessarily) irrational
- intentional	- ingredient in self-deception

Table 3. Davidson: self-deception vs. wishful thinking.
Distinctions from Davidson (1989, pp. 85-89).

Third, the kind of evidence, needed to self-deceive, is important for Davidson's account. Davidson (1986) criticizes Kent Bach for holding that the self-deceiver does not believe in the weight of the contradictory evidence (pp. 90-91). I think that, to acknowledge contradictory evidence, but, still, believe to the contrary, is a strong personal level requirement for self-deception. Interestingly, Davidson (1998) weakens this requirement by suggesting that such personal level activities as *vivid imagining* (p. 13) or *acting as if* (p. 15) could be means of acquiring self-deception.[6] These activities imply that the attitude,

[6] Davidson (1998) argues also that *cognitive dissonance* could also play a role in acquiring self-deception (p. 17). For more on cognitive dissonance and self-deception see setion 3.1.1.

acted out by them, is not taken as seriously by the subject as other attitudes he might possess (see 1.2.4 for Gendler's pretense model of self-deception as imagining).

That the question on how causes and reasons interact in self-deception is still an issue, can be seen on the example of Bagnoli's (2012) account who argues that self-deceivers are not taking evidence as reasons. The cause of such behavior is that the self-deceiver aims to protect "the coherence and stability of her emotional and epistemic system" (pp. 110-111), which leads her to omit *taking the available evidence as reasons* that should have led her to change the belief acquisition process (p. 107). Another little elaborated question is that on the kind of accounts of beliefs to be favored. Bortolotti (2010) has interpreted Davidson as accepting a dispositional account of belief: "dissonant attitudes can be ascribed as beliefs if these are not simultaneously activated (maybe 'stored' in different mental compartments)" (p. 78). A dispositional account of beliefs is often assumed in accounts of self-deception, yet, so far it has been more of an impediment than has led to insight in explaining self-deception (for explanatory problems connected with it see the "product" debate, for the critique of the dispositional account of beliefs and an alternative constructivist view see Michel, 2014 in 1.2.3).

Interim conclusion: personal level dispositional accounts of self-deception as a certain kind of irrational belief are difficult to argue for and do not allow to distinguish self-deception from other related irrational phenomena (more on demarcation criteria in 1.2). Davidson's compartmentalization strategy into the (personal) level of reasons and the (subpersonal) level of causes demonstrates that one constraint on self-deception is to explain *disunity* in *unity*. Unity level is in this case the one of the reasons governed by logic and rationality and disunity is achieved by compartmentalization on the subpersonal level.

1.1.1.2 Pears: subsystemic agency view

Pears' account is similar to Davidson's in that both explain self-deception in terms of a division of the belief set out in at least two subgroups. While Davidson guards himself from postulating separate centers of agency, Pears (1991) defends the given hypothesis (p. 394). The question is, thus, whether *subpersonal agency* is a smart move in explaining self-deception. To make the distinction between Davidson and Pears more vivid: the claim that certain kinds of motivations have *caused* the reasons during explicit deliberation (personal belief-forming process) to change, or their evidential value to change, is to be distinguished from the claim that it is not subpersonal causes, but rather subpersonal centers of agency that intentionally bias the belief-forming process of another center of agency. Pears (1986) argues that there can be at least two kinds of self-deception (with possible cases in between, making self-deception a gradual phenomenon):

- the *extreme form of self-deception*, which follows from the "full surface connotation" of the term self-deception is to believe both *p* and not-*p*;[7]
- the *open and unstable self-deception*, in which the self-deceiver is merely aware of the irrationality of the belief obtained, as well as its underlying motivation, but not of the contradiction itself (pp. 66-68).

[7] There are different degrees to which one can claim that *p* and not-*p* is believed: "Here we can arrange the mistakes on a scale of increasing badness. It is a mistake to believe *p* firmly when the evidence is equally distributed between *p* and *not-p*, a worse mistake to believe *p* in the teeth of inductive evidence for *not-p*, and an even worse mistake when logical or mathematical necessity are defied. Now the most extreme form of this last mistake is to believe both *p* and *not-p*, which is what is required by the full surface connotation of the word 'self-deception'" (Pears, 1989, p. 60).

Instability and tension as the phenomenological components indicating inappropriateness of the belief-forming process of the self-deceiver will be also emphasized in other accounts of self-deception (e.g., 1.1.2.3 and 1.1.2.4). To explain the extreme form of self-deception, Pears (1986), similarly to Davidson and drawing an analogy to Freud's theory, postulates two subsystems: subject of deception (*S*) and object of deception (*O*) -, which are temporary, functionally insulated, rational, brought about by an "ordinary wish" and, supposedly, only one of these subsystems enjoys the benefit of its beliefs being conscious[8] (pp. 76-77). Pears (1986) holds that the other subsystem (the deceiver S) is preconscious, instead of unconscious, because the preconscious is rational, while the unconscious, according to Freud, is not. Moreover, he defends Freud from Sartre's circularity criticism[9] by stating that *S* produces the belief that *p* not in itself, but in *O* (pp. 73-74). I will not go into the discussion of Freud's theory any further. The purpose of this passage was to show that postulations of motivated divisions of the belief set might be coupled to certain psychodynamic connotations. Though not in such a strong form, psychodynamic connotations and the conscious/unconscious distinction is up until now evoked to aid in the explanations of self-deception (e.g. see Lockie's, 2003, approach in 1.2.1).

Returning to the need to postulate subpersonal intentions, one of the reasons for Pears to do this is to avoid that motivation in self-deception is seen as mere *automatic* goal-influence. Thus, Pears not only argues for the presence of additional, non-truth-attaining *goals* in self-deception,[10] a claim that I agree with, but strengthens his claim by postulating *intentional* biasing of beliefs in self-deception (1991, p. 393). In doing this, he differentiates himself from Johnston's (1988) account of subintentional biasing due to a mental tropism (1.1.2.1), where a mental tropism is "a characteristic, non-accidental, and nonrational connection between desire and belief – a *mental tropism* or purpose-serving mental mechanism" (p. 67). Pears draws a parallel between conscious intentional processing of information and that in self-deception, which he thinks may ground on the equally efficient similar kind of intentional processing[11] such that the subsystem S exerts "a *continuous*

[8] For Pears, consciousness is a property that can be propagated to functionally connected beliefs: "In any case, it is necessary to treat lack of consciousness [in self-deception] as a special kind of *functional insulation*" (Pears, 1989, p. 76; my emphasis).

[9] Sartre's criticism of Freud is according to Pears as follows: "Sartre's main criticism of Freud misses this point. He argues that Freud has to explain how the self-deceiver can form the belief that *p* while still believing that *not-p*, and that his explanation, which is that the belief that *not-p* is repressed in the unconscious, will not work, because the same conflict between the two incompatible beliefs will then break out in the unconscious" (Pears, 1989, pp. 73-74). This is to say that a property of consciousness per se is not a *mechanism* that add and takes away this property. Further it is doubtful that unconscious beliefs *could* conflict at all. An alternative would be that only those attitudes that successfully passed the barrier of consciousness can stand in conflict.

[10] The assumption is that there are two different kinds of goal representations, truth-conducive and not: "It is an important fact that self-deception always needs a further goal to override the usual goal, truth, or, at least, rationality, which gives us the best chance of attaining truth" (Pears, 1989, p. 63). This assumption is also accepted in psychological theories about the workings of motivation, see 1.3.3.

[11] The doubtful assumption here is that goal-directed unconscious processing has to possess the attributes of goal-directed conscious processing: "The suggestion is not that simple tropisms – still less, a single simple tropism – will explain all cases of self-deception. What is being suggested is that the conscious processing of information may be paralleled by an equally effective kind of processing which does not require a separate centre of consciousness, but which is, nevertheless, sufficient to support the concept of a separate centre of agency (or, at least, to justify Davidson's view, that the sub-system acts on the principle of intentional action)" (Pears, 1991, p. 400). I will review psychological literature that such an agency-overhead is not necessary for unconscious processing in section 2.2.1.

influence on the main system's thinking of its own thoughts" (Pears, 1991, p. 403; my emphasis). This elaboration makes vivid the problem of identifying the kind of motivation in self-deception: more than *automatic* causation, but less than *conscious* agency, and if possible, without the postulation of *subsystemic agency*. It is in question whether self-deceiver's phenomenology will aid in solving this problem. As Pears (1991) notes, self-deception is different from paradigmatic cases of intentional action in which an intention is coupled with its explicit avowal (p. 399) and is tied to explanation of actions (p. 401):

> The attribution of intentions to the sub-system lacks the support of the *self-deceiver's explicit avowal* at the time of the notion and we have no detailed understanding of the basic mental acts of the sub-system. If we still adhere to intentionalism, either in Davidson's version or mine, it will be in spite of these deficiencies. (Pears, 1991, p. 404; my emphasis)

Another point worth keeping in mind is the notion of unstable self-deception, or the one characterized by a certain phenomenology of uneasiness. As we will see later, the question about the phenomenology of self-deception, and the question about the necessity of the feeling of uneasiness in any kind of self-deception, is still one of the central points of the discussion today.

Interim conclusion: Doing justice to the *complexity* of self-deception (Pears, 1991, p. 398), while staying truthful to Occam's razor is a goal for every theory that aim to explain this phenomenon. Thus, *parsimony*, along with *disunity in unity*, is an important *constraint* on explanations of self-deception. Accepting agency as a characteristic of self-deception leads to unattractive view of agency as a subpersonal phenomenon, or else paradoxes of self-deception occur. One has to resort to another level of explanation, not governed by rules and rationality, to explain self-deception, else Davidson's paradox of irrationality would occur. And this another level is the subpersonal one.

1.1.1.3 Fingarette: personal identity account

Fingarette (2000[1969]) argues that self-deception is intentional, but is about the constitution of personal identity, instead of belief-acquisition, in the first place. This account is worth commenting, because it goes into the direction of Bortolotti's (2010) assumption that I am inclined to accept, namely that hypothesis testing and the construction of a coherent self-narrative cannot be (easily) disentangled from each other (1.2.6). If further a constructivist view on belief-forming is accepted (see Michel, 2014 in 1.2.3), then what might be followed is that at each point in time attitudes are formed on the basis of the available context, part of which is the self-narrative.

Let me look at Fingarette's (2000[1969]) definition of self-deception. He holds it to be intentional[12] - essentially purposeful (p. 16) - to distinguish it from "commonplace inconsistency in beliefs" (p. 15). Self-deception is said to involve a conflicting state of partial belief and partial disbelief (p. 24; see Schwitzgebel's in-between-belief view in 1.2.3) and the acknowledgement of adverse evidence (p. 23). The gist of Fingarette's (2000[1969]) alternative (to the belief-forming) explanation is as follows: self-deception is usually being described in *cognition-perception terms*, which builds a contrast to *volition-action terms* (p. 33) and it is the latter kind of account that he favors (p. 34). Fingarette

[12] Interestingly, self-deception is argued to be a *specific kind* of wishful thinking (p. 19). Hence, while for Davidson (1.1.1.1) wishful thinking was a part of self-deception, for Fingarette their relationship is the other way around: self-deception is a kind of wishful thinking. This is just an indication of how difficult it is to differentiate self-deception from other related phenomena.

reinterprets the meaning of being explicitly conscious of something, which figures centrally in the explanation of self-deception (p. 34), to mean *spelling out* certain features of the engagement of the person with the world (p. 38), thus, consciousness to be a skill (p. 43), that is being exercised in linguistic or similar form (p. 42). On the assumption that in the majority of cases the individual does not spell out his engagement (p. 39), self-deception encompasses situations in which there is compelling reason not to spell out the given engagement (p. 42). As such, the presence of strong emotions or surprise in cases of public spelling-out confirms the presence of self-deception (p. 56). The puzzling aspect in the disavowal of the self-deceiver is that it is "insincerely sincere" [13] (p. 50):

> Hence (1) he [the self-deceiver] says nothing, (does not spell-out the truth); (2) he gives us the impression that, "in some sense", he could if he would; but (3) he also gives us the impression that he has somehow rendered himself incapable of doing it. (Fingarette, 2000[1969], p. 47)

Such disavowal aids in determining *personal identity*, instead of being just a certain case of faulty belief-forming process (p. 66). Personal identity is understood here as one's own identity self-defined through avowals, acknowledgement of something as his for him, in contrast to individual identity which does not imply avowal (pp. 68-70). Though the framework that Fingaratte constructs is different from that of Davidson and Pears (see the previous sections), there are common elements. Fingarette (2000[1969]) argues that disavowal is constituted by three dimensions – isolation, non-responsibility and incapacity to spell-out (p. 73). The isolation from everything that is avowed builds on the assumption that rationality is central to the self (p. 80). The self as a coherent unity is being constructed by the individual throughout his life where avowals link him to his engagements (p. 82) and self-deception enables the defensive use of particular roles played by the individual for cover purposes (p. 89). According to Fingarette (2000[1969]), disavowal is the "'dynamic' essence of defence" in the Freudian sense[14] with defense serving to reduce anxiety (p. 130-132; for other accounts emphasizing anxiety see Johnston in 1.1.2.1 and Barnes in 1.1.2.2). It is the third option of comporting oneself with respect to engagements inconsistent with one's cause and aims, the first two being that an engagement is either abandoned or avowed (pp. 137-138). Disavowal has the characteristics of the lack of authority with respect to expressing the engagement in question, rejection of personal responsibility and irrational persistence in its pursuit (p. 140). While Fingarette states that self-deception has a *defensive*

[13] The criteria for sincerity ascription are, according to Fingarette (2000[1969]), as follows: "(1) It is not the case that there is an intentional difference in the way the individual spells his engagement out to others and the way he spells it out to himself; (2) The way he spells the engagement out to himself reflects the engagement correctly and aptly. (3) He has not been unintentionally wrong in the way he came to express the engagement [thus, he is purposely wrong]" (p. 51).

[14] Interestingly, Greenwald (1988) thinks that Fingarette reintroduces the paradoxes of self-deception as an entity similar to Freud's censor: "This reintroduction occurred in the form of an unnamed mechanism that analyzes the true (threatening) import of circumstances and, on the basis of the knowledge so obtained, purposefully prevents the emergence into consciousness of both the threatening information and the defense against it. Fingarette's unnamed mechanism is capable of inference and intention in a way that requires sophisticated symbolic representation, yet Fingarette assumed that this mechanism operates *outside* of the ordinary machinery of inference and symbolic representation – that is, outside of consciousness." (Greenwald, 1988, p. 116)

function to avoid unpleasant truth,[15] he leaves open the possibility of unconscious[16] knowledge in self-deception (p. 64).

Interestingly, the psychological means of self-deceiving are argued to be the changes in the focus of one's attention in the face of the limited capability of humans to process information (p. 168): "suspension of disbelief" (p. 175) is taken to be "a form of sincerity, of authentic belief and emotional reaction, even while the person is taking account, in the background, of the real world and its contrary features" (pp. 174-175). As for the neural level, Fingarette considers the dissociations present after brain damage that constrains the information flow between the left and the right hemisphere as indirect support for his theory of self-deception, arguing that the distorting of the communication between the hemispheres by psychological means would fit his definition of disavowal (p. 157; see 3.1.2.1 for Ramachandran's hemispheric specialization account of self-deception).

There is a more recent account, that of Rachel Brown (2003), that is similar to Fingarette's. According to her, self-deception is a kind of constructed self-narrative or "the story one narrates to oneself about one's motivations and projects, both short-term and long-term; one groups the events of one's life thematically, in order to understand, assess, and monitor oneself" (p. 282). The function of a self-narrative may be prospective or retrospective (p. 287). In self-deception, the guidelines of constructing the narrative collide - the truth orientation of the narrative, on the one hand, and one's commitments, on the other (p. 289). Curiously, the distinction that she makes between her account and Fingarette's is that the narrative construction in the latter is not sufficiently intentional (Brown, 2003, p. 294), though this does not seem to be so according to Fingarette.[17] More precisely, Brown holds that the changes in attentional focus, in the way suggested by Fingarette, lead to the *avoidance* of unfavorable evidence, while, according to her, the self-deceiver *notices* his fears, desires and beliefs, feels anxious and, figurally speaking, "obscures her unwanted pink elephant by quickly constructing a story according to which it does not exist" (p. 294).

Interim conclusion: there are at least two different ways to explain self-deception – a narrative construction or as belief formation (more on this in 1.2.7). Yet, even if self-deception is about the construction of a coherent self-narrative, instead of belief-forming as such, its explanation in intentional terms still remains enigmatic. I will consider different kinds of intentions in section 1.1.1.5.

[15] He has in mind defense performed by the agent or personal-level defence: "If our subject *persuades* himself to belive contrary to the evidence *in order to evade*, somehow, the unpleasant truth to which he has already seen that the evidence points, then and only then is he clearly self-deceived" (Fingarette 2000[1969], p. 28).

[16] Fingarette (2000[1969]) argues that his account fits those of Sartre, Kierkegaardand Freud, yet is more specific and informative (p. 98). The distinction between acknowledging engagements as one's own and accepting responsibility for them is made neither by Sartre, nor by Kierkegaard according to Fingarette (pp. 103-104).

[17] The following passage demonstrates that Fingarette (2000[1969]) holds self-deception to be intentional and differ in this respect from wishful thinking: "If having a passionate wish that something should be so brings it about unintentionally that one believes it so, this may with some propriety be called self-deception, though I believe we are much more likely to call it prejudice or wishful thinking. To the extent that we suppose the belief to be *intentionally cultivated* because of the wish to believe, we distinguish the case as one of self-deception, a less innocent and philosophically far more interesting kind of wishful thinking" (p 19, my emphasis).

1.1.1.4 Rorty: irrational implications of a non-unitary self

While Fingarette explains self-deception by means of the unifying concept of 'personal identity,' Amélie Rorty explains it by the disunity of the self-concept. Her (1994) article on self-deception has been reprinted in Martin's collected volume on the topic of deception (Rorty, 2009), suggesting that her views still have not lost their actuality. Her understanding of self-deception is very broad and some of her ideas are similar to those developed in recent accounts of self-deception (see below). Her understanding of intentionality is also very broad. Rorty (1988) holds self-deception to be intentional, yet, understands intentionality as a gradual matter,[18] beginning with preconscious discrimination and ending with justification of certain beliefs (pp. 19-20). She defines (at least one variety of) self-deception as a (traditional) personal level account of belief formation and justification:

> X is self-deceived about p when
> (1) X believes that p at t (where t covers a reasonable span of time);
> (2) Either (a) X believes not-p at t; or (b) X denies that he believes p at t;
> (3) X *recognizes* that p and not-p conflict;
> (4) X denies that his beliefs conflict, advancing an improbable ad hoc reconciliation, making no attempt to suspend judgement or to determine which belief is defective.
> (Rorty, 2009, p. 244, footnote 1; my emphasis)

Similar to Davidson, Rorty (1988) holds that self-deception is an instance of "patterned and persistent forms of irrationality" (p. 14) whose explanation necessitates the change to the classic picture of the self as unitary in gradual way to different degrees (p. 15-16). According to her, the two pictures of the mind – unitary and non-unitary - are ineliminable and complementary (1988, pp. 22 - 25).

Classic view of the self
1. *Unity* dominated by rationality.
2. *Transparency*[19] which is accessibility of mental states to each other.
3. *Truthfulness* which is transparency's aim to minimize error; the implication would be that psychological functions have the form of propositions.
4. *Reflexiveness* which is that the criteria for rationality are themselves subject to critical evaluation.

Table 4. Rorty: description of the classic view of the self
Distinctions from Rorty (1988, pp. 14-16).

What is special about Rorty's (1988, 2009) account is that she does not focus on a certain property or a mechanism of self-deception, but offers a variety of them, that can be encountered also in other accounts:

- not episodic, but a "continued and complex pattern of perceptual, cognitive, affective, and behavioral dispositions" (Rorty, 2009, p. 248);
- sustained through social support, needn't result in a false belief (for the discussion of the belief-question see section 1.2);

[18] The following passage demonstrates that for Rorty rationality is a matter of degree: "Intentionality begins with relatively simple preconscious discrimination, and ranges through increasingly complex forms to self-consciously and systematically justified clusters of propositionalized beliefs. Since these routines of intentional activity can occur relatively independent of one another, intentionality can be a matter of degree" (Rorty, 1988, p. 19).

[19] Since this is the first use of the term 'transparency' is this review chapter, but they are further to come in sections 1.2.3 and 1.3.1, I want to note that, starting from chapter 2, I will use this term only in the sense of Metzinger (2003) as inaccessibility of earlier processing stages.

- its range of domains needn't be restricted to important matters, it needn't center on the self (most accounts of self-deception hold the reverse, e.g. see Holton, 2001, that self-deception is about oneself, or Taylor's positive illusions elaboration in section 3.1.3.1);
- its motivation can be a disposition or a *habit* instead of a desire or a wish (the assumption of such a weak kind of motivation, if ever, as habit, is unusual for accounts of self-deception);
- it needn't be harmful (Rorty, 2009, pp. 248-250; for the reverse, that self-deception may have immunological and psychological costs, see Trivers, 2011);
- adaptive to survival due to the need to function in highly diverse environment (Rorty, 1988; see section 3.2 for evolutionary functions of self-deception);
- occurs in *cases of indeterminacy* (see Sloman's experiment that is supposedly shows that self-deception requires vagueness in 1.3.2.3) and involves an interest of generating a *self-fulfilling prophecy* of accomplishing our goals (p. 17).

The mechanisms,[20] by which self-deception is being accomplished, involve salience, bootstrapping in the situation of indeterminacy (enhancement) with the goal of generating a self-fulfilling prophecy, inertia of belief in the face of counterevidence (a characteristic that is also argued by Van Leeuwen to aid in the generation of self-deception, see 3.2.3.1), changes in the level of generality of descriptions (compare to Greve & Wentura's self-immunization discussed in 3.1.2.2), rationalization (1988, pp. 18 – 19; 1986, pp. 125-126; 2009, pp. 250-253) which aid rational actions and, consequently, is the by-product of functional structures and strategies that attempt to integrate different subsystems of the self[21] (1988, p. 21).

Interim conclusion: Rorty's account demonstrates the *breadth* of phenomena that could – and have been – taken to be cases of self-deception: from non-motivational habit over self-fulfilling prophecies and self-enhancement to rationalization. As we will see, the question whether self-deception is a certain *kind* of phenomenon and how to narrow down the use of the term, still haunts every attempt of an explanation. It has made its last appearance in Van Leeuwen's (2013b) criticism of Triver's usage of the term – supposedly too broad to be scientifically useful (3.2.2.1). I label Van Leeuwen's point of critique as 'usefulness problem' and categorize it as a subproblem of the *demarcation* problem. To see the list of the problems, one can refer oneselves to section 1.1.3.

1.1.1.5 Bermúdez: three kinds of self-deceptive intentions

So far, four accounts have been presented. Three dispositional one's (Davidson, Pears and Rorty) and one constructionist (Fingarette). Two constraints have been identified – parsimony and disunity in unity. Demarcation constraint has been mentioned, but will be elaborated in 1.1.2 and 1.1.3. The account to be presented is a dispositional one that solves the disunity in unity issue by denying disunity and arguing that the *content* of the intention can be relaxed to be congruent with one's set of beliefs. Bermúdez (2000b) argues that it is a necessary condition that a self-deceptive belief is brought about intentionally (p. 310) and differentiates three kinds of intentions that a (potential) self-deceiver[22] could possess (see figure 1):

[20] Interestingly, Rorty (1989) holds that self-deception needn't always be motivated, but can also be a matter of habit.

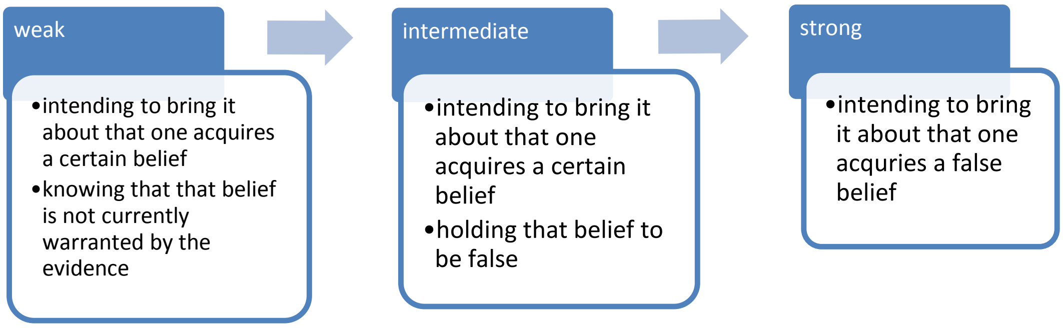
[21] The reader is encouraged to compare Rorty's (1986) claim that "[s]elf-deception and *akrasia* are by-products of psychological processes that make ordinary rational action possible" (p. 119) to Van Leeuwen's spandrel view that is discussed in section 3.2.3.1.

[22] Interestingly, Bermúdez (2000b) argues that wishful thinking – and any other similar kind of motivationally biased beliefs - is different from self-deception insofar as (p. 312):

- the self-deceiver a) wants it to be the case that p (as in wishful thinking); b) wants it to be the case that he believes that p;

(1) S believes that *not-p* but intends to bring it about that he acquires the false belief that *p*.
(2) S believes that *not-p* but intends to bring it about that he acquires the belief that *p*.
(3) S intends to bring it about that he acquires the belief that *p.*
(Bermudez, 2000b, p. 310)

According to Bermúdez (2000b), the given differentiation of types of self-deceptive intentions makes sense on the basis of the assumption that "intentions are not closed under known logical implication" (p. 311) which imposes the same requirement on knowledge ascription and means that: "I can know that x entails y and intend to bring about x without *ipso facto* intending to bring about y" (Bermúdez, 2000b, p. 311). All the three types of self-deceptive intentions mentioned above have, according to Bermúdez (2000b), in common that the self-deceiver is intentionally manipulating his own belief-forming mechanisms (p. 312). Weak kind of intention is similar to deflationary attempts to explain which kind of inconsistency is present in self-deception (inconsistency between the self-deceptive belief and the belief that there is a significant change that the former belief is false; see Mele's account in 1.1.2.3). As such, the weak kind of intention is also susceptible to the criticism that can be applied to deflationary approaches, namely that such kind of inconsistency is too weak for the phenomenon to deserve the label self-deception. Bermúdez' suggestions on how the strong and the intermediate kinds of intentions might avoid the paradoxes of self-deception (see table 5) are not parsimonious (postulation of unconscious intentions), explanatory unhelpful (it is unclear *how* inferential insulation or losing touch with an intention should occur) or again too weak (in case of one belief gradually changing into another).



weak

- intending to bring it about that one acquires a certain belief
- knowing that that belief is not currently warranted by the evidence

intermediate

- intending to bring it about that one acquires a certain belief
- holding that belief to be false

strong

- intending to bring it about that one acquries a false belief

Figure 1. Bermúdez: types of self-deceptive intentions. Distinctions from Bermúdez (2000b, p. 310-312).

Answer to the static paradox (threatens self-deception with the intermediate and the strong kind of intention)	**Answer to the dynamic paradox (threatens self-deception with the strong kind of intention)**
1. inferential insulation in case of simultaneously contradictory beliefs 2. weakening of belief in not-p that is inversely proportional to the strength of belief in p	1. unconscious intentions are possible 2. losing touch with the conscious intention in the process of implementing it

Table 5. Bermúdez: answers to the static and dynamic paradoxes Distinctions from Bermúdez (2000b, pp. 313-314).

- self-deception is intentional.

Nelkin and Funkhouser (see section 1.1.2.5) emphasize the importance of the desire to believe for self-deception. This is yet another possibility for how wishful thinking and self-deception might be related.

Apart from descriptive possibilities that intentional and deflationary accounts offer, what is the justification to favor one over the other? According to Bermúdez (2000b), explanation of self-deception in intentional terms is an *inference to the best explanation* and this being the case implies that one only needs to show that garden-variety instances of self-deception are intentional (p. 315-316). Instead of accomplishing this task though, Bermúdez argues against deflationary accounts of self-deception by identifying the problem to which the latter ones – and especially Mele's influential account (1.1.2.3) - are susceptible, namely the selectivity problem. The *selectivity problem* consists in the fact that a desire that *p* is insufficient as a motivation to initiate cognitive biases in favor of acquiring a given belief.[23] Intentions and motivated biases are, thus, two different solutions to the selectivity problem. The preferred solution determines the camp one finds oneself in – intentionalist or deflationary one. Cold biases are cognitive distortions or errors that occur in random situations such that an ascription of a certain motivation for their occurrence cannot be made (Bermúdez, 1997, p. 108). Hot biases[24] are those distortions for which motivation ascriptions can be made. Mele's account makes use of cold biases by stating that they can be motivated by desires of the individual. This can be understood as a claim that whatever mechanism is responsible for the distortions that occur in random situations, those can be activated by motivational attitudes. According to Bermúdez, it is not always the case that desires trigger those mechanisms, but intentions do. Mele, on the contrary, doubts that intentions satisfactorily explain the reason why one might self-deceive in one kind of situation, but not in another (1.1.2.3). In short, the selectivity problem is based on the assumption that in two identical situations it is hypothetically possible to either self-deceive or not and that it is a certain kind of attitude that can distinguish between these two cases. For intentionalists it is intention, for proponents of deflationary accounts – motivation, usually a desire. But what if there is no sharp distinction between the situation and one's motivational attitudes? If motivation is understood broadly to also encompass emotions, then Jackendoff's (2012) proposal that thoughts possess character tags that are experienced as feelings (p. 220) might be used as support for the claim that the distinction between "hot" and "cold" biases is a gradual one. Duncan and Barrett (2007) present a similar argument:

> Any thought or action can be said to be more or less affectively infused, so that there is no ontological distinction between, say, affective and non-affective behaviours, or between 'hot' and 'cold' cognitions. (Duncan & Barrett, 2007, p. 1202)

Interim conclusion: first, more sophisticated kinds of self-deceptive intentions do not make an explanation of self-deception any easier, and, second, the avoidance of the selectivity problem is yet another constraint on a theory of self-deception. Avoiding disunity could not be achieved, since, as mentioned in the introduction to 1.1.1, the awareness of the content of the intention and the causal connection between it and the self-deceptive belief

[23] According to Bermúdez, a lower acceptance threshold is not a *sufficient* condition to self-deceive: "It is simply not the case that, whenever my motivational set is such as to lower the acceptance threshold of a particular hypothesis, I will end up self-deceivingly accepting that hypothesis" (Bermúdez, 2000b, pp. 317-318). Curiously, intention seems to be for intentionalists such a condition, highly likely because of the *control* and *free will* that comes for granted when agency is involved.

[24] Bayne & Fernández (2009, pp. 8-9) offer two possibilities for distinguishing affect and motivation from cognition: phenomenology (hot states as vivid and intense) and functional role (hot cognition as directed at goal achievement).

would undermine the justification of the latter. Thus, also on this kind of account there is a certain disunity, namely the one between the content of the intention and the set of acquired beliefs, or maybe, there is an absence of the causal connection between the two.
Selectivity problem is, as the usefulness problem mentioned in the previous section, a subtype of the demarcation issue (for the list of constraints, see 1.1.3). Nelkin (2002) has rephrased the selectivity constraint as follows: The kind of motivational attitude, that causes self-deception, has to be *distinctive* of self-deception: broad enough so that it is present in all cases of self-deception, yet narrow enough to exclude cases of irrational belief-formation that are not those of self-deception ("content dilemma," p. 393). I see the difference between selectivity and content problems in that the first is about the distinctiveness of the *mechanism* (that reliably brings about self-deception and only self-deception) and the second is about the *motivation* (that reliability brings about self-deception and only self-deception). Again, see section 1.1.3 for the summary of constraints.

1.1.1.6 Talbott: utility maximization account

William Talbott (1995) has offered a Bayesian account of self-deception that explores the possibility of a *personal level* self-deceptive Bayesian belief-forming process in a unified self. The presentation of his account is useful as a contrast to the predictive coding which is a *subpersonal* Bayesian model that will be presented in the fourth chapter. Talbott (1995) argues that one does not need to model self-deception as interpersonal deception and because such kind of modeling is the only reason for postulating the belief that not-*p* in the subject, abandoning the interpersonal model leads to abandoning the condition of two contradictory beliefs (p. 30) and relinquishing this condition leads to the abandonment of divisionist approaches. Talbott (1995) holds Rorty's theory of self-deception to be an instance of robust divisionism approach, of Pears – moderate divisionism approach and of Davidson weak divisionism approach. *Divisionist* approaches are defined by him as the ones that posit a certain partition of mental life to explain self-deception where this partition is an *extra kind of division* that has not been used to explain non-self-deceptive phenomena (p. 29). Postulation of such kinds of extra divisions are explanatorily not parsimonious.

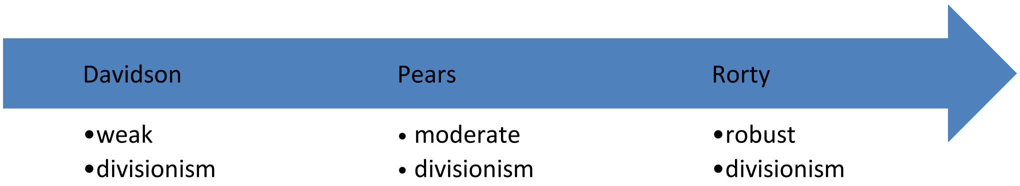


Figure 2. Talbott: characterization of divisionist approaches to self-deception Distinctions from Talbott (1995).

By abandoning divisionism and the contradictory belief requirement Talbott does not have to face the static paradox anymore, but still has the dynamic paradox to explain, namely[25] how the self-deceiver can acquire the self-deceptive belief given the desire to believe it regardless of its truth-value (p. 34). Talbott further holds to offer not only an *anti-divisionist*, but also an *intentionalist* account[26] (p. 32). According to the three kinds of

[25] Talbott (1995) states that it is a higher-level *static* paradox that he has to explain (p. 34), but I argue that it is the *dynamic* paradox on which he elaborates, because the focus is on the strategy by which the self-deceiver can succeed and not on the relationships among acquired beliefs.

[26] Talbott's theoretical reasons in favor of the argument that an account of self-deception has to be intentional is that he discards the three possible non-intentional explanations of the selectivity problem – why self-deception occurs in some cases in which motivation that p or

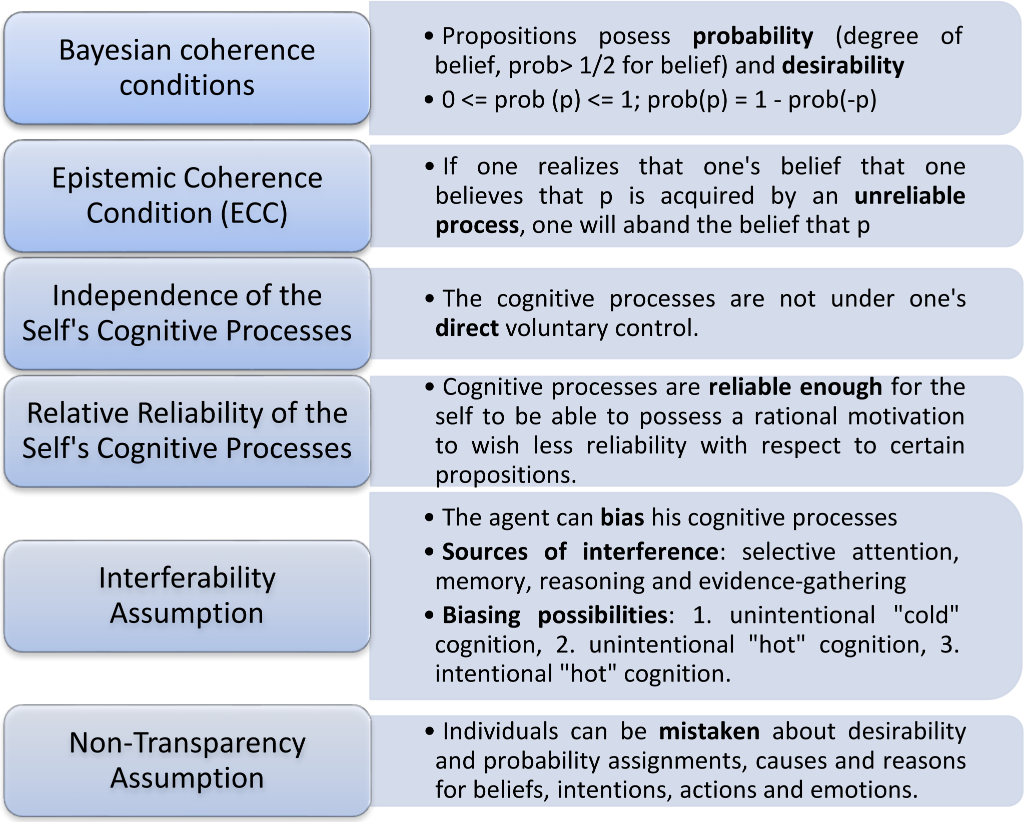
intentions differentiated by Bermúdez (see the previous section) it would be an even weaker intention than the weak one (in which the self-deceiver at least knew that the belief he wishes to acquire was unwarranted by the evidence). For Talbott, an acquisition of a belief regardless of its truth value is already self-deception. This emphasizes once more the explanatory tension between the requirement that a contradiction should be available to the agent and the impossibility to explain the acquisition and maintenance of such a contradiction in a unitary self:

> On my account, self-deceptive belief that *p* does not involve lying to oneself about p. It instead 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. (Talbott, 1995, p. 30)

A Bayesian agent is practically (not epistemically) rational (p. 37, footnote 16) in that it has according to Talbott (1995) the intention to maximize expected utility (p. 48, footnote 26). Since propositions possess not only probabilities with which the agent assumes that they are true, but also *desirabilities*, the latter can take the role of desires (p. 48, footnote 26). The constraints that Talbott (1995) formulates on an ideally coherent human self (see figure 3) are argued to allow for the possibility of intentionally biasing the agent's beliefs. Their interaction can be explained as follows: Agents reason by maximizing expected utility which means that they assign probability and desirability to propositions (Bayesian coherence condition), but they can be mistaken about the latter (non-transparency). Though the reasoning process is to a certain degree reliable (relative), the agent cannot believe something at whim (independent of cognitive processes) and if unreliability is recognized, the belief would be abandoned (ECC). Agents can nevertheless bias their reasoning process, because non-transparency allows for certain kinds of interference, e.g selective attention.[27] In ordinary circumstances cognitive processes are still reliable due to them being evolutionary selected for (Talbott, 1995, p. 34).

the motivation to believe that p is present, but not in others (see the previous section for more on the selectivity problem): 1. innate differentiation (natural selection); 2. conditioning; 3. non-intentional mechanism for expected utility maximization (pp. 62-63). He discards the first two as implausible and the third as not parsimonious, because it would lead to the postulation of two mechanisms for expected utility maximization – one intentional and one not (p. 63). His practical reason, moreover, is that, even if it seems implausible to assume such a complex reasoning process of which the agent is unaware, but "in fact, almost all of one's reasoning goes on without of one's being aware of it" (p. 57). Talbott's proposal is, then, that the demarcation between self-deception and other related psychological phenomena is due to the *intentionality* as a criterion (p. 65). For example, insofar as repression and denial are argued to be brought about by non-intentional mechanisms, they are not instances of self-deception (p. 65). The fact that, according to Talbott (1995), the subject can be unaware of his intentions (p. 33), makes it difficult to prove that those intentions were present in the first place and this, as a consequences, throws a shadow onto intention as a proposed demarcation criterion.

[27] The strategies that, according to Talbott (1995, pp. 34 - 36), enable self-deception are 1) selectivity of memory; 2) division between the intentional self and *sub-intentional or sub-personal* (in Dennett's sense) psychological mechanisms, implying that the details on the processes that accomplish the biasing are not inherent in the self-deceiver's intention (p. 35); 3) non-transparency or fallibility of (at least some) mental processes; 4) division of mental states into conscious (about which the individuals possesses accurate beliefs) and unconscious (about which the individual is ignorant or possesses inaccurate beliefs).



Bayesian coherence conditions

- Propositions posess **probability** (degree of belief, prob> 1/2 for belief) and **desirability**
- 0 <= prob (p) <= 1; prob(p) = 1 - prob(-p)

Epistemic Coherence Condition (ECC)

- If one realizes that one's belief that one believes that p is acquired by an **unreliable process**, one will aband the belief that p

Independence of the Self's Cognitive Processes

- The cognitive processes are not under one's **direct** voluntary control.

Relative Reliability of the Self's Cognitive Processes

- Cognitive processes are **reliable enough** for the self to be able to possess a rational motivation to wish less reliability with respect to certain propositions.

Interferability Assumption

- The agent can **bias** his cognitive processes
- **Sources of interference**: selective attention, memory, reasoning and evidence-gathering
- **Biasing possibilities**: 1. unintentional "cold" cognition, 2. unintentional "hot" cognition, 3. intentional "hot" cognition.

Non-Transparency Assumption

- Individuals can be **mistaken** about desirability and probability assignments, causes and reasons for beliefs, intentions, actions and emotions.

Figure 3. Talbott: Defining assumptions for an ideally coherent self
Distinctions from Talbott (1995, pp. 48 - 52).

Talbott (1995) argues that human reasoning is *selective*,[28] because in trying to explain certain evidence it is impossible to test every hypothesis.[29] The selectivity is achieved by *fixed background beliefs* that limit the range of hypotheses to be tested (p. 41; see figure 4). Talbott draws, in this respect, the analogy between this kind of hypothesis testing and Tversky and Kahneman's *framing effect*: certain framing concepts invoke schemata for cognition and action, thus, predetermine the latter two to a certain degree (p. 41), for example students expecting a warm/cold guest lecturer will interpret him accordingly (p. 42 footnote 20). The disanalogy lies in the fact that framing is "hot" in the case of self-deception (p. 41). In this claim Talbott borrows Mele's idea (1.1.2.3). Importantly, Talbott (1995) argues that the (possible) strong emotional reaction to attacks on the truth value of protected beliefs is not due to the recognition, that the opposite is true at some level, but rather it is the phenomenological "accompaniment"[30] to the recognition that they *could* be true which would be destructive for the protected beliefs (p. 45). The shift from the cognitive to the emotional domain is important not only as a phenomenological requirement on self-deception, but also as weakening of the irrationality of the self-

[28] Talbott (1995) holds that self-deception can be understood as a kind of *doxastic immunization* against failing to believe some proposition (p. 37; notice the similarity to Greve & Wentura's self-immunization account explored in section 3.1.2.2).

[29] See Van Leeuwen's account in section 3.2.3.1 for a similar idea. In fact, apart from the finitude of the human mind as an argument for the appropriateness of selectivity in reasoning, he and Talbott share other similar elements, such as the emphasis on practical rationality and an attempt to explain self-deception on a personal level in a unified self.

[30] Talbott (1995) argues that desires need not have any phenomenological feel (p. 33, footnote 10).

deceiver: no contradictory beliefs, but a feeling evoked by the inappropriateness of the belief-forming process (for more see, Proust, 2013, 2014 in 2.1.3). This so called tension requirement (feeling of uneasiness) will also become important for deflationary positions, because it will be used by intentionalists against deflationary accounts. The claim will be that deflationary accounts cannot explain tension, because there are no inconsistent attitudes, but only causally induced false beliefs (1.1.2). The reference to framework beliefs is also interesting, because delusions have been claimed to be framework beliefs too (see Bortolotti, 2010 for a critique of this hypothesis). Thus, it might be an explanatory move for attitudes resistant to change despite contradictory evidence that their resistance is said to be due to them not being in need of justification themselves.[31] An alternative, explored in section 2.2.3, will be that self-deceptive attitudes acquire the property of realness and lose their belief status.

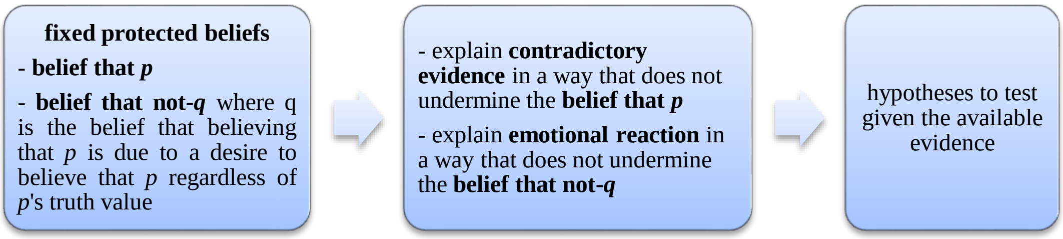

fixed protected beliefs
- **belief that *p***
- **belief that not-*q*** where q is the belief that believing that *p* is due to a desire to believe that *p* regardless of *p*'s truth value
- explain **contradictory evidence** in a way that does not undermine the **belief that *p***
- explain **emotional reaction** in a way that does not undermine the **belief that not-*q***
hypotheses to test given the available evidence

Figure 4. Talbott: biased mechanism for selecting hypotheses
Distinctions from Talbott (1995, pp. 43-44).

Last, I want to mention the way in which Talbott (1995) solves the *selectivity problem* presented in the previous section. According to him, not every instance of a motivated bias is an instance of self-deception. During the reasoning process, the expected utility of biasing is compared to the expected utility of not biasing and self-deception occurs only when there is a certain preference ranking of outcomes. For it to occur, the top first position in the preference ranking must be "*p* and I believe that *p*" and the top second position must be "not-*p* and I believe that *p*" (p. 54; p. 55, footnote 34). In other words, the self-deceiver wants to believe the truth and if not the truth, then the favorable proposition, even if it is not the case. Basically this idea is similar to Mele's (2001) error minimization account. Mele assumes that desires can influence the false rejection ("I do not believe *p* and *p* is the case") and false acceptance errors ("I believe *p* and not-*p*") and that calculation of the given errors influences the threshold of belief acceptance/rejection.

The calculation of utility that determines the preference ranking can be accomplished without awareness despite its complexity (Talbott, 1995, p. 57) and human hierarchies of self-deceptive intentions are argued to be finite, yet complex enough and not necessarily in awareness despite their complexity (pp. 57-58). Talbott (1995) argues that self-deception is testable, even if the given tests should be more complex than simply asking the subjects about their cognitive processes being intentionally biased (p. 58) and might contribute to psychological health, which is determined by its fit to the personality and circumstances (p. 66). The difficulties with its testability will be a central topic of the section 1.3 and its function – of the third chapter.

Interim conclusion: Talbott's *Bayesian* account is interesting for two reasons. First, it is interesting as an attempt to explain self-deception in a rationality preserving way. This is

[31] See coherentist and foundationalist theories of belief justification: Coherence among beliefs would be violated by the contradictory belief requirement in self-deception, but those being at the foundation might explain their resistance to change.

because probabilistic Bayesian reasoning is an alternative account of rationality to the logical one[32] (Oaksford & Chater, 2009). Unfortunately, Talbott's account does not go above the claim that can be found in Mele, namely that self-deception is accompanied by a certain preference ranking. The key tool of probabilistic accounts is not that of a preference ranking, but that of *uncertainty*. Logic specifies absolutely certain inferential relations, while Bayes – inferential relations of various degrees of uncertainty (Oaksford & Chater, 2009, p. 70). The second one is to be preferred, because it does justice to the fact that cognitive systems have to possess a model of the environment in order to solve problems in a rational manner. As Oaksford & Chater (2009) put it, "[t]he core objective of rational analysis, then, is to understand the structure of the problem *from the point of view of the cognitive system*, that is, to understand what problem the brain is attempting to solve" (p. 72).

This leads directly to the second reason why Talbott's account is interesting, namely as a predecessor of predictive coding (see chapter 4) – a recent computational theory whose main claim is that the brain attempts to minimize free energy and that it is a "probabilistic inference machine." That this is the case has been doubted by Oaksford & Chater (2009, p. 83), as they argue for a weaker version that "the mind is a qualitative probabilistic reasoner, in the sense that the rational analysis of human reasoning requires understanding how the mind deals qualitatively with uncertainty" (p. 84; see 4.3 for different types of Bayesian explanations). *Levels* of uncertainty are used in predictive coding explanations of psychopathology in general and delusions in particular (e.g., Hohwy, 2013a; Friston, Stephan, Montague, & Dolan, 2014; for the connection between delusion and self-deception see 1.2.7). Different *kinds* of uncertainty are also distinguished, e.g. uncertainty about the states of the world (perceptual uncertainty), on the one hand, and uncertainty about how those states of the world change (environmental volatility), on the other (Mathys et al., 2011). More generally, predictive coding also tries to preserve the intuition that psychopathology can be explained by *optimal* Bayesian inference, namely "optimal inference under an agent's generative model of the world," which is argued in an article with the expressive title "Optimal inference with suboptimal models" (Schwartenbeck et al., 2015, p. 110). Are generative models of the self-deceivers optimal? I will come back to this question in 1.1.2.3 when discussing Mele's error-minimization account. Last, referring to the disunity in unity constraint, disunity has been argued to be avoided, but is also present here in the form of the absence of the causal connection between one's utility calculations and the acquired self-deceptive beliefs.

[32] Johnson-Laird's mental model view is according to Oaksford & Chater (2009) a kind of a logical view and, thus, different from the probabilistic solution (p. 71).

1.1.2 Deflationary positions

To sum up the results so far, in section 1.1.1 I presented six intentionalist accounts of self-deception and identified two kinds of constraints on explanations of self-deception: parsimony and explaining disunity in unity. To satisfy the latter constraint, one, first, has to decide which *kind* of unity self-deception violates and which *kind* of disunity violates it. The accounts presented in the following section will be the deflationary ones: Instead of an intention, different kinds of motivation (mostly desires) will be argued to be causally relevant in inducing self-deception. These accounts do not invoke paradoxes, but leave the question open whether the phenomenon they describe merits the label 'self-deception' and how to distinguish it from other kinds of biases (subpersonally caused influences on the belief-forming process of the self-deceiver). Two more constraints will be introduced now: *need of demarcation criteria* and *congruency of the phenomenological description* of the self-deceiver to the mechanism that is argued to explain it. The demarcation problem consists of three subproblems, as I already mentioned in the introduction to section 1.1.1, namely the usefulness (see section 1.1.3 and 3.2.2), selectivity (see 1.1.1.5, 1.1.2.3, as well as 1.1.3) and content problems (see 1.1.2.5). The demarcation problem is defined as the one constraining the concept of 'self-deception' in the right way. This may be either by restricting the application of the concept of 'self-deception' only to certain kinds of *phenomena* such that this concept possesses scientific usefulness (usefulness problem), restricting the *mechanisms* of self-deception such that conditions when self-deception arises can be clearly stated, e.g. how to explain that self-deception not always arises when there is motivation for it to happen, and restricting the *motivation* of self-deception such that it is characteristic only of self-deception. Though I have categorized the subproblems in this manner, it is not a given that there is necessarily a mechanism characteristic only of self-deception, or that there is a motivation characteristic only of self-deception. In other words, it is not a given that *all* of the subproblems of the demarcation problem have to be solved, but that it is also in question *which* is to be solved, or what the characteristic, unifying cases of self-deception, is. In section 3.2 I will introduce Trivers' evolutionary theory of self-deception according to which a unifying characteristic of self-deception phenomena is its *function*, namely to deceive others. I leave here the demarcation problem, which deflationary accounts have been argued to be susceptible to and which is to be revisited in the summary of constraints in section 1.1.3.

Apart from the demarcation constraint, the second constraint to be introduced is the *congruency to the phenomenological level* which consists in the necessity of congruency between the postulated mechanism of self-deception and the phenomenological description of the phenomenon. The deflationary accounts to be presented have been argued to be susceptible to the *tension problem*, or that they cannot explain the feeling of uneasiness that arises in the presence of contradictory beliefs, because according to deflationary accounts there are not contradictory beliefs present at all (apart from accounts in this section, see 1.2.2 for the tension problem). I categorize this tension problem as a special case of the congruency to the phenomenological level constraint.

Let me introduce the schema of deflationary account explanations. Deflationary explanations cast self-deception in terms of a folk-psychological (personal level) description of the self-deceiver's supposed (conscious) reasoning process such that the evidence and knowledge, available to her, have been skewed by subpersonal biasing processes. In the following, I will present six deflationary accounts: Johnston's, Barnes', Mele's, Noordhof's, Nelkin's and Funkhouser's. Johnston and Barnes agree that it is an *anxious* desire that is characteristic for self-deception. Barnes and Mele agree that *biases* are the mechanisms by which self-deception is accomplished, while for Johnston it is a mental tropism (a nonaccidental causal regularity). For Noordhof, it is a certain kind of *attention*, or absence of attention, that leads to self-deception. Finally, Nelkin and Funkhouser argue that a distinctive characteristic of self-deception is a *desire-to-believe* that something is the case and not the desire that something to be the case, but deviate in whether the resulting self-deceptive attitude is a first- or a second-order belief. Each of these account, thus, solves the demarcation problem (for some it is a certain kind of mechanisms, for others – a certain kind of motivation) and the problem on the adequacy of the phenomenological description in a different way.

The elaborations of these six accounts will serve three purposes. First, deflationary accounts will serve as a comparison to the intentionalist accounts already introduced. Particularly, I want to argue that the tension between intentionalist and deflationary accounts of self-deception can be described as the search for an optimal kind of *personal level recognition* of the faultiness of belief acquisition or maintenance. The argument is a two-step one. For intentional accounts, personal level recognition has been achieved by dual-belief requirement. A weaker form such as dual-rationality has been argued by Michel & Newen (2010). The tension problem has its roots in the contradictory beliefs requirement on self-deception (the requirement states that self-deceivers believe their self-deception, on the one hand, but also that they believe the truth, on the other, see section 2.1.2 and Lynch, 2012 for more on the different uses of 'tension' in self-deception) and 'beliefs' are personal level concepts that are governed by the rules of logic. Thus, the tension debate is actually more about personal level recognition than about the phenomenology.

Second, the discussion of deflationary accounts will aid in creating a third- and first-person level description of the self-deceiver that is to be given in chapter 2. There I will argue that on the phenomenological level, *tension* (feeling of uneasiness) and *insight* characterize self-deception. On the third-person level of description, it is *inconsistency* and *justification* of one's self-deceptive beliefs that round up the categorization.

Third, elaborated accounts will give some background for the discussion of the self-deceptive process, namely one on the role of biases (e.g. section 1.1.2.3) and attention (e.g. section 1.1.2.4). Deflationary philosophical accounts of self-deception have gained popularity in the psychological literature, particularly Mele's account. This is due to the introduction of *biases* as mechanisms by which self-deception is accomplished. Section 1.1.2 will, thus, serve as a foundation for section 3.2.2 where I will discuss Trivers' evolutionary account of self-deception that accepts biases as *proximal* mechanisms of self-deception. Van Leeuwen (2013b) has then criticized Trivers for widening the application of the concept of 'self-deception' to the ubiquity of biases, because this would question the possibility of the scientific investigation of self-deception (*usefulness problem*).

Noordhof and Demos (1.1.2.4) argue for different types of attention as an explanation for self-deception. Further, biases may be argued to serve a certain kind of selectivity, e.g. selective evidence-gathering (see Mele's categorization of self-deceptive biases in section 1.1.2.3). Selectivity is, thus, important for self-deception, but it is unclear which *kind* of selectivity is needed. I distinguish in section 1.1.2.4 between personal and subpersonal selectivity and argue that different kind of control may underlie both. In section 1.1.3 I will resume the selectivity issue by arguing that since self-deception is a cluster-concept, it is probable that there is more than one kind of selectivity characterizing it. In section 1.2 I will argue that self-deceptive selectivity might lead not to a certain kind of an epistemic agent model, but to a changed world- and/or self-model. In section 2.2.1 I will then distinguish more kinds of selectivity, apart from personal/subpersonal.

To sum up, two new constraints have been introduced – demarcation constraint and phenomenological congruency constraint. Six deflationary accounts are to be presented that will focus either on a kind of motivation (desire) or a kind of process as a demarcation criterion for self-deception.

1.1.2.1 Johnston: anxiety-reducing mental tropism account

I have already mentioned Johnston's (1988) account in the review of Pears' view (see section 1.1.1.2). For Johnston, self-deception has an anxiety-reducing function and can be explained by a mechanism called 'mental tropism.' Pears has criticized Johnston for labelling something 'self-deception' that can be caused quasi *automatically* (without the agent's intentions) by this kind of subpersonal mechanism. This disagreement shows the first point of concern for this section – degree of personal level recognition necessary for self-deception. This is because Johnston seems to deny the necessity of any and *demarcates* self-deception by the presence of certain *internal* states. The settlement of the presence of these states is important for a *third- and first-person characterization* of the self-deceiver. Thus, Johnston's account will help to shed more light on two of the three questions I have posed in the introduction to section 1.1.2, namely those on personal level recognition and third- and first-person characterization, as well as demarcation criteria.

Let me tackle the personal level recognition issue on the example of criticism that Johnston voices against other accounts. In short, he criticizes intentionalist accounts as too complex and paradoxical. Against intentionalist theories in general Johnston (1988) points out that those lead to the *paradox of repression*:[33] Intentionalist explanations of the mechanism of self-deception involve either repression or a temporal strategy of forgetting, because "[n]o project or action plan can satisfy the condition of simultaneously including awareness and ignorance of the repressed material" (p. 77; see static and dynamic paradoxes). Johnston (1988) further criticizes what he calls *homuncularist approaches* to self-deception – those postulating distinct subsystems (p. 63) - with respect to the *regress of subsystems*[34] and argues that the over-rationalization of mental processes that are "purposive but not intentional" leads to the paradoxes of self-deception and to the requirement that self-deception should be intentional in the sense of being a "process initiated and directed by an agent because he recognizes that it serves a specific interest of his" (p. 65). In particular, Pears' account is found unsatisfactory,[35] because he postulates a complex protective

[33] Johnston (1988) attributes this paradox to Sartre, issued against Freud (pp. 75-76).

[34] Regress of subsystems consists in the fact that the self-deceiver and the self-deceived needs to be isolated one from each other, in order to be able to influence one another: "Again, how does the deceiving system engage in an extended campaign of deception, employing various stratagems to alter the beliefs of the deceived system, without the deceived system's somehow noticing? If the deceived system somehow notices then the deception cannot succeed without the collusion of the deceived system. However, to speak of the collusion of the deceived system in its own deception simply reintroduces the original problem. The deceived system is now both (partial) agent and patient in the deception. Must we now recognize within the deceived system a deceiving subsystem and a deceived subsystem?" (Johnston, 1988, p. 64)

[35] Johnston summarizes Pear's answers to all the paradoxes that Johnston's postulates as following: "The **surface paradox** is solved by having distinct subsystems play the respective roles of liar and victim of the lie. The **paradox of wishful thinking** is solved by having the protective system altruistically set out to allay the main system's anxiety that not-*p* by inculcating the belief that *p* in the main system. The **paradox of repression** is solved by having

system that is pseudo-altruistic – it is supposed to aid the main system to cope with anxiety, yet produces the effects worse than anxiety,[36] - and whose influence on the main system is indirect (pp. 82 - 85). Davidson's explanation of self-deception by mental causes that are not reasons Johnston's thinks to be unsatisfactory, because he interprets Davidson as implying that it still has to be *rational* chains of causes that lead to a self-deceptive belief[37] (pp. 80-81). As already mentioned, rationality and logic are the rules governing the personal level of description. These arguments lead Johnston to reject the intentionalist view (p. 85) in favor of a mental tropism:

> mental process of self-deception by which anxiety that one's desire that *p* will not be satisfied is reduced by one's acquisition of the belief that *p* (wishful thinking) and one's ceasing to acknowledge one's recognition of the evidence that not-*p* (repression). This process is not mediated by intention; rather, processes of this kind persist because they serve the end of reducing anxiety. Here I speak of a mental tropism, a characteristic pattern of causation between types of mental states [that is not rational] […]. (Johnston, 1988, p. 86)

Summing up this short presentation of Johnston's criticism of intentionalist accounts: an explanation in terms of subpersonal agency or personal agency (rational reason-giving) is rejected in favor of a certain kind of subpersonal mechanism. If a subpersonal mechanism changes the personal level beliefs that one possesses without the need for personal level *justification* for the transition of the belief set before and after, then I think that there is no personal level recognition of self-deception at all. And this is the kind of a mechanism that Johnston proposes. But this also simplifies an explanation of self-deception a lot. Let me now come to demarcation criteria. Self-deception is argued by Johnston to be brought about by a *mental tropism*, defined as a "nonaccidental mental regularity" that arises when an anxious desire that *p* leads to accept the given belief that *p* (p. 66). Self-deceiver is further argued to recognize contradictory evidence (p. 67) and as such be "a *species* of wishful thinking: it is motivated belief in the face of contrary evidence" (p. 67). This characterization follows from Johnston's (1988) distinction between two kinds of wishful thoughts – positive thought and purposive but non-intentional thought (p. 74). The former kind involves a self-fulfilling meta-belief that acquiring a certain belief will more probably than not lead to success, for example in the context of high-performance sport competition (pp. 69-70). Such kind of meta-belief is argued to be absent in self-deception (p. 70). This distinction is interesting, because (also self-fulfilling) unrealistic optimism has been argued to be a kind of self-deception (see 3.1.3.1). Johnston provides an *internal* distinction between the two – a certain meta-belief. Thus, a competitor in a sports competition without the meta-belief will be self-deceiving himself, even if he wins, which is counter-intuitive. An alternative distinction between wishful thinking and self-deception would be an *external* one: something which looked like self-deception might turn out not to be as such. To sum up, one demarcation criterion for self-deception is argued to be the presence of a

the protective system altruistically set out to allay the main system's anxiety that *p* by <u>distracting it from its anxiety-producing belief that not-*p*</u>. The **paradox of irrationality** is solved by modeling self-deception (and wishful thought) on interpersonal testimony" (p. 81; bold and underscore emphasis added).

[36] Here there is an implicit assumption of Johnston, that self-deception's effects are neither helpful, nor healthy (p. 83).

[37] To remind the reader, in section 1.1.1.1 I already asked the question about the connection between (subpersonal) causes and (personal) reasons. I doubt that rationality might be ascribed to the former, because rationality is a property of *agents*. See more on the personal/subpersonal distinction in section 1.3.4.

certain internal states. The self-deceiver per definition has no knowledge of self-deception, when engaging in self-deception, so, the only way to judge something as self-deception is either retrospective, or by *ascribing* self-deception from the third-person point of view. Yet, I do not think that our *ascription* practices of self-deception are best explained by reference to certain mental states of the self-deceiver, as long as the self-deceiver does not confirm to have them. For ascriptions of self-deception, behavior and phenomenology are the only kinds of information available. For reasons of parsimony, demarcation of self-deception has to refrain from the postulation of untestable internal mental states. Alternatively, a way has to be proposed that the self-deceiver possesses exactly these kinds of internal mental states.

What I have left unemphasized, but what nevertheless is a common theme to many accounts of self-deception, is Johnston's claim that is has a defensive function to *reduce anxiety*: "a mental mechanism or tropism by which a desire that *p* and accompanying anxiety that not-*p* set the conditions for the rewarding (because anxiety-reducing) response of coming to believe that *p*" (Johnston, 1988, p. 73). That self-deception is about certain mental contents that threaten the stability of the self-concept or self-esteem is common to many psychological explanations of self-deception (3.1). Even Robert Trivers' proposal that self-deception has an evolutionary function to deceive others still acknowledges that proximal mechanisms by which self-deception is accomplished may be defensive (3.2). Some may see this as remnants of a psychodynamic approach to self-deception (1.3.1), but I think that the main reason for the postulation of a defensive function of self-deception is that it is this function and not the postulation of certain mental states, e.g. intentions or desires, that has so far served as a *demarcation* criterion for what counts as self-deception: self-deception seems to be about contents that are unbearable to the self-deceiver and, which *kinds* of contents these are, has to be determined idiosyncratically. A certain kind of circularity about the role of anxiety in self-deception should also be noted. On the one hand, self-deception can be argued to reduce anxiety, as Johnston does. On the other hand though, anxiety is the *result* of self-deception due to contradictory beliefs or some other kind of inconsistency (I mentioned this in the introduction to section 1.1.2 and referred to Lynch, 2012), as Noordhof argues (1.1.2.4). I will call this the *self-deceptive vicious circle of anxiety*. The reader is encouraged to consider that circular causality may be a property of affectively-laden information processing in general and not self-deception in particular as e.g. "fear is both a cause and a consequence of (predator) perception" on the assumption that affect influences perception (Pezzulo, Barca, & Friston, 2015, p. 38).

Interim conclusion: Johnston postulates a new concept – mental tropism – whose usefulness is questionable, given that it is not explanatory parsimonious to postulate an entity just for an explanation of a phenomenon and such a postulation does not make it intelligible *how* exactly self-deception is accomplished. Absence of personal level recognition or import of any contradictory evidence makes, on the one hand, an explanation easier, while on the other, the doubt remains that one has satisfactorily explained self-deception, because the amount of personal level recognition is the measure of goodness for a theory of self-deception. For parsimony reasons I denied postulation of certain kinds of mental states as demarcation criteria for self-deception. I also introduced the notion of a self-deceptive vicious circle of anxiety that will play a role particularly for the phenomenology congruency constraint of self-deception, or that an explanation of self-deception has to do justice to its phenomenology. It is to be noted that, according to predictive coding, a relatively new theory about perception, action and cognition, what the brain is trying to learn, are causal regularities at different temporal and spatial scales (Clark, 2013a). Barnes develops an account similar to that of Johnston, but substitutes the notion of mental tropism by that of *bias* – a notion well-known in psychology as a deviation from the expected functioning of a certain reasoning or decision process. I will thus explore Barnes' account in the following section, but still want to point out again, as mentioned in the introduction to section 1.1.2, that the use of the notion of bias in self-deception may have created a slippery slope that has led to the over-categorization of everything that contains biases as self-deceptive – Van Leeuwen's accusation of Trivers' account discussed in 3.2.

1.1.2.2 Barnes: anxiety-reducing bias account

In the previous section I discussed Johnston's account that denies the importance of personal level recognition in self-deception and postulates mental tropism or nonaccidental causal regularity to explain it. This may be seen as the weakest kind of a deflationary self-deceptive account, if personal level recognition is taken as a measure of goodness. Annette Barnes (1997) emphasizes personal level recognition. In short, her answer to the disunity in unity problem is that the unity on the personal level of reasons and justification is achieved by means of disunity in the complex inferential relations by which self-deceptive belief is brought about. The main reason for the priority of personal level recognition that I can think of lies in the weight that she gives to the fact that self-deceivers *justify* their self-deceptive beliefs: if they justify them, then some congruent reasoning must be going on the personal level and no solution that would merely explain how by subpersonal means certain personal level beliefs might "pop-out" on the surface of our reasoning will be enough.
I will shortly present her definition of self-deception. Barnes (1997) considers self-deception as a *sui generis* form of deception (p. 142), independent of the deception of others (pp. 125-133), and gives the following individually necessary and jointly sufficient conditions for being self-deceived:[38]

> 1. One has an anxious desire that *q* which causes one to be biased in favor of beliefs that reduce one's anxiety that not-*q*. This bias or partiality operating in one's acting or thinking or judging or perceiving etc. causes ["in the right way"] one to believe that *p*.
> 2. The purpose of one's believing that *p* is to reduce one's anxiety that not-*q*.
> 3. One is not intentionally biased or partial.
> 4. One fails to make a high enough estimate of the causal role that one's anxious desire that *q* plays in one's acquiring the belief that *p*. One believes (wrongly, when condition 1 is met) that one's belief that *p* is justified. (Barnes 1997, p. 117)

First, her reason for choosing a deflationary kind of account is that she takes intentionalist positions to reintroduce a kind of a dynamic paradox. Self-deception according to intentionalist accounts is modeled on intentional interpersonal deception. Intentional deception of others is usually taken to imply that 1) the deceiver knows or truly believes that *p* is false; 2) the deceiver intentionally gets the deceived to believe that *p*. This is not the decisive problem that Barnes (1997) sees though. She argues that intentional interpersonal deception does not need to involve neither the stronger assumption that the deceiver knows or truly believes that *p* is false (where *p* is the belief that the deceiver wants to deceive about) nor the weaker assumption that the deceiver knows or truly believes that

[38] Barnes (1997) argues that such examples as examples of so-called *temporal* intentional self-deception given in the literature are not ones of self-deception, because "[t]he anxiety that causes the belief that reduces a later occurrence of anxiety is not itself reduced by the belief that it causes" (p. 114).

something is false. The dynamic paradox that Barnes sees in modeling self-deception according to interpersonal deception is this (pp. 16-17):

> As self-deceiver I must introduce into my situation something such that I believe there is a real possibility that that something will cause me to believe that *p*, and I take that something to provide *neither adequate evidence for p nor direct sensory awareness that p*. (Barnes 1997, p. 16; my emphasis)

In other words, the problem is a similar one to that already discussed in Davidson's and Johnston's accounts: for the self-deceiver to accomplish self-deception on the personal level one has to use reasons in the belief-forming process that one at the same time does not recognize as valid reasons. More generally, the tension between intentionalist and deflationary accounts of self-deception can be described as the search for an optimal kind of personal level recognition of the faultiness of belief acquisition or maintenance.

Let me now take a look at the indirect inferential relations by means of which she solves the disunity issue. First, which kinds of inferential relations are too direct for self-deception to be possible? Intentionalism is for Barnes too strong, particularly given her definition of intentional action as the one that has a reason of which the individual can be *non-inferentially* aware (p. 88). Self-deceiver's intentions are at best *inferentially* recognizable (p. 93), else self-deception would not succeed given the dynamic paradox. As a consequence, self-deception must be not intentional.[39] Even if intentional kind of self-deception would succeed, according to Barnes, memory of intentions would be present when one is aware of the belief that one acquired via those intentions: this is evident in her critique of Davidson's account as such that "requires that the self-deceiver intentionally try to promote a belief that *p* in himself, and that he be aware of intending to do this when he is aware that he believes that not-*p*" (p. 32). Too strong would be for Barnes also a direct causal connection between beliefs with conflicting contents. This is her critique of Davidson and Johnston: she takes both accounts to be implausible (pp. 34-35). As for Davidson, if *p* - the self-deceptive belief acquired – is the infidelity of the subject's wife, *belief that not-p* would sustain the belief that *p*, thus the belief that the wife is faithful would sustain the belief that she is not. In Johnston's case, the self-deceptive belief would reduce *anxiety that not-p*, thus, anxiety that the wife is faithful. Too weak is a partitioning strategy, if it is understood as something akin forgetting, because there would be no violation of the self-deceiver's own standards of rationality, for which awareness of contradictory evidence is necessary.[40] But if one is aware of such evidence, then, at least on the personal level, one could not acquire a self-deceptive belief.

Second, why does, despite a certain indirectness, the disunity have to be settled on the personal level? This is because self-deceivers justify their beliefs. Self-deceivers are, according to Barnes (1997), *epistemically* irrational, but not *deeply epistemically irrational* (p. 143) and *prudentially* neither rational, nor irrational (pp. 135 - 157). Self-deceivers are

[39] To Talbott's criticism of deflationary accouts that it is not parsimonious to postulate two mechanisms for maximizing expected utility, an intentional and an unintentional, Barnes answers that "there is, I believe, no presumption against a plurality of kinds of mechanisms for maximizing expected utility," especially in the light of the inability of self-deceivers to be non-inferentially aware of their self-deceptive intentions (p. 96).

[40] Rationality as restricted to the domain of the conscious: "We saw that if the self-deception is to succeed, then at some point in the self-deceiving process, the self-deceiver's desire to avoid what RTE [Requirement of Total Evidence] counsels must cause RTE to be walled off; relevant information must no longer be present to the self-deceiver. But if the self-deceiver is not aware of the contrary evidence at the time he comes to believe that *p*, then he does not knowingly violate his own norm of rationality" (Barnes, 1997, p. 30).

not *deeply epistemically irrational*, because they do not have contradictory beliefs or believe "that the totality of his evidence favors not-p," else they would not *justify and defend her self-deceptive belief* (table 6). Barnes' argument against Davidson is that the self-deceiver cannot *recognize* the totality of the evidence, because else the self-deceiver would not be able to always answer the question why he believes something that the evidence does not warrant and give reasons for believing it: "even if a self-deceiver momentarily recognizes that the totality of his evidence favors not-*p*, this recognition is most parsimoniously viewed as a *epiphenomenon*; it is not part of what self-deception essentially involves" (Barnes, 1997, p. 146; my emphasis). Self-deceivers are said to be *epistemically irrational* though, because the self-deceptive belief is "not well founded on reasons the person has" (p. 138), meaning that the reasons are inadequate, insufficient, contextually ill-applied etc. (pp. 138-142). Here we see the comportment of the self-deceiver being brought up as an argument in favor of a weaker kind of irrationality that self-deceivers are guilty of. Barnes further characterizes the self-deceiver as neither *prudentially rational*, nor *prudentially irrational* in the sense of knowingly achieving the preferred outcome (p. 151), because self-deceivers do not acquire the self-deceptive belief *in virtue of the belief that it will relieve their anxiety* (p. 155).

Evidence available to the self-deceiver	Self-deceivers comportment
1. Evidence for not-*p* is **stronger** than for *p*	Give **reasons/grounds** for the belief that *p*
2. Evidence for not-*p* is stronger **to some extent** → It is **inconclusive**	

Table 6. Barnes: inconsistency between assumed evidence and behavior
Distinctions from Barnes (1997, p. 148).

Third, given that self-deceivers justify their beliefs (point two), but also given the problems with direct inferential relations (point one), which indirect inferential relations are good enough for self-deception? Davidson's account being unsatisfactory, Barnes turns her attention to Johnston's anxiety-reducing mental tropism account, which she takes as a basis for developing her own account.[41] She agrees with Johnston that the causal role in triggering the self-deceptive belief belongs to the *anxious desire* and disagrees in this respect with Mele who supposedly argues that it is just a desire (p. 37). She disagrees with Johnston, though, with respect to the *content* of the anxious desire. According to her, it is not the anxious desire that not-*p*, but some other anxious desire that triggers self-deception where the self-deceiver needn't be aware of having this anxious desire[42] (p. 38, footnote 15):

> The self-deceptive belief that *p* can reduce anxiety that not-*q* because the subject believes that if *p* then *q* (or, perhaps, if p then probably q). That is, on what I call the two-variable formula for self-deceptive belief, the self-deceptive belief that *p*, together with the belief that if *p* then (probably) *q*, reduces anxiety that not-*q*. (Barnes, 1997, p. 36)

[41] In this context Barnes (1997) assumes: "While Johnston's account seems intended as a fully general account of self-deception, Davidson acknowledges that he is dealing with one sort of self-deception" (p. 35).

[42] Barnes (1997) makes a distinction between self-deception and wishful thinking on the basis of their phenomenology: "What shows that self-deceptive belief cannot be assimilated to wishful belief is the fact that in wishful belief that *p, the believer cannot have a strong felt desire that not-p and lack a desire, all things considered, that p,* while in possible cases of self-deceptive belief *that p, the believer does have a stronger felt desire that not-p and lacks a desire, all things considered, that p*" (p. 52).

Thus, the change in how Barnes defines the content of the anxious desire gives more space for *indirectness* by considering chains of inferences. Despite this indirectness, a causal connection between desire and self-deceptive belief has to be there according to her. It is one of the criteria for an *anxiety-reduction function*[43] of a self-deceptive belief that it is not only the case that its purpose is to reduce anxiety, but also that there is a causal link between the desire and the self-deceptive belief (p. 59). For the causal link to hold there has to be some *bias* or partiality of belief caused by the anxious desire (p. 65). The notion of bias have become ubiquitious in the self-deception literature. Interestingly, Barnes (1997) argues that even though it is generally possible for self-deceivers to resist biases, it needn't always be the case.[44] The anxious desire can play its biasing role to begin with, because of the underestimation of its role by the self-deceiver.[45] Though in cases of such anxiety reduction desires often are not satisfied, since it would require that *p* is the case, in some cases self-deception might serve not only to reduce anxiety, but also to satisfy the desire associated with the self-deceptive belief (if the self-deceptive belief happens to be true despite the biasing).[46] Summing up the discussion on the personal level recognition in this section: Barnes in difference to Johnston, argues for a certain kind of indirect *personal* level inference relation, because self-deceivers justify their beliefs. Also Nelkin (1.1.2.5) argues that there has to be a causal link between self-deceptive desire and resulting attitude. This is her reason for ascribing responsibility to self-deceivers: Were they aware of the link, they would recognize the insufficient justification for the acquisition and maintenance of their self-deceptive beliefs. My point is not about responsibility though, but about the personal level recognition: as Barnes shows, *direct* inferential relations are too strong, but indirect ones may be constructed. As Nelkin emphasizes, personal level recognition in self-deception is at best a *potential* one, else, as Barnes holds, self-deceivers would be in difficulty justifying their self-deceptive attitudes. But inferential relations that are too weak might, on the contrary, be not enough for *self-deception*, but only for some weaker kind of phenomenon. This is because, since the personal level is governed by rules of logic and

[43] Barnes (1997) argues that it is an open question whether self-deception due to the function of reducing anxiety has been selected evolutionary (p. 38, footnote 13).

[44] Epistemic responsibility should, according to Barnes, not be ascribed to self-deceivers failing to resist the temptation: "While self-deceivers are required, if they are to be epistemically responsible, to try to resist their bias for anxiety-reducing beliefs, some self-deceivers who try to resist their bias nonetheless fail to resist it. And it may not be plausible to regard those who try to resist, but fail, as epistemically responsible" (Barnes, 1997, p. 87). The underlying assumption is probably that temptation is not something that one can always control, e.g. sometimes the desire to eat a chocolate is just beyong me. What you see is that connection of self-deception to weakness of the will that already Davidson pointed to (see section 1.1.1.1).

[45] The assumption is that only then a desire threatens self-deception, when its influence is recognized by the self-deceiver: "What I shall call the lack-or-a-high-enough-estimation condition requires that either the self-deceived underestimate how much his anxious desire that *q* contributes to his believing that *p*, or the self-deceived fails to make any estimate at all about the role his anxious desire that *q* plays in his coming to believe that *p*" (Barnes, 1997, p. 99).

[46] The following is the example where both is given: "Or suppose that a man anxiously desires that *q*, I not leave my wife. He is uncertain whether he will leave because he is uncertain whether *r*, she is unfaithful, having seen a woman looking very much like his wife having intimate conversations with Cesar, a notorious womanizer. If his wife is unfaithful, he will leave her. He deceives himself into believing that the woman he has seen with Cesar is not his wife. By believing that *p*, my wife is not having intimate conversations with Cesar, he believes that his wife is faithful. In fact his wife is faithful; the intimate conversations she had with Cesar were not lovers' conversations. The man is not self-deceived in believing that his wife is faithful, he is self-deceived in believing that she is not having intimate conversations with Cesar. Believing the latter allows him to believe that she is faithful, and this belief allows him to satisfy his desire not to leave his wife" (Barnes, 1997, p. 69).

rationality, disunity is achieved not by compartmentalization, as on intentionalist positions, but by weakening of inferential relation. But weakening of inferential relations also constrains the strength of the (possible) personal level recognition.

Such kind of weakening also has an influence on the explanation on how self-deceptive phenomenology arises. As we have seen, Barnes' account of self-deception is weak with respect to the recognition of contradictory evidence: reasons for the self-deception beliefs are argued to be not well-founded, but this also implies that they are not implausible, the totality of the evidence is not recognized. In order to justify that this phenomenon is one of self-deception, Barnes makes a similar move to that of Talbott: while Talbott argues that there is a phenomenological accompaniment that self-deceptive beliefs could be false, Barnes in a similar vein makes an emphasis on uncertainty and context. Barnes' (1997) argues that using the term *anxious desire* that not-*q* is redundant in the sense that being anxious about not-*q* implies having a desire that *q* and being *uncertain* whether *q* or not-*q*, but nevertheless explanatory useful.[47] The uncertainty could be a result of *suspicion* or some stronger cognitive attitude, yet neither the belief that not-*q*, like Davidson argues, nor the recognition of the impact of the evidence against *q*, like Johnston does (p. 40). Interestingly, according to her, it cannot be determined a priori which belief would be acquired given a certain anxious desire. The *context* is decisive with respect to the question why the anxious desire led to the acquisition of exactly that and not another self-deceptive belief:

> While it is necessary in self-deception that there be an anxiety-reducing belief that the self-deceiver acquires, there is no anxiety-reducing belief such that it is necessary that the self-deceiver acquire it. That the self-deceiver acquires the particular belief that he does depends upon further facts about him and the world. (Barnes, 1997, p. 47)

Apart from personal level recognition and the phenomenology, the last topic to discuss in this section is the one about the *scope* of self-deception: changing the world or one's beliefs. Self-deceptive beliefs are, according to Barnes (1997), not only ill-justified, but may also be *behaviorally inert*. This makes sense if one imagines self-deceiver as somebody who wants to escape from reality and whose aim is, thus, changing one's own beliefs. A response she gives to the selectivity problem is that universal dispositions to protect oneself from harm can override the ones to self-deceive in life-threatening situations, so that world-directed actions can be taken (p. 82). I take the implicit assumption here to be that self-deceptive beliefs do not lead to action directed at the world, but only at changing one's belief system: self-deceivers acquire some false[48] and (often) self-enhancing content (p. 82). World-directed actions though would require that there is some correspondence between what one takes the world to be and what the world is (see section 1.1.2.5 for more on whether self-deception concerns a self- or world-focused desire). As I already mentioned in section 1.1.1.6, there is a new computational theory about perception,

[47] Those are Barnes' conditions for being anxious about something: "When a person is anxious that not-q, the person (1) **is uncertain whether *q* or not not-*q*** and (2) **desire that *q***. So a simpler analysis is correct: one has an anxious desire that q just in case on is anxious that not-q. But the redundant analysis has the advantage of making explicit the desire that enters into anxious desire" (Barnes, 1997, p. 39). Uncertainty is indeed very important, not only to explain self-deception, but also different other kinds of phenomena (e.g. see section 4.4).

[48] There is a difference between the process of self-deception and its end result: "The process of being self-deceived can lead to a true belief, but the state of being self-deceived requires that some false belief has been acquired" (Barnes, 1997, pp. 118-119). Such is the structure of the first chapter of this thesis: first I speak about how different kinds of motivation influence the process, and then (in section 1.2) about the product of self-deception.

cognition and action – predictive coding, which holds that the brain builds a generative model of the environment, such that top-down predictions of causes of our sensations are compared to those suggested by bottom-up prediction errors, in order to either change the model via perceptual[49] inference, if prediction errors are too big, or to look for further evidence for the model via active inference (Seth, 2015b). If predictive coding is applied to self-deception, then the question whether self-deceptive motivation is a self- or world-focused desire can be reformulated as a question whether in self-deception perceptual or active inference is at work. Not surprisingly, self-deceivers not only change their belief-system, but also show avoidance behavior (1.1.2.5), which suggests that both perceptual and active inference play a role in the explanation of self-deception. It is also to be noted that, if self-deceptive beliefs were behaviorally inert, then no self-fulfilling beliefs could be self-deceptive, e.g. there could be no self-deceptive beliefs in sports that could actually help to achieve better results.

Interim conclusion: bias, context, uncertainty, desire satisfaction, the phenomenological level and with it the degree of recognition of the evidence, justification of self-deceptive beliefs by the self-deceiver, irrationality, defensive function of anxiety reduction – these all are elements still often emphasized in contemporary accounts. Potential personal level recognition (of evidence or the causal connection between motivation and self-deceptive attitude) comes at varying degrees and is used as a measure of goodness for an explanation of self-deception: the more personal level recognition a theory can explain, the stronger the kind of self-deception will be explainable. Yet, this idealized case is difficult to achieve, because, as seen on the example of Barnes' account: the more personal level recognition one requires, the weaker kind of self-deception one will get, because personal level is governed by the rules of logic such that sufficiently indirect inferential connections will need to be postulated. Such weakening will also change an explanation for anxiety as one of the phenomenal characteristics of self-deception. In the previous section I already mentioned the vicious phenomenal circle of self-deception: anxiety as both the cause and the result of self-deception. At this point, I want to add that it can also be seen as a *mechanism* by which self-deception is accomplished. Dennis & Halberstadt (2013), for example, have tested that a *belief in the danger of negative emotions* (measured via a questionnaire) selectively biases attention towards neutral and positive facial cues. They hypothesize that the belief, that negative emotions are dangerous, on the one hand, and anxiety, on the other hand, are similar in that both involve a conflict between vigilance, as attention towards the threat, and avoidance, as attention away from it (p. 4, 16). The difference is, according to them, which of the two overpowers the other: in the case of the belief it is avoidance and in the case of anxiety – vigilance.

1.1.2.3 Mele: biased error-minimization account

In the previous two sections I presented Johnston's and Barnes' accounts that focus more on the phenomenology of self-deception (anxiety), but differ in the import of personal level recognition. In this section I will present Mele's account that has gained a lot of popularity, especially in the psychological literature, since he uses popular empirically tested psychological constructs in the explanation (error minimization, biases). The discussion of Mele's theory might be seen as the further elaboration of the question, posed at the end of

[49] So far, no additional kind of cognitive inference has been distinguished in the predictive coding literature, but only perceptual, active, interoceptive and interpersonal inferences (see section 4.2).

the discussion of Talbott's Bayesian account in section 1.1.1.6 about the optimality of generative models of self-deceivers. Mele (2012) gives the following set of conceptually sufficient conditions for self-deception:

> *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 ~*p* than for *p.*
> 5. *S* consciously believes at the time that there is a significant chance 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. (Mele 2012, p. 12)

The *first condition* is according to Mele (2001, 2012) a "purely lexical point" that highlights the difference between deceived in believing and deceived into believing: one can be deceived *into* believing something true (the acquisition process), yet in this case one is not deceived *in* believing that it is true (the fact that it is true). The *second condition* is the one that specifies the way in which the reasoning process of the self-deceived is different from the non-self-deceptive one. Mele's way to explain self-deception is to argue that:

1. Humans reason in a certain way specified by the FTL model (Friedrich-Trope-Liberman hypothesis testing model, see below).
2. Biases can influence this process and the biases of self-deceivers are motivated.

To 2: Mele (2001) differentiates between cold or unmotivated bias and hot or motivated bias (p. 26) where bias is a process that precludes the impartial investigation of data:

> We may have a tendency to believe propositions that we want to be true even when an impartial investigation of readily available data would indicate that they are probably false. A plausible hypothesis about that tendency is that our desiring something to be true sometimes exerts a biasing influence on what we believe. (Mele, 2001, p. 11)

Mele (2001, 2012) argues that desires influence hypothesis-testing by misinterpretation and selection of information and on the assumption that cold biases could be motivated[50] he identifies certain mechanisms of cold biasing by which such misinterpretation and selection can occur. These cold biases are the vividness of information (personal interests enhance the vividness of information), availability heuristic (the formation of beliefs is dependent on the objects and events that are accessible during the process) and confirmation bias (the hypothesis that people test tends to be accepted, rather than rejected, due to search for confirming instances). This list of motivated biases is not exhaustive (Mele, 2001, p. 51) and the biasing strategies can be divided into internal-biasing strategies, as well as input-control strategies (Mele, 2001, p. 60).

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." (p. 26)
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." (p. 26)

[50] The following quotation demonstrates that position that 'hot' bias = 'cold' bias + motivation: "The main point to be made is that, although sources of biased belief can function independently of motivation, they may also be triggered and sustained by motivation in the production of particular *motivationally* biased beliefs" (Mele, 2001, p. 29).

Selective focusing/attending	"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*." (p. 26)
Selective evidence-gathering	"Our desiring that *p* may lead us both to overlook easily obtainable evidence for ~*p* and to find evidence for *p* that is much less accessible." (p. 27)

Table 7. Mele: Ways in which desire influences hypothesis-testing.
Distinctions from Mele (2001, pp. 26-27).

To 1: Mele takes it that the challenge for deflationary views of self-deception which are susceptible to the dynamic puzzle is "to provide an alternative account of the mechanism(s) by which desires lead to motivationally biased beliefs." (Mele, 2001, p. 14) The FTL model of hypothesis-testing, that Mele embraces, is the one that combines Friedrich's and Trope & Liberman's accounts that I discuss in section 4.2. Thus, I will only consider Mele's interpretation of the given theories here. The way people test hypothesis depends on the errors they want to minimize (which is the motivation) and the errors depend on the costs associated with them. These costs determine the threshold on which acceptance/rejection of a hypothesis depends.

> The *acceptance threshold* is the minimum confidence in the truth of a hypothesis that [one] requires before accepting it, rather than continuing to test it, and the *rejection threshold* is the minimum confidence in the untruth of a hypothesis that [one] requires before rejecting it and discontinuing the test. (Trope & Liberman, 1996, p. 34; my emphasis)

The acquisition of confidence requires gathering of information. This can be costly, because it requires effort. Thus, the two thresholds depend on the cost of false acceptance and on the cost of false rejection, but both are relative to the cost of information (Mele, 2001, p. 34). The motivation of the hypothesis-tester is the one to minimize errors and it is a personal kind of motivation. Mele criticizes Friedrich's PEDMIN theory insofar as the latter does not make a distinction between the motivation to be accurate and the motivation to reach a certain conclusion. Mele (2001) disagrees that the motivation to find out the truth is the motivation to reduce errors to which one is indifferent and holds further that those two *different* kinds of motivations – to be accurate and to reduce errors – might be simultaneously at work (pp. 39-40).

The *third condition* of the jointly sufficient ones (that about biasing being a nondeviant cause of the acquisition of self-deceptive belief) reflects according to Mele (2001) "a familiar problem for causal characterizations of phenomena in any sphere" (p. 51): a case that looks like causation, but is not. The *fourth condition* serves to fulfill the requirement of believing against available evidence. Mele (2001) argues though that it is an inference to the best explanation that there need not be two contradictory beliefs in human mind (p. 77). The motivation for self-deception can be sustained not only by an undesirable belief, but also by a belief that there is a significant chance of the undesirable belief[51] (Mele, 2001, p. 71-72). That way, static and dynamic paradoxes of self-deception can be avoided and no intentionality is needed for self-deception. It is also argued not to be necessary to assume tension in self-deception due to the inference to the best explanation (p. 52).

The *fifth* and the *sixth conditions* were newly presented in Mele (2012). The fifth (about believing that there is a significance chance that the self-deceptive belief is false) has been

[51] Interestingly, though Mele explicitly criticizes Gur & Sackeim's conditions for SD (Mele, 2001), Emily Balcetis (2008) holds that both Mele and Gur & Sackeim require knowing and not-knowing the truth at the same time as a condition for SD.

added as a result of the tension-debate: Is it necessary for self-deception that there is a certain tension, a feeling of uneasiness, to be present in self-deception? The sixth has been added as an answer to Scott-Kakures (2002) claiming that cats cannot be self-deceived, but only humans possessing sufficient rational capacities can. Mele (2001) argues that self-deception defined by these criteria cannot be explained by a purely cognitive account, because impartial cognitive peers would fail the *impartial observer test*:

> As self-deception is commonly conceived, if *S* is self-deceived in believing that *p*, and *D* is the collection of relevant data readily available to *S*, then if *D* were made readily available to *S*'s impartial cognitive peers (including merely hypothetical people), those who conclude that *p* is false would significantly outnumber those who conclude that *p* is true. (Mele, 2001, p. 106)

Mele (2001) further argues that the advantages of his account are that it can explain *twisted self-deception*, because his account is unifying to the extent that "in all cases of self-deception, straight and twisted alike, a tendency to minimize errors that are costly, given the person's current motivational profile, plays a central explanatory role" (p. 98).
Twisted self-deception, which has also been acknowledged by Barnes[52] (1997) to be a necessary explanatory target for a satisfactory theory of self-deception, is the one in which the self-deceiver acquires an unwelcome belief:

> in twisted cases of *S'*s being self-deceived in believing that *p*, *S* desires that ~*p* and desires something, *x*, associated with ~*p* but is motivated to believe that *p* because so believing is conducive to *S*'s bringing about *x*. (Mele, 2001, p. 96)

Noncognitive elements such as emotions and other kinds of motivation are argued to play a similar causal role in self-deception (Mele 2000, 2001) and are responsible for self-deceivers failing the impartial observer test. This is important, because though most philosophical accounts emphasize *desires* as motivational elements, emotions and feelings might play an explanatory role that is as important. Mele (2000) denies that anxious feelings are necessary for self-deception (and with it that Barnes' account is distinct from his own), even if Mele (2000, 2001) agrees that self-deceivers are *uncertain* about the matter at hand prior to the acquisition of a self-deceptive belief. If belief-forming processes are accompanied by certain kinds of feelings (see section 2.1.3 for the elaboration of this claim) and if there is no distinction between hot and cold cognition (Duncan & Barrett, 2007, p. 1202) insofar as every thought possesses an affective component (Jackendoff, 2012, p. 220), then Mele would be not justified to ascribe such a minor role to feelings and emotions.
Mele (2001) answers the selectivity problem (why does one self-deceive in one case and not in another, despite the presence of a desire in both) by stating that intentionalist positions have a selectivity problem on their own (p. 66) and by giving the following argument:

> If on one of the two occasions Don has a biasing intention whereas on the other he does not, then, presumably, some difference in the two situations accounts for *this* difference. But if there is a difference, *D*, in the two situations aside from the difference in intention that the argument alleges, an argument is needed for the claim that *D* itself cannot

[52] An example of twisted self-deception is this: "In a familiar enough type of self-deception, the self-deceptive belief seems not to be welcome. A parent deceives himself into believing that she is to blame for her child's death, although the child died of leukemia" (Barnes, 1997, p. 34).

> account for Don's self-deceptively biasing data in one situation and his not so doing in the other. Given that a difference in intention across situations (presence in one vs. absence in the other) requires some additional difference in the situations that would account for this difference, why should we suppose that there is no difference in the situations that can account for Don's biasing data in one and not in the other in a way that does not depend on his intending to bias data in one but not in the other? (Mele, 2001, p. 63)

Certain critique points have been brought against Mele's account that concern the general question of how rational humans actually are and which level – personal or subpersonal – a satisfactory explanation of self-deception requires. Noordhof (2009) argues that Mele's explanation is unable to capture the theoretical irrationality of self-deception, as it implies that biased reasoning belongs to the standards of theoretical reasoning and as such is rational:

> If it [Mele's model] determines how we *ought to form* our beliefs, then its explanation of self-deceptively formed beliefs entails that they are *rational*. When a self-deceived subject's beliefs depart from what we suppose might be supported by the evidence, this just demonstrates that they have different confidence levels and, with regard to those confidence levels, form the beliefs they ought to form. It is one thing to allow that self-deception may, on occasion, be beneficial. It is quite another to explain it in such a way that, by the agent´s lights, it not only is always beneficial but also reflects the appropriate influence of *standards of theoretical reasoning.* (Noordhof, 2009, p. 52; my emphasis)

Noordhof's (2009) reasoning goes as follows: If Mele's model reflects how we *ought to* reason, then it is rational, if it reflects how we *do* reason, in spite of the evidentialist reasoning taken as a standard, then the puzzle remains to explain why we do not reason how we should (p. 52). I think Nordhoof oversees here a distinction between the *epistemic norms* of belief formation and *procedural norms* of belief formation:

> it is implicitly assumed that the procedural norms of belief formation and the epistemic norms of belief formation must converge – that is, an account of epistemic norms will double as an account of the procedural norms of belief formation. On this view, a human being who fails to believe only so far as the evidence allows will also manifest abnormalities of belief formation. This picture should be resisted. (Bayne & Fernández, 2009, p. 5)

In the light of this heavy emphasis on the rationality in explanations of self-deception it might be useful to mention that apart from the idealized view of rationality as unbounded and human reasoning as consisting of heuristics and biases there are some other alternatives (see table 8). Van Leeuwen, for example, argues for practical rationality defined similar to Gigerenzer's (2006) second option – rationality optimized under constraints (see section 3.2.3.1). It may be the case that for Trivers' evolutionary explanation of self-deception ecological rationality may fit best (3.2.2). My point is that it is questionable how and to which degree humans are rational and arguments against theories of self-deception based on certain assumption about human rationality should be treated with caution.

Unbounded rationality	*Optimal* strategy that maximizes a given criterion under assumptions of *omniscience* and *unlimited computational resources.*
Optimized under constraints	*Optimization* without *omniscience* under constraints of limited mental or environmental resources; optimal cost-benefit stopping point.

Heuristics and biases	The reference point for the evaluation of heuristics is still the norm of logic and probability.
Ecological rationality	The reference point is the success in solving problems posed in natural environment. Problems are solved by means of the *adaptive toolbox* which consists of *fast* (time-limitations) *and frugal* (limitations on available information) *heuristics.* Two factors influence those heuristics: *embodiment* (they are "anchored in the evolved brain," p. 120) and *situatedness* (dependence on environment).

Table 8. Gigerenzer: kinds of view on human rationality
Distinctions from Gigerenzer (2006).

Michel & Newen (2010) criticize Mele's account for being unsatisfying, because it cannot explain the counter-evidence requirement of self-deception. Data and evidence should be distinguished according to them (p. 738). Whereas data are by themselves no evidence, its *evaluation* gives them their evidential quality. Michel & Newen argue that Mele's biasing processes (or at least the input-control part of them) happen on the level of data evaluation, which precludes data to become evidence and thus precludes the self-deceived subject's awareness of counter-evidence:

> Motivationally biased perception of data, as characterized by Mele, is situated on the level of data-evaluation. **S** may collect all the relevant data, but non-intentional mechanisms of biased evaluation of data will highlight p-supportive data and feed them in **S**'s belief-formation process while data in support of not-p are left aside and downplayed. Therefore, S can access every single datum, while misperceiving the evidential tendency of all information. As a consequence, the evidential basis on which **S** establishes p-confidence is already the *result* of the biased treatment of data. When being biased, the world as **S** accesses it is a result of biased treatment of data, i.e. **S** fails to access the actual evidential value of data due to sub-intentional manipulation. (Michel & Newen, 2010, p. 738)

Michel & Newen's (2010) solution is that for matter of subjective importance, dual standards of rationality apply: if self-deceivers want to preserve the applicability of a certain trait to them, they change the trait description according to the available evidence. They claim, though, that "[d]ual rationality will typically be observable in *'automatic' or unreflected adaptation* and this is one reason for criticizing **S**'s activity from a rational point of view" (p. 741; my cursive emphasis). Whether self-deceivers' irrationality is *first-person* one, or only the one ascribed from the third-person perspective, thus the one we as observers would ascribe to the self-deceiver, is subject to debate:

> Additionally, and perhaps most importantly, in most cases of self-deception we will in fact find that there is sufficient evidence to support the belief that p, from the self-deceiver's perspective, if only we look more carefully. (Patten, 2003, p. 245)

I think that Michel & Newen's point concerns the distinction between causes and reasons (see section 1.1.1.1 for this distinction): Mele's explanation in terms of biases is one of causes, but Michel & Newen think that an explanation of self-deception should be given by means of reasons that self-deceivers weigh in a personal level belief forming process. Nordhoof (2009), similar to Michel & Newen, criticizes cold biasing, which Mele takes as the basis for defining the motivational bias, as exercising brute causality:

> One caveat regarding the appeal to the mechanisms involved in *cold biasing* is that they don't, in fact, identify an alternative explanatory scheme in which desire may figure to supplant the appeal to an agency explanation […]. While it is extremely doubtful that an

> agency explanation should be given for why certain information is vivid to us, we have nothing in its place except *brute causality* between various mental elements. At best, we just have more details as to how brute causality may operate. (Noordhof, 2009, p. 51; my emphasis)

Last, one should note that though Mele's characterization of self-deceptive process is non-intentional biasing, Talbott (1997) for example, argues that the mechanism responsible for the emergence of self-deception is *intentional* biasing, where an action is intentional if "it is based on agents' beliefs and desires in a way that is responsive to evidence and reasoning, including reasoning about how to achieve what they desire (none of which need be conscious)" (p. 127). The author's way to combine the two is as follows:

> The resourcefulness and ingenuity of self-deceivers in avoiding the belief that *r,* that is, in avoiding coming to the conclusion that their belief that *p* is caused by a desire to believe it regardless of whether it is true, is what most inclines me to think that there are genuine strategies of self-deception. These are not strategies in the sense of consciously premeditated plans, but in the weaker sense of intentional biasing of one's cognitive processes, both internal biasing and input-control biasing (biasing in favor of *p* and in favor of not-*r*). (Talbott, 1997, p. 127; my underscore emphasis)

Though I agree that the reason for arguing in favor of the presence of intention ("resourcefulness and ingenuity") is to be accounted for in an explanation of self-deception, positing the presence of intentional biases – a notion itself in need of an explanation – by no means solves the issue. Further, Talbott's (1995) elucidation of the notion that in the case of self-deception the mechanism that accomplishes the biasing is *decoupled* from self-deception (p. 35), makes the insistence that intention is at work in self-deception even more mysterious.

Interim conclusion: Mele's account has gained popularity in the self-deception literature, combining biases with error minimization in the explanation of self-deception, critique coming mainly from it not doing justice to the phenomenology of the self-deceiver (either by ignoring tension or recognition of evidence) or the irrationality of self-deception. I will talk about *personal* level error-minimization in section 4.2, where I will discuss the two theories (Friedrich's and Trope & Liberman's) that Mele endorses. Here I want to come to another point – that of rationality and optimality – that I first talked about in section 1.1.1.6 when discussing Talbott's Bayesian account. As has become clear in the discussion of this chapter, a kind of theory about self-deception, that one endorses, depends, not least, on the answer to the questions whether self-deception is a result of optimal kind of processing or a breakdown, which kind of optimality it could be and which implications for a theory of the workings of the mental in general this would have. Bowers & Davis, 2012, (p. 398) offer two kinds of ways to reconcile suboptimal behavior with Bayesian optimality.

1. Behavior is optimal only if it can be deduced from rational analysis, which means that the environment and the task to be solved, "combined with Bayesian probability theory, determine what an optimal solution should look like" (p. 398)
2. Behavior is assumed to be optimal, such that assumptions about priors and likelihoods have to be made to fit to the exhibited behavior.

Applied to cases of psychopathology like self-deception, the question is whether self-deception stems from an optimal model and if yes, how it could lead to suboptimal behavior. Schwartenbeck et al. (2015) argue that in the case of predictive coding explanations of psychopathology, *inference* is optimal, but the model is not, because inference cannot be defined independently from the specific parameters of the model, e.g. aberrant prior beliefs (p. 110). This solution resembles the second option, because specific parameters of the model are being looked for, that explain behavior that *looks* suboptimal, but is not, if the generative model is taken into account:

> Put differently, Bayesian inference based on individual models of the world will look suboptimal in a manner proportional to the differences between these models and our beliefs about the true structure of the world. (Schwartenbeck et al., 2015, p. 110)

But even if subpersonal inference is Bayes optimal, how does it relate to personal level biases that self-deceivers might exhibit as Mele argues? For example, if different kinds of such personal level biases need to be explained by different kinds of suboptimal models, what does it mean for the concept of self-deception: can self-deception be explained by only *one* kind of model? I will come back to this question in chapter 4. This adds another demarcation possibility to self-deception: by the kind of a model, instead of a kind of motivation, process, internal states, or certain types of phenomena that fall under this umbrella term that the concept of 'self-deception' obviously is.

1.1.2.4 Noordhof: instability account

> Fifth, and finally, it is often noted that consciousness has something to do with self-deception, although others have located its importance in different places.
> (Noordhof, 2009, p. 67)

In the previous section in Mele's account, the focus was more on the mechanisms by which self-deception is accomplished (biases) and not the personal level recognition and phenomenology that might result from it. For this lack of emphasis, Noordhof criticizes Mele and presents a similar account, but wants to give a personal level account of self-deception and for this introduces a questionable distinction between different kinds of consciousness. Such an account solves the disunity in unity problem by actually dividing the personal level. This is because the personal/subpersonal and conscious/unconscious distinctions are actually tightly connected. In case of attitudes at least, conscious attitudes would also be those that are personal-level ones, even if this implication is not always true the other way around. In the light of Noordhof's solution of the enigma of self-deception, the main topic of discussion in this section will be the question how *attention* and, more generally, *selection*, contributes to self-deception.

Noordhof (2009) holds that his account is different from Mele's with respect to the "centrality of motivational factors and instability," because for Noordhof motivational factors alone do not determine the presence of self-deception, but it is *instability* which is necessary (p. 70). Instability is understood as anxiety present in the self-deceiver. Noordhof (2009) differentiates between the phenomenal impact in consciousness and *attentive consciousness* (p. 62). He adds that attentive consciousness can reach different levels, because one can attend in a brief or sloppy manner (p. 63). Thus, for him self-deception takes place when, on the one hand, contradictory evidence or its motivation are not processed with attentive consciousness, while, on the other hand, the product of self-deception is consciously attended to. Self-deception is undermined when contradictory evidence or motivation for self-deception becomes attentively conscious. Moreover, he holds that biased reasoning can be controlled, if brought to attentive consciousness (p. 63-64). In this he distances himself from Mele, for whom, according to Noordhof (2009), cold biasing, even in absence of motivation, would be present in the impartial peers of the self-deceiver (p. 63).

> S is self-deceived that p if and only if S cognitively endorses that p and[53]
> (a) The subject, S, fails to *attend consciously* in a certain way, W, to either the evidence that rationality clashes with p, which she believes, or some element of the psychological history characteristic of the self-deception behind the cognitive endorsement that p.
> (b) If the subject were to attend consciously in way W to both p and either the evidence that rationality clashes with it or the psychological history (whichever applied from clause (a)), the subject would no longer cognitively endorse p.
> (c) (a) holds because of S's motivational state or emotional state that p or cognitive schema that favors the cognitive endorsement that p. (Noordhof, 2009, p. 64; my emphasis)

The benefits of his account, according to Noordhof (2009), are firstly that it captures the *irrationality* of self-deception. If attentive consciousness to the biasing mechanisms undermines self-deception, then this is the case because the agent does not live up to his own ideals of reasoning (p. 65). Secondly, this account does not require *intention* as a necessary component (p. 65). Noordhof illustrates this point with reference to Amélie Rorty's case of Dr. Androvna, whose behavior contradicts her statement that she does not have cancer. According to Mele, the contradictory behavior can be explained through recognition by the self-deceived of the fact that there is a *significant chance* of cancer. According to Noordhof (2009), the *absence in attentive consciousness* of the significant chance of cancer is crucial (p. 66):

> For some reason – perhaps she is in a heightened state of anxiety – the belief that there is a significant chance that she has cancer threatens to result in the belief that she does have cancer […]. To avoid this, Dr. Androvna is not *attentively conscious* to the mental history that gives rise to the belief that she does not have cancer and/or not attentively conscious to her belief that there is a significant chance that she has cancer. (Noordhof, 2009, p. 66; my emphasis)

On the one hand, Noordhof (2009) criticizes deflationary accounts which try to explain self-deception by "brute causality" (p. 51), but on the other hand, he denies the necessity of intention and simultaneously emphasizes attentive consciousness (p. 65-66). Therefore, I think that some elaboration of the notion of non-intentional guidance of attentive consciousness should be carried out:

> It might be thought misguided to appeal to attentive consciousness to explain some mental phenomenon when it is so clearly in need of further explanation itself. I do not share this rather strict view. (Noordhof, 2003, p. 97)

Third, according to Noordhof (2009), his account solves the problem of the *ascription of beliefs*. He states that functional roles of belief are always relative to circumstances. Thus, he omits the need to establish the importance of different functional roles of belief in general through tying down the importance-ranking to concrete circumstances of a given case (p. 67). On the one hand, this solution allows for the ascription of contradictory beliefs *p* and not-*p*, while, on the other hand, his description of belief ascription seems to be susceptible to Noordhof´s critique about "brute causality," because it is so vague:

[53] Noordhof (2003) has already enumerated these conditions, similarly formulated, in his prior article on self-deception, noting: "The conditions cover cases in which a subject, who believes that p, systematically avoids situations in which he or she might obtain evidence against p, because of a desire to believe that p. I think, *intuitively*, we would count these as cases of self-deception" (p. 88; my emphasis).

> If we are convinced by the reasons, then we come to believe that p. It doesn't follow from this that we don't also have the belief that not-p. In being convinced by p**,** *there may be ways* in which it doesn't sink in so that we still believe that not-p. The point is simply that while instability can explain how self-deception need not require that we have both beliefs, at the same time, the means by which we can explain this reveals that sometimes the self-deceived will have both beliefs. (Noordhof, 2009, p. 67; my emphasis)

In comparison, I would like to review the account of Raphael Demos (1960), which is said to have set the stage for the debate about self-deception (Clegg & Moissinac, 2005, p. 97). After all the accounts presented, the reader is now sensitized for the question, what *advances* haven been made in the explanation of self-deception over the decades. According to Demos, self-deception is characterized by

- *an inner conflict*: Interestingly, Demos (1960) assumes that the necessity of the inner conflict follows from the dual belief requirement[54] and that the absence of the inner conflict is the difference between delusion and self-deception (p. 590). All accounts presented above assume either a contradiction or an inconsistency present in the mental states of the self-deceiver: the explanatory solutions range from the dual belief requirement (e.g., Davidson, Pears) to the belief that *p* and the belief that there is a significant chance that not *p* (Mele).
- *being unpleasant.* The feeling of uneasiness and anxiety is still an explanatory issue (see above).
- *being driven by impulse or passion:* "A man wishes (or is inclined) to believe p, contrary to what he knows to be the case." (Demos, 1960, p. 589). This is a characteristic that intentionalist theories hold to be important (see section 1.1.1).
- *responsibility*, because the self-deceiver "knew what he was doing, and he could have done otherwise" (p. 589). Remember that especially Nelkin (2002) emphasizes that there has to be a potential knowledge of the causal connection between motivation and self-deceptive belief in order to explain responsibility of the self-deceiver.
- *"I knew it all along"* or the description of the mental state after the self-deception is uncovered that one has known that one were self-deceiving oneself.

Demos (1960) argues that two possible explanations for self-deception, temporary one (believing *p* at one time and not-*p* at the other) and the divisionist one (believing *p* consciously and not-*p* unconsciously), are not applicable to self-deception, because of the 'nagging doubt': "Here both the belief and the doubt are simultaneous and both seem to be in the conscious mind" (p. 591). Demos' (1960) solution is thus postulating two levels of awareness,[55] "one is simple awareness, the other awareness together with attending, or noticing" (p. 593). Here, the similarity is to be emphasized between Demos' (1960) and Noordhof's (2009) accounts in the characteristics of instability and the distinction between

[54] It is interesting that Demos formulates the dual belief requirement in terms of the *experience* of a contradiction: "In short, self-deception entails that B believes both p and not-p at the same time. Thus self-deception involves an inner conflict, perhaps the existence of a contradiction" (Demos, 1960, p. 588). Obviously an experience of a contradiction cannot be explained by rational means, because the rationality assumption implies that there is no such experience at all.

[55] In the previous footnote I stated that according to Demos there may be something like the experience of the contradiction, but his analysis fails to do justice to this possibility, because he insists to preserve the law of contradiction for descriptions of the means by which one could self-deceive: "My own analysis of self-deception follows a similar line. As with *akrasia*, there is an impulse favoring one belief at the expense of its contradictory; and the person who lies to himself, because of yielding to impulse, fails to notice or ignores what he knows to be the case. Such an analysis 'saves' the phenomena while at the same time conforming to the requirements of the law of contradiction" (Demos, 1960, p. 594).

different kinds of consciousness – one attentive and one not. The new characteristic not mentioned in the accounts before is the "I knew it all along" effect.

Summing up, the main characteristic of Noordhof's and Demos' account is that the disunity of self-deception is identified at the personal level. I think that there are three main reasons for this. First, one can, similarly to Barnes' indirect inferential relations solution, preserve a certain kind of personal level recognition in this way, namely *potential* recognition: Contradictory evidence is *available*, but not focused on. Second, this is a way to explain the anxious phenomenology of self-deception: Evidence is available and, thus, creates tension simply in virtue of the rules of logic by which the personal level is governed. The problem regarding the first two points (recognition and phenomenology) is that if all needed to recognize self-deception is, figurally speaking, so near to the surface, then why does the self-deceiver not recognize her own self-deception in situations in which she has to *justify* her self-deception in front of others? This indicates that the offered solution needs to be performed by some other mechanism explaining the *maintenance* of self-deception.

Let me now discuss the question about the role of attention and selectivity in self-deception. The third reason for the proposition of such an attentive mechanism is that cognitive agency in self-deception seems to be preserved this way. This is the case on the assumption that this kind of self-deceptive attention involves cognitive agency. Since self-deception cannot be result of *exogenous* attention, which is evoked by the stimuli themselves, attention that Noordhof and Demos have in mind must be some kind of *endogenous* attention, or the one initiated by the agent (for the exogenous/endogenous distinction see Hohwy, 2012). Thus, endogenous attention per definition seems to involve cognitive agency. Cognitive agency is characterized by high-level cognitive *self-control* (Metzinger, 2013a). What personal level accounts of self-deception in general, and such attentive accounts in particular, try to preserve, according to me, is control over one's higher order cognitive processes. Though whether such a control is present, or whether only the *feeling* of control is present, or neither of them at all, is not clear. Feelings of control would explain ascriptions of responsibility from the first-person point of view. Control would explain ascriptions of responsibility from the third-person point of view. In the discussion of personal level recognition in Barnes' section (1.1.2.2) I mentioned that on Nelkin's view (1.1.2.5) the potential to recognize the causal connection between motivation and self-deception serves to explain ascriptions of responsibility in self-deception. The effect of Barnes' indirect inferential relations could be seen in the same light as Noordhof's and Demos' self-deceptive attentive mechanism. The weakness of such kinds of accounts is to explain how the self-deceivers justifies his beliefs, because during such a justification process the inferential relations, or also attention, would be drawn by others to our own weaknesses in the belief-forming process. Due to the rules of logic and rationality governing the personal level (Bermúdez, 2000a), if the self-deceiver would know that he has controlled the acquisition and maintenance of a self-deceptive attitude, he would abolish it in virtue of being the rational being he is. Since this does not happen, the self-deceiver does not have this feeling of control. Control over what actually? At this point the matter gets complicated. Self-deceivers do not seem to report absence of control over their reasoning processes and justify their attitudes. What they must be missing is the feeling of control over the *selectivity* they exhibited. Attention is selection. Biases also lead to a certain selection. Generally, I think that it is a viable thesis to defend that every kind of a self-deceptive mechanism to be proposed can be described as a specific selective process. The only way to preserve the feeling of control over the selectivity would be to justify that selectivity on rational grounds. But how? For Michel & Newen (2010) self-deceivers possess dual-rationality standards. But then again, this could be pointed out to them by observers. A simpler alternative that is based on the assumption that self-deceivers also do not seem to report a *missing* feeling of selectivity is that selectivity itself is a subpersonal process, e.g. *subpersonal* biases. This is how I see Noordhof's criticism of Mele about brute causality: Noordhof interprets Mele to imply that the selectivity is subpersonal, which for him is not enough. Hence, Noordhof himself postulates agentive selectivity – attention. That self-deception involves a certain kind of selectivity is a given. Which kind of selectivity it is, is in question. And even whether only one kind of selectivity is involved, is also not clear. The latter two questions depend on the answer to the demarcation constraint, namely whether we distinguish self-deception by a certain kind of process, or some other characteristic. I will propose my take on the matter in sections 1.1.3 and 2.1.2.

Since I have already started talking about responsibility, as well as selection and control, I want to briefly answer the question whether I hold self-deceivers to be responsible. Responsibility is for me a matter of control. In the light of the distinction between personal (involving agentive control) and subpersonal selection, as well as control and feeling of control, it can be said that self-deceivers would hold themselves responsible, if they had the feeling of control during self-deception. Observers would hold self-deceivers responsible, if the actual agentive control was there. A subpersonal selection mechanism would preclude such kind of control. Thus, only certain types of explanations of self-deception, namely the personal level selection ones, would be able to incorporate responsibility into the theory of self-deception. Moreover, on another level, it is not only that self-deceivers might be responsible or not, but that they might escape responsibility *via* self-deception, or as Tenbrunsel & Messick (2004) put it, "[s]elf-deception acts as an ethical bleach, removing the ethical colors from the decision" (p. 233).

Interim conclusion: Noordhof (2009) claims that his account can explain the difference between delusion and self-deception through *instability* present in self-deception, while also admitting that his appeal rests on an intuition (p. 71). The instability that Noordhof (2009) speaks about is nothing else than a variation on the tension requirement: "In being convinced that p, there may be ways in which it doesn't sink in so that we still believe that not-p" (p. 67). Noordhof focuses on a special notion – that of *attentive consciousness* – in explaining self-deception. His account is in this respect also similar to that of Demos' (1960) who makes a distinction between simple and attentive awareness. The characteristics that he holds to be important in self-deception are the instability, indeterminacy of belief ascription and irrationality. I argued that what underlies the question how attention contributes to self-deception is which kind of selectivity one holds to underly it. The answer depends on the personal level recognition and demarcation issues and will be pursued in section 1.1.3. The following section will be concerned with the desire-to-belive view that demarcates self-deception by the presence of a certain kind of motivation.

1.1.2.5 Nelkin & Funkhouser: desire-to-believe account

> It is the nature of deception to aim at changing someone's mind, not the world.
> (Funkhouser, 2005, p. 298)

In the previous section I argued that the question whether or how much personal level recognition is present in self-deception depends on the demarcation criteria one chooses – whether it is a mechanism, function, or a certain kind of internal state present in the self-deceiver. Johnston (1.1.2.1) argued that a certain kind of meta-belief might distinguish self-deception from wishful thinking. I was skeptical about this, since the only available characteristics of self-deceivers are their behavior and phenomenology, from which the presence of such internal states would have to be deduced. In this section I will present two more accounts that defend the presence of certain internal states as a demarcation criterion for self-deception, namely the desire-to-believe as a motivation for self-deception. The result of the discussion of this section will be that, since the behavior of the self-deceiver serves as a basis for such demarcation, on the one hand, but is open to interpretation, on the other, I deny that the presence of certain internal states should be taken as a demarcation criterion for labelling something 'self-deception.'

The deflationary accounts presented so far have argued that certain *first-order desires* are motivational elements of self-deception. Dana Nelkin (2002, 2012) and Eric Funkhouser (2005) argue for another view, namely that the motivation characteristic for self-deception is the *desire-to-believe*. Funkhouser (2005) even argues that the product of self-deception is not a first-order, but a *false higher-order belief*: "self-deceivers (falsely) believe that they believe as they desire" (p. 305). The main reasons for this is that self-deceivers are argued to not wish to change the world, but only their own mental states, since they are behaviorally inert in certain situations, e.g. when self-deceptively denying cancer. It should also be noticed that such a view might follow, if one understands self-deception to have a *defensive* function. Defense implies that there is a boundary between what is to be defended, in this case the self-concept, and the aggressor – in this case the world and the information that it offers. The discussion of Nelkin and Funkhouser's accounts will shed more light on the understanding of the relationship between the self-deceiver and the world.

Nelkin (2012) argues that among the intentionalist and deflationary accounts, the latter one is to be preferred[56] and among the two possible kinds of motivationist (deflationary) accounts which are the desire-to-believe[57] account and the desire-account (Mele's) it is the former that is to favor. One of the reasons is that a desire-to-believe can explain why the following case that *seems* to be a case of self-deception is *not* one of self-deception:

> Or consider the case of Otis, who has a desire to have an opinion about everything that he is asked about. On a particular occasion he is asked whether he believes that the 1991 Braves would have prevailed over the 1999 Yankees in a seven-game series. Before being asked, Otis had not considered the matter, and *does not care in the least about what the correct answer is*. But now his *desire to form an opinion* causes him to give undue weight to a particular set of statistics concerning Braves' pitching that come first into his mind. As a result, he *immediately forms* the opinion that the Braves would have won, even though further reflection would have shown him that he had ample justification for concluding that the Yankees would have won. Although Otis's case is a clear case of irrational belief acquisition, it seems that it is not a case of self-deception. (Nelkin, 2002, p. 385; my emphasis)

In particular, Nelkin (2002) criticizes Mele's account because he postulates a content-unrestricted desire to motivate self-deception where the *content-unrestricted* desire is defined as "a desire that causes one to treat evidence in a way inconsistent with its actual

[56] Nelkin (2002) argues that intentionalist theories face a dilemma: "Either intentionalist models of self-deception lead to paradox, or they lead to the postulation of unnecessary theoretical entities, or both" (p. 386). Deflationary theories suffer from being too inclusive and "leaving out what is characteristic of self-deception" (p. 387).

[57] Nelkin (2002) differentiates "cold" from "hot" biases by stating that the former are "error in inferential reasoning due to inattention or from the use of innate heuristic "short-cuts" in reasoning" (p. 385).

evidential value" (p. 389). According to her, non-self-deceptive "cases of hot irrational belief formation" (p. 391) could satisfy an account, such as the case of Otis quoted above. One of her reasons for drawing the distinction between self-deception and other motivated irrational beliefs lies in that "[t]he category of motivated and biased belief seems, *intuitively*, larger than that of self-deception" (Nelkin, 2012, p. 122; my emphasis). The other reason is that if a jealous husband acquires a false belief that his wife is guilty of cheating by being motivated by a desire to maintain close relationships, then it is not a case of self-deception, because there is (presumably) no knowledge of *causal connection* between the two (Nelkin, 2002, p. 392). If self-deceptive desire is a desire that *p*, which she calls Mele's restricted desire account, and not a desire to bias the evidence, then this account can no more incorporate cases of twisted self-deception. She calls it the *content dilemma*:

> Either the motivation condition is so inclusive that it cannot form part of a set of sufficient conditions for self-deception, or it is so exclusive that it cannot capture even central cases of self-deception. Call this the "content dilemma." (Nelkin, 2002, p. 393)

Nelkin's (2002) solution to the content dilemma is the restriction of the content of the desire in another way – the desire-to-believe and holding that function of self-deception is *desire satisfaction* (p. 396). Although the desire that *p* can also contribute to self-deception, it is the desire to believe which is according to her essential (p. 396) The desire-to-believe account is based on the assumption that knowledge of the causal connection between desire and self-deceptive belief should be (potentially) available to the self-deceiver, but leaves it open whether one is aware of the given desire itself:[58]

> On the Desire to Believe Account, not only must the agent see something desirable in believing that *p*, but the self-deceiver must, by *mere reflection* on his or her mental states, be able to see that believing that *p* will result in the particular gain of desire satisfaction. (Nelkin, 2002, p. 404, footnote 35; my emphasis)

Nelkin (2012) holds that (potential) knowledge explains why we ascribe responsibility to self-deceivers on the assumption that the criterion for responsibility is that one "should have known"[59] (p. 127). Please note that it is a requirement that stems from an *internalistic* conception of self-knowledge that the connection between evidence and justificatory object of evidence has to be grasped by the agent (Boghossian, 2000, p. 483). I argue that neither Nelkin's hypothetical example, nor her empirical argument for the desire-to-believe account is convincing:

1. Her case of Otis is supposed to be one in which there is motivated irrationality, but no self-deception, because there is no desire-to-believe. The question is whether this is the only difference between the case of Otis and cases of self-deception. If there are other differences

[58] The main idea, as in previous accounts that I have explored, of how self-deception is possible is that, whatever the mechanism be, as long as the agent does not recognize how the mechanism has affected her, she can self-deceive: "for the desire to succeed in causing a biased treatment of the data, it is not necessary that the agent be unaware of it. (However, it might be psychologically necessary for the agent to be unaware of its *role in the process* in order for the biasing process to occur.) Thus, the Desire to Believe Account leaves it open whether a self-deceived agent can be aware of possessing the operative desire." (Nelkin 2002, p. 395)

[59] Recognition of the motivating role of the desire is a step towards responsibility ascription: "Where 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." (Nelkin, 2012, p. 128)

available, then Nelkin cannot use the case of Otis in her argumentation in favor of the desire-to-believe account. As a matter of fact, there are other options available for why Otis' case could not be a case of self-deception, apart from the desire-to-believe:

   a. The matter about which he is wrong is *unimportant* for him. If pointed at his calculation mistake, Otis would probably admit his mistake and relinquish his false belief, because he does not care about the answer.
   b. There is no contradictory evidence or no weighting of the evidence during the hypothesis-testing (Nelkin speaks about him *immediately* forming the belief).

2. Nelkin argues that her desire-to-believe account is compatible with the empirical evidence on the assumption that a desire-to-believe plays the *same causal role* as any other desire in biasing the threshold. She argues that her account fits with the experimental data, because it fits those mentioned by Mele: "we can easily understand how it [desire to believe] could cause us to set differential thresholds" (Nelkin, 2002, p. 398).
   a. Using Ockham's razor, why then postulate a desire-to-believe?
   b. If the mechanisms of self-deception are those already specified by Mele (p. 395), how can one empirically distinguish a desire account from a desire-to-believe account?
   c. Nelkin's chief assumption for her theory that there should be potential knowledge of the connection between the desire and the acquired belief for somebody to be self-deceived is itself a requirement in need of justification.
   d. Even if ascription of responsibility is facilitated by Nelkin's account, I do not see how this constraint can justify the choice of mental entities possessed by the self-deceiver.

Interim conclusion: Nelkin argues that the motivation for self-deception is a desire-to-believe and that the product of self-deception is according to her the false first-order belief that *p*.[60] She thinks that such an account solve the selectivity problem and explains the fragile nature of self-deception, namely via the causal connection between the motivation (desire-to-believe) and the self-deceptive belief. However, there is a recent account (Lynch, 2013) that, on the one hand, is in agreement with Nelkin's narrative that the content of the self-deceptive desire has to be "essentially or thematically" (p. 1342) connected to the self-deceptively acquired belief, else self-deception would be indistinguishable from stubborn belief;[61] but on the other hand, prefers Mele's account. Whether the postulation of a desire-to-believe is necessary (and explanatory parsimonious) to explain the causal connection between motivation and self-deceptive belief is in question.

Funkhouser (2005) argues that the motivation is a desire-to-believe, but that the product is a false higher-order belief.[62] Thus, self-deceivers, contra Nelkin, do not get what they desire and the function of self-deception is not desire-satisfaction (p. 299). According to Funkhouser (2005), there are two types of desires – world-focused desires (a desire that *p*) and self-focused-desires (a desire to believe that *p*). Self-deception is guided by the latter

[60] Nelkin's view in short is: "Putting the pieces together then, we have a view according to which a person if 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*" (Nelkin, 2012, p. 124).

[61] An example Lynch (2013) gives is the one of a stubborn believer: "Thus we can describe the culpable desire/aversion motivating his stubbornness while eliminating any reference to what the belief is about: as, for instance, the desire to be right about matters which one has staked one's professional reputation on and spent one's career defending, or the aversion to having one's life's work invalidated. That the matter concerns Napoleon in this case is a relatively incidental point, and in another close possible world, where Flanagan specialized in something else, this same desire could have caused him to stubbornly cling to a belief with a different content" (pp. 1343-1344).

[62] Funkouser later (2009) changes his mind with respect to the product of the belief. He argues now that the notion of the belief is not fine-grained enough to serve as a product of self-deception and that one has to consider different regarding-as-true stances. I will discuss this possibility in the "product" section.

desire (p. 296). He advocates that the avoidance behavior, indicating the absence of motivation to find out the truth (pp. 297-298), and the analogy of self-deception to interpersonal deception support this choice for the motivation (p. 298).[63] The avoidance behavior shows that a certain non-self-deceptive belief can be attributed to the self-deceiver. Otherwise (if non-self-deceptive belief is absent) it is a case of delusion and not self-deception (Funkhouser, 2005, p. 303):[64]

> Indeed, doubting, suspecting and believing can be seen as three parts of one spectrum. Still, the *sophistication and apparent well-informedness of the avoidance behavior* of self-deceivers point toward attributing full-fledged true beliefs. More importantly, when the stakes become too high for self-deceivers and it becomes important that they act on their sincere beliefs, we see that self-deceivers do act in a way that shows they believe the truth. (Funkhouser, 2005, p. 310, footnote 15; my emphasis)

Funkhouser (2005) suggests that there are four functions of beliefs – reporting and regarding as true, on the one hand, (which belongs to contemplation) and being used in theoretical and practical reasoning, on the other hand, (which belongs to action). Out of them self-deceptive second-order beliefs – let me assume they are false - employ only the first two functions, which correspond to the public and private avowal (pp. 305-307). A true first-order belief can still be attributed to the self-deceived, because of the importance ranking of behavioral dispositions:[65] one could prefer nonlinguistic behavior over agent's phenomenology and verbal assertions in the attribution (Funkhouser, 2005, p. 307).

Interestingly, according to Funkhouser (2005), the appeal to consciousness in elucidating self-deception can be explained through the first- vs. second-order belief opposition. Second-order beliefs are more associated with consciousness than first-order beliefs, not least because of their reflective nature (p. 305):[66]

> We should add that not only must self-deceivers have a false second-order belief, they must also *lack* the true second-order belief that not-*p*. This would explain why the false claim is before **consciousness**, but the true claim is not. (Funkhouser 2005, p. 308; my bold emphasis)

[63] Funkhouser (2005) formulates this analogy as follows: "When person A deceives some other person B about *p*, A desires that B believe that *p*. Person A certainly need not desire that *p*. The self-focused account preserves this form of motivation even in cases when A = B" (p. 298). Funkhouser (2005) does not exclude the possibility of self-deception motivated by a world-desire. He proposes that in addition to *straight* and *twisted* self-deception (Mele, 2001), *indifferent or apathetic* self-deception presents the third type (p. 298).

[64] According to Funkhouser, Mele presents cases that are examples of self-delusion and not self-deception: "When Mele presents examples of (what he calls) self-deception, he invariably presents cases of what we call self-delusion. Such examples include a survey which shows that 94% of university professors think they are better at their jobs than their colleagues" (Funkhouser, 2005, p. 303). Similar example of the self-enhancement bias as the one used in the quotation above are employed by Mele (2001) and von Hippel and Trivers (2011a, b).

[65] Funkhouser's (2005) argumentation for the priority of linguistic cues in belief ascription is twofold (p. 300):

1. Beliefs and desires, in conjunction with their interactions among each other explain behavior.
2. The explanation of behavior is not simplistic, such that the whole belief-desire network is needed for this kind of explanation.

[66] Funkhouser (2005) claims to accept at least partly the higher-order belief account of consciousness, because he thinks higher-order theories to be most appropriate for explaining self-consciousness (p. 306).

It is essential to Funkhouser's argument, which conclusions we could draw from the avoidance behavior of the self-deceived. Pedrini (2012) interprets Funkhouser (2005) as claiming that the self-focused desire to belief that *p* is primary in self-deceivers, while the world-focused desire that p is contingent, disagrees with the second part of the claim and calls Funkhouser's argumentation an "(largely undeclared) inference to the best explanation about the motivation for avoidance behaviour" (p. 146). Pedrini's argument for the necessity of the world-focused desire in the self-deceiver is the fact that "they try to justify their convictions epistemically" (p. 149) or, in other words, that they bring arguments in favor of their self-deceptive beliefs. The conclusion is that self-deceptive behavior does not seem to warrant a decision between the desire and desire-to-believe account. Moreover, if the first-order belief corresponds to the evidence obtained, but second-order belief does not, then one needs to explain how this is possible, especially if one accepts the premise that each kind of misrepresentations presupposes weighting of the evidence, as it does Proust (2013):

> In the real world, metarepresenting that one believes that *p*, even in the shallow sense described in the ascent routine, presupposes that one has already formed the corresponding decision to accept *p*, which presupposed weighing *p* against not *p.* (Proust, 2013, p. 69)

Interim conclusion: Two questions were in the center of this section, namely whether the motivation for self-deception is a desire or a desire-to-believe and whether the self-deceptive attitude is a first or a second-order belief. For Funkhouser it is desire-to-believe and false higher-order belief. For Nelkin it is desire-to-believe and a false first-order belief. Accordingly, Nelkin's account might be said to be an 'error-about-justification' account, in which self-deceiver is said to err about his reasons for having a certain belief, and Funkhouser's account might be categorized as an 'error-about-belief' account, in which the self-deceiver acquires a false meta-belief[67] (for the distinction between 'error-about-justification' and 'error-about-belief' accounts see Fernández, 2013, p. 397). Both the question about the nature of the self-deceptive desire and that about the self-deceptive attitude depend on what one takes to be the goal of the self-deceiver: change one's belief-system or change the world. The desire-to-believe account suggests that it is rather the former than the latter. These two options need not be exhaustive. Scott-Kakures (2009) argues that self-deceivers are motivated by the desire to find out the truth (3.1.1). The difference between the first two options and the last one is that if one asks whether self-deceivers desire to change their belief system or the world, one assumes that self-deceivers take into account the causal connection between the self-deceptive attitude and their actions and, thus, see the self-deceptive attitude as a *means* to achieve another goal. If self-deceivers desire to find out the truth, then their self-deceptive belief is already the goal. If self-deception is motivated by other desires, then self-deception is a *means* to achieve something else, e.g. reduce anxiety as Johnston has proposed (1.1.2.1). This can be a desire to change the world or oneself. If the desire to change one's belief system is present in at least some cases of self-deception, then we need to ask what this means for a predictive coding account of it: is goal-driven change of one's belief system a kind of active inference

[67] Fernández (2013) presupposes the bypass model according to which "a subject forms a belief that she has a certain belief on the basis of grounds that she has for that first-order belief" (p. 391). A self-deceiver might for example believe that she does not have cancer despite evidence that would suffice for a contradictory first-oder belief (pp. 396-397).

or a kind of hypothesis testing?[68] Namely, the concept of active inference is so far restricted to involve classical reflex arcs:

> In brief, active inference can be regarded as equipping standard Bayesian updates schemes with classical reflex arcs that enable action to fulfil predictions about (hidden) states of the world. (Adams et al., 2015, p. 2)

This question only makes sense though, if self-deception is restricted to only *one* of these options. And here we are back to the question about the demarcation criteria for self-deception: desire vs. desire-to-believe? Change the world or change one's internal states? Holton (2001) has argued that self-deception is about oneself. This is the case simply because we think we were not biased when we are self-deceived, while in fact we are; this is a mistake we make about ourselves according to him (p. 60). Cases of self-deception that one finds in the literature are construed such that it is not only the case that self-deceivers never act on their self-deception, but also that their actions may be ambivalent. This does not tell us which *desire* was present, but that self-deception obviously might also change one's view of the world. The relationship between self-deception and the world is a complex one and will be further discussed in section 1.2. Let me in the following section summarize the constraints that I have distinguished so far.

1.1.3 Constraints on a satisfactory theory of self-deception

In this section I will have a threefold aim: I will shortly summarize the variety of different intentionalist and deflationary positions, as well as the criticisms voiced against these two and set constraints on a convincing concept of self-deception.

I have introduced four constraints on a theory of self-deception: parsimony and disunity in unity (1.1.1), as well as phenomenological congruency and choice of demarcation criteria (1.1.2). To recapitulate, an explanation of self-deception consists in choosing a personal level concept, e.g. self (Rorty), personal identity (Fingarette), one's belief set (Davidson and Pears) and showing how certain kinds of disunity can be generated in it, e.g. compartmentalization (Davidson and Pears) or relaxing the content of self-deceiver's intentions such that there is a disunity between the intention as cause and self-deceptive attitude as produced effect (Bermúdez). Extent of personal level recognition of evidence, or one's being self-deceived, varies in intentionalist and deflationary accounts. With it also changes the selectivity that one holds to produce self-deception: it may be either personal level selectivity of endogenous attention (Noordhof), or subpersonal level selectivity. One can interpret biases (Mele) or compartmentalization (Davidson, Pears) to produce the latter. In the case of biases, it is selectivity of information or evidence such that only certain kinds of information/evidence take part in the conscious reasoning process. In the case of compartmentalization, it is selectivity of a belief set that one consciously accepts at a certain moment in time. For parsimony reasons, no artificially complex disunity should be created. The phenomenological congruency constraint consists in creating an explanation that can also explain the phenomenology of self-deception. So far, weight has been on explaining self-deceptive anxiety, but in 2.1.2 I will argue that insight is another phenomenological characteristic of self-deception. Last, but still very important, varying explanations of self-deception lay explanatory weight differently on either a distinct motivation, or a distinct process. Thus, a theory of self-deception has to state demarcation

[68] This was one of the question Wanja Wiese and me have asked in our poster "Towards a predictive processing account of mental agency" at KogWis14.

criteria for what are the distinctive characteristics of self-deception, e.g. a certain internal state (motivational content dilemma), mechanism (selectivity problem), or a restricted set of phenomena (usefulness problem).

Intentionalist theories are mostly concerned with locating the disunity at the personal level and proposing an agentive selectivity mechanism that explains self-deception. They posit that the motivation for self-deception is an intention while the most popular deflationary alternative is that it is a desire. Both are personal level concepts. Van Leeuwen (2007a) does mention desires as well as goal-directed practical attitudes as alternatives to intention, yet, the position, that the motivation for self-deception are goal representations, has not received sufficient attention, at least in the philosophical literature (for more on goal representations see 2.3.1). The underlying intuition that drives the debate between intentionalists and proponents of deflationary positions is that self-deception is not something that happens *automatically.* This is best seen in Pears' (1991) and Johnston's (1988) mutual critique. Pears argues that there is some center of agency that upholds self-deception, while Johnston proposes that a non-intentional, but desire-guided mechanism – mental tropism – is responsible for the emergence of self-deception. While Johnston holds Pears' account as too *complex*, Pears holds Johnston's account as not doing justice to the *continuous* and *selective* influence of the motivation in self-deception. Given that intentions are usually ascribed to agents and positing subconscious intentions is not parsimonious, it is comprehensible that Mele calls intentionalist positions the *agency* view:

> Consider the following two bold theses about motivationally biased beliefs.
> 1. The agency view: all motivationally biased beliefs are intentionally produced or protected. In every instance of motivationally biased belief that *p*, we try to bring it about that we acquire or retain the belief that *p,* or at least try to make it easier for ourselves to acquire or retain the belief.
> 2. The antiagency view: no motivationally biased beliefs are intentionally produced or protected. In no instance of motivationally biased belief that *p* does one try to bring it about that one acquires or retains the belief or try to make it easier for oneself to acquire or retain the belief. (Mele, 2001, p. 13)

Yet, positing conscious intentions leads to the *dynamix paradox* of self-deception which is that, as agents, it is impossible to intend to acquire a belief one holds to be false, as well as to the *static paradox* of the impossibility to possess contradictory mental states. *Dissociations* have been postulated to solve the latter, though such a solution is not necessarily parsimonious. Nevertheless, there are still contemporary positions advocating something similar to Davidson's and Pears' position, namely that *agents* can create dissociations in their own systems (e.g., different kinds of dissociations are emphasized in von Hippel & Trivers 2011b; see section 3.2.2.3). Let me consider one example of how dissociations might be defined:[69]

[69] Interestingly, Braude (2009) argues that dissociation does not encompass self-deception and cognitive dissonance, because they do not satisfy his conditions for dissociation. He holds that self-deception and cognitive dissonance satisfy the following description:
- Third-person description: "ongoing behaviors or perceptions that are inconsistent with a person's introspective verbal reports";
- First-person description: "simply failing to grasp that simultaneously held beliefs are inconsistent" (p. 33).

Most probably, it is the third condition for dissociation, or *erecting* a barrier (an agentive process), that Braude (2009) thinks is not the case in self-deception. For Braude (2009), as far as I can see, dissociation is to be understood intentionally, but self-deception – not. The latter is the case because he falls victim to deflationary understanding of self-deception that is contemporary the favored position. Braude (2009) further defines "repression as unconscious

> (1) *x* is an occurrent or dispositional state, or else a system of states (as in traits, skills, and alter identities) of a subject *S*; and *y* is either a state or system of states of *S*, or else the subject *S*.
> (2) *y* may or may not be dissociated from *x* (i.e., dissociation is a nonsymmetrical relation).
> (3) *x* and *y* are separated by a phenomenological or epistemological barrier (e.g., amnesia, anesthesia) erected by *S*.
> (4) *S* is not consciously aware of erecting the barrier between *x* and *y*.
> (5) The barrier between *x* and *y* can be broken down, at least in principle.
> (6) Third- and first-person knowledge of *x* may be as direct as (respectively) third- and first-person knowledge of *S*'s nondissociated states. (Braude, 2009, p. 32)

At this point it may merit to explain why I actually did not choose to call the *disunity* in unity constraint a *dissociation* in unity constraint. This is because I use disunity in the sense of point 3 of Braude's definition, though I am not sure that the list of barriers (phenomenological and epistemological) is exhaustive. For example, how should one call the barrier between the implications of one's intentions on one's acquired attitudes? In a certain sense one might call this an epistemological barrier, but in another one might say that there is no barrier at all, though there is a certain kind of introduced disunity. A barrier may also be not only between states, but also between contents of states, or maybe something else. To recapitulate from section 1.1.1.5, the intuition that there is *no boundary in between conscious states* has so far precluded personal level accounts of self-deception to succeed: every intention, phenomenal state, consequences of behavior have to be considered by the agent when she decides what she believes, because all personal level states are connected by the rules of logic and rationality. Intentionalists have tried to solve this problem by weakening the kinds of intentions that operate in self-deception. Bermúdez (2000b) holds that there is a difference between intending to bring a *certain* belief, or a beliefs one thinks to be *false* and a difference between intending to bring a contradictory beliefs *while knowing* otherwise and just *intending* to bring about that one believes something (p. 310). The weakest possibility would thus be just intending to bring about a certain belief which one does not think to be false, but which is improbable given the evidence. This would be a position very similar to Barnes' deflationary one about indirect inferential relations (1.1.2.2), apart from the constraint that it has to be an *intention*. Yet, if it was not for the possibility that one would lose touch with the intention while one implements it (Bermúdez, 2000b), which would make self-deception, if ever, a very special kind of intentional action, I do not see how an intentional account *to deceive oneself* can be upheld. Postulating special kinds of intentions is not parsimonious though.

Not only the *content* of intentions, but also how *direct* their influence is, has been offered as an explanation of the possibility of self-deception. The underlying intuition regarding the directness is that the more indirect the influence of intentions on self-deception, the easier is it to self-deceive in virtue of not realizing that it is the agent's intention that is

denial, dissociation as subconscious denial, and suppression as conscious denial" (Braude, 2009, p. 32). Interestingly, he argues that, as in the case of self-deception, repression is not an instance of dissociation, because the 6th condition is violated, which means that repressed states can be accessed only indirectly, whereas for dissociated states the access is direct: "third- and first-person knowledge of dissociated but not unconscious—states can be as direct as (respectively) third- and first-person knowledge of nondissociated states" (Braude, 2009, p. 31). Strangely, knowledge acquired by means of hypnosis is said to be direct or comparably direct ("requiring little more than assumptions about behavior-reliability," p. 31). Indirect knowledge means for the author the knowledge that "must be inferred from its possibly distorted or primitive cognitive, phenomenological, or behavioral by-products." (p. 31) responsible for the acquired belief. Thus, if the intention is *otherwise directed*,[70] its connection to the self-deceptive belief might not be realized by the agent. Galeotti's (2012) account is an example of this kind. She argues that an *invisible hand explanation* is applicable to self-deception: the single self-deceptive steps are intentional, yet the aim is brought about non-intentionally. The steps of the self-deceiver are "directed at reconsidering evidence and forming a true judgement" (p. 58), yet a "cover story" is formed (= the resulting explanation for the acquisition of the self-deceptive belief from the first-person perspective) and hung on to under the influence of emotion.[71] Galeotti (2014) seems to combine Mele's FTL model with her "invisible hand" explanation: "SD is actually the response to negative evidence just in case costs of inaccuracy are not too high and/or the threat is so much beyond control that SD is worthwhile as an, at least temporary, soothing thought" (p. 7).

Instead of weakening contents of intentions, or introducing indirect inferential relations between mental states, one might also argue that there is no *synchronous* disunity what so ever, but a diachronous one. Scott-Kakures (2012) has argued that there cannot be two diametrically opposed conscious intentions at the same time, i.e. the intention of deceiving oneself and the other aimed at discovering the truth. This leads him to suppose that intentional self-deception is only possible as intentional causing of unintentional deception, illustrated by the case of *temporal self-deception*: Mathematician Sammy, knowing of the high possibility of him having an Alzheimer's disease later on, decide to fake evidence that he has led a rich social life, so that he can do math, while still young (pp. 21-22). Scott-Kakures (2012) argues that in this case the aim to deceive oneself in the young Sammy (the deceived) has been altered to be settling the question in the old Sammy (the deceiver). The author further generalizes that intentional self-deception is possible only for such a model:

> 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*. (Scott-Kakures, 2012, p. 37; my emphasis)

I agree with Scott-Kakures (2012) that a single agent can have only *one* intention, to find out the truth *or* to deceive oneself. I do not see an explanatory value in intentional explanations of self-deception along the lines of "intentionally ignoring, blocking out, and

[70] In section 1.1.1, I argued that Davidson and Pears' explanations are *personal* level, not subpersonal. Smith (2014) argues that they are subpersonal in that parts of agents possess beliefs and intentions. My reason was that intentions are personal level concepts and that even if beliefs are *stored* in different compartments, still only one congruent belief set can then be consciously retrieved.

[71] Again, also in Galeotti's view the main factor enabling self-deception is that, whatever the mechanism be, as long as the agent does not recognize what the mechanism lead to, she can self-deceive: "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" (Galeotti, 2012, p. 58). As for the mechanis, Galeotti argues it to be non-intentional on the distinction between purposeful and agentive behavior: "Similarly, also in SD there is a *purpose* set by the wish 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 money or language, the seemingly purposive outcome needs neither a teleological model nor a causal one to be made sense of" (Galeotti, 2014, p. 3).

engaging in activities that will lead them to forget evidence that *p*" (Perring, 1997). I also do not see in how far intentional activities otherwise directed are of explanatory value, if it is deflationary attitudes (desires and emotions) that have to bear the explanatory weight to explain the emergence and success of such intentions in accomplishing self-deception. In general, diachronic disunity relays explanatory weight on subpersonal mechanisms and, thus, makes the use of 'intentions' as descriptors unnecessary. To sum up, the main categories of intentional accounts of self-deception are (Mele, 2001; Smith, 2014):

1. Intentional activities "engaged in as part of an *attempt* to deceive oneself" (Mele, 2001, p. 18);
   a. Reflexive version (e.g., temporal self-deception) – personal self-deception: an agent as a whole has the intention to self-deceive, which evokes paradoxes of self-deception
   b. Partitive version (e.g. Davidson and Pears) – subpersonal, but not subintentional: some parts of an agent can have intentions to deceive, which is empirically implausible.
2. Intentional activities otherwise directed, e.g. to relieve anxiety.

Being "unwittingly moved by desire" during the belief-forming process (Scott-Kakures, 2002, p. 600) is the popular alternative to intentionalist accounts. Deflationary, in difference to intentionalist, accounts do not focus on the kind of disunity that explains self-deception, but on the kind of motivation or process that has led to it. Typically, a defended kind of motivation is also seen as the only kind of motivation characterizing self-deception which means it is a demarcation criterion for labelling something 'self-deception', e.g. anxious desire vs. desire (Barnes vs. Mele) or a desire vs. a desire-to-believe (Mele vs. Nelkin/Funkhouser). In arguing that it is an anxious desire one restricts the possible function of self-deception to a defensive and protective one. Further, so far the argumentation in favor of the desire about a first- or a second-order state has been based on the interpretation of the behavior of the self-deceiver and, up until now, there is no way to empirically distinguish between the two (1.1.2.5). The main different kinds of desires motivating self-deception have been summarized in figure 5. There is also an account that argues that self-deception is not motivated at all, or motivated only in a sense of possessing a *general desire to make coherent sense of the world and our actions* (Patten, 2003, p. 236). I have not put this option into the figure, because the given desire is too unspecific to be characteristic of self-deception (and cannot also serve as a demarcation criterion).

Patten (2003) accepts Bem's self-perception theory, according to which we infer our beliefs and desires from our actions, like we infer them about the mental states of others. Self-deception along these lines is a mistake about one's inferred states and reasons for acting or the difference between our first- and second- order beliefs (Patten, 2003, p. 230). The second-order beliefs meant in this case are, according to Patten (2003), the preconceptions, we have about ourselves, that are part of our self-schema and that influence the inferences we draw from behavior about our first-order states. I think that cognitive dissonance research (Festinger, 1957; see section 3.1.1), as well as metacognition research (Proust, 2013; see section 2.1.3) would argue against Bem's self-perception theory, but there is also a predictive coding account of the acquisition of self-representations that is based on Bem's self-perception theory (4.4). I think that the desire to make coherent sense of the world as such is too general to act as a specific motivation in each case, though it might be the case that this desire certainly aids self-deception via its contribution to the construction of a coherent self-narrative (for the importance of the latter for self-deception see 1.2.6 and, more generally, the debate about the "product" of self-deception).

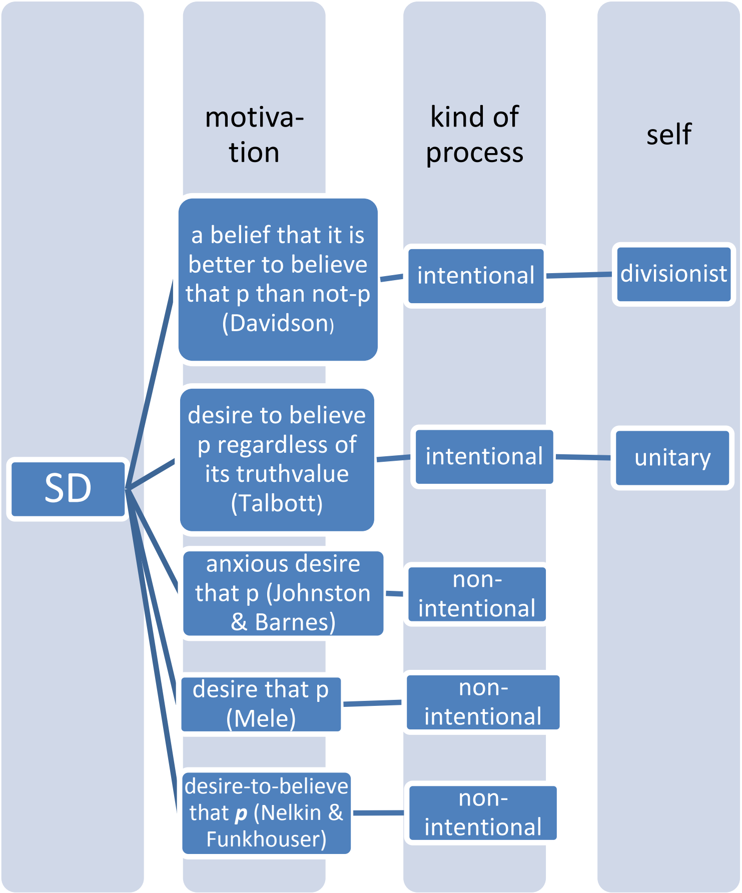

motiva-
tion
kind of
process
self
SD
a belief that it is better to believe that p than not-p (Davidson)
intentional
divisionist
desire to believe p regardless of its truthvalue (Talbott)
intentional
unitary
anxious desire that p (Johnston & Barnes)
non-intentional
desire that p (Mele)
non-intentional
desire-to-believe that *p* (Nelkin & Funkhouser)
non-intentional

Figure 5. Different kinds of motivation
Distinctions from Barnes (1997, p. 79), Nelkin (2002, 2012), Funkhouser (2005).

To sum up so far, intentionalist positions are mainly concerned with the question of how to find the right kind of personal level disunity to explain self-deception. Deflationary positions, on the contrary, are more concerned with the choice of a right kind of desire that causes self-deception. Since deflationary accounts are non-agentive, I think that the implicit assumption is that the selective process leading to self-deception is subpersonal and this is the reason why the disunity question is not as relevant, as finding the right kind of motivation or subpersonal process. Different kinds of problems have been argued to haunt intentionalist and deflationary positions (see table 9). For intentionalist positions, the best discussed one's are static and dynamic paradoxes and for deflationary positions – the tension problem. Both kinds of account have been argued to be susceptible to the *selectivity problem* and also to the need to explain *twisted self-deception* (1.1.2.3).

Arguments against intentionalism	**Arguments against deflationarism**
Kind of motivation: intention	Kind of motivation: Non-intentional motivational element makes use of the "cold" biasing procedures/"hot" biasing is at work.
1. **selectivity problem**: It is possible to imagine case in which one would not be able to self-deceive despite the intention to do so. Thus, one cannot always self-deceive, when one has the intention to do so (Mele, 2001, p. 66).	1. **selectivity problem**: It is not always the case that motivation leads to the acceptance of certain self-deceptive hypotheses (Bermúdez, 2000, p. 317).
2. **problem of negative/twisted cases:** If the formation of the irrational belief in self-deception always is caused by a desire to form it, then negative cases – Mele's twisted self-deception - cannot be explained (Lazar, 1999, p. 275).	2. **problem of negative/twisted cases**: Not every kind of desire can incorporate twisted cases, or the acquisition of self-deceptive attitudes one does not desire to be true (Nelkin, 2002).
3. **problem of crazy choices:** If the choice of the self-deceptive belief is reasonable, then appeal to practical reasoning is possible, even in the face of a breakdown of theoretical reasoning, if not – then either way of explanation is not possible (Lazar, 1999, p. 274). Lazar's example is that if one's intention is to stay healthy, then to self-deceive in the face of evidence to the contrary will not aid in achieving this aim.	3. **tension problem**: There is a certain kind of tension – feeling of uneasiness - in self-deception that can only be explained if either an inconsistency or a contradiction is present in the belief set of a self-deceiver (Lynch, 2012).
4. **static paradox**: Believing two contradictory propositions at the same time (Mele, 2001)	4. **specificity/usefulness problem:** Galeotti (2014, pp. 3-4) calls it the *specificity* problem. Making the category of self-deception too wide precludes its scientific investigation and also its characterization as a unified class (Van Leeuwen, 2013a).
5. **problem of how it is done/dynamic paradox:** Either intentionalists have to explain the lack of awarenes of the given intention by appeal to non-thematic explanations (general facts and principles pertaining to information processing) which is unplausible given a relatively complex nature of the intention to believe, or they are caught in the circular argument (by what means an agent intentionally keeps the intention away from consciousness) (Lazar, 1999).	5. **content dilemma:** The motivation for self-deception has to be not too inclusive to encompass non-self-deceptive cases, but also not too exclusive to leave cases of self-deception out (Nelkin, 2002).

Table 9. Summary of the problems for intentionalist and deflationarist positions
Distinctions from Lazar (1999, pp. 268-280), Bermúdez (2000b), Lynch (2012), Mele (2001), Nelkin (2002), Van Leeuwen (2013a), Galeotti (2014).

Galeotti (2014) holds, apart from specificity and selectivity, ascription of *responsibility* to be another important criterion, because she thinks self-deception to be preventable by character-building or pre-commitment, the latter being an explicit agreement *ex ante* between self-deceiver and his person of trust that the self-deceiver will use the person of trust as a referee concerning his motivated false beliefs (p. 4). Nelkin (2002, 2012) also emphasizes responsibility as a criterion for estimating the goodness of the explanation of self-deception. I will not go into detail in discussing this criterion, however.

To return to the four constraints for a good theory of self-deception that I established (parsimony, disunity in unity, phenomenological congruency and demarcation criteria), I promised at the end of section 1.1.2.4 to answer the question which kind of selectivity, personal level recognition and demarcation criteria to be characteristic of self-deception. I want to pursue this now. From the problems listed in table 9, the *demarcation problem* encompasses the *content dilemma* and the *usefulness problem* and the *selectivity problem*, because these problems are connected among each other. The explanation of self-deception has to be scientifically useful (usefulness problem), for this, it has to *precisely* circumscribe the phenomenon, not too narrow or too wide (content dilemma). Of course, there is also a selective aspect: which kinds of internal mental states are characteristic only of self-deception, so that when they are present, self-deception occurs? Or is it maybe a certain kind of process (and not a certain kind of motivation) that guarantees the occurrence of self-deception? Maybe a combination of a certain kind of motivation with a certain kind of a process? Or maybe it is better to restrict self-deception by restricting kinds of phenomena that fall under this term? To remind the reader, Rorty's account demonstrates the *breadth* of phenomena that could have – and have been – taken to be cases of self-deception: from non-motivational habit over self-fulfilling prophecies and self-enhancement to rationalization. Is such breadth of an application of the concept of self-deception justified and scientifically useful, as Van Leeuwen asks? The answer to the demarcation problem depends on the ascriptions of self-deception by others and retrospective ascription of self-deception by oneself, since one by definition cannot recognize self-deception while being self-deceived. Such ascriptions are based on the third- and first-person characterization of the self-deceiver, which will be the topic of chapter 2. I will argue there that from the third-person point of view self-deceivers are inconsistent, but justify their inconsistency. From the first-person point of view, they experience anxiety, while being self-deceived, and insight at the time when self-deception is abandoned. I have subsumed the truthfulness to the first-person description under the phenomenological congruency, while truthfulness to the third-person description – under the disunity in unity constraint.[72] This is because self-deceivers justify their beliefs on the basis of a unified personal level beliefs governed by logic and rationality, while some kind of disunity explains the inconsistency of behavior. As for the phenomenological congruency constraint, the tension problem is a part of it. Yet, I think that this is only a part of the picture. In full, one should speak of a self-deceptive phenomenological vicious circle (1.1.2.1): How can anxiety be the cause and the result of self-deception and self-deception still be maintained? I will propose my take on the questions about motivation, process and self-deceptive attitude in section 2.3. I want to note here already that since ascription practices demarcate the phenomenon of self-deception, most certainly no *single* process or motivation will be characteristic of self-deception.

If there are no unequivocal demarcation criteria for what we call 'self-deception', then it may be a *cluster concept*. The term 'cluster concept' stems from Wittgenstein:

> Wenn aber Einer sagen wollte: „Also ist allen diesen Gebilden etwas gemeinsam, - nämlich die Disjunktion aller dieser Gemeinsamkeiten" - so würde ich antworten: Hier spielst du nur mit einem Wort. Ebenso könnte man sagen: es läuft ein Etwas durch den ganzen Faden, - nämlich das lückenlose Übergreifen dieser Fasern. (Wittgenstein, 2011[1953, postum], p. 58)
>
> If, however, one would want to say: "There is thus something that all those creations have in common, namely the disjunction of all the similarities" – then I would respond: Here you are playing with words. One could also say: there is something that combines all the threads, namely the seamless blending of all the threads. (my translation)

[72] I think that knowledge of the truth, advocated for by Bayne & Fernández (2009), is too strong a criterion for self-deception: "There are three constraints on a theory of self-deception: (a) that it avoids the two classical paradoxes of self-deception, (b) that it accounts for the fact that the subject appears to know the truth at some level, and (c) that it imply that the self-deceived subject is epistemically negligent" (p. 12).

Self-deception may be a cluster-concept for several reasons. First, given Van Leeuwen's criticism of Trivers: more and more different phenomena are subsumed under it (see section 3.2.2.1, as well as the usefulness problem discussed above). If every two pairs of these phenomena share a characteristic, then the set of such phenomena is connected, though no unified necessary and sufficient conditions for self-deception can be formulated. More, a slippery slope can then arise, since a new phenomenon can always be added in virtue of its similarity to a chosen another one that is in the set. For example, first cases of self-deception were the ones of contradictory *knowledge* (intentionalist accounts). Tension would result from contradictory knowledge (premise in virtue of the personal level being unified and the introduced disunity as presence of contradictory knowledge). Tension can be explained by weaker kind of disunity – possible contradictory knowledge (deflationary accounts, more on possible knowledge in section 3.2.2.3), e.g. a significant chance that the self-deceptive belief is wrong (Mele's solution). In virtue of tension and inconsistent behavior, cases of cognitive dissonance can in certain cases (presence of certain internal states) be seen as those of self-deception (3.1.1). In virtue of inconsistent behavior, biases can also be seen as self-deceptive. Second, self-deception may be a cluster-concept in virtue of its result being a certain kind of *attitude*. Michel (2014) has argued that attitudes are cluster-concepts, on the premise that attitudes possess more or less fine-grained dispositional profiles, the conditions for the expression of which can be more or less precise (p. 231; see section 1.2.3 for his view). For the first option, self-deception would be a cluster-concept because of the variety of phenomena subsumed under it. For the second, it would be because of the resulting self-deceptive attitude being a cluster-concept.

Despite the absence of clear demarcation criteria for self-deception in terms of phenomena subsumed under it (usefulness problem), process leading to it (selectivity) and internal states characteristic of it (content dilemma), there is a tendency, namely to regard those cases that allow for more personal level recognition as stronger cases of self-deception (see introduction to 1.1.2 and particularly the difference between accounts in sections 1.1.2.1 and 1.1.2.2). This is because if an explanation for a case can be thought about that allows for a disunified personal level, despite it being governed by logic and rationality that unify it, then it would be an explanation of a stronger paradox than if subpersonal mechanisms just changed the way we perceive the world or think. For such an explanation, there would be a need of a personal level process – some sort of *agentive* selection and, as a result of such a selection, agentive control (for a personal/subpersonal level selection distinction in self-deception see 1.1.2.4). Deflationary accounts prefer subpersonal selection – less paradoxical, but also easier to handle, e.g. self-deceivers misperceiving evidence as a result of certain biases. Think of two different cases: first, self-deceiver who samples the environment in a search for evidence as input to her belief-forming process, has the phenomenology of controlling her sampling and will justify it; second, a self-deceiver who has the phenomenology of sampling and of control, but whose sampling was not controlled by the agent, but by some other subpersonal mechanism. I think that psychological biases fall under the second category: the weighting of the evidence or the evidence itself has been changed without the agents' knowing so that on the personal level, what the agent is confronted with, is an already biased evidence. To use the distinction *conscious\unconscious* instead of *personal/subpersonal*, e.g. something not being in the focus of attention (Noordhof, Demos), would only redescribe the underlying problem in personal level terms, unless one could propose a *personal level unconscious* mechanism that could lead to self-deception, but would happen at the same time as the personal level conscious reasoning process. This is because there is always some thinking, imagining, or some other high-level process going on in our minds on the personal level. Two distinct personal level processes happening at the same time, one conscious and the other not, are hard to imagine for me. This is clearly not an argument for the impossibility, but an appeal for empirical science to find out. In any case, even if there is such a possibility, there might be another subpersonal process leading to self-deception as well.

How the personal and subpersonal level interact is an important issue for the clarification of self-deception (1.3.4), more important than the motivation or the resulting attitude. The debate on the self-deceptive attitude will lead back to the question about the process of self-deception – narration or belief-forming, which helps to clarify what is going on the personal conscious level during self-deception: Is self-deceptive process one of evoking a somehow skewed epistemic agent model (Metzinger, 2013a), which is the transparent conscious self-representing of capacity for controlling one's epistemic processes, or is it one of intact epistemic agent model but skewed phenomenal world- or self-model? The difference is that a phenomenal world- and self-model constrain such an epistemic agent model in virtue of rules of logic, consistency and rationality governing the personal level. To bring an example, an epistemic agent model with the intent to clarify the question whether one possesses hands would be unexpected, if one indeed possessed hands. In case of *anosognosia* though (which has been argued to be one of self-deception, see section 1.2.7), an epistemic agent model questioning the presence of one's hand would be expected and is tried to be evoked by doctors, but only with partial success. Another example is *denial of pregnancy*. Denial of pregnancy does not exhibit such characteristics of self-deception as *third-person* inconsistency and, hence, no need for justification. No tension is present either, but there is something described as subpersonal "knowledge" by the autonomic system of the pregnancy that is absent at the personal level (Sandoz, 2011, p. 784). The reason for such an assumption is a silhouette effect: the figure of those denying pregnancy does not change, since the foetus expands in vertical direction (that of diaphragm; p. 783). When insight into their body condition comes upon a medical examination, there are cases when almost instantly the position of the foetus changes and the silhouette changes on the eyes of a doctor. As Sandoz (2011) puts it, there is a "complete physical metamorphosis that had taken place so quickly because of her psyche" (p. 783). An explanation given for the silhouette effect by the author is *reactive homeostasis* defined as reflexive immediate escape from emergent situation (Sandoz, 2011, p. 784). Emergent situation is said to be that of denial and is described as a case of "persistence of paradoxical realities" (p. 784). The author wonders which kind of informational pathway between semantic knowledge of pregnancy provided by the doctor and the autonomic system might have led to an immediate appearance of pregnancy signals. In case of denial of pregnancy, there is no personal level recognition at all and even in reverse, autonomic signals are changed by something to uphold a certain model of reality, which would obviously preclude certain epistemic agent models, e.g. those about possible pregnancy. No case of a conscious belief-forming process about pregnancy needs to take place, no tension and even others are fooled by the absence of physical signals of pregnancy. This example demonstrates two things. First, I think that the reader would agree that denial of pregnancy comes close to being a case of self-deception, but some typical characteristics are lacking. Difficulty of stating demarcation criteria for self-deception is demonstrated. Second, there is no personal level selectivity processing going on. It is not like a case in which a silhouette of a woman starts changing, she wonders whether she is pregnant, finds the thought aversive and miraculously becomes slim again for up to the birth date. No rumination or something similar was there from the starts. All the epistemic agent models during pregnancy were not about it either. Yet, the paradox of contradictory personal and subpersonal information (I will not call it knowledge to reserve this term for the personal level only), that has been caused by a motivated unknown subpersonal psychic mechanism, is a description that makes denial of pregnancy very similar to paradigmatic cases of self-deception, because there is motivation and contradiction at some level. The difference is that denial of pregnancy can be empirically tested, insight into the bodily condition can be evoked by a doctor at a determined moment in time too. It should be noted that denial of pregnancy also demonstrates the disruptive effects on health that self-deception can have. I will say more about it in chapter 3.

Interim conclusion: In this section I summed up different intentionalist and deflationary solutions. I have also summed up the constraints set out by me on an explanation of self-deception in section 1.1.1 and 1.1.2, namely parsimony, disunity in unity, phenomenological congruency and demarcation criteria. Earlier, the question was about the demarcation criteria of self-deception. The simplest and most unclear answer would be that it is a cluster-concept. It is the simplest because this kind of answer only relays explanatory weight to another unclear concept. It is the most unclear, because weight relay in itself does not explain anything. I then asked whether phenomena labelled 'self-deception' can be ordered according to a certain criterion, namely amount of personal level recognition. I pointed out that the amount of personal level recognition is connected to the choice of a selection mechanism – agentive or not – that is responsible for self-deception. An implication of such an ordering function is that the interaction between the personal and subpersonal level is important for an explanation of self-deception in general and the clarification of the process of self-deception in particular. The process of self-deception need not be one of belief-forming as an instance of an epistemic agent model, or need not involve a construction of an epistemic agent model at all, it could involve a more basic construction of a world- and self-model, as the example of denial of pregnancy shows. What the latter example also shows is that motivated contradiction may be the only characteristic unifying different empirically testable cases of self-deception (see 1.2.7 for more on the relationship between self-deception and delusion that points into this direction). The following section will discuss different answers to the question which kind of attitude results from self-deception and during this discussion the recurring themes will be whether self-deception is belief-forming or narrative construction, as well as dispositional or constructionist. Typically self-deception is said to be dispositional belief-forming. As I have tried to sketch above though, self-deception may lead to changes in the world- and/or self-model, instead of the epistemic agent model. This depends on the kind of selection that is at work in self-deception. I will refine the chategorization of kinds of selection that may be responsible for self-deception in 2.2.1. The purpose of the following section will be, first, to show all the different possibilities, given in the literature for the resulting self-deceptive attitude; second, to argue that the assumption underlying them all is that the self-deceptive process is that of creating a certain epistemic agent model; third, to open up the possibility of changes in the world- and/or self-model as self-deceptive results.

1.2 “Product” or kind of misrepresentation debate

> The tropisms do not give us apparently conflicting behavioral patterns, where it is as if the self-deceived agent believes both that *p* and not-*p*. But if there is not such conflicting patterns of behavior to be explained, then why at all resort to the notion of “self-deception”?
> (Borge, 2003, p. 9)

The structure of this section is similar to the previous one on the intentionalist/deflationary accounts. I will first introduce different kinds of solutions (section 1.2.1 – 1.2.5) and then summarize the results (section 1.2.6 and 1.2.7). I will argue that the product debate demonstrates the inadequacy of the folk-psychological description of the result of a self-deceptive process as belief or another similar folk-psychological notion. Belief forming and narrative construction might be two sides of the same coin so that the acquisition of self-deception is not a case of explicit deliberation, but of some subpersonal process. The maintenance (or justification) of self-deception, on the contrary, may be the case of explicit narrative construction. It should be underlined, though, that the term ‘narrative construction’ is ambiguous, because it may denote the construction of a certain epistemic agent model (I as an agent am constructing a story about myself) or the construction of a world- and self-model itself can over time be seen as a consistent narrative construction. It is the latter possibility that have been so far left unexplored in accounts of self-deception and that I will return to in section 2.2.3.

Let me, in short, classify the types of accounts that are to come and then sketch how I will enrich them in section 2.3.3. First, a very short general introduction to the “product”-problem in general. In the previous section the focus was on the motivation of self-deception, the present section discusses the “product” of self-deception, which has been defined by Van Leeuwen (2007a) as the attitude that results from self-deception (p. 419). Given, on the one hand, that self-deceivers do behave in a contradictory manner and we ascribe to them first-person authority (see table 10), as well as peculiar and privileged access to their mental states and, on the other, that ascriptions of contradictory beliefs provide little as explanations: If subjects in general and self-deceivers in particular are to preserve a certain minimal sense of rationality, then the mental attitude of the self-deceiver may have to be redescribed as something other than a belief in order to coexist with another beliefs, that have contradictory content. This way the behavior can still be seen as a consequence of the given attitude, but the contradiction between this attitude and a certain belief still can be preserved. Here, the implicit questionable assumption is that only attitudes of the same kind clash if they have contradictory contents, but not different kinds of attitudes.

Aspects of full self-knowledge (p. 12, 14):	**(Independent) properties of self-knowledge (p. 15):**
1. *Self-reference/ownership* acquired via *attribution* of the attitude to oneself	1. *Epistemically peculiar access*: There is an epistemic asymmetry in that access to our own mental states is different from the access to it of others.
2. *Attitude-type* (e.g., belief, desire) acquired via *classification* of the attitude	2. *Epistemically privileged access*: Our access to our own mental states is *superior* to that of others. Background assumption: “if externalism is true, it is compatible with privileged access to thoughts and attitudes” (p. 15)
3. *Mental content*	*“First-person authority” (FPA)*: Self-ascriptions of our own mental states enjoy an authority.

Table 10. Michel: characteristics of self-knowledge to explain
Distinctions from Michel (2014).
Michel himself argues for an incorrigible FPA, peculiar and privileged epistemic access (p. 21).

Two questions have been asked with respect to the results of the self-deceptive process (Van Leeuwen, 2007a):

1. *How many* attitudes are present in the self-deceived: only the true attitude, only the self-deceptive attitude, both or none? [73] (Van Leeuwen, 2007a, p. 421)
2. What *kind* of an attitude is the product of self-deception?

If the answer to the first question does not leave much space for options, the answers to the second one are more complex, ranging from belief and second-order belief (see section 1.1.2.5) to recurrent thought and avowal (section 1.2.1), regarding-as-true stances (section 1.2.2), pretense (section 1.2.4), emotion (section 1.2.5) and analog form of representation (see below).[74] In 1.1.2.5 I have already mentioned that Funkhouser (2005) held the self-deceptive attitude to be a false higher order belief – a false belief that one believes that *p*. The other options are to be discussed in this section.

Let me now, second, classify the mentioned self-deceptive attitudes. Two kinds of distinctions will be helpful in categorizing the proposed self-deceptive attitudes: belief-formation/narrative construction and dispositional/constructionist (see table 11). That self-deceptive process is that of narrative construction has been argued by Fingarette (1.1.1.3) and Gendler's account (1.2.4) holds in the same vein as well. Most self-deceptive belief-forming accounts are dispositional (see section 1.2.1). Yet, inconsistent behavior of self-deceivers is difficult to explain in that way (1.2.2). Michel (1.2.3), thus, argues that self-deceptive beliefs are constructionist instead of dispositional – they are not stored and retrieved, but constructed at every moment in time from the available evidence.

	Dispositional	**Constructionist**
Belief-formation	(probabilistic) belief, avowal, regarding-as-true stances (sections 1.2.1 and 1.2.2)	belief (section 1.2.3)
Narrative construction	personal identity construction, pretense (sections 1.1.1.3 and 1.2.4)	temporal self-deception: disturbed memory case (see section 1.1.3)

Table 11. Types of self-deceptive attitudes

On the assumption that self-deceivers actually memorize their narratives, I have identified Fingarette's (section 1.1.1.3) and Gendler's (section 1.2.4) accounts as *dispositional*, instead of constructionist narrative. Notice that self-deceptive constructionist belief-forming accounts will need to explain the diachronic discontinuity of their attitudes, because even if attitudes are not stored and retrieved, the agent would notice one's own diachronic discontinuity of attitudes on the assumption that the personal level is not only *synchronously*, but also *diachronously* consistent. The example of Sammy in section 1.1.3 who has a memory disease and intentionally writes falsehoods into a diary to acquire a certain belief about oneself later can be seen as *constructionist* narrative construction, because a narrative about oneself is constructed on the spot from the diary.

Two solutions that do not fit into the table 11 are that self-deception is about emotion (1.2.5) and that it is not the kind of attitude, but the way of representing that allows for the presence of inconsistent representations. The second is Hirstein's analog representation solution that I want to shortly mention. Hirstein (2005, p. 229) advocates the solution of the paradox of

[73] If one holds that only one attitude is present, as Mele for example, a question about tension in self-deception arises. How is it possible to explain the tension in self-deception, if no true attitude is present somewhere in the mind? Van Leeuwen (2007a) calls this sort of account the one-belief view (p. 422).

[74] Van Leeuwen (2013a) mentions that the second-order belief-view is advocated by Funkhouser, the recurrent thought view by Bach and the pretense view by Gendler. He also mentions the avowal-view in his article, though he does not mention Hirstein's analog representation-solution or Funkhouser's shift to the regarding-as-true stances view (see below).

two contradicting beliefs by differentiating between conceptual and analog forms of representation:

> The information that p is already represented in analog form. There is conflict in the person's mind, but the conflicting information is represented in two different forms, conceptual and analog. This is different from holding two contradictory beliefs in full conceptual form. What *may be happening* is that the brain has a way of preventing certain types of analog information from being represented in conceptual form, from being explicitly thought and believed. (Hirstein, 2005, p. 229; my emphasis)

Hirstein's solution demands a more thorough elaboration, because some questions remain unanswered. Hirstein does not answer the question of how these two forms of representations are connected with each other. Does all information, first, enters the mind in analog form and only afterwards is processed to conceptual forms? Can information enter the mind directly in conceptual form? Is there a reverse order of transformation (from conceptual to analog form) also possible? Apart from the mutual influence-question, it remains to be answered how they are connected to action. Can I act on an analog form? Self-deception does apparently involve acting, namely acting contrary to one's evidence (at least so states the widespread intuition). Hence, if the analog form cannot explain the connectedness to action, then it can't be used for the explanation of self-deception. The reader is encouraged to ask analogue question also about accounts that will be presented below.

After the categorization of the accounts that are to be presented, let me explain how they will be extended in section 1.2.6 and 1.2.7. Section 1.2.6 will focus on Lisa Bortolotti's account of ways in which delusions might be irrational, because, first, self-deception may be irrational in the same ways, so her account is a nice summary and, second, I will transfer her argument to self-deception that irrationality is not a distinctive characteristic of delusion. In other words, irrationality can not serve as a demarcation criterion, which is a criterion for distinguishing a phenomenon from all the other related ones, for either delusion, or self-deception. In fact, they are similar in their irrationality. In section 1.2.7 I will then continue the comparison of delusion and self-deception. They have been claimed to be similar not only in their irrationality, but also in motivated nature and the resulting belief-like attitudes. Here again the claim, which is true for delusion, is also true for self-deception, namely that, as there are accounts arguing that delusion changes the nature of our *experiences* and not the nature of our *beliefs*, the same may be true for self-deception. In how far is this an extension in comparison to the categorization of accounts presented in table 11? Those accounts describe different kinds of epistemic agent models. 'Narration' can not only denote a mental action though, it can also denote the formation of our world- and self-model that is transparent to us in the sense that we do not experience the processing stages of the formation of such a model, but only the end result (Metzinger, 2003). Gerrans (2015) classifies narrative self-theory, according to which the function of the self-concept is to communicate one's autobiography, as one of four main theories on self-awareness (p. 3). Here 'narration' is most probably not used in the sense of an epistemic agent model construction, which would be a case if we had the feeling that what we denote as self is constructed at every moment of time by us. Then we would also feel control over our construction of our self, but it would be not clear who would be doing the construction, since the self-model, then, cannot be both the constructing and the to-be-constructed part. In short, the use of 'narration' is ambiguous with reference to the self-concept. In delusions narrative formation may refer to the construction of self-model and not to the construction of the epistemic agent model which would already have the transparent self-model as its part, namely as the one controlling one's epistemic processes. In section 1.2.7 I will refer to some accounts of delusion that regard it as changes in experience/reality and not belief-formation. The same may be true for self-deception. See for example Ying-Tung Lin (2015) in a commentary on Gerrans:

> When inconsistency and non-veridicality is detected and such certainty [about the veridicality of a mental autobiography] is lost, the mental autobiography will be modified to re-create a new *subjective reality* – a *new story about ourselves* with more or less difference. (Lin, 2015, p. 3; my emphasis)
>
> However, if the predictive coding framework is correct, the clear distinction between experience and rationalization assumed in the traditional discussion does not exist: Perception, cognition, and action are now considered continuous and highly integrated (Clark, 2013b; Hohwy & Rajan, 2012). (Lin, 2015, p. 10)

In the first quotation above, 'narration' is used in the sense of constructing a *transparent* self-concept, and not an opaque epistemic agent model. In the second quotation, predictive coding as a theory that explains our perception, cognition and action, might be interpreted not to distinguish between those. I think that adding one more nuance is useful: Yes, according to predictive coding, the explanatory tools would be the same for those three, but since there is a phenomenological difference in the experience of the transparent self-concept as such and in the experience of the transparent self-concept as an epistemic agent, e.g. a thinking agent, there has to be an explanatory difference in the predictive coding models as well. Summing up the first important point that I will come back to in chapter 2: Self-deception need not involve belief-formation, but narrative construction, not only in the sense of an epistemic model construction, but also in the more basic sense of changing the construction of the transparent self-model.

Keeping in mind these two senses of narrative formation, as well as belief-formation, the dispositional/constructionist distinction may be argued to avoid the static paradox of self-deception: instead of *stored* contradictory beliefs, there are changing constructions of beliefs whose stability stems just from the auspicious fact that our basis for belief-formation does not change that often over time (see Michel's view in 1.2.3). But if we think about us as narrative constructs (in either sense), then usually there is a *continuity* of such constructs. One can generalize that the personal level is not only characterized by *consistency*, but also by *continuity*. One of the third-person characteristics of self-deceivers is the lack of such continuity, or an inconsistency of behavior. Van Leeuwen has termed it 'flackiness' (1.2.1). Lynch has argued that self-deceptive beliefs change in the degree of uncertainty, because self-deceivers are plagued by *doubts* (1.2.2). Interesting to compare, anosognosics, who have also been argued to be self-deceived (1.2.7), exhibit inconsistency of behavior in the absence of doubt, e.g. regain insight into their illness after caloric vestibular stimulation, but loose it shortly after (3.1.2.1). Phenomenology of uneasiness and uncertainty (tension) in self-deception may be seen as reflecting not only the presence of contradictory beliefs, but to be the accompaniment of doubts. Despite tension though, self-deceivers do not relinquish their self-deceptive attitudes. An answer to this question is important and may be related to the answer to the question how such personal level diachronic inconsistency, *without* doubt, is possible in *delusions*, as a comparison. In section 1.2.5 I will categorize the role that affective states – feelings and emotions – play in self-deception (to cause self-deception, to result from inconsistent self-deceptive attitudes, to be part of the self-deceptive process and to indicate faultiness of the self-deceptive process). I will further hypothesize that self-deceptive anxiety causes doubts which themselves, then, trigger renewed hypothesis testing and postpone the answer to the question how doubts disappear to sections 2.1.1 and 2.2.3.

Embedding this section into the subsequent argumentation flow, I will first present the accounts on self-deceptive attitudes (1.2), then switch to the empirical side and present psychological experiments testing self-deception (1.3), in order to, last, present my own model of self-deception in chapter two. I will revisit the question of doubt in section 2.2.3, after discussing the role of intuitions in section 2.1.1, third- and first-person description of self-deceivers in section 2.1.2, types of selection in section 2.2.1, interaction between personal/subpersonal levels in section 1.3.4. Finally, I will discuss characteristics of self-deceptive attitudes in section 2.2.3, one of them being psychological uncertainty (=doubt), evoked by psychological and epistemic uncertainty.

1.2.1 Belief (Van Leeuwen) vs. avowal (Audi)

> Self-deception is interesting *because*, among other reasons, it forces us to ask what a belief is.
> (Van Leeuwen, 2007a, p. 436)

Van Leeuwen (2007a) argues that the self-deceptive attitude is a belief and that there is no contradictory true belief, because a mere suspicion would be enough to explain the inconsistency (p. 427). In this Van Leeuwen accepts Mele's solution (1.1.2.3). Audi argues that it is an avowal and that there is a contradictory true belief. I find the notion of an avowal, which is behaviorally inert assertion, not explanatory parsimonious, because it has been evoked to explain only self-deception, and is not explanatory useful in that it covers only cases in which the self-deceiver acts in one way and ascribes beliefs in another, but not cases in which *both* self-deceivers attitude ascriptions and behavior are inconsistent at different points in time. Since this is the case, I will focus more on Van Leeuwen's arguments in favor of the view that the self-deceptive attitude is a belief.

Van Leeuwen gives two arguments for his claim: the first one concerns belief-attribution and the second – the cognitive role of beliefs. There are two main sources for *belief-attribution*: *verbal behavior* and *action*. If both are consistent, belief attribution is easily accomplished, but as soon as verbal behavior and action become inconsistent, the matter gets complicated. What is more important, verbal behavior or action (see section 1.2.4 for more on this question)? If the inconsistency is not general, but specific to some context, the same question (of importance) should be answered in that specific context, as well as an additional question of how many such inconsistent specific contexts are allowed for the attribution of a belief. Van Leeuwen (2007a, p. 432) proposes an account for the inconsistency between verbal behavior and action – *flakiness* - which he defines as talking and acting in contradictory ways.[75]

One question is the belief attribution; the other is an appropriate account of the *cognitive role of belief*. What distinguishes beliefs from other cognitive attitudes? Van Leeuwen, who supports the view that belief is the product of self-deception, suggests the following criteria for belief:

[75] Similar view may be found in Rorty (1986): "One theory is that self-deception concerns beliefs while *akrasia* concerns actions. But self-deception can be behavioural and non-propositional […]. The manner of his actions undermines their apparent direction and intent […]. Of course that evidence will not have the force of certainty; but ascriptions of behavioural self-deception are no more unsound than are attributions of psychological states in opaque contexts" (p. 127-128).

> Motivation for the belief view was found by applying independent criteria for belief to cases of self-deception. Those criteria were (1) that beliefs cognitively govern the other cognitive attitudes and (2) belief is the default for action relative to other cognitive attitudes. (Van Leeuwen, 2007a, p. 436)

Van Leeuwen's first criterion is supposed to mean that beliefs make other cognitive attitudes possible, like hypotheses and imagining. This is the case because hypotheses are not possible without beliefs about situations that are to be explained and concept of objects are necessary for the process of imagining (Van Leeuwen, 2007a, p. 433). The claim that in order to *imagine* an apple, I have to have *beliefs* about that apple, is questionable, yet the topic of imagination is not my focus. The second criterion emphasizes the link between belief and action: actions are accomplished on the basis of certain beliefs and the latter encompass at least *several* contexts.[76] This is a dispositional view on beliefs as mental entities that are stored, retrieved and acted upon in appropriate contexts determined by the content of the given belief. In the following section I will agree with Michel (2014) that dispositionalism is not an appropriate view on self-deceptive beliefs. Despite the shortcomings that I notice in his argumentation, I agree with Van Leeuwen for the reasons mentioned at the beginning of this section that if self-deception is to be described on the folk-psychological level, its resulting attitudes should be named 'beliefs,' not least because it fulfill the folk-psychological criteria for belief ascription:

> First, the product of self-deception can play the role of belief in the causation of action. Second, the product of self-deception seems to cognitively govern the other cognitive attitudes in the manner of a belief. Third, the product of self-deception has the role of being the default for action relative to the other cognitive attitudes. (Van Leeuwen, 2007a, p. 436)

Audi (1997), as opposed to Van Leeuwen, postulate that self-deception involves a self-deceptive *avowal* or a "virtual belief" that differs from a belief by its absent/diminished *connection to action* (Van Leeuwen, 2007a, p. 426). Van Leeuwen argues that there is no independent motivation for the avowal view, except for the motivation to solve the paradox of self-deception (simultaneously believing in contradictory beliefs) and maintain the account of tension present in self-deception. If one supposes that the paradox can be solves by a deflationary account (Mele and partly Van Leeuwen), then the account of tension can be maintained by making the requirement of only an uncomfortable suspicion about truth rather than the true belief.[77] Thus, no independent motivation for the avowal view remains (Van Leeuwen, 2007a, p. 429-431). Audi is not convinced by Mele's deflationary account. According to him, sincerely avowing is a main component in believing. It catches one's eye that the present discussion is not only about self-deception, but primarily about belief.

[76] Beliefs are argued to be the default for action, because they are *context-independent*: "non-belief cognitive attitudes require specific contexts in order to function as the background of deliberation for the constitution of action […]. Beliefs, however, are the default for the constitution of action; their role in the causation of action is not limited to specific types of situation" (Van Leeuwen, 2007a, p. 434). In section 1.2.3, I will present Michel's view according to whom beliefs *are* context-dependent.

[77] True belief is used here only as an opposite to the self-deceptive belief, because, as it is recognized in the self-deception literature, even if it happened that the self-deceptive belief is true, while the other one is false (due to different circumstances), then it would still count as self-deception.

Different premises on the nature of belief exert influence on how one defines the product of self-deception:

> My positive suggestion here is that what is missing (above all) is a certain *tension* that is ordinarily represented in self-deception by an avowal of *p* (or tendency to avow *p*) *coexisting* with knowledge or at least true belief that not-*p*. […] Since sincere avowal of *p* does not entail believing *p,* I can agree with Mele that self-deception does not require having incompatible beliefs; but because **sincerely avowing *p*** (or being disposed to avow it sincerely) is a **main element in believing,** this account captures something Mele is here omitting: the apparently dissociational phenomenon of sincere avowal – "virtual belief," one might almost say – together with knowledge that things are otherwise. (Audi, 1997, p. 104; my bold emphasis)

I agree with Van Leeuwen insofar as I do not recognize the usefulness of the artificially constructed notion of avowal and, thus, do not consider this kind of account any further. In the next two sections I will discuss the question of which kind of account of beliefs – dispositional or constructivism is more appropriate for an explanation of self-deception.

Before concluding this section though, I want to briefly comment on the solution that in self-deception there are contradictory beliefs on different levels of consciousness. First of all, it is important to consider the scope of the application of the notion of belief. Frankish's (2012) argues that the term "belief" is used for non-conscious, passive behavioral dispositions ("level 1 belief") and conscious controlled commitments ("level 2 belief"). In case of self-deception, the unconscious belief is also ascribed on the basis of behavioral dispositions. Yet, there are two interpretations of the unconscious – the cognitive and psychodynamic[78] (Drayson, 2012). Lockie (2003), Billon (2011) and Marraffa (2012) argue for a psychodynamic interpretation of the unconscious in self-deception. Lockie (2003) argues for the following distinction between cognitive and psychodynamic unconscious. While the first one is just "not conscious" (p. 128), the psychodynamic unconscious possesses the following characteristics (pp. 127-130):

- ✓ person is made up of parts of some kind;
- ✓ which have their own motivational interests;
- ✓ that can remain hidden from the "person as a whole";
- ✓ these parts can actively hide knowledge from other parts, for example through deceiving (thus, the active system of interacting parts is "dynamic");
- ✓ uncovering activities of these parts may require psychological "detective work".

Lockie's (2003) argument in favor of applying the psychodynamic approach to self-deception is that the motive to self-deceive can be accomplished only by some kind of agency that has that motive (pp. 134-135). Billon (2011), grounding his account of SD on Block's (1995) distinction between P-consciousness (phenomenal) and A-consciousness (access), defends that self-deception is explicable by intentional repression: the undesired proposition is only P-conscious, but not A-conscious and, thus, can be repressed without invoking a contradictory belief-paradox. Maraffa (2012) characterizes psychoanalysis as "personal psychology that is masked as subpersonal psychology" (p. 228), but still argues

[78] McKay, Langdon, and Coltheart (2007) argue that due to the conceptual similarity bewettn delusion and self-deception, the psychodynamic, defensive interpretation of delusions is also possible: "In brief, theories of the first type [motivational in comparison to deficit theories] view delusions as serving a defensive, palliative function; as representing an attempt (however misguided) to relieve pain, tension and distress. Such theories regard delusions as providing a kind of psychological refuge or spiritual salve, and consider delusions explicable in terms of the emotional benefits they confer. This approach to theorizing about delusions has been prominently exemplified by the psychodynamic tradition with its concept of *defense*, and by the philosophical notion of *self-deception*" (p. 933).

for a modified account of dynamic unconscious in which “defense mechanisms are the very structure of the mind” (p. 237) and the fundamental self-deception is the mind’s appearance as transparent and unitary on the personal level in order to protect a person’s self-image (p. 239). Smith (2002) defines the subject of psychoanalysis to be unconscious processing of emotionally significant information which is related to conflicts and the underlying conflict arises out of the working of evolved psychological capacities (p. 526). Smith (2002) sees, further, his framework to be able to incorporate Trivers’ idea that self-deception evolved in the service of other-deceit (see section 3.2.2). I doubt the ascription of motivational interests to subpersonal units and, thus, refuse a psychodynamic interpretation of self-deception. An overly simplistic version of the conscious-unconscious distinction also leads to implausible conclusions. Nachson (1997), for example, argues that an implicit-explicit dissociation and “unawareness of one’s mode of functioning” (p. 293) is characteristic of self-deception and along these lines that prosopagnosia, conceptualized as the “dissociation between a (largely modular) face-recognition system and the (central) conscious awareness system” (pp. 294-295) is an instance of self-deception. I would agree that unconscious *representations* do exist, but would deny ascribing unconscious *beliefs*, which is a personal level attitude.

1.2.2 Belief decomposed: regarding-as-true-stances (Funkhouser) and degrees of belief (Lynch, Porcher)

In this section I will explore explanations of self-deception that still assume a dispositional account of beliefs, but provide modifications to it by decomposing the notion of belief. I will discuss Funkhouser (2009), Lynch’s (2012) and Porcher’s (2012) accounts. What they all share is the insight that the folk-psychological notion of belief reaches its limits in the explanation of self-deception.

To recapitulate Funkhouser’s[79] (2005) view (see section 1.1.2.5): He defines tension as “suspicions, coupled with avoidance behavior” (p. 300). Thus, he holds that those, who self-enhance, are self-deluded (believe what they want), if they do not possess tension regarding the matter in question (p. 303). The resulting mental state of self-deception is, for Funkhouser (2005), that “the self-deceived (falsely) believe that they believe that p” (p. 305). Thus, the self-deceiver desires “the phenomenology associated with the deceived belief” and the failure of self-knowledge would be a false higher-order belief (p. 306).

The assumption on which Funkhouser (2005) builds his account is that non-linguistic behavior is to be preferred when determining what someone believes (p. 300). In his later article on the topic, Funkhouser (2009) relinquishes this assumption. As the result, he redefines what he called self-deluded cases of self-deception as one (tensionless) kind of self-deception and calls the kind of self-deception, he is concerned with, “deeply conflicted cases of self-deception” (p. 4), thus allowing cases of self-deception without tension. He further redefines the mental state resulting from self-deception not as beliefs, but as *regarding-as-true stances* (p. 13). He (2009) differentiates between subpersonal and qualified regarding-as-true stances[80] (p. 6). While the first stance is ascribable to

[79] As has been stated above, Funkhouser (2005) agrees with Nelkin (2002) against Mele that the desire of the self-deceiver is the one to believe that p and not the desire that p (Funkhouser, 2005, p. 297). This is said to allow a unified account of self-deception for straight and twisted cases (p. 297). Moreover, it is argued to explain of tension in self-deception.

[80] *Sub-personal* regarding-as-true stances are those attributable to a subsystem of a person, for example to the visual system, while *qualified* regarding-as-true stances are attributable to the whole person, but only with respect to a certain aspect, such as a driving ability (Funkhouser, 2009, p. 6).

subsystems, the second one is ascribable to certain aspects of the persons: to their theoretical (p. 6) and practical (p. 7) reasoning, to the internal (p. 7) and external (p. 8) reports, emotion and perception (p. 8). These different regarding-as-true stances depend on the context (p. 9). They serve as belief-indicators and can stand in conflict with each other (Funkhouser, 2009, p. 6). Moreover, Funkhouser (2009) claims that "belief reduces to, or is nothing over and above, these regarding-as-true stances" (p. 9). There is no reason for a privileged weighting of one of the regarding-as-true stances in establishing the beliefs of the self-deceived[81] (pp. 10-12).

Since Funkhouser thinks of belief as constituted by regarding-as-true stances and since he states that in conflicts between belief's more specific constituents no privileged weighting is allowed, the indeterminacy of belief ascription present in self-deception indicates the limits of the application of the notion of belief (Funkhouser, 2009, p. 11): "Rather than desiring belief *simpliciter*, the self-deceived might be motivated to acquire specific values for one or more of the component regarding-as-true stances" (p. 13). Thus, according to Funkhouser (2009), self-deception does not have a common product or a common motivation; its motivation and product for deeply conflicted cases has to be inferred in every case from the conflicting regarding-as-true stances (p. 15).

There are other authors who defend similar accounts insofar as they agree with Funkhouser (2009) that the notion of belief is not fine-grained enough. Let me consider their solutions to the product problem. Lynch (2012) argues that the product of self-deception can be characterized by degrees of conviction. His account is different from Funkhouser's as he accepts the assumption that non-verbal behavior is to prioritize over verbal behavior in cases of belief ascription that Funkhouser (2009) rejects (Lynch, 2012, pp. 443-444). His argument is that the cost and gain of belief-consistent behavior is not equal: the aim of verbal behavior is the display in front of other people, while non-verbal behavior inflicts the cost of inappropriate action (pp. 444-445). Lynch (2012) differentiates between two kinds of self-deception, the deeply conflicted case which is the one where the self-deceiver "knows the truth deep down" and avoids it and the kind of self-deception where the self-deceiver is merely skeptical towards the truth to an unwarranted degree (p. 446; Pears has also made such a distinction, see section 1.1.1.2). Lynch (2012) refuses to count deeply conflicted cases as the cases of self-deception, arguing that they are cases of escapism, because

1. else the term self-deception would be ambiguous,
2. only the case of a skeptic is similar to interpersonal deception, while "(1) avoiding reflecting on and confronting an unpleasant truth that one knows about" (p. 446) is not, precluding their subsumption under the same kind of a phenomenon,
3. there is an already existing term for such avoidance cases – escapism (p. 446).

Lynch (2012) further accepts Funkhouser's distinction between behavioral (conflict between behavior and belief) and cognitive (doubts, insecurities, instability) tension and claims that "these philosophers [those postulating mental tension] think that such tension

[81] Self-deceiver's conflict need not restrict to the one between behavior and verbal report and neither of them is to favor in belief ascription: "First, I falsely assumed that the conflict would be between behavior and avowal. But, this is not always the case. We have already imagined situations, which I assume are also psychologically real, in which the conflict is between, say, *emotional responses* and theoretical reasoning (here, assume that the behavior is neutral). In order for there to be a determinate fact of the matter concerning belief there would have to be a way to settle all such conflicts among the regarding-as-true stances. Second, I no longer accept my earlier claim that non-linguistic behavior, particularly in high stakes context, trumps avowals and determines what a person really believes" (Funkhouser, 2009, p. 11). Borge is the one who focuses on emotional reponses and whose account is the focus of section 1.2.5.

is the experiential accompaniment for those cases in which behavioral tension is present or liable to occur" (p. 435). His solution to the product question that satisfies the tension requirement[82] is to speak of *degrees of conviction* in a proposition, instead of beliefs, as "[i]n a state where one's confidence level ranges between wholehearted belief and disbelief, it would be natural to expect the aforementioned tensions to appear" (p. 440). Having a belief that *p* means in this context having a high degree of confidence in *p* (Lynch 2012, p. 439). Self-deceivers, according to Lynch (2012), have unwarranted degrees of conviction in a proposition insofar as this degree deviates from those their impartial cognitive peers would have if they considered the evidence[83] (p. 439). Lynch (2012) differentiates himself from Mele insofar as Mele posits that the self-deceiver acquires the unwarranted belief, not explaining how nagging doubts and the stakes in the matter influence the mental state of the self-deceiver. According to Lynch, uncertainty and a stake in the matter lead to attempts to justify the unwarranted position that do not succeed completely:

> With regards to one's phenomenology, we could here expect "mental tension" to arise too, if this is to be understood along the lines previously mentioned ("doubts, qualms, suspicions, misgivings, and the like," "recurring and nagging doubt," etc.). It is true that uncertainty in itself does not cause mental tension. Many propositions we are uncertain about give rise to no such experience, but the idea here is that uncertainty *combined with* the fact that one has a *stake* in the issue makes for the difference between merely having doubts about something, and *feeling plagued* or *nagged* by those doubts. The self-deceiver struggles to justify and find reasons for her favored position through her biased thinking, but she does not entirely succeed in countering the unwelcome evidence to her own satisfaction. Because of the stake she has in whether p, her doubts as to whether p are a source of *worry* for her, which they would not be for someone with those same doubts but without such a stake. They plague or nag her, but not a non stakeholder in the same doxastic position. (Lynch, 2012, pp. 440-441)

Lynch emphasizes that the product of self-deception is having an unwarranted degree of conviction into a certain proposition. Yet, for his account to explain tension, understood as contradictory behavior of the self-deceivers, the degree of conviction should oscillate over time, because a *fixed* degree of conviction would not lead to contradictory behavior.

It is the oscillation of degrees of conviction that Porcher (2012) emphasizes. He (2012), like Lynch holds instability (tension) to be in need of an explanation in self-deception (p. 70). This term comprises, according to him, some kind of recognition or suspicion by the self-deceiver of the truth, as well as the presence of avoidance behavior against the avowed belief (p. 70). His solution is to conceptualize self-deception as an *in-between belief* in Schwitzgebel's sense (p. 78; see the following section for more on Schwitzgebel). If, as it is the case in self-deception, our practices of belief attribution force us to attribute contradictory beliefs to the self-deceiver, then it is our *practices of belief attribution* that are at fault and should be changed:

[82] Lynch (2012) accepts the tension requirement, because humans are generally responsive to evidence and "this is not a contingent truth about self-deceivers" (p. 442).

[83] Lynch (2012) describes the impartial observer test as follows: "S's degree of conviction in the proposition that *p* will be unwarranted if it deviates to a noteworthy degree from that which her ICPs [impartial cognitive peers] would form on the basis of considering the same information that S was acquainted with, and deviates in the direction of what S wants to be true" (p. 441). Mele (2001) proposed the impartial observer test as the one that differentiates cases of self-deception.

> This *shifting back and forth* [of the degree of confidence in the belief that p] 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. (Porcher, 2012, p. 77; my emphasis)

Summing up the three presented accounts, Funkhouser doubts that the folk-psychological notion of belief can explain the representation resulting from self-deception and opts for a lower-level explanation: regarding-as-true stances. Lynch offers an explanation that preserves the applicability of the notion of belief, but emphasizes that this notion is a continuum. Porcher agrees with Lynch and adds that it is the shifting along this continuum that explains the signs of self-deception available from the third-person perspective. I want to add to this view what I think is responsible for the oscillation, namely the recurrence of the hypothesis testing with each new piece of evidence that is available. If taken from this perspective, self-enhancement is a kind of self-deception which is characterized by a small number of hypotheses testing cycles, because it is often about domains where the criteria for the ascription of an attribute are not fixed:[84] Whether one is skilled or knowledgeable in a certain domain is pretty much subject to debate. Thus, contrary evidence is rare. In such cases as infidelity of a partner though, it is plausible to assume that the amount of contradictory evidence is greater and with it the number of hypotheses testing cycles. Of course, first, I would have to argue that self-deceptive process is that of hypothesis-testing. I will do it in section 4.5.
Interim conclusion: Whether regarding-as-true-stances is a useful notion is questionable, but particularly the idea of subpersonal beliefs should be avoided. It is a view advocated by Frankish (2009) that subpersonal beliefs differ from personal ones, because they are graded ("corresponding to subjective probability assignments"), "formed passively" and typically unconscious (pp. 103-104). I would rather avoid the application of the personal level term 'belief' to subpersonal states, as Frankish (2009) does:

> Beliefs are states that serve as premises in reasoning, and we can distinguish two broad types of belief, depending on whether the reasoning in question is subpersonal or personal. The former will be subpersonal states of the cognitive system, whereas the latter will be behavioral dispositions of the whole person. (Personal reasoning is an activity, and to have a personal-level belief is to be disposed to conduct this activity in a certain way, *taking the content of the belief as a premise*). […] Subpersonal beliefs are operative at a non-conscious level, whereas personal ones are entertained as premises in episodes of conscious reasoning. (This is not to say that we are not aware of possessing our subpersonal beliefs; we may, for example, infer their existence from our own behavior; but we do not employ them in our conscious reasoning.) (Frankish, 2009, p. 103; my emphasis)

The uprising computational account of cognition – predictive coding – also makes the same mistake as Frankish in using the term 'belief' for both personal and subpersonal states (see chapter 4). All in all, Lynch's and Porcher's recurrent oscillations in the degree of conviction might as well fit a non-dispositional, constructivist view on the nature of beliefs, which I will discuss in the next section.

[84] There is an article of Sloman, Fernbach, and Hagmayer (2010) called "Self-deception requires vagueness," whose title speaks for itself. I will describe its results in section 1.3.2.3.

1.2.3 Dispositionalism (Bach, Schwitzgebel) vs. constructivism (Michel)

> Accordingly, what matters in self-deception is not the belief that p *per se* but the occurrence of the thought that p, especially on a sustained or repeated basis.
> (Bach, 1981, p. 354)

In this section I will first briefly present Kent Bach's (1981) view that a *belief* that *p* and a *thought* that *p* might be the two attitudes present in the self-deceiver, which is the view on self-deceptive attitudes that follows directly from the dispositional view on beliefs. Then, I will present Schwitzgebel's in-between-beliefs view that allows for a more appropriate description of the mental states of the self-deceiver, but I will agree at the end with Michel (2014) that a constructivist view on beliefs should be accepted, at least in the explanation of self-deception.

Bach (1981) offers another possibility to omit the static paradox of self-deception concerning believing and not-believing *p* at the same time. He distinguishes *belief* that *p* from the *thought* that *p*,[85] basing this distinction on the two senses of belief – dispositional and occurrent accordingly (Bach, 1981, p. 354). He claims that believing that *p* does not necessary involve thinking that *p* (Bach, 1981, p. 361) and otherwise, even if normally it is the case that "if he believes that *p*, then whenever he thinks of p he thinks that *p*" (Bach, 1981, p. 355). Bach illustrates the difference between thinking and believing by the case of an indecisive thinker (thinking that *p* and that not-*p* and not knowing what to believe about p) or a phobic (Bach, 1981, p. 357). Using this distinction he states that a self-deceptive state differs from the normal belief state insofar as the tendency to think what one believes is inhibited in the self-deceived. Consequently, self-deception is for him a matter of avoidance (Bach, 1997, p. 105). Thus, the ways in which self-deception happens, according to Bach, are rationalization[86] (explaning away of the evidence; Bach, 1981, p. 358), evasion (avoidance of the thought of *p* as a means to avoid the thought that *p*; Bach, 1981, p. 360) and jamming (considering evidence to the contrary of what one believes and acting-as-if; Bach, 1981, p. 361):

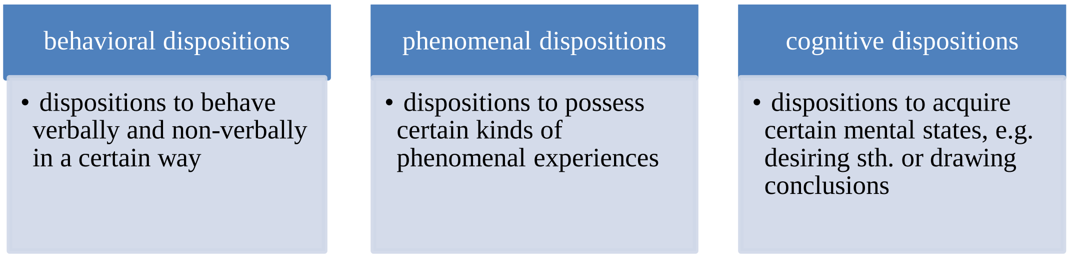
> The self-deceiver, believing that p while desiring that not-p, need not, on my view, try to get himself to believe that not-p. That is neither his objective nor essential to it. It is enough that he not (sustainedly or repeatedly) think what he believes, for what matters is what occurs to him. Self-deception need not (though it can) lead to change in belief - it is the thought that counts. (Bach, 1981, p. 357)

[85] Bach also distinguishes the *thought of p* from the *thought that p*, insofar as the thought of p includes the thought that p, the thought that not-p or the thought about indecisiveness about p (Bach, 1981, p. 354).

[86] Bach (1981) defines rationalization as follows: "In psychological contexts rationalization is understood as a person's makeshift justification of an action in terms of motives that *seem <u>to others</u> not to be his genuine* ones" (p. 358; my cursive and underscore emphasis). It is remarkable, insofar as von Hippel & Trivers' (2011) evolutionary account of self-deception states that self-deception goes undiscovered and that makes her evolutionary useful. The question whether self-deception is discovered from the third-person perspective will be discussed later.

A similar view has been offered by Frankish (1998) who argues on the basis of Dennett's distinction between beliefs and opinions (dispositions to behave in a certain way vs. occurent beliefs or episodes of active thinking) for a dual system theory of cognition: Opinions as personal attitudes that we consciously consent to supervene on low-level Bayesian graded beliefs. Clearly, Bach' recurrent thought-view, as well as the avowal view, seems to take out one of the functions of belief – verbal confirmation (internally: thinking, externally: avowal) – and considers it independently from belief, contrasting the two, though for Audi avowal is a disposition, whereas for Bach thought is an occurrence. The avowal view has been criticized in the last section, insofar as the only difference between belief and avowal – connection to action by the former but not by the latter, – cannot be maintained. If it is maintained, then the difference between belief and avowal vanishes. The same critique can be applied to Bach's recurrent thought view. Namely, Bach's view also implies that thoughts, like beliefs, actually are connected to action. This poses the question, whether the notion of a sustained and recurrent thought is really different from belief and whether it is explanatory parsimonious and useful to postulate such a notion.

A *dispositional* view that most appropriately describes the self-deceptive attitude is Schwitzgebel's in-between-view (Porcher, 2012 has argued for self-deceptive attitude to be an in-between-belief). There are three main categories of dispositional properties belonging to belief stereotypes (see figure 6):[87] behavioral, phenomenal and cognitive (or drawing conclusions from the belief in question, p. 252). The closer the fit to a stereotypical dispositional profile, the more appropriate it is to describe the person as having that belief, but no disposition is necessary or sufficient (p. 252). Vagueness and context-dependency can be incorporated by specifying excusing conditions (p. 253). Further, social interactions are dependent on us fitting other people into such stereotypical dispositional profiles (p. 262), because it makes people predictable, if we can fit them either into the stereotypical profile or a familiar pattern of deviation from the stereotypical one (p. 263). The latter includes, according to Schwitzgebel (2002), modularized believing (procedural vs. declarative knowledge), unconscious believing (and self-deception as its instance), low confidence, ignorance, partial forgetting etc (pp. 263-266). Schwitzgebel's accounts fits better than other dispositional accounts because it focuses on the practices of belief ascription more than on the postulation of additional mental attitudes.



behavioral dispositions

- dispositions to behave verbally and non-verbally in a certain way

phenomenal dispositions

- dispositions to possess certain kinds of phenomenal experiences

cognitive dispositions

- dispositions to acquire certain mental states, e.g. desiring sth. or drawing conclusions

Figure 6. Schwitzgebel: phenomenological, dispositional account of belief Distinctions from Schwitzgebel (2002).

Recently, however, Michel (2014) has proposed to view attitudes in general and beliefs in particular as evaluations that result from a context-sensitive *constructive* process, instead of the dispositional view on beliefs (p. 138, 147). He sees dispositions only as explanatory

[87] Schwitzgebel (2002) argues also that personality traits are characterized by the possession of a certain dispositional stereotype (p. 251) which is a cluster of properties associated with it (p. 250). The more agreement there is on a property belonging to a stereotype, the more central this property is (p. 251-252).

cluster-constructs[88] (p. 227, 233). This is because he rejects the distinction between attitudes as trait-like dispositions and situational evaluations as varying with context (p. 218). Attitudes are, according to Michel, *constructed* every time depending on context, instead of being trait-like stable dispositions that are *accessed*: “In the constructionist picture, what appears as a retrieval of a stored attitude is in fact an update or re-evaluation” (p. 260). Attitudes as evaluations are person-level concepts (p. 257) and still depend on prior experience (p. 253) via “mental residues” or “tendencies to evaluate” (p. 251), but there is no epistemic gap between evaluations and attitudes (p. 234). An important property of attitudes is according to Michel (2014) the transparency which maps values to attitudes (p. 146): “If the attribution of a value or set of values is sufficient for an attitude, the attitude can be called a transparent attitude” (p. 184). He counts beliefs, desires, hopes and fears as transparent attitudes. More specifically, for a true second-order belief judging *p* true is necessary and sufficient (p. 40).

Cross-contextual or intra-contextual dissociations that motivate the distinction between dispositional attitude and avowal (Michel, 2014, p. 239) are explained as cases of *attitude-shift*. Importantly, stability of an attitude across contexts is said to be the result of the stability of evaluative tendencies. The shifting, but not possession of contradictory evaluative tendencies at once, is also present in Schwitzgebel’s in-between attitudes: “She is not in-between in the sense of being subject to an in-between *state*, but in the sense of a dynamic pattern of attitude due to contradicting evaluative tendencies” (p. 247).

Similarly to Funkhouser (1.1.2.5), Michel (2014) evokes an object vs. self-directed distinction: first-order attitudes are argued to be object-directed and metacognitive, while second-order attitudes - self-directed and metarepresentational.[89] Metacognition is according to Michel enough to provide the right kind of control over our actions (p. 87). He labels the procedure of gaining self-knowledge cognitive ascent or the classification of attitudes in folk-psychological terms (p. 173):

Cognitive Ascent Model of Attitudes (CAM)
(i) *o* has *V*.
(ii) I represent *o* as having *V*.
(iii) *V* is associated with attitude-type Φ.
(iv) If I represent *o* has having *V*, I Φ *o*.
(v) I Φ *o*. (Michel 2014, p. 183)

[88] Those cluster-concepts need to be updated every time we encounter a case in which the dispositional profile has been met but the behavioral pattern does not follow: “We incorporate problematic behavioral data into a contextually fine-grained dispositional profile by adding constraints for the contextual conditions under which an attitude expresses itself in a certain behavioural pattern” (Michel, 2014, p. 231).

[89] “My attitudes are available to me via the values they ascribe by representing the evaluative relation and classifying it in folk-psychological terms. When *S* represents *o* as having *V*, first-personal metarepresentations are derivable from the object-value. When we form a second-order belief, we transform what is represented as an objective value-property into an explicit representation of the evaluative relation between ourselves and the object.” (Michel, 2014, p. 150)

Michel (2014) applies, on the one hand, Proust’s distinction between metacognition and metarepresentation and favors it over the implicit/explicit distinction (p. 82), but, on the other hand, claims object-evaluative cognition to be implicitly reflexive, but not indexical: “It is further distinctive of *S*’s visual representation of the red apple that *S* is representing the apple in a *self-related* way, namely, as lying right in front of *him* (without having to do so explicitly by indexical self-reference)” (Michel, 2014, p. 167). For Proust metacognition is indexical (see section 2.2.3).

The third and fourth steps achieve the conceptualization of an attitude in folk-psychological terms by means of background knowledge (Michel, 2014, p. 183) and an inference from (ii) to (iv) is conceptually true for *transparent* attitudes (p. 183). The aim of self-indicating attitudes via cognitive ascent is to widen the range of inferences and actions (p. 165). Michel's view is explanatorily parsimonious in that self-deceptive attitudes and other kinds of attitudes are generated in the same manner – constructed at every point in time, but in one case the context is stable and in another one it is not. What provides the context, on which attitudes are constructed – is it our model of the world or some kind of fantasized model of the world? This will be the topic of the following section.

1.2.4 Pretense (Gendler)

Michel's (2014) view might explain how self-deceptive attitudes are *constructed*, but not how they are *retained* in the face of criticism from others that leads to the need to justify oneself. Some sort of process must lead to a selective construction of reality that provides the context of the attitudes. Such selectivity is a characteristic of such processes as imagination. Recently, Tamar Szabó Gendler (2007) has developed an account of self-deception as pretense.[90] She distinguishes pretense from belief, though acknowledging that in self-deception pretense plays the role of belief (introspective vivacity and connection to action), without necessarily being misidentified as belief (Gendler, 2007, p. 249, footnote 9):

> Self-Deception as Pretense: A person who is self-deceived about not-P pretends (in the sense of *makes-believe* or *imagines* or *fantasizes*) that not-P is the case, often while believing that P is the case and not believing that not-P is the case. The pretense that not-P largely plays the role normally played by belief in terms of (1) introspective vivacity and (ii) motivation of action in a wide range of circumstances. (Gendler, 2007, pp. 233-234)

According to Gendler, her account can capture the distinctive features of self-deception: tension, instability, irrationality[91] and its belief-like manner, - while allowing to omit the static and the dynamic paradoxes of self-deception. Holding such a view presupposes a certain account of belief. Gendler holds that the crucial feature that distinguishes belief from other mental states is not the subjective vivacity (traditional account), nor the connection to action (dispositional account) (Gendler, 2007, p. 236), but the *telos of truth* (or reality-sensitivity). Moreover, she states that identifying a distinguishing mark of belief is wrong. However, I think that she comes in for the same critique, because, according to me, the telos of truth can be seen as a distinguishing mark of belief. Talking about the traditional and dispositional accounts she stresses that:

> Each identifies some sort of *mark* by which belief can be distinguished from other mental states, and then suggests that that mark is criteria of some attitude being a belief. But such a commitment misrepresents something crucial about our mental lives: namely, the numerous ways in which belief can obtain without its normal manifestations, and the numerous conditions under which other cognitive attitudes can bear them in its stead. (Gendler, 2007, p. 236)

[90] Pretense does not need to be a conscious phenomenon (Gendler, 2007, p. 253, footnote 39).

[91] Gendler gives the following definition of irrationality (referring to Gregory Currie): "it is a state where something imaginary inappropriately comes to play the cognitive role of something real" (Gendler, 2007, p. 234).

It has mostly been acknowledged in the literature that reality-sensitivity (which includes evidence-sensitivity) is a weakness of self-deception. If Gendler defines belief as having the telos of truth, then belief cannot simply be the product of self-deception for conceptual reasons. If one admits that manifestations of belief indicate neither lack of belief, nor its presence (Gendler, 2007, p. 237), in addition to the presupposition that other mental states besides belief can also play a role both in theoretical and practical reason (Gendler, 2007, p. 238), the question remains why to choose pretense from the rich repertoire of other mental states as responsible for self-deception.

Pretense is, according to Gendler, "acting as if" either in imagination, or in action. She distinguishes two arts of pretense (which, I think, mirror the two manifestations of belief – verbal statements and actions) – the imaginative and the performative pretense respectively.[92] According to her, the motivated shift of attention[93] triggers imaginative and performative pretense. Imaginative pretense gradually gains the vivacity usually present in belief. This vivacity in connection with the habitual adaptation assists the performative pretense in acquiring the central role in guiding one's actions.

That way Gendler hopes to omit the static and the dynamic paradoxes. The static paradox is omitted according to the following rationale: belief and pretense are two different attitudes and it is a tacit assumption present in the self-deception literature that two different attitudes can hold contradicting propositions without contradicting themselves. It is doubtful though that the dynamic paradox is avoided. Some underlying mechanism needs to be introduced that would explain how pretense can gradually and habitually acquire the role of belief (which can, according to Gendler, happen consciously or not).

For the sake of completeness I will illustrate how tension, instability and irrationality are explained by the pretense account of self-deception. Conceptual reasons show in which ways pretense is *irrational*:

1. Irrationality is defined by Gendler as state when imagination unsuitably plays the role of something real (Gendler, 2007, p. 234).
2. Imaginative pretense plays the role of belief which has the telos of truth.
3. Thus, pretense plays the role of something real.

Tension arises, because one attitude – projective – is used in the context of another – receptive (Gendler, 2007, p. 242), though once more (see the dynamic paradox) the mechanism that underlies this change of attitudes is yet to be understood. If receptive attitude (belief) is distinguished only through the telos of truth from the projective attitude (pretense), then it seems like a repostulation of the paradox of self-deception at another level with another termini. Self-deception is not susceptible to evidence in the same way that belief is, this is the fundamental intuition about it, but how does this explain the paradox? *Instability* is also explained by the telos of truth. Gendler distinguishes between the *topic-neutral* and *topic-specific* reason. The first one is a general reason that demands the reflection of truth in one's cognition, because in this way worldly desires could be accomplished (Gendler, 2007, p. 242). It is evidenced in the illusion of objectivity – the illusion that one has been impartial in reaching the desired conclusion, even if that

[92] Imaginative and performative pretense is respectively pretense in the sense of imagining as if something were the case and non-believingly acting as if something were the case (Gendler, 2007, pp. 239-240). According to Gendler *empathy* is required for pretense: "… imaginative pretense – also requires the capacity for successfully comprehending the possibility of a perspective other than my own, and for recognizing the role that that certain features of the world play in causing me to take the things to be one way rather than another" (Gendler, 2007, p. 240).

[93] *Selective attention* plays an important role also in other account of self-deception: that of Mele, Van Leeuwen, von Hippel & Trivers.

conclusion has been the result of motivated cognition.[94] Topic-specific reason is present in self-deception (understood by Gendler as pretense) and is reality-indifferent. Together with the available evidence topic-neutral and topic-specific reasons create tension and instability. It is them that determine the conditions under which self-deception is abandoned:

1. *motivational occlusion* – deceasing of motivation,
2. *evidential override* – the amount of evidence surpasses the threshold of reality-insensitivity, necessitating the telos of truth,
3. *trumped incentive* – a goal other than the motivation for self-deception becomes more important (Gendler, 2007, p. 243).

Michel & Newen (2010, pp. 736-737) acknowledge the existence of pretense as a phenomenon of auto-manipulation,[95] but they deny its equation to self-deception. Pretense view, according to them, cannot explain the confidence of self-deceivers to defend *p* against counter-evidence and against challenges by others or, in other words, it cannot explain the belief-like role of the product of self-deception[96] (p. 737). Triandis (2009), on the contrary, offers a definition of self-deception similar to Gendler's as "a special case of fantasy, in which individuals construct a belief on the basis of needs, wishes, hopes, or desires, rather than because it corresponds to reality" (p. xi).

Interim conclusion: Gendler holds the self-deceptive attitude to be similar to imagination. Such a comparison is useful insofar as it points out that self-deception involves a selective construction of a context. Though it is more parsimonious to assume that the context on the basis of which one type of attitudes is constructed is the same and that in the case of beliefs it is the model of the world, the *means* by which such a construction is accomplished might share some similarities with imagination in the case of self-deception, where the attentional focus might deviate. Another virtue of Gendler's account is that she takes the phenomenological level into account. Subjective vivacity is a kind of feeling that comes in varying degrees in beliefs and pretenses.

1.2.5 Emotional micro-takings (Borge)

> The self-deception literature reveals two facts. First, *there is no consensus* on how to answer these questions or what strategy to take. Second, positions that appear to have died can come back.
> (Van Leeuwen, 2007a, p. 421)

The claim that self-deception does exist, appears to have been on a number of occasions refuted only to come back again with vengeance.This is the position defended by Haight (1980) and Borge (2003). I will argue, though, that Borge's position is such that one is self-deceived about one's emotion. This is the reason for discussing his account in this section. But starting off with Haight (1980), she has argued that self-deception is in reality

[94] Illusion of objectivity is described by Ziva Kunda who is cited by Gendler in a footnote (Gendler, 2007, pp. 253- 254, footnote 42).

[95] Auto-manipulation is a more general phenomenon for Michel & Newen than self-deception and pretense, thus comprising both of them (Michel & Newen, 2010, p. 732, 736).

[96] Pretense and self-deception are argued to differ in that self-deception involves some changes of internal representations, while pretense is not: "The 'pretense'-model remains a static picture in which evidence and motivation go separated ways. **S**´s actual beliefs are not subject to motivational influence by definition. 'Pretense', if it is different from belief in any significant way, neither alters the subject´s confidence in not-p, nor does it influence confidence in p either, since p is never believed. Hence, in the 'pretense'-view, self-deception is a folk-psychological myth from the very beginning" (Michel & Newen, 2010, p. 737).

pretense:[97] "People who pretend not to know the obvious – including 'real' self-deceivers, the ones who trouble us – are typically either playing a game with other people or trying to" (Haight, 1980, p. 117).

The only reason according to Haight (1980) that the notion of self-deception exists is that it is about human's state of mind which cannot be easily inferred. Thus, it leaves space for interpretation, not least because of the outward similarity between the believer and the pretender (Haight, 1980, p. 118-119).[98] Some time ago Borge (2003) has as well argued in his work that self-deception as a phenomenon does not exist. While Haight (1980) has conceded a failure of understanding of one's internal life from an observer's point of view, Borge (2003) has insisted on inability to understand one's emotional life by the person himself (p. 2). Yet, while Borge argues that the *notion* of self-deception should be abandoned, he does not argue that there is no *phenomenon* underlying the description of self-deception (p. 1). Thus, his account can be interpreted as the one that solves self-deception by introducing the inconsistence between emotional micro-takings and beliefs of the self-deceiver.

Initially, I will shortly summarize Borge's understanding of self-deception and then his solution. The two defining features of self-deception are according to him:

> (1) The agent must somehow believe that *p* and that not-*p*.
> (2) The belief that the agent would assent to or sincerely assert is somehow the result of or is otherwise upheld in some manner by the other belief that the agent must be said to have according to (1). (Borge, 2003, p. 3)

The evidence for the first point is argued to be the behavior of self-deceiver, while the second one distinguishes a self-deceiver from "a victim of unfortunate cognitive circumstances" (p. 3). Further, Borge (2003) assumes that self-deceivers are rational to a certain degree, because he claims that "in the allegedly self-deceptive agent any conclusive counterevidence against the belief that not-*p* that leads the agent to believe that *p* would destroy the former belief […]" if it is recognized by the self-deceiver (p. 4).

Borge (2003) rejects intentionalist accounts, because he does not see how a divisionist account, like that of Davidson, could explain self-deception, because there has to be a subunit that sees "the need for the boundary and the paradox of self-deception re-emerges" (p. 3). He rejects Johnston's deflationary mental tropism account because of two reasons:

1. It is not clear how an agent could be responsible for something he did not have control over: "[…] where does the self-element of self-deception get its bite?" (p. 7).
2. Mental tropism would not lead to the conflicted pattern of behavior which is characteristic of self-deceivers (p. 9).

Borge (2003) agrees with Lazar that beliefs can be affected by emotions and develops his own account on this assumption (p. 10). It bases on the following assumptions about emotions:

- Emotions are "a special kind of cognition, which carries information" (p. 13);
- Emotional microtakings[99] (= "subdoxastic states that have cognitive content" that "cannot misrepresent, but they can nevertheless mislead and be misinterpreted") can play the role of

[97] Haight does not have the same pretense in mind that Gendler does. Gendler understands pretense as a projective attitude directed primary towards oneself, while Haight conciders pretense as deception that is directed only at others.

[98] Compare this view to von Hippel & Trivers, who state that inward deception (self-deception) has evolved because of the existing differences between the believer and the pretender.

[99] Borge (2003) defines emotional microtakings in reference to subdoxastic discriminations offered by Daniel Dennett (p. 14). He denies agreement with Dennett's theory of consciousness

unconscious motivations which can evoke the sense of uneasiness[100] if they contradict the beliefs of the self-deceiver (pp. 14-15);

- Emotions can influence other cognitive processes despite not "being a conscious part of practical reasoning leading up to our volitions" (p. 15).

Borge (2003) declares that with respect to self-deception our standpoint is equal to the observer's (p. 18). According to him, his emotional account can best explain not only the common definition of self-deception, but also the implicit phenomenological feel of uneasiness present in it (p. 19). Borge (2003) states that self-deception does not have an explicit phenomenological quality, as opposing to deception and weakness of the will (p. 11). This is the case because acknowledging the state of self-deception would undermine being self-deceived and, thus, change the experienced phenomenological quality. This does not mean that there is no phenomenology of self-deception at all, but only that there is no phenomenology of self-deception as self-deception (p. 12). He wants the notion of self-deception to be abandoned not only for reasons of simplicity,[101] but also for its connotation to blameworthiness (Borge, 2003, p. 17).

I would like to criticize Borge on a number of accounts. Firstly, it is questionable that the notion of self-deception should be abandoned. Borge's strategy of explaining this paradox is to claim that this is not a belief-contradictory belief state, but a belief-contradictory emotion state. As there have been many, like Audi or Bach for example, who have maintained the notion of self-deception and denied the "belief-contradictory belief" condition (see the belief-chapter), some additional reason has to be brought in for abandoning the notion of self-deception. Secondly, Borge's reasoning that the notion of self-deception should be abandoned because it is blameworthy (Borge, 2003, 17) does not stand criticism. It seems true that, from the *rational point* of view, self-deception may be blameworthy and best be to avoided, because it does not seek truth. However, from the *evolutionary standpoint* (see Trivers account) self-deception should be preserved to a degree in a human being. Thirdly, an emotional account of self-deception in general is not necessarily contradictory to other contemporary accounts of self-deception. Mele (2001) argues that our knowledge of the relationship between emotion and cognition is unsatisfactory (p. 100), but he does not deny that emotions play some role in self-deception. He concedes that this role may be direct[102] (Mele, 2001, p. 116) and claims that a hybrid account in which desires *and* emotions play a motivational role is compatible with his own (Mele, 2001, p. 118). This is slightly different from Borge's view, insofar as for Borge the

in whole, but reaches agreement with Dennett´s picture of multiple drafts (Borge, 2003, p. 26, footnote 27).

[100] Difference between beliefs and emotions is argued to lead to the sense of uneasiness (tension property in self-deceivers): "There can be a sense of uneasiness accompanying the decisions that are being made or the beliefs that are being formed due to the lack of awareness of certain *emotional micro-taking***,** which nevertheless exercises an important influence on us. The emotional micro-takings carry information that goes against our own belief and this jarring between these two realms of the mental can be felt by the agent in question, without ever being made fully explicit" (Borge, 2003, p. 14).

[101] Emotional cognition is argued to explain the ascription of self-deception: "The emotional cognition is responsible for there being a behavioral display that validates attributing a belief to the agent that is in conflict with what the agent would consciously and verbally assent to, where these beliefs are connected in a way that makes us prone to say that the agent assents to *p* because of his other belief that not-*p*. On grounds of theoretical economy the notion of 'self-deception' ought to be abandoned" (Borge, 2003, p. 17).

[102] Mele also acknowledges that emotions might possess the (single) biasing role in self-deception: "If, as I suggest, an emotion can play a direct biasing role in self-deception, the door is open to the possibility that an emotion may contribute to an instance of self-deception that has *no* desires as significant causes" (Mele, 2001, p. 117).

role of emotions is not only the biasing one, but that emotions are those attitudes standing in conflict with beliefs and by virtue of this explaining tension. Yet, the point is that if the notion of self-deception is compatible with it being about emotion and not necessarily belief.

Borge’s view offers a nice bridge for the discussion of the roles of emotion in self-deception. This question is important to clarify the explanatory weight that emotions have for self-deception. The following roles of emotions in self-deception have been argued for:

- Cause/motivation (as desires) (e.g., Mele, 2000)
- Phenomenological accompaniment or tension (e.g., Noordhof, 2003)
- Functional role, e.g. reduce anxiety (e.g., Johnston, 1988; Barnes, 1997)
- Content about which one self-deceives (e.g., Borge, 2003; Damm, 2011)
- Mechanism of self-deception: emotional coherence satisfaction (Sahdra & Thagard, 2003)

In section 1.1.2.1 I have already introduced the phenomenological vicious circle of anxiety. This consists in anxiety *causing* and *resulting from* self-deception. Causing, because unfavorable evidence might be aversive and lead to the acquisition of the self-deceptive belief, but if the true belief is also in the belief set, then there will be anxiety (tension) resulting from the inconsistency between these two beliefs (or other kinds of attitudes). The ascription to self-deception of the functional role to reduce anxiety implies according to me that anxiety is also, at least partly, the cause/motivation for self-deception, but also that self-deception cannot result in anxiety. Thus, at some point there should be no tension, else the condition for the success of self-deception would not been achieved. An implication of such a devious circle is that anxiety cannot be the cause and the result of self-deception at the same time. In such a form the implication is too strong though, since one could think of at least two ways in which anxiety can be kept: the first is that the resulting anxiety is *misattributed* not to be related to self-deception (see section 3.1.1), the second is that anxiety is only *temporally* relieved for short periods in time in order to return again. On the assumption that constant anxiety is psychologically unhealthy, misattribution should also result in a temporally relief of anxiety. The interesting question is exactly how such relief is accomplished. One should note that such cause-effect relationship can in the case of self-deception be constructed not only regarding self-deceptive *attitudes*: anxiety triggering acquisition of self-deceptive attitude or anxiety being alleviated by a self-deceptive attitude. On the premise that each attitude possesses not only content, but also valence (and maybe intensity), self-deceptive process can be argued to consist in valence of attitudes influencing the kind of attitude that is acquired (Sahdra & Thagard, 2003). Here, an affective state would change the process itself. Or an affective state (e.g., feelings or emotions, because both possess valence and intensity) can be the result of a certain process in order to indicate the appropriateness of the latter (Proust, 2015b).

To sum up, affective states, which are those possessing valence and intensity, e.g. feelings and emotions, may either cause a self-deceptive process as such, be triggered by a self-deceptive attitude in virtue of its inconsistency with some other attitude, be part of the self-deceptive process itself, or indicate the rightfulness of the self-deceptive process. In the latter case, like in the cause of attitudes, the question is why the self-deceiver does not relinquish self-deception despite being phenomenally aware of the fact that the process has been faulty. Faultiness can be misattributed, but what anxiety triggers is *doubt* (see Lynch’s view in section 1.2.2).

Doubt is the absence of *psychological* certainty which arises, when the subject is convinced in the truth of some proposition (Baron, 2011). Psychological uncertainty may have its roots in *epistemic* certainty, which reflects the epistemic status of attitudes (for the latter see Baron, 2011). In self-deception literature, those two have not been clearly distinguished yet, probably because of the assumption that epistemic certainty (justification of beliefs) would also lead to psychological certainty. My reason for imposing such kind of assumption on theories of self-deception is precisely the emphasis of tension as a characteristic of self-deception. If inconsistencies in the belief set of self-deceivers lead to felt anxiety, then this has to be the case in virtue of some sort of connection between the epistemic status of self-deceivers' beliefs and his acknowledgement of this epistemic status. If ever, this assumption can be set only for *conscious* decisions (Vlassova, Donkin, & Pearson, 2014, p. 16217). The psychological-epistemic certainty assumption shares with the assumption, that contradictory attitudes would evoke tension, that a certain relationship between attitudes on the personal level has been assumed, namely that personal level attitudes are connected. I think that the function of doubt is to trigger renewed sampling of evidence in order to either affirm or disconfirm the proposition that one is in doubt of. Contradictory behavior of self-deceivers might stem from different cycles of hypothesis testing so that at one point in time the proposition was affirmed and at later points in time – disconfirmed. To sum up, self-deceptive anxiety should evoke doubt, which is psychological uncertainty because of epistemic uncertainty regarding the self-deceptive attitude and which would lead to repeated hypothesis testing.

Interim conclusion: Borge, like Gendler, lays more weight on the phenomenal level – granted that emotions *are* parts of this level. I have categorized the role of affective states – feelings and emotions – into four types: causing self-deceptive attitudes and resulting from (contradictory) self-deceptive attitudes, being part of the self-deceptive process or on a meta-level indicating that the self-deceptive process violates belief-forming criteria usually employed (for the latter see Proust, 2013 in 2.1.3). It is to be stressed that those roles of affective states may allow different kinds and degrees of *control*. In all these cases, the role of anxiety is to evoke doubt and to trigger renewed hypothesis testing. The interesting question is when the cycle is broken so that no anxiety is generated by the hypothesis testing process. I will answer this question at the end of the following chapter, after I have provided a more precise description of the phenomenology of self-deceiver in 2.1.2 and types of selection, that could be employed in self-deception, in section 2.1.1, as well as the relation between the personal and subpersonal level in self-deception in section 1.3.4.

1.2.6 Self-deceptive process: belief formation or narrative construction?

The basic assumption on which this debate centers is that our beliefs determine our actions. Thus, given that self-deceivers exhibit inconsistent behavior, there should be inconsistent beliefs, or at least other inconsistent attitudes that fulfill a belief-like role in triggering action. The two basic kinds of solutions are either postulating that the contradictory beliefs are on different levels (conscious vs. unconscious) or postulating new kinds of attitudes that fulfill the belief-like role.[103] The proposed attitudes to fulfill the belief-like role are avowal (Audi, 1997), sustained and recurrent thought (Bach, 1981) and pretense (Gendler, 2007). Yanal (2007) entertains a similar idea to Gendler's and Bach's insofar as he argues for the analogy between self-deception and fiction and explains self-deception by arguing that a certain belief is inactivated (pp. 118-119). While occurrent beliefs are those that the person is at the moment thinking about, *active beliefs* are those that have effects on the rest

[103] I have not placed emphasis in this section on the discussion of the question which of the attitudes – belief or the other one – is that which is supported by the evidence and which is not, as well as on the view that does not explain tension, namely that there is only one belief, either "true" or "false". For reference see Nicholson (2007, p. 48), Van Leeuwen (2007a).

of a person's mental and somatic states (for example by causing intense anxiety; pp. 116-117). Yanal (2007) holds the motive of the self-deceiver to be the maintenance of the positive self-concept.

From the solutions that postulate two different kinds of attitudes with inconsistent content the account of Gendler is most useful since it makes the connection between self-deception and imagination, from which it is just a stone's throw to the connection between self-deception and the experience of realness that I will consider in the following chapter. Due to the inconsistency of different attitudes, tension or the phenomenology of uneasiness is said to be explained. I doubt the success of such kinds of solution, though, in the explanation of tension. On the one hand, contradictory attitudes are said to coexist, precisely because they are different kinds of attitudes. On the other hand, they are said to explain tension, because their content is contradictory. Such an argumentation is inconsistent, because if the contents of attitudes clash, then the attitudes also do so and vice versa. There has also been a proposal to decompose the notion of belief into regarding-as-true stances (Funkhouser, 2005), as well as the one to emphasize the oscillation of degrees of confidence in a belief (Lynch, 2012; Pedrini, 2012). The latter solution emphasizes the *dynamic* aspect of the available evidence or that amount and kind of processed evidence by the self-deceiver changes over time and leads to changes in her belief-system. The limits of folk-psychological notion of belief is demonstrated by these accounts. Thomason's (2010) metaphor of comparing beliefs with a gentleman's tailor makes vivid the restriction of a general-purpose attitude of beliefs. According to Thomason, beliefs are tailor-made in the sense of being fitted to every occasion, with reasoning being a manufacturing process. In a similar vein, Michel (2014) proposes a constructionist view on beliefs. I think that it is a useful perspective on self-deceptive attitudes, yet what needs to be emphasized is that viewing self-deception as a process of personal level belief-formation that results in irrational beliefs will not aid in distinguishing it from other irrational phenomena (demarcation problem; see section 1.1.3).

I will use Lisa Bortolotti's (2010) claim that delusions are *beliefs* despite the deficits in procedural, epistemic and agentic rationality (see table 13) in order to transfer her claim that delusions cannot be demarcated by their irrationality to the claim that self-deceptive attitudes also cannot be demarcated by their irrationality. Bortolotti's description of the ways in which deluded are procedurally, epistemically and agentively irrational can also be transferred to self-deception and, thus, the summary of her account will serve a further aim to summarize the kinds of irrationality that may also characterize self-deception. Her line of reasoning is that if other kinds of irrational beliefs (that we still call beliefs) show qualitatively similar deficits (p. 77), then both delusions and other kinds of beliefs may be irrational ("the claim that humans are rational is a philosophical myth," Bortolotti, 2010, p. 7), but still are beliefs. This is because belief-ascription codetermines the nature of beliefs (p. 2) and mostly rational behavior is enough for belief ascription of also belief-like states violating some rational standards (p. 20, 63). Importanlty, she argues that it is not irrationality that can distinguish delusions from other kinds of beliefs (p. 22, 242). The function of explaining and predicting behavior is the same in ascription of delusions and other irrational beliefs (p. 8). The consequence is that the consideration of the doxastic dimension is not enough to explain delusion (p. 27), e.g. phenomenology, disruptive effects on health and life are also important (pp. 24-27). Most interesting is here Bortolotti's (2010) thesis that, since we often do not form attitudes via *explicit deliberation*[104] (p. 183) and

[104] That we might have attitudes that have not been acquired by the construction of an epistemic agent model will be important for the dolphin model of cognition (see section 2.2.2.3): "The

reason-giving, prone to change the attitude (p. 214), what it is useful for, is the construction of a *coherent self-narrative* (she cites Gerrans in this respect; p. 221, 226).

Philip Gerrans (2013) argues that the mechanism responsible for the generation of delusions is the default cognitive processing which is unsupervised by decontextualized cognitive processing (p. 84). The main argument in favor of the doxastic approach to delusion (which states that delusions are beliefs) is that delusions occupy the functional role of belief,[105] while the contra-argument focuses on the ambivalence and subjectivity of delusions (Gerrans, 2013, pp. 84-85). The *ambivalence* can take an explicit or a tacit form which delusions have in common with self-deception:

> *Tacit ambivalence* is manifest in behaviour. It can take the form of defensiveness, evasiveness or confabulatory rationalisation provoked by requests for justification or, in some cases, a compartmentalising or partitioning of delusion from disconfirming evidence and background knowledge. In these respects delusions resemble other attitudes such as motivated irrational beliefs and self-deception that are problematic for doxastic theories. (Gerrans, 2013, p. 85)

The second non-doxastic attribute of delusion – *subjectivity* - accounts the fact that instances of delusions do not undergo testing according to intersubjective standards of evidence and argument[106] (Gerrans, 2013, p. 85). The ambivalence and subjectivity of delusions are according to Gerrans personal level redescriptions[107] (p. 87). Their vindication is, according to the author, only possible in terms of the cognitive architecture (p. 87) – in terms of the anticorrelation between the default mode network (ventromedial circuitry) and decontextualized cognitive processing (dorsolateral circuitry)[108] (pp. 90-91). Gerrans (2013) holds the first kind of circuitry to enable simulations of "affective and motivational response *in the absence of the stimulus*" (pp. 90-91), while the second one to enable consistency and coherence among representations.[109]

upshot is that there is not *agential route* to first-person authority: reason-searching is merely a heuristic and not a deliberative process, given that the content of the reported attitudes is established independent of the process of reason-giving and is not affected by it" (Bortolotti, 2010, p. 220).

[105] The functional role of belief is to think and act according to that belief: "The debate between doxastic and non-doxastic theorists starts from the fact that delusions are generated in response to perceptual and sensory information and combine with other psychological states in an intelligible though irrational (by prescriptive standards) way, to cause behaviour" (Gerrans, 2013, p. 84).

[106] Gerrans (2013) understands Kant to say that irrationality is ignorance of intersubjective standards: "Kant makes the point that compromised reality testing does not merely render the delusional patient or dreamer irrational but, since rationality consists in the application of intersubjective standards, *subjective*" (p. 86).

[107] From the point of view of the subject, an intersubjective criterion is difficult to establish though: "Compromised 'reality testing', 'changed framework propositions' and the idea that delusions represent a distinctively 'subjective attitude' to experience are perspicuous redescriptions. The point is not that these descriptions are wrong. At the folk psychological level at which they are pitched, deciding between them is really more a matter of emphasis on different aspects of the phenomenon combined with pretheoretical commitment to assumptions about the nature of personal level explanation (Functionalist, Wittgensteinian, Phenomenological)" (Gerrans, 2013, p. 87).

[108] Decontextualized processing as in opposition to affective simulation: "What they [patients with lesions to ventromedial structures] have lost is the ability, provided by ventromedial structures to simulate affective and motivational response *in the absence of the stimulus* while they retain the ability to process information in a decontextualized way" (Gerrans, 2013, pp. 90-91).

[109] Decontextualized processing is argued to be a control structure testing for consistency: "Whether in fact a default thought can be tested for consistency and coherence however

Personal level explanation	
Delusion: doxastic vs. non-doxastic	
Functional role of belief	ambivalence, subjectivity (irrationality)
Cognitive architecture and its characteristics	
Decontextualized cognitive processing	consistency, coherence
Default network ("densely connected to motivational and affective system")	stimulus-independent, subjective adequacy
Reflexive system	automatic, inflexible, stimulus-independent

Table 12. Gerrans: hierarchical cognitive architecture
Distinctions from Gerrans (2013).

Dreams and delusions are, according to Gerrans (2014), similar insofar as both are characterized by deficient "reality testing" (detection and correction of misrepresentations). The argument is that DMN – active in dreams, as well as in delusions, but also in daydreaming and mind wandering – provides the narrative context and allows us to simulate alternative personal histories that include emotional and motivational states. Thus, in Gerrans' words, DMN is responsible for goal-directed evaluation of actual and possible experiences. The question for him is then, why the given misrepresentation is justified by experience, but decoupled from the wider context of other background representations possessed by the subject. His answer is that the activity of DMN that is not supervised by decontextualized processing leads to delusion/dreaming being treated as a *narrative element*, instead of the hypothesis that can be confirmed or disconfirmed by the evidence. Where Bortolotti (2010) says to disagree with Gerrans is in the relationship between hypothesis testing and construction of a narrative: for her, both cannot be distinguished from each other that easily. Further, revisions of beliefs depend on their *emotional significance* (p. 263).

> I see as entirely sympathetic to Gerrans' proposal but perhaps occupying a different position in the dialectic, would be that *for some thoughts* there is no clear-cut distinction between the stance of the scientist weighing up possibilities on the basis of past evidence, and the stance of the autobiographer trying (at the same time) to provide an accurate description of significant life events and to integrate these events into a coherent narrative. (Bortolotti, 2010, p. 250)

> It is important to distinguish between the processes of *hypothesis generation and evaluation* that can happen at the subpersonal, pre-conscious level and the behaviours typical of subjects with delusions that are manifested in the subjects' attachment to the delusional state and in their level of *commitment* to the content of the delusion (these are the features of endorsement we discussed with respect to the notion of agential rationality). (Bortolotti 2010, p. 34; my emphasis)

Bortolotti's ideas about the belief-forming process might also be applied to explanations of self-deception (see table 13). First, self-deception has also been characterized as procedurally, epistemically and agentially irrational in the literature. Self-deceivers are *procedurally irrational* in that the self-deceptive attitude is not well-integrated with other attitudes.[110] Even more, not only acquisition, but justification of self-deception is puzzling.

depends on whether the required decontextualized processing can be activated." (Gerrans, 2013, p. 92)

[110] To conclude that delusions are similarly procedurally irrational to other beliefs, Bortolotti (2010) argues against two claims, namely that (1) procedural irrationality precludes belief

Self-deceivers belong to those non-deluded people whom Bortolotti claims not to be *always disposed to restore well-integration* when they are made aware of the irrationality[111] (p. 95). Bortolotti (2010) argues that rationality as following *epistemic norms* is to be distinguished from rationality as *intelligibility* (p. 99). Only *synchronic* inconsistency precludes intelligibility (p. 101). This is reminiscent of Davidson's *paradox of irrationality* that explanations of self-deception are susceptible to (1.1.1.1): If we can make the behavior intelligible, there would be no irrationality anymore. Are self-deceivers irrational or is it the scarcity of the relevant information, available to the interpreter that makes applications of heuristics in explaining behavior difficult[112] (that the latter is the case see Bortolotti, 2010, p. 98, 102)?

Procedural rationality	**Epistemic rationality**	**Agential rationality**
Well-integrated with other beliefs	Supported and responsive to evidence	Supported by reasons and acted upon
Procedural irrationality	**Epistemic irrationality**	**Agential irrationality**
Bad integration: 1. Bad integration that is revised if discovered. 2. Bad integration that is *explicable* and can be excused (pp. 90-91).	1. Lack of support/insufficient evidence (feature of belief acquisition) 2. Unresponsiveness to evidence (feature of belief maintenance)	1. Failure of action guidance[113] 2. Failure of reason-giving **Two kinds of reason-giving:** - Deliberation (during belief formation) - Justification (if questioned)

Table 13. Bortolotti: Rational features in beliefs.
Modified from Bortolotti (2010, p. 16, 56).

Self-deceivers are also *epistemically irrational* in their handling of the evidence. Bortolotti (2010) claims that it is difficult to distinguish between *competence* (mechanism) and *performance* (output) failures, not least in the case of delusions (p. 129), and that delusions are characterized by similar *biases* as normal cognition: "asymmetries between 'me' and others', and the use of double standards or self-serving biases" (p. 138). This is reminiscent of Noordhof's argument against Mele's theory (1.1.2.3) that if human reasoning is generally bias-prone, then biases belong to the standards of theoretical reasoning. Also with respect to self-deception the question might be asked whether it is a result of normal or abnormal functioning of belief-forming processes. I particularly like Bortolotti's analogy

ascription to deluded and that (2) procedural irrationality precludes belief ascription with a *determinate* content.

[111] Bortolotti (2010) also denies that compartmentalization is a useful solution when there is no emotional or personal investment (p. 88): "People who endorse conflicting attitudes in reasoning tasks are fiercely resistant to reviewing their answers even after they have been explained the principles of reasoning relevant to the tasks. When the attention of the subjects is drawn to the conflict between the attitudes they have endorsed and they do not revise their attitudes even if they have *no particular emotional or personal investment* in the tasks, neither of the ready 'excuses' for bad integration seem to justify their procedurally irrational behavior" (Bortolotti, 2010, p. 90; my emphasis).

[112] Interpretation of behavior proceeds on the restricted and idiosyncratic knowledge of the interpreter: "When we interpret others we use our background knowledge constituted by some folk-psychological generalizations and by what we know about the subject to be interpreted and the surrounding environment" (Bortolotti, 2010, p. 106).

[113] According to Bortolotti (2010), self-deception is not restricted to a contradiction between behavior and assent in virtue of the influence that assent has on behavior: "Even when the policies we verbally commit to […] are confabulatory reconstructions with limited evidential bases, they do affect our behavioural dispositions and ultimately our actions – not all our actions are hostage to unconscious and passive states" (p. 170).

that links delusional content and content of certain visual illusions. The analogy consists in being *encapsulated* in the same way (p. 121):

> The experience of recognising the voice of the spouse on the phone, but not recognising it when the spouse is in front of them, should have an effect on the Capgras patient that is similar to the effect of viewing the straw out of the glass after having been subject to the perceptual illusion that the straw was bent or broken. (Bortolotti, 2010, p. 123)

This analogy makes vivid our puzzlement also with self-deceivers that the evidence is right at hand, but is still ignored by the self-deceiver (and it also shows that delusion and self-deception share many similar characteristics, for more see section 1.2.7). Self-deceivers are finally *agentially irrational*. Endorsement of seemingly inconsistent attitudes is procedurally irrational, while inconsistency between attitudes and behavior is agentially irrational (Bortolotti, 2010, p. 162). As we have seen, these two kinds of inconsistencies are often emphasized with respect to self-deception. Self-deceivers may even be not only agentially irrational, but also fail to acquire authorship over the self-deceptive attitude. Bortolotti's distinction between these two kinds of endorsements is the following:[114]

- *Authorship*: endorsing contents of a belief on the basis of one's reasons that one recognizes to be one's *best* reasons[115] (p. 180).
- *Agential rationality*: endorsing contents of a belief on the basis of intersubjectively good reasons (pp. 176-177).

In the light of all these flaws in the belief-forming process, that are ascribed to the self-deceiver, and particularly in agreement with Bortolotti (2010), that the formation of attitudes via explicit deliberation might not be as ubiquitous as our rational ideals might entice us to think,[116] I find the application to self-deception of Bortolotti's claim worth

[114] According to Bortolotti (2010), failures of authorship, but not those of agential rationality, undermine self-knowledge (p. 210).

[115] How can the relationship between *reasons (personal level)* and *mechanisms (subpersonal level)* be characterized? According to Bortolotti, those two have to coincide, in order for authorship of the attitude to occur: "This account of attempted justification where the reasons for the reported attitude do not necessarily coincide with the psychological causes of its formation, does not compromise authorship of the attitude, unless the subject's reasons are not what she takes to be her best reasons." (Bortolotti, 2010, p. 187) In the same vein, Bortolotti (2010) describes Nistett & Wilson's experiment that demonstrated that subject choose right-hand item out of identical items on a display and justify their choice: "Notice that this is an excellent example, because the charge of agential irrationality can be made on the basis of two distinct versions of agential rationality. People endorse their preferences on the basis of reasons that have nothing to do with the psychological mechanism that is responsible for those preferences; moreover, they provide reasons for their attitudes that are not intersubjectively good reasons." (p. 199)

[116] Bermúdez (2000a) holds that personal level explanations come at a cost, if inferential rules that govern them – those of rationality or consistency – are not taken into account[116] (p. 71). Reasoning abilities of humans differ from demands of rationality and, thus, the latter cannot predict the former[116] (p. 72). Bermúdez (2000a) proposes three strategies to deal with this problem:

1. view agents as striving to satisfy the ideals of rationality,
2. revise rationality in terms of "bounded rationality",
3. abandon the idea that norms of rationality can fulfill the explanatory/predictive function and accept that that function is prescriptive (Bermúdez, 2000a, p. 72).

He criticizes the first by claiming that human reasoning strategies are in the most cases not applications of rationality techniques at all. Bermúdez (2000a) also criticizes the second, because human reasoning still can live up to rational standards (p. 75). Thus, he approves the third strategy and gives an example of skilled behavior as the one in which personal level explanations are not fine-grained enough and subpersonal states make a part of personal level explanations (p. 79): "No explanation proceeding purely in terms of personal level events will considering, that hypothesis testing and narrative construction might not differ. Most researchers on self-deception explain it as a kind of aberrant belief formation (e.g., Davidson, Mele), some distance from hypothesis testing and see self-deception as a kind of narrative construction (Fingarette, Brown), but those two may be just two kinds of description for the process that is not available at the personal level. Explanations of self-deception as personal level belief-forming processes have been found wanting. Those in terms of narrative construction have also been applied to delusion. Thus, if self-deception involves narrative construction, the latter is not its distinctive feature. Narrative construction is susceptible to Davidson's paradox of irrationality, as belief forming is: if I can explain my utterances and behavior in a coherent manner, then those are not paradoxical or contradictory anymore. A certain paradox of self-deceptive narrative construction can also be made: if beliefs are constructed at every moment in time on the basis of the current (selected) view of reality (see section 1.2.3) and in the case of self-deception alternate *a lot* (see also section 1.2.2), why does the self-deceiver not notice the oscillations of his views in the self-narrative that connects his attitudes in time? I will make a more thorough comparison between delusion and self-deception in the following section and explore the possibility of the application of the non-doxastic conception of delusion to self-deception.

1.2.7 Non-doxastic conception of self-deception

In the previous section I transferred Bortolotti's (2010) argument for delusion to self-deception: Irrationality of self-deception cannot distinguish it from other irrational phenomena. In this section I will substantiate this claim by making a comparison between wishful thinking, optimism and delusion in folk-psychological terms. After that I will explore the non-doxastic explanation of delusion and the possibility of its transference to self-deception.

Wishful thinking has been often compared to self-deception with inconclusive results (see section 1.1). Triandis (2009) argues that wishful thinking is the "essense of self-deception" (p. 31), but Scott-Kakures (2002) stands for an account that would distinguish between wishful thinking and self-deception (p. 591). Those who make a distinction between the two often ground it in the relationship to evidence: "The wishful thinker believes on too little evidence, the self-deceiver 'in the teeth of the evidence'" (Haight, 1980, p. 1). This is also the view that Szabados (1973) argues for. Some, though, argue for a distinction between the two in terms of different desires, e.g. Bermúdez (2000b) argues that wishful thinking is motivated by the desire that *p*, while self-deceptive motivation also contains the desire to believe that *p* (p. 312). But there are also those who argue against the difference between self-deception and wishful thinking (Mele, 2001). Mele (2001) states that insofar as wishful thinking is understood as wishful believing ("motivationally biased, false believing"), the two might overlap (pp. 73-74). Thus, a distinction between wishful thinking and self-deception as two irrational beliefs is difficult to make. I do not think that a distinction between wishful thinking and self-deception can be drawn sharply in the light of both having converged over time to denote special cases of motivated cognition against the evidence. Recently Smith (2014) argued that the difference between self-deception and

be sufficiently fine-grained to explain why that shot was played then — to explain why my opponent extended his racket precisely *that* distance at precisely that *angle"* (p. 79).

wishful thinking is in that the latter has no property of success-aptness, it is not something that one can succeed or fail to do:

> Suppose that Lois notices a lump in her breast. Although initially concerned that she may have cancer, she comes to believe, in defiance of her own epistemic standards, that the lump is nothing to worry about and that she need not consult a physician about it. Lois adopted a belief that she regarded as untrue. She wanted it to be the case that the lump was no cause for concern, and she somehow made herself believe this. Put differently, Lois *succeeded in misbelieving* that the lump was nothing to worry about, and if she not acquired this misbelief she would have *failed at misbelieving* it. Now, contrast this with a case of wishful thinking. Lois badly wants Peter to phone her. The phone rings, and when Lois answers it she momentarily mistakes the caller for Peter. In this example, unlike the previous one, it would be inappropriate to say that Lois succeeded in misbelieving that the caller was Peter. And if she had identified the caller correctly from the start, it would be peculiar to say that she had failed at misbelieving that he was Peter. (Smith, 2014, p. 186)

In the example above, self-deception is a case in which there is some reasoning going on, while wishful thinking is something where we become aware only of the result, not the process and, thus, there is no epistemic agent model in the latter. I think that the intuition, determining this version of the distinction, is the one underlying intentionalist positions, namely, that self-deception is an activity of the agent with success conditions, while wishful thinking is something that merely happens to us. Intentionalist positions are hard to maintain though (see section 1.1.3 for the summary of the criticisms applicable to both intentionalist and deflationary positions). Further, as I mentioned in the introduction to section 1.2 and particularly in the light of the denial of pregnancy example at the end of section 1.1.3, self-deception need not be only about the construction of a certain epistemic agent model, but may result in an aberrant world- or self-model, as it has already been claimed for delusions. After briefly discussing unrealistic optimism as a supposed case of self-deception, I will compare delusions and self-deception in greater length to substantiate the above transference claim.

Unrealistic optimism is one of the positive illusions, along with self-enhancement and illusion of control, argued to be self-deceptive by Taylor (1989; see section 3.1.3.1). As wishful thinking, optimism can be described in similar terms to self-deception as bias maintained in the face of counter-evidence: “an optimism bias is maintained *in the face of disconfirming* evidence because people update their beliefs more in response to positive information about the future than to negative information about the future” (Sharot, 2011, p. R943; my emphasis; compare to Ditto’s QOP). Yet, if the extent of optimism correlates with the degree of *uncertainty* (Johnson & Fowler, 2011; Sharot, 2011), then it is unclear how it could still be against the evidence (this is also a problem for Sloman et al’s study that will be discussed in 1.3.2.3). It is also unclear whether the uncertainty is reflected in the phenomenology of optimism, as it is in the phenomenology of self-deception. It may be the case that, while in the case of optimism the *circumstances* may be uncertain, in the case of self-deception the circumstances point to a certain unfavorable conclusion and its refutation is reflected in the uncertainty in *phenomenology*. Thus, the classification of optimism as self-deceptive hinges on the evaluation of evidence as sufficiently against the optimistic belief that has been acquired. A characteristic that precludes the identification between self-deception and optimism is that the scope of the latter might be only future-directed, if optimism is defined as the difference between the expectations and the outcome (Sharot, 2011). The future-oriented character links optimism to a self-fulfilling prophecy (Johnson & Fowler, 2011). Now, the decisive question is: Can a self-fulfilling prophecy be a kind of self-deception, or is the contradictory evidence in optimism enough to warrant the label 'self-deception'? Moreover, a meta-analysis conducted by Krizan & Windschitl (2009) indicate that the evidence in favor of optimism being motivated (=optimism being desire-driven, a phenomenon called by the authors wishful thinking) is rather limited in the cases in which outcomes are uncontrollable. Stankov & Lee (2014) similarly advocate that overconfidence (difference between confidence in solving a task and the actual accuracy afterwards) is a result of cognitive, but not motivational bias in the cases where the outcome was controllable: Participants had to guess the next number for a sequence of numbers and the statistical analysis indicated that it is accuracy that negatively correlated with overconfidence, but not confidence. According to the rationale that correlation between overconfidence and accuracy would imply the dependence of overconfidence on cognitive abilities (p. 10), the authors concluded that overconfidence is a cognitive phenomenon ("people lack competence to evaluate their own incompetence," Stankov & Lee, 2014, p. 3). It is also interesting that across world regions, confidence was comparably similar, so that overconfidence resulted from differences in accuracy scores at solving the task (p. 13). All in all, as in the case of wishful thinking, it is difficult to distinguish optimism and self-deception by the degree of contradictory evidence. As a consequence, Bortolotti's (2010) claim that delusions cannot be differentiated from other irrational beliefs by irrationality can also be applied to distinctions between other kinds of irrational beliefs, or, in this case, the impossibility to make a distinction along this criterion.

In the following I will show how similar delusions[117] and self-deception are, not only in that they share irrationality as a feature. Three properties of delusions have been argued to be applicable in the case of self-deception (see Bayne & Fernández, 2009), namely that delusions are pathologies of beliefs (see section 1.1.1 for the discussion in case of self-deception), that affect & motivation might influences them (for self-deception see section 1.1.2) and also that products of delusion might be belief-like states that are not beliefs (which has been explored in section 1.2.1 – 1.2.5).

First, self-deception and delusion have been argued to be *pathologies of belief*, since they might violate standards of epistemic rationality and, possibly, mechanisms of belief formation (Bayne & Fernández, 2009, pp. 3-7). The difference between self-deception and delusion with respect to *epistemic rationality* (evidence-sensitivity) has been argued to be one of degree, with delusion being at the extreme end of the continuum (the amount of contradictory evidence is almost insurmountable) and self-deception at a more moderate position along the continuum.[118] Yet, persistence to the contrary in the face of the available adverse evidence might also more generally be seen as a characteristic of self-validating belief systems to which self-deception and delusion belong.[119] As with the violations of

[117] In the (English) literature often the first definition of delusion is the one taken from the *Diagnostic and Statistical Manual of Mental Disorders* (as done by e.g. McKay, Langdon & Coltheart, 2009; Bayne & Fernández, 2009; Mele, 2009). In it delusions are characterized as false beliefs despite evidence to the contrary. Frith & Friston (2013) draws a distinction between delusions and hallucinations, insofar as delusions are false beliefs and hallucinations – false perceptions.

[118] Both can be seen as ordinary erroneous judgements that differ only in the degree of being ordinary: "Together with the heterogeneity to be found among delusional states, centuries of using the aberrant thinking of those who are delusional to marginalize them should encourage us to at least start with inclusion, and to recognize that many or perhaps most delusions differ only by degree from ordinary erroneous judgments" Radden (2013, p. 127).

[119] Boudry & Braeckman (2012) describe as self-validating belief systems whose beliefs are on the one hand inconsistent with the available evidence, but on the other hand resistant to correction. Thus, adherents of these belief systems, due to certain cognitive features utilize epistemic defence mechanisms in order to uphold these systems, resulting in the dissemination

epistemic norms, one might try to make a distinction between self-deception and delusion in terms of the *procedural norms* of belief formation or “how a psychological system ought to function” (Bayne & Fernández, 2009, p. 5). One might say that self-deception does not violate the procedural norms, because e.g. self-enhancement is pervasive (Bayne & Fernández, 2009), yet, I agree with Bortolotti (2010) that violations in the *mechanism* of belief formation would be difficult to prove (p. 129) and that deluded are procedurally irrational in the sense that their delusions are not well-integrated with other beliefs (note that Bayne & Fernández and Bortolotti use the term procedural norms of belief formation differently; see section 1.2.6).

Second, *affective & motivational factors* have been argued to play the same role as in self-deception in that “the subject’s motivational and affective states have led him or her to flout certain norms of belief formation” (Bayne & Fernández, 2009, p. 2). “[I]n those (putative) cases where motivational factors figure in the formation of delusions” (McKay, Langdon & Coltheart, 2009, p. 173) delusion and self-deception might overlap (p. 172). They have been said to be distinct phenomena though, insofar as delusion can occur without motivation (p. 173) and self-deception may involve contradictory evidence not strong enough to warrant the ascription of delusion (p. 173). Among those who agree that motivated delusions are instances of self-deception is Mele (2009). Mele’s argumentation for it relies on the application of the *impartial observer test* which he developed for determining whether something is an instance of self-deception. The impartial observer test states that if motivational and affective factors are removed, given the evidence the self-deceiver would come to the same conclusion as his peers (Mele, 2009, p. 63). If it is the case, the deluded cannot pass the impartial observer test, then this would be because the deluded has some *deficit* in the belief forming process, while his peers do not. Even if the cognitive peers share the deficit (the possibility pointed by Davies, 2009), they are per construction of the test devoid of motivation. According to Mele (2009) then, if it was not some deficit, but motivation that caused the delusion, then this delusion would be an instance of self-deception. Similarly, Bortolotti & Mameli (2012) also argue that motivation is an overlap area between delusion and self-deception, but that for delusion also “[o]ther factors (e.g., perception failures, brain damage, cognitive deficits, reasoning biases) need to be in place” (p. 209). Davies (2009) disagrees with Mele (2009) though in that he states that the impartial observer test is only a necessary, but not a sufficient condition for judging whether somebody is self-deceived, because there could exist a motivation[120] which is impartial, like Kruglanski’s notion of the need for closure or the desire to get a definite answer (pp. 82-83; see section 3.1.1). All in all, the authors agree that the presence of motivation makes delusion similar to self-deception, yet disagree about the degree of that similarity.

In certain kinds of delusions, the *kind* of motivation has also been argued to be alike to the motivation often posited in self-deception: protection of self-esteem and/or negative affect,

of the latter that can reach cultural level. The mentioned cognitive features are: “(a) our proficiency at ad hoc reasoning and rationalization; (b) the motivation to reduce cognitive dissonance; (c) the persistence of the confirmation bias; and (d) the psychological premium placed on being rational and free from bias” (Boudry & Braeckman, 2012, p. 355). Whether the adherents’ utilization of the defence mechanisms is intentional and whether they are aware of this utilization, does not undermine the possibility of their holding given weird beliefs sincerely (Boudry & Braeckman, 2012, p. 355). The latter – sincerely holding a weird belief – seems to be the authors’ criterion on whether these adherents are self-deceived or not.

[120] Interestingly, for Davies (2009) motivational bias is an “intuitively, a personal-level phenomenon” (p. 75). This is interesting because deflationary positions proposes biases as an alternative to intentions, thus I would say that biases are intuitively subpersonal.

e.g in persecutory delusions (Bayne & Fernández, 2009, p. 13). At this point it might be fruitful to shortly review how motivation might be embedded into explanations of delusion. McKay, Langdon & Coltheart (2009) consider how a two-factor framework can be changed to incorporate motivational factors. The two-factor framework explains delusions by combining an aberrant experience and a deficit in belief-formation. McKay, Langdon & Coltheart (2009) point out that historically two kinds of approaches have been applied to delusion: the psychodynamic (motivation of defense) and the deficit model (pp. 166-168). Their motivation to embed motivation into the two-factor framework is the "reverse Othello syndrome" (patients develop delusional beliefs about the fidelity of their partners, e.g. a quadriplegic man developing the belief that his romantic partner is still with him although the latter has abandoned him after the injury). This syndrome poses a problem the two-factor account, because there is both a deficit *and* a plausible (defensive) motivation available in this case (pp. 170-171). Their solution to how motivation might play a role in delusion formation is the following:

1. Two-factor framework: The first factor of the two-factor framework can have multiple sources: aberrant experiences or motivation (McKay, Langdon & Coltheart, 2009, p. 174). Moreover, motivation may feature in the second factor (p. 175) as a second set of constraints, except for the belief-related ones – idea proposed by Westen (McKay, Langdon & Coltheart, 2009, p. 175) and used by Sahdra & Thagard (2003) who model self-deception in a constraint satisfaction framework (see section 4.1).
2. Ramachandran's theory of hemispheric specialization (left hemisphere = defense, right hemisphere = "discrepancy detector"). Here the kind of motivation assumed is the defensive, psychodynamic one according to the authors (McKay, Langdon, & Coltheart, 2009, pp. 175–177). Here it should be noted that Ramachandran takes anosognosia to be an instance of self-deception (see section 3.1.2.1). Levy[121] (2009) agrees with Ramachandran that anosognosia is an instance of self-deception, because it is motivated (patients hold the belief that they are healthy despite the testimony of doctors, eyesight, cf. p. 338) and both the belief that *p* and that not-*p* can be *attributed* to the anosognosic on the basis of their behavior (e.g., partly avoiding bimanual tasks, acknowledging paralysis under vestibular stimulation, cf. pp. 336-337).

That delusion and self-deception are irrational, motivated phenomena makes them interpretable (Bortolotti & Mameli, 2012). Delusions in particular have been argued to be interpretable, because either they are motivated and thus, understandable from the folk-psychological perspective, or they arise as a result of unusual experience and consequently, understandable from the folk-psychological perspective (p. 214). But in being interpretable, both are susceptible to Davidson's paradox of irrationality (1.1.1.1). This is in accordance with the hypothesis that irrationality, but also the presence of motivation, will not distinguish between different kinds of irrational beliefs, because explanations in these terms are a matter of interpretation.[122]

[121] Levy (2009) argues contra Hirstein that the truthful information is available at the personal and conscious level to the anosognosics and not only represented in the brain by stating that the behavior of the anosognosic indicates a certain degree of *availability* of the truthful information (p. 336). Yet I disagree here with Levy in that showing that certain medical procedures trigger the avowal of truthful information (e.g. vestibular stimulation), or that in forced-choice situations anosognosics like blindsight patients can use truthful information in decision tasks, or that *implicit* processing of truthful information has happened (anosognosics being shown a house half in flames and indicating that they do not want to live in that house), or even that anosognosics deny failures in bimanual tasks after completion of such a task, all this is not to show that the information was available at the *conscious* level.

[122] On the other side, delusions have been also argued to be characterized by *incomprehensibility* ("sheer incomprehensibility" in the case of bizarre delusions and "mundane incomprehensibility" in the case of mundane delusions, Langdon & Bayne, 2010, p. 321-322), incorrigibility and subjective certainty (p. 321).

Finally, since the products of delusion and self-deception depart from *paradigmatic beliefs* where it is often assumed that it is "constitutive of belief that one employs the content of what one believes in practical and theoretical reasoning"[123] (p. 7), there have been attempts to label the deluded attitude other than 'belief' – a certain kind of *imagination*. As elaborated above, Gendler (2007) sees self-deception also as a kind of imagination. Egan (2009) proposes as a criterion for comparing the products of delusion and self-deception that both are intermediary attitudes between belief and some other attitude (imagining in case of delusion and desire in case of self-deception). Thus, the product of delusion is, according to Egan (2009), a "bimagination" and the product of self-deception is a "besire." His argumentation centers on the description of the kind of a role the cognitive attitude plays.[124] Delusional subjects are said to exhibit imagine-like (and not belief-like) behavior, insofar as they are unresponsive to the evidence and the role of the product of delusion is more circumscribed than the role that belief-like attitudes play (pp. 266-267):

1. Inferences are often not drawn: e.g. a delusion about the wife being an impostor may be independent from the general world-view that could incorporate such an attitude.
2. They often do not act on their delusional beliefs.
3. Affective responses also may be absent that would be congruent with having this particular belief, e.g. that the wife is an impostor.

Egan (2009) argues that *tension* felt by delusional subjects indicates that they are to some degree responsive to the evidence and draw some inferences from the delusional attitude. Moreover, they sometimes exhibit behavior congruent with the delusional attitude. Just as Egan (2009) thinks of the product of delusion to be an intermediate attitude between belief and imagining, he thinks that self-deceptive product is an intermediate attitude between beliefs and desires, where the difference between beliefs and desires is that given a behavior-planning system, beliefs determine "*start* states" and desires – "*goal* states" (p. 264). Egan (2009) hypothesizes that in the case of self-deception, a representation might have first exhibited desire-like role, but due to belief-like vividness and fantasizing may

[123] The strategies Bayne & Fernández (2009) identify as answers to the problem are: 1. division of the self into partitions that contain contradictory beliefs, 2. ascribing to beliefs different triggering conditions, 3. positing some other attitudes *sui generis* being at work, e.g. imaginings. We have already explored the first in the motivational debate and the second and the third in the product debate.

[124] "Believing, desiring, imagining, and the bearing of propositional attitudes in general are a matter of having a representational item with the right kind of content, which plays *the right kind of role* in our cognitive economy." (Egan, 2009, p. 264; my emphasis)
Egan (2009) argues that establishing of roles that representations play should not proceed in a manner such that first the representations are categorized and then their role determined (Egan calls this the *restrictive* view). This is because the categorization of the representation is dependent on the roles this representation plays (*permissive* view, cf. pp. 270). This means that the role that a representation plays depends on the *context* and strength of connections between representations (Egan, 2009, pp. 270-274). More specifically, in the case of beliefs, connections among representations can guide the behavior in one case, but not in another and sometimes an inconsistency in the belief set is detected even to the point that "the elimination of inconsistency and drawing of inferences comes easily or automatically", but in some cases this is not the case. The empirical example Egan brings is the study of Ditto and colleagues (1998), described here in the subchapter on the QOP, in which male participants had to determine what the female confederate thinks of them, given that she had a choice or no choice to choose among a favorable and unfavorable evaluation. Ditto's result was that participants made the correct attribution in the case of negative evaluations, but attributed positive evaluations to choice in both the condition where the female confederate was said to have chosen freely or to have been constrained. Egan (2009) interprets the result as an absence of inference between the belief that the female confederate was constrained and the belief they avowed that the female confederate had a favorable impression of them (p. 273).

have acquired a belief-like role (p. 276). But he also admits that given such a description of the process, the product of self-deception may also be the "bimagination" and not a "besire" (p. 276).

Similarly to Egan's bimagination, Sass & Pienkos (2013) have proposed a phenomenological approach to delusions that regards them as not concerning *beliefs*, but *modes* of experience: "What is altered is not an isolated belief or framework proposition but the overall lived background or horizon of the whole life-world, thought-world, and lived body" (p. 643). The authors emphasize that delusional experience is more about *unreality,* than *reality* (Sass & Pienkos, 2013, p. 650). The possibilities to choose from in altering one's mode of experience are according to them:

- *poor reality-testing*: "the patient takes the imaginary for real, that he *believes* in his imaginary objects with essentially the same form of beliefs as we address to our surrounding world (i.e., adopting the doxastic or "natural attitude")" (p. 650);
- *double bookkeeping*: "the patient recognizes the essential unreality of his delusional world and thus of the distinction between the imaginary and the real" (p. 650);
- *double exposure*: "merging of two perspectives on reality" (p. 650).

The traditional perspective would be to accuse self-deceivers of double bookkeeping, but what if the phenomenology is the one of double exposure? One of the reasons for assumed double bookkeeping by self-deceivers is the assumed phenomenology of uneasiness. The distinction between the explanationist and the endorsement account of delusions reflects the supposed difference in phenomenology between thinking and perceiving (Langdon & Bayne, 2010, p. 321):

- *"explanationist,"* labelled by them "reflective" – the experiential content of delusion is coarse-grained and underdefined, e.g. something is wrong (p. 328); delusion arises out of inferential processes explaining underdefined experience;
- *"endorsement,"* labelled by them "received" – the experiential content of delusion is fine-grained and precise, e.g. experiencing a beloved one *as a stranger* in Capgras delusion (p. 328); delusion arises as unconditional acceptance of weird experience.

Langdon & Bayne (2010) note that an endorsement account better explains subjective certainty and incorrigibility of delusions: "If other people were to challenge your belief and to tell you that the sky was a lovely shade of green, you would likely think them crazy, at least initially" (p. 332). The other, reflective, end of the continuum, is characterized according to the authors by "gradual crystallisation of subjective certainty" (p. 332), "awareness of delusional belief *as* an explanation, and hence requiring justification" (p. 332) and that the "[d]elusional conviction grows more gradually; perhaps feed by rumination and [hot and cold] attentional biases" (p. 335).

How could one test what phenomenology of the self-deceivers really is like? It may be the case that self-deceivers' *experience* is different and, when they *justify* the beliefs ascribed to them, they do what confabulators do - "inventive plausible story-telling" (Langdon & Bayne 2010, p. 323). Langdon & Bayne (2010) give such examples for confabulation as the nylon-stockings experiment by Nisbett & Wilson (1977) in which participants that had to choose among stockings which were in truth identical and had to justify their choice were found to post hoc rationalize their choice, as well as anosognosia (p. 323). These are two examples also favored as those of self-deception. Further, the difference they make between *spontaneous*[125] *and provoked confabulation*, where the latter arises out of direct

[125] Langdon & Bayne (2010) define spontaneous confabulation as "a mistaken memory which is generated spontaneously (with no intention to deceive) and which is adopted uncritically as true and with a degree of subjective conviction that is unwarranted in light of general knowledge and/or the evidence to hand (either or both of which ought to confer doubt)" (pp. 324-325).

questioning in memory testing (p. 324), is reminiscent of the difference between acquisition and justification of self-deception. Langdon & Bayne's (2010) idea is that provoked confabulation does not involve pathological belief-forming processes, as well as the degree of subjective conviction that spontaneous confabulation possesses and, hence, it is not delusionary. This implies that it is pathological belief-forming processes that cause subjective certainty: "it [spontaneous/provoked distinction] reflects the presence versus absence of pathological belief formation processes (in an individual patient) which associate phenomenologically with the presence versus absence of an unwarranted subjective conviction on espousal of the confabulation" (Langdon & Bayne, 2010, p. 324). But it is not only results of a belief-forming process that may possess the feeling of subject certainty (Picard, 2013; Picard & Kurth, 2014). Ecstatic seizures may also evoke such a feeling with respect to certain conscious representations such that these representations "were absolutely immune to any rational doubt," "have many of the same features as the most justified, and rationally derived beliefs that a person could possibly hold" (Picard, 2013, p. 2496). To sum up, I have argued that it may be the case that the acquisition of self-deception results in a certain experience that upon questioning has to be incorporated into the self-narrative in a coherent manner. The question posed at the end of the previous section remains, though – why do self-deceivers do not experience the relatively often *changes* in their narrative? Langdon & Bayne (2010) have supposed that *(phenomenological) fixity* is a feature that distinguishes delusions from confabulations: delusions are consistent whereas confabulations are labile in that they change over time (Langdon & Bayne, 2010, p. 326). The parallel to self-deception is straightforward: Self-deceivers are *inconsistent* in their behavior and, maybe, justification (upon the acquisition of new evidence) which leads to the assumption about the *malleability* of their phenomenology.

Interim conclusion: Delusions are said to be similar to self-deception in their motivation, in the belief-forming process that violates epistemic rationality, in their resulting cognitive attitude and in their folk-psychological interpretability. The acquisition of self-deception may be more a matter of experience than belief and if self-deceivers are questioned, they engage in the construction of a coherent narrative. Concluding section 1.2, apart from the review of the self-deceptive attitudes, the main issue was the question whether self-deception is a certain kind of belief-formation or narrative construction and if it is narrative construction, whether it is narrative construction in the direct sense of the creation of an epistemic agent model, or narrative construction in the figurative sense as changes in the world/self-model. The given distinction will enrich a model of self-deception, particularly because different kinds of selectivity may underlie the first or the second kind. For example, since world/self-model is transparent (no previous processing stages are available), no feeling of control will be available and the kind of selection will be subpersonal.

To recapitulate, in section 1.1 and 1.2 I have introduced four constraints on a convincing theory of self-deception: parsimony, disunity in unity, phenomenological congruency and demarcation constraint. Demarcation of self-deception is difficult, because since it is a cluster concept, more than one kind of selectivity might be at work. Self-deception requires a certain kind of disunity to be introduced. For this to be the case, there has to be a certain kind of unity in the first place. Personal level governed by logic and rationality is often assumed to be the unifying element. A certain kind of phenomenology, e.g. uneasiness or anxiety, may be argued to be an indicator for such a disunity and evoke doubts (Lynch, 2012). More precisely, I have argued in section 1.2.5 that self-deceptive anxiety can play four different roles: motivate self-deceptive process, be an indicator of inconsistent attitudes that themselves have been generated by the self-deceptive process, be part of the self-deceptive process itself, be an indicator of the faultiness of the self-deceptive process. How self-deception in general and the presence of personal level recognition and anxiety in particular has been empirically tested will be the topic of the next section.

1.3 Classic psychological experiments testing self-deception

In sections 1.1 and 1.2 I have discussed philosophical theories of self-deception. I have argued that two important constraints on self-deception are the disunity in unity constraint, as well as the phenomenological congruency constraint. In this section I will move from the theoretical to the empirical perspective and show how psychological experiments have been constructed to do justice to those constraints. Parsimony and demarcation constraints have not played such a great role in the empirical literature.

I will start with presenting two psychological definitions of self-deception, as well as questionnaires designed to measure it – from Gur & Sackeim and Paulhus. The main question will be whether these paradigms and questionnaires achieve their aim, namely to test and measure self-deception. Section 1.3.1 on definitions and questionnaires will show that the intentional measurements of questionnaires developed to test self-deception are open to interpretation. The implication thereof is that if certain qualities are ascribed to a subgroup of persons on the basis of them having got a certain score on the SDQ or BIDR questionnaire, then this is per se is not enough to justify the claim that *self-deceivers* are characterized by this quality. This will be relevant for chapter 3, where I will cite more studies using BIDR and drawing certain consequences on its basis with respect to self-deception.

In section 1.3.2 I will then discuss two classic paradigms for testing self-deception - Gur & Sackeim's voice recognition experiment and Quattrone & Tversky's pain endurance paradigm. The personal/subpersonal distinction (section 1.3.4), as well as the quality/quantity of processing (section 1.3.3) distinction will be in the focus. The fact that the second paradigm has been explained by the view that *quantity* and not *quality* of processing is modified by motivation restricts the amount of the discrepancy that the pain endurance paradigm allows. Particularly in the Paulhus' study (section 1.3.2.4) that is developed according to the same paradigm there is only a weak kind of discrepancy on the personal level such that participants can justify their experimentally acquired biased attitudes. Both paradigms use physiological (subpersonal) measures to provide a psychological (personal) explanation of self-deception. Which personal level *states* might be ascribed on the basis of physiological measures and how *content* of those states is fixed will be the topic of the last summarizing section 1.3.4. I will argue that self-deceivers (participants) fix content in a way different from observers (experimenters). This evokes the need for experimenters to develop additional measures to check how participants fix the content of their attitudes in order to be able to follow from their experiments that there really is a certain kind of discrepancy on the personal level. This is relevant for fulfilling the disunity in unity constraint on a satisfying theory of self-deception. To remind the reader, I have argued in section 1.1.1 that an explanation of self-deception requires a certain kind of disunity to be postulated which by definition means that another kind of unity has to be present. Usually, the personal level serves as such kind of unity. If the disunity is between the representations on the personal and subpersonal levels, then, if ever, such kind of disunity would lead to a very weak kind of self-deception, because no rules of logic or rationality exists between these levels, while they do between representations on the personal level. Considerations of content fixation are the possible means of specifying the disunity in unity constraint.

1.3.1 Questionnaires testing self-deception: SDQ (Gur & Sackeim) and BIDR (Paulhus)

In this section I want to first present two psychological definitions of self-deception from the authors of questionnaires measuring self-deception in order to show that the personal/subpersonal level is confused. I will then discuss these questionnaires with the aim to show that their intentional measurement objects are not clear. The implication is that scoring high/low on such questionnaires per se without additional *conceptual* arguments is not enough to prove that self-deception is present/absent in participants.

Psychological definitions of self-deception to be presented demonstrate, first, that also in the psychological literature there is not unity on how to define self-deception and, second, that *discrepant beliefs* (Gur & Sackeim's definition) or *discrepant representations* (Paulhus' definition) are argued to result from self-deception and their existence to be provable in experiments. It is to stress that beliefs differ from representations in that 'beliefs' belong to the personal level of description, while the term 'representations' is more general in that respect. I will argue in the following section then that the personal and subpersonal levels are confused in the voice recognition experiment.

McKay, Langdon & Coltheart (2009) hold that Sackeim & Gur's definition of self-deception is "arguably the most widely accepted characterization in the psychological literature" (p. 172). Gur & Sackeim (1979) claim that their necessary and sufficient conditions for self-deception arise out of a logico-linguistic analysis (p. 147; see table 14 for the conditions). Sackeim & Gur (1979) argue that a nontransparency of consciousness can explain the paradox of self-deception (= the possession of contradictory beliefs at the same time, p. 146-148). Self-deception, as well as repression, of which self-deception is a general category (p. 142), are characterized to have in common[126] that

1. The state of consciousness is not transparent (= not "all contents of consciousness are capable of reaching awareness"), as well as that
2. Cognitions one is not aware of are "stored in an unconscious" (pp. 140-142).

The third condition (= not being aware of contradictory beliefs) is said to be added in virtue of explaining the first two (= simultaneous possession of contradictory beliefs). Concerning the last condition (= motivated nature of self-deception), Sackeim & Gur's (1979) definition of motivated act is that "such acts lead to differential outcomes for the individual" (p. 150). It is important, because the presence of motivation is tested in this way: If subjects react differently to two experimental setups, in which the only difference is the story that subjects have been told about the purpose of the experiment and the desirability of the expected outcome, then variations in behavior congruent with the story are said to point to the presence of motivation (compare e.g. Quattrone & Tversky's in the next section or Ditto's study in section 1.3.3).

[126] The difference between self-deception and repression is argued by Sackeim & Gur (1979) to be that the former does not require intentionality or the nonunity condition (= consciousness is "divided into separate and independent functional structures"), while the latter does: "The ascription of repression requires the additional claim that the belief that is not subject to awareness is stored in an unconscious. The unconscious is a functionally independent control system, capable of intentional influence on behavior. In this respect, the concept of repression entails that consciousness is not only nontransparent but also nonunitary" (p. 151).

Paulhus & Buckels (2012) define self-deception in a similar manner to Gur & Sackeim – as a defensive phenomenon that involves the maintenance of contradictory representations. Paulhus & Buckels (2012) distinguish in a recently published brief overview over the psychological research on self-deception between two general uses of the term self-deception:

- *"soft" self-deception*: e.g., setting one's watch some minutes ahead, procrastination (delay of unpleasant tasks and subsequent rationalization of such delay) (p. 365);
- *"deep" self-deception*: maintenance of biased belief under recurrent confrontation (p. 366).

They (2012) argue that though there is a "diversity of competing but overlapping and intertwining concepts" (p. 363), the term self-deception should be "reserved for cases where strong psychological forces prevent us from acknowledging a threatening truth about ourselves" (p. 365). Thus, the "soft" kind is not a kind of self-deception, but a "crude coping mechanism" (p. 365). Moreover, self-deception is said to differ from motivated cognition insofar as discrepant representations are present (p. 367), being an extreme motivated bias.[127]

In accordance with this definition, they argue that the dual processing approach of affect and cognition, or that emotion and cognition are processed differently (p. 367), can explain self-deception (p. 363, 366). The authors do not disambiguate the term dual processing though. On the one hand, they speak of emotional system blocking information and quote Greenwald's "junk-mail" analogy which posits that processing of unfavorable information stops at some point. This use of the term would be incongruent with their requirement of discrepant cognitions on self-deception. On the other hand, they state that approaches on repression, defensiveness, narcissism, implicit and explicit self-esteem supports "the conclusion that some individuals more than others manage to maintain multiple representations of the same information – and that one of those representations induces sufficient distress to *minimize its full availability to conscious awareness"* (p. 374; my emphasis). The latter quotation seems to imply that unfavorable information has been fully processed and *then*, it is made unavailable on the personal level.

Gur & Sackeim's, as well as Paulhus' definitions are stronger than deflationary ones in that they require presence of contradictory representations, which is not the case in deflationary accounts. Recently though, deflationary positions – especially that of Mele (2012) - have gained more weight in psychology, leading to a slippery slope of equating self-deception with optimism, self-enhancement and motivated cognition in general (for comparison of definitions see table 14). Nevertheless, the questionnaires based on Gur & Sackeim's definition (SDQ and BIDR) are still in use. One possibility now is that proponents of Mele's deflationary definition might use these questionnaires in their experiments. This scenario is possible because psychological questionnaires, once created, are tested for reliability, not least by testing the presence of correlation with other questionnaires measuring conceptually similar phenomena. If found reliable (if this test repeatedly leads to the same results), such questionnaires are used because of their reliability and not the congruence with the researcher's own definition of self-deception that she wants to test. This happens

[127] What is interesting is that Paulhus & Buckels (2012) argue that contradictory representations have to be *activated*: "The notion of self-deception implicates a deep-seated psychological process that eventuates in a distorted self-perception. As in the case of other motivated biases, the victim possesses the information to draw the correct conclusion but, for emotional reasons, does not do so. Self-deception goes further to suggest that both the accurate and inaccurate representations remain active. To regulate dangerous information most effectively, one must, at some level, recognize and manage it. As such, it may be seen as the most *extreme version of motivated bias*" (p. 374; my emphasis). In section 2.2.1, I review the different degrees of influence that have been argued for to occur in a network of representations.

because there is enough freedom for different interpretations for what the results of self-deceptive questionnaires actually measure. Thus, intentional objects of measurements will be the topic of discussion in the remaining part of the section.

Gur & Sackeim: **contradictory beliefs**	**Mele:** **one-belief**
"Accordingly, we have offered the following criteria as necessary and sufficient for ascribing self-deception to any given phenomenon: 1. The individual holds two **contradictory** beliefs (that p and not that 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; bold emphasis added)	"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 ~p than for p. 5. S **consciously believes** at the time that there is a significant chance 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." (Mele 2012, p. 12; bold emphasis added)

Table 14. Gur & Sackeim and Mele: psychological definitions of self-deception. Distinctions from Gur & Sackeim (1979) and Mele (2012).

Interim conclusion: In the two definitions presented – that of Sackeim & Gur, as well as Paulhus – presence of discrepant representations is one of the central characteristics. In the following section I will present Gur & Sackeim's voice recognition experiment which attempts to prove the definition of self-deception here presented. Discrepant representations have been argued there to be participant's verbal response, on the one hand, and skin conductance response, on the other. I will argue that skin conductance and verbal response belong to different levels of description – subpersonal and personal respectively. Thus, additional assumptions are needed if one wanted to argue that skin conductance is present on the personal level and also *in which form* it is present.

In Sackeim & Gur (1979) the authors argue that participants committing voice-identification errors scored higher on the SDQ (p. 213). Moreover, SDQ was negatively correlated with certain scales measuring psychopathology (BDI, EPI-N, MSQ), leading to the interpretation that "the more likely individuals are to engage in self-deception, the less likely they are to *report* psychopathology" (my emphasis; p. 214). Does SDQ offer an appropriate measurement of self-deception? One should notice that SDQ (see table 15) is based on a psychodynamic take on self-deception. Greenwald (1988) comments the results of the voice recognition study as showing the "motivated blocking of conscious voice recognition that is initiated by a knowing observer operating outside of consciousness" (Greenwald, 1988, p. 117). An intentionalist explanation of self-deception is yet difficult to sustain (see section 1.1.1 and 1.1.3). Another worry, as mentioned at the beginning of section 1.3 is whether such a questionnaire may be used in testing self-deception based on deflationary definitions. Taylor (1989) substantiates her claim that "positive illusions" are adaptive by stating their link to self-deception (= "Negative information must be recognized for what it is and simultaneously kept from awareness as much as possible," p. 157) and by mentioning the negative correlation between high scorers on the SDQ and depression (p. 159). Here interesting is that adaptivity of "positive illusions" – self-enhancement, optimism and enhanced beliefs in one's control – phenomena that per se do not have a defensive connotation, are tied to a questionnaire measuring defensive reactions or responses to universal but threatening facts. In the following I will comment on Paulhus' BIDR that can be called an improved version of SDQ.

SDQ (Self-Deception Questionnaire)	Other-Deception Questionnaire
"The Self-Deception Questionnaire (SDQ) consists of 20 items, the positive endorsements of which were judged to be **universally true but psychologically threatening**, for example, "Have you ever enjoyed your bowel movements?" and "Have you ever made a fool of yourself"? Items are rated on 1 to 7 Likert-type scales for frequency or intensity of the behavior in question, and ratings of **1 or 2** on items are scored as instances of self-deception." (Sackeim & Gur, 1979, p. 213; my bold emphasis)	"The ODQ contains 20 positively keyed items culled from various lie scales. Ratings are made on 1 to 7 Likert-type scales, with ratings of **6 and 7** scored as instances of other-deception." (Sackeim & Gur, 1979, p. 214)

Table 15. Sackeim & Gur: SDQ vs. ODQ
Dinstinctions from Sackeim & Gur (1979).

Paulhus (1984) has performed a factor analysis on different social desirability scales, among them the SDQ and ODQ, to discover the dependencies among the items from different scales. The aims was to find the latent variables with which those items correlate best. The result was that the items loaded on two factors and that self-deception and impression management best describe these two factors. For us, the factor of self-deception is of interest. Thus, I will review the arguments in favor of interpreting the factor as that describing self-deception.

1. First, it is important to understand that labeling such a factor is a matter of interpretation. After the statistical computations are completed, the items that load the highest on the factors, are looked at and *interpreted*. In the case of self-deception factor, Paulhus (1984) argues that the fact that best loading items concern personal threats and that many of the items considered belonging to the factor were those from the SDQ,[128] is an argument in favor of interpreting the given factor as a self-deceptive one.
2. Another argument in favor of interpreting the factor as describing self-deception is that there was a difference in how the subjects answered the items in public and anonymous conditions, such that "impression management scales changed significantly more than the self-deception scales" in the direction of socially desirable responses (Paulhus, 1984, p. 605, 606), allowing the conclusion that it is "an unconscious defensiveness that underlies self-deceptive responding." (Paulhus, 1984, p. 607)

Let me consider these points more thoroughly. Paulhus' factor analysis has led to the discernment of two factors which were then, on the basis of the content of the items assigned to either factor, interpreted to measure one or the other phenomenon. In the given case, self-deception is interpreted as a defensive phenomenon. Then, when the interpretation is established, the items are again looked at, keeping in mind which scale they came from and which phenomena those scales correlate with. New predictions are derived with respect to the newly established factors and old scales. The interesting question is how to interpret the correlations of the self-deception questionnaire with other questionnaires, not only with the ones of socially desirable responding. While statistical

[128] The content of the SDQ questions: "Many of the high-loading items refer to sexual and parental conflicts and other deep personal concerns. These kinds of conflicts play a primary role in the psychoanalytic conceptions underlying Sackeim and Gur's view of self-deception. The item results, along with the fact that the SDQ was the best overall marker of the factor, argue strongly for a self-deception interpretation of the first factor" (Paulhus, 1984, p. 601). Whether a questionnaire testing for such content-related questions is fitted to test self-deception towards other kinds of content is in question.

methods measure the relationship between items, it is the *interpretation* of the factors discovered by clustering similar items that can be questioned. Let me look at some of such interpretations. Paulhus (1984) argues for example that given that the Marlowe-Crown scale loaded on both self-deception and impression management factors and given that this scale is shown to exhibit "behavioral correlates more clearly than other social desirability scales," self-deception and impression management may be "necessary for an individual to display need-for-approval behavior" (p. 606). Another connected hypothesis made by Paulhus (1984) is that self-esteem, repression, desirable responding may be the result of the "same underlying mechanism," because they are difficult to measure independently.[129] Millham & Kellogg (1980), like Paulhus (1984), establish the correlation between Marlowe-Crowne Social Desirability Scale (MC-SDS) which contains items high in social desirability, but low in the probability of occurrence and vice versa, and self-deception, measured by the score obtained by applying another need for approval scale (Jacobson-Kellogg Scale or J-K) under bogus pipeline condition (p. 449). MC-SDS and J-K are said to be equivalent measures of the need for approval (p. 449). Thus, Millham & Kellog (1980) assumed a "defensive-denial" interpretation of the MC-SDS (p. 447), asked participants to fill out the MC-SDS once and J-K twice, this second time being one week later when the participants were connected to a device called a "lie detector." The authors assumed that those participants who obtained a high score on J-K under the lie detector-conditions were self-deceiving. Other-deception was measured as the difference between the first and the second application of the J-K (p. 450). A diminished recall of socially-undesirable self-descriptive traits (Millham & Kellogg, 1980, p. 454) has been shown in self-, but not in other-deceivers, supporting the interpretation of self-deception and MC-SDS that correlated with it as defensive-repressive (p. 455). The crucial point though is that even if one were to agree that there is a defensive kind of self-deception about socially desirable attributes, since self-deception has been measured by a social desirability questionnaire from the beginning, no conclusion can be drawn that *all kinds* of self-deception are defensive or about socially desirable attributes, which would be a circular conclusion. The second point is that here the criterion for the ascription of self-deception is only its assumed nature and nothing else: no contradictory information, no justification by participants etc. These two points show that it is not statistical analysis per se, but the *interpretation* of statistical analysis that leads to a decisive role in characterizing a certain phenomenon.

By methods described above Paulhus has developed a questionnaire often used now for the measurement of self-deception - BIDR. Paulhus (1984) has developed it on the basis of Sackeim & Gur's SDQ. To remind the reader, Sackeim & Gur's Self-Deception Questionnaire (SDQ) and Other-Deception Questionnaire (ODQ) test for participants' judgments on "universally true but psychologically threatening" and "socially desirable but statistically infrequent behaviors" items respectively (Paulhus, 1984, pp. 599-600). It has been moreover argued that a further difference between the SDQ and the ODQ items is that for SDQ items only the respondent could know the truth value of the responses, while for

[129] The following quotation demonstrates the difficulty to distinguish certain measures of (supposedly) related phenomena: "Intermixed with these items [the core items of the self-deception factor which are the deep personal threat items of the SDQ] are other items reporting low anxiety and high self-esteem. It is well-known that standard measures of anxiety, self-esteem, repression, and social desirability are difficult to tease apart psychometrically. Even worse, these measures are hard to distinguish from accurate self-reports" (Paulhus, 1984, p. 607).

ODQ it is not the case (Paulhus & Reid, 1991, p. 308). Paulhus' refinement consisted in rephrasing the questions:[130]

> Paulhus (1984) addressed a number of psychometric deficiencies while developing his new instrument, the Balanced Inventory of Desirable Responding (BIDR). The new subscales, termed the Self-Deception scale (SDS) and the Impression Management scale (IMS), were improvements in several respects: (a) The keying direction was balanced, (b) items referring to adjustment were deleted, (c) items with low part-whole correlations were replaced, and (d) nonpsychoanalytic items were added (e.g., "I could easily quit any of my bad habits if I wanted to"). The BIDR has been *used successfully* in a number of studies […]. (Paulhus & Reid, 1991, p. 308; my emphasis)

How the newly-won scales are to be interpreted, has changed several times. One issue was the relationship between self-deception and denial. The self-deception-impression management scales where disambiguated with respect to the attribution-denial (or enhancement-denial) distinction: Confounding self-deception with *denial* of socially undesirable attributes and impression management with *confirmation* of socially desirable attributes is, according to Paulhus, being avoided by the above changes to the SDQ and ODQ (Paulhus, 1984, p. 598, 601). In doing so, Paulhus divides the SDS in two subscales: self-deception enhancement and self-deception denial.

Further analysis using the new BIDR questionnaire challenges the assumption though that self-deceptive enhancement and self-deceptive denial belong to the same factor. Paulhus & Reid (1991) argue that self-deception enhancement is linked to *self-esteem* and that enhancement items of the SDS and denial items of the SDS do fall closer to *different* factors (p. 309). Thus, self-enhancement, on the one hand, and impression management *and* denial, on the other hand, are said to characterize the two factors. Moreover, "[c]orrelations with adjustment (self-esteem, trait anxiety, and neuroticism)" are measured to be higher for enhancement than for denial items of the SDS (Paulhus & Reid, 1991, p. 313). The interpretation that is close at hand is that denial is a case of impression management, instead of self-deception.[131] Another one is that it is the items measuring denial that are inappropriate to tap self-deception, but not denial per se (Paulhus & Reid, 1991, p. 315). To draw an example, "I enjoy my bowel movements" is said to possess "public embarrassment value," (Paulhus & Reid, 1991, p. 315) confounding the measurement of self-deception with impression management. Paulhus & John (1998) propose another interpretation of the relationship between impression management and denial as "conscious and unconscious versions of exaggerated moralism" (p. 1036), on the basis of the fact that

[130] The rephrasing concerns the formulation in terms of (affirmative) statements: "In the present study, the items from the SDQ and ODQ were rewritten so that (a) all items were worded as statements rather than questions, (b) all statements were worded as trait *affirmations* (I am nice); *negations* (I am not nice) were eliminated; and (c) equal numbers of attribution and denial items appear on each scale. For instance, the ODQ item, "I am honest," was changed to "I sometimes tell lies if I have to." The SDQ items, "Is it important to you that others think highly of you," was changed to "It's alright with me if others happen to dislike me." To get a perfect score on either scale, the respondent must now endorse 10 socially desirable attributes and deny 10 socially undesirable attributes. The overall set of 40 items was labeled the Balanced Inventory of Desirable Responding (BIDR)" (Paulhus, 1984, pp. 602 - 603).

[131] The relationship between *scales* is argued to determine the relationship between the *phenomena* they measure: "It is intriguing that the SDS denial items fall close to the impression management factor. This phenomenon was presumably masked in previous research because the enhancement and denial items were not scored separately. This location of the denial items suggests the provocative possibility that our subjects are *faking good* on the denial items rather than *self-deceiving*" (Paulhus & Reid, 1991, p. 315; my emphasis).

neither self-deceptive enhancement nor self-deceptive denial "is responsive to audience effects." This interpretation brings denial again in a conceptual proximity of self-deception, instead of impression management. 'Egoism' ("tendency to exaggerate one's social and intellectual status", p. 1025) and 'moralism' ("tendency to deny socially deviant impulses," p. 1026) are labels that the authors give to the two factors identified in previous papers and described at that time as self-deception and impression management. Paulhus & John (1998) argue that there is both egoistic and moralistic self-deception which arise out of need for power and need for approval respectively (p. 1045): "The egoistic and moralistic biases are both self-deceptive styles in that they operate *unconsciously* to preserve and amplify a positive self-image" (p. 1041; my emphasis). Paulhus & John (1998) introduce a new interpretation of the two factors. Instead of unconscious self-deception and conscious impression management factors, they describe the two factors as *agentic and communal factors* which differ with respect to the domain of distortion – oneself or relationship with others, but can both be either conscious or unconscious (p. 1048). Thus, the explanation of the correlation between denial and impression management is now that both have a certain characteristic in common – they belong to the same factor, but that denial and self-deception are both unconscious. What the debate about whether denial is kind of self-deception or impression management shows is that it is still in question what the items of the questionnaire which is said to measure self-deception do really measure.

Since the relationship between self-deception and self-enhancement will become important in chapter 3, I will comment here on Paulhus' (1998) analyzes of the relationship between self-deceptive enhancement (SDE) and self-enhancement ("tendency to overestimate one's positivity relative to a credible criterion", p. 1197). The results of studies conducted by him show that the trait self-enhancement can take three forms (pp. 1198, 1204):

a. self-peer evaluation discrepancy;
b. self-deceptive enhancement (measured by SDE of BIDR);
c. narcissism (measured by NPI).

This "conceptual overlap of the three constructs" is considered by Paulhus (1998) as evidence in favor of self-enhancement being a *trait* (p. 1205). He confines the usefulness of this trait[132] to "positive self-attitudes," but not to "harmonious interpersonal relations" (p. 1205), because people possessing this trait are evaluated negatively by others over time. The positive relationship between "ego enhancement" and adjustment, on the contrary, has led Paulhus & Reid (1991) to hypothesize that "ego enhancement" is useful as an "alternative tactic" compared to "ego defense" in the cases where the latter is futile by building a positive buffer and that it may be explained by the terror-management theory (TMT, p. 315). The general conclusion of Kurt & Paulhus (2008) is though that self-enhancement is maladaptive. They compared the two measures of self-enhancement – social comparison and criterion discrepancy – with respect to their connection to self- vs. peer-rated adjustment. Social comparison self-enhancement was operationalized by asking

[132] Self-deception is also brought into relation to Taylor's "positive illusions" in general, and not only self-enhancement as one of their types. Self-enhancement and exaggerated sense of control which are according to Taylor (1989) two types of positive illusions, are equated with each other by Paulhus & Reid (1991):
"The highest correlating items from the large MIB (Miscellaneous Indexes of Bias) inventory were items such as "I am always honest with myself," "my first impressions are usually right," "I could easily quit any of my bad habits," and "when I criticize someone, it's only for their own good." These items have in common an *exaggerated sense of control and confidence in one's thinking powers* – almost a cognitive narcissism. We suspect that this form of bias, rather than the indiscriminative claiming of positive attributes, is central to the *enhancement construct* […]." (Paulhus & Reid, 1991, p. 315; my emphasis)

participants to “rate themselves relative to the average college student” (p. 842), while criterion discrepancy self-enhancement was operationalized by computing the difference between self-ratings and peer-ratings. Personal adjustment concerned “general well-being including happiness, emotional stability and (lack of) depression and alienation,” while interpersonal one – “harmonious (vs.) antagonistic relationships with others” (Kurt & Paulhus, 2008, p. 843). They found that except for the link between discrepancy self-enhancement and *self-perceived* personal adjustment, chronic (meaning trait-like tendencies of) self-enhancement is maladaptive (p. 851). Elaborated studies supposes that both self-enhancement and self-deception are defensive, generally maladaptive phenomena, which is, as we have seen, partly a matter of definition.

Last, I want to mention which biases and personality traits have correlated with self-deception. Paulhus & Reid (1991) point to the differences in correlation of enhancement/denial subscales with scales measuring various biases, but also caution that these differences “were an exploratory attempt to assess with direct self-reports various defenses and biases” (p. 311). The correlation results are the following (one should pay special attention to the link between enhancement and self-fulfilling prophecy):

> Note that several indexes correlate more positively with the enhancement subscale than with the denial subscale: dogmatic thinking, lack of procrastination, lack of parental conflict, illusion of control, and self-fulfilling prophecy. In contrast, the denial subscale correlates higher than the enhancement scale with the following indexes: denial of hostility, denial of sexuality, rejection of criticism, undesirable acts, use of suppression, hindsight bias, just-world belief, and belief in prayer. In general, the indexes that correlate with *denial* also correlate with *impression management*. (Paulhus & Reid, 1991, p. 311; my emphasis)

With respect to personality traits, agentic domain/self-deceptive enhancement is shown by statistical methods to associate with extraversion and openness to experience, while the communal domain/impression management – with agreeableness and conscientiousness (Paulhus & John, 1998, p. 1030, 1043; Kurt & Paulhus, 2008, p. 843). Yet, this analysis of Paulhus & John (1998), at least partly, relies on their new measurement tool - the *self-criterion residual (SCR)*, which measures the deviation between the self-report of the self-deceiver and some objective measure. Paulhus & John (1998) argue that this tool fulfills Sackeim & Gur’s criterion of possessing “two conflicting representations of the same information” (p. 1051):

> the bias index calculated as the residual variance that remains when a self-report measure is regressed on a corresponding criterion measure of that same variable. For example, scores on a self-report Agreeableness scale are regressed on a set of peer ratings of the same items and the self-report residual is isolated as a separate variable. Because all the self-report variance shared with the peer ratings has been removed, the residual represents only self-report inflation: High scores indicate overclaiming agreeableness (and low scores indicate underclaiming) relative to the peer–rating criterion. These self-favoring bias scores can then be correlated with any other variable in the data set. (Paulhus & John, 1998, p. 1032)

Paulhus & John (1998) also quote a study in which narcissists, shown their own performance, still self-inflate, yet the conclusion that they draw from this seems unsatisfactory, namely that “narcissists are not merely overlooking objective criterion information but actively distorting it” (p. 1051) insofar as

1. It is unclear which criteria the narcissists used in evaluating their own behavior, thus the presence of contradictory evidence is unclear;

2. Even if in this study one tried to confront the assumed self-deceiver with contradictory evidence, the SCR criterion does not require this to be the case. In cases, in which it is not so, it is a far stretch that the self-deceiver possesses the contradictory evidence.

Summing up, neither the relationship between denial and self-deception, nor that between self-deception and self-enhancement can be clarified by statistical analyses alone, but only by their *interpretation*. Self-deception has been interpreted by the authors cited in this section to be an unconscious, defensive phenomenon of overemphasizing one's socially desirable characteristics. This is a very narrow characterization that makes one wonder how to interpret the results of studies using the self-deceptive enhancement scale from BIDR (The relationship of the self-deceptive denial scale to self-deception *is* questionable, as we have seen):

1. If someone uses BIDR to test a kind of self-deception that does not fit the interpretation given by BIDR, how could one then interpret the results?
2. For the participants scoring low on BIDR, could one say there was no self-deception, or just that an unconscious defensive socially desirable responding was absent, but other kinds of self-deception could be present?

Hagedorn (1996), for example, measured the correlation between Life Satisfaction Research questionnaire and social desirability measured on the basis of Paulhus' (1984) paper that introduced BIDR. The results were that self-deception did positively correlate with life satisfaction and happiness, as well as that its function may not be defensive after all:

> If the purpose of self-deception is to preserve self-esteem when it is threatened, then it would be expected that people high in Self-Deception who are dissatisfied with their lives would blame their Circumstances, while their happier counterparts would claim credit for having Made their lives satisfying. The results failed to confirm this prediction, but this may have been due in part to the fact that less satisfied self-deceivers still had such high satisfaction ratings that there was no adequate-sized group of unhappy self-deceivers. (Hagedorn, 1996, p. 158)

Uziel (2014) argues that BIDR-IM does not measure deception, but *self-control* that is directed towards communication partners. His reasons are that participants high in IM that had to judge their actual and ideal personality were associated with a large actual-ideal self discrepancy. The author interpreted this as proof that IM does not measure self-enhancement, else the discrepancy should have been small, rather than large. Uziel (2014) found also that external observers confirmed participant's high IM rating; self-control predicted IM ratings and external observers also confirmed high IM participants to possess high self-control. BIDR-SDE scale showed the opposite results to the BIDR-IM: large actual-ideal self discrepancy among high SDE individuals and poor confirmation by external observers, though self-control was also the strongest unique predictor of SDE. Pompili et al. (2011), on the other hand, equate the level of unconscious self-deception with the SDE scores and conscious overly positive description with IM scores (p. 28). The results of their study, given this fixation of the intentional objects of measurements are that self-deceivers possess lower degrees of helplessness.

Interim conclusion: In this section I have mentioned several alternatives for the intentional objects of BIDR measurements and have sketched the reasons for the change from one object to another, but the question about the intentional objects that BIDR measures remains open. I agree with the limitation that Farrow et al. (2015) mention, namely that though the aim of their study was to determine the neural correlates of self-deception and impression management, what one actually has tested may be only the neural correlates of BIDR:

> Though we gave carefully worded instructions to encourage ecologically valid imagination of impression-management and self-deception situations, we have no direct measure of the success of this and it is therefore arguable that we have not studied the neural correlates of self-deception or impression-management per se, but rather, the correlates of the BIDR performance. (p. 171)

Since Farrow's et al study is the first to my knowledge to try to discern the neural correlate of self-deception, I want to conclude by briefly describing their results. Procedure of the study was as follows: in an fMRI scanner participants completed BIDR under two conditions – "faking good" and "faking bad," which means that they tried to convey a favorable or unfavorable impression of themselves by answering the questions either in a best possible or a worst possible fashion. Interestingly, self-deception scores were harder to manipulate (Farrow et al., 2015, p. 169). The conclusion was that "although self-deception and impression management are dissociable concepts, they utilise very similar underlying brain processes" (p. 173).

All in all, the fact that different intentional objects of measurements have been ascribed to the scales of the BIDR questionnaire is not an argument for or against using BIDR in establishing self-deception per se, but it shows that this questionnaire is insufficient without additional conceptual arguments for establishing the presence of self-deception. As a result, all the connections established between BIDR and other questionnaires that would indicate which other qualities characterize participants scoring high on BIDR, e.g. different thinking styles or different kinds of biases, cannot automatically be ascribed to self-deceivers. The connection between BIDR and self-deception merits further empirical analysis after conceptual clarification of the explanandum of self-deception has been provided. My take on the latter will be given in the second chapter.

1.3.2 Two classic testing paradigms

Gur & Sackeim's voice recognition experiment and Quattrone & Tversky's pain endurance study represent two classic paradigms for testing self-deception in the psychological literature. The weakness of the voice recognition experiment is in locating the discrepancy between verbal reports, on the one hand, and skin conductance response, on the other. Pain endurance study, in difference to the voice recognition experiment, does not focus on the presence of a certain discrepancy, but on motivation. When motivated, participants hold their hands longer in the cold water. Two further similar studies that use the pain endurance paradigms will then be mentioned: Sloman's dot-tracking study and Paulhus' cheating test study. I will then argue in section 1.3.3 that the pain endurance paradigm allows for a weaker kind of discrepancy on the personal level and in section 1.3.4 that both paradigms confuse personal and subpersonal levels of explanation.

1.3.2.1 Gur & Sackeim: Voice recognition experiment

> We do not seem to be aware of our affective responses in the great majority of cases, and sometimes only skin conductance can reveal whether the response is present.
> (Bortolotti, 2012, p. 184)

Gur & Sackeim (1979) confirm their hypothesis by voice-recognition experiments in which galvanic skin responses are taken as an indication for the presence of two contradictory beliefs (p. 156). As an answer to Mele's criticism that a physiological reaction is not sufficient for belief ascription, Sackeim & Gur (1997) state that if ascription of belief on the basis of a physiological reaction (which belongs to the category of a behavioral reaction) is not possible, then as the only possibility of belief ascription remains self-report which is absent in self-deception by definition. Thus, they conclude that Mele's criticism precludes empirical tests on self-deception and the search of evolutionary mechanisms for self-deception (p. 125). Sackeim & Gur (1978) test self-deception via voice recognition on the basis of the results of research on self-confrontation. They summarize these results as follows:

1. The degree of holding "discrepant cognitions about the self" is different in individuals;
2. Self-confrontation leads to heightened self-awareness of these discrepancies;
3. This results in "increased physiological arousal, negative affect."[133]
4. Attempts to reduce cognitive dissonance occur under conditions of perceived responsibility and of being evaluated. Self-confrontation may also produce changes in self-esteem (pp. 168-170).

Self-deception according to Sackeim & Gur (1978) can be tested by the self-confrontation paradigm exactly because the former involves the *avoidance* of the state of self-consciousness, while the latter leads to it.[134] As we have seen, avoidance behavior is a characteristic often mentioned also in philosophical research on self-deception. Sackeim & Gur's (1978) voice recognition experiment can be described as follows: Confrontation with one's own voice is a kind of self-confrontation and it leads to greater changes in skin conductance. Confronted with one's own voices and voices of others, subjects had to choose whether it was their own voice. According to the authors, the cases in which subjects' reports differed from the behavioral measures are the ones of self-deception: false positive and false negative (Sackeim & Gur, 1978, pp. 183-184; Gur & Sackeim, 1979, p. 157). Here a parallel can be drawn between an error-minimization account in general and Mele's account in particular, according to which self-deceivers set different thresholds depending on their motivation to minimize the one or the other error. Sackeim & Gur (1978), though, see the results of their experiment as evidence that "the concept of self-deception requires a dynamic view of consciousness" in Freud's sense (p. 182). They argue that their experimental setup fulfills their four criteria of being self-deceived:

1 + 2) Simultaneous contradictory beliefs are indicated by self-report on the one hand and behavioral measures on the other hand;

3) The non-awareness criterion was confirmed by comparing the reaction times in different conditions;

4) The motivated nature is argued to be confirmed by:

a) comparing the results of voice recognition with the results on paper-and-pencil tests on "cognitive discrepancy" (Sackeim & Gur, 1978, p. 179) which is "the dissatisfaction with aspects of self" (p. 144): "individuals who gave false positive responses scored lower on measures predicting the aversiveness of self-confrontation than individuals who did not make false positive responses" (Sackeim & Gur, 1978, p. 180);

b) comparing the results of voice recognition with Sackeim & Gur's self-deception questionnaire, the results of which point according to the authors to the fact that self-deception is a "generalized response set, or characteristic defense, the frequency of its use

[133] Sackeim & Gur (1978) quote different researchers that claim that anxiety is a consequence of discrepancies in the self-concept (p. 168).

[134] Avoidance of a metacognitive state of self-confrontation is argued to indicate self-deception: "We have argued that the concept of self-deception is the superordinate category of instances of motivated selective transparency in consciousness. If self-deception exists, it might be demonstrated in examples of avoidance of the *metacognitive* state of the self-consciousness produced by self-confrontation" (Sackeim & Gur, 1978, pp. 171-172; my emphasis).

varying among people" (Sackeim & Gur, 1978, p. 184) and not a "stimulus-bound response";[135]

c) the fact that experience of success/failure which enhances/lowers *self-esteem* can influence the direction of self-deception (Sackeim & Gur, 1978, pp. 183-184; Gur & Sackeim, 1979, p. 162).

The inferences that the authors draw from their elaboration of self-deception is that cognitive dissonance, as well as self-serving biases, may be explained "in terms of self-deceptive acts" (Gur & Sackeim, 1979, pp. 166-167; see also section 3.1.1 and 3.2.2.3) and that differential hemispheric specialization may account for a mechanism for the existence of self-deception (Sackeim & Gur, 1978, pp. 184-189; see also Ramachandran's theory in section 3.1.2.1). It can be countered, however, that activity of the autonomic nervous system does not find its reflection in conscious experience (Davies, 2009, p. 73). Another point is that even if skin conductance response is somehow reflected in experience, it need not be an indication for a presence of a *belief*. De Neys et al. (2010) have conducted an experiment in which participants had to judge the logical validity of believable and unbelievable syllogisms[136] to measure the level of SCR in conflicting trials (believable-invalid, unbelievable-valid). The result was that SCR in the interval after the presentation of the conclusion and before the judgment of validity was significantly higher. The authors' interpretation is that high SCR indicates a "gut feeling" - *conscious* experience of conflict between believability (intuitive response) and logic (De Neys et al, 2010, p. 209). They claim that dual processing theory lends support to the idea that arousal, resulting from conflict detection, is conscious, but are cautious about this interpretation (p. 215). Thus, an inference from SCR to the phenomenal level is still in need of further experimental support. Gur & Sackeim's definition seems to be widespread in the literature, even though, according to Trivers (2010) no follow-up work has been conducted[137] (p. 379).

To sum up, in virtue of their intentionalist conception of self-deception, Gur & Sackeim argue that a skin conductance response is an indication of one belief and participant's avowals – of another. I think that they confuse a personal with subpersonal level of explanations: Avowals are personal level entities. Skin conductance responses are subpersonal level entities. Even if skin conductance responses correlate with a certain phenomenology, it is unclear which content this kind of phenomenology possesses. If they do not correlate with any phenomenology or awareness at all, then it is even more puzzling to call such kind of physiological responses beliefs. Further, if goal conflict, i.e., a prolonged, unresolved internal competition between inconsistent goal-representations, leads to higher levels of skin conductance (Huang & Bargh, 2014b, p. 129), then there is ambiguity about whether higher skin conductance represents the possession of inconsistent

[135] Errors in voice recognition are argued to correlate with scoring high on the self-deception test: "Questions on this inventory were specifically worded so that they centered on statements judged to be universally true and also psychologically threatening to subjects. It was found that subjects who committed false positive or false negative errors [in voice recognition] scored higher on this measure of self-deception than subjects who did not make these errors" (Sackeim & Gur, 1978, pp. 179-180).

[136] Participants are known to exhibit the belief bias in these cases: They hold believable syllogisms to be valid (De Neys et al., 2010, p. 208). An example of a syllogism: "All birds have wings. Crows are birds. Therefore, crows have wings" (p. 209).

[137] Trivers (2010) argues for Gur & Sackeim's expriments to be clearer than the recent ones: "Although we have few examples of self-deception with the simple clarity and power of Gur and Sackeim (1979), there is, in fact, an enormous literature on the subject, crossing several disciplines" (p. 383).

beliefs (1.3.2.1), or prolonged activation of inconsistent goal representations, or some kind of non-conceptual gut-feeling.

1.3.2.2 Quattrone & Tversky: The pain endurance study

In this section I will present three studies that follow another paradigm argued to confirm self-deception: participants are given different motivations to work on the task and those influence the *duration* in a manner that participants work longer if they expect to confirm a favorable trait that they possess, e.g intelligence or health.

Quattrone & Tversky's (1984) cold water experiment is aimed to show that people confuse causal and diagnostic contingencies or, with other words, that people confuse actions that *cause* desired outcomes with those that merely *indicate* desirable outcomes. Thus would mean that subjects also might possess skewed *probabilities* for outcomes conditional on having selected the action (p. 237), as diagnostic actions do not make desirable outcome more probable. The authors hold to fulfill Gur & Sackeim’s criteria for self-deception which they interpret as having two contradictory beliefs and being unaware of holding one of them. The two contradictory beliefs are said to be “I purposefully engaged in the behavior to make a favorable diagnosis” and “I did not purposefully engage in the behavior to make a favorable diagnosis.” The subjects are held to be unaware of the first belief due to the motivation for the desirable outcome to be true (p. 239). The procedure of the experiment itself is to try holding the hand in a cold water container as long as possible, prior and after pedaling an exercycle. The exercise is needed to uphold the cover story that the purpose of the experiment is to measure “rapid effects of rapid changes in temperature on heart rate after exercise” (p. 241). After the exercise and prior to the second tolerance test subjects were administered to a story about type 1 and type 2 heart – a “unhealthy” and a “healthy” type. They have been told either that tolerating the pain of cold water would indicate the possession of a healthy or an unhealthy type (pp. 240-241). During the tolerance test, participants heard the letter of the alphabet being recited after equal periods of time (5s), so they could calculate how long they held their hands, as well as compare this to their first tolerance trial. Now the critical question is whether this is self-deception, because, as Quattrone & Tversky (1984) mention themselves, it is probable that the participants held the belief that they *cannot control* their physiological mechanisms of pain and, thus, that trying harder on the tolerance test would not have had any effect (p. 243). The authors hold this to indicate that people would self-deceive easier about “actions (incorrectly) believed to be uncontrollable” (p. 243). Yet, another conclusion one could draw would be that it is not self-deception at all, if people believe not to control their actions, because if they are asked to try their hardest and they think not to be able to control the outcome, than the results just show their will-power. Quattrone & Tversky (1984) also mention that it is possible that participants really felt less pain, because of a placebo-similar influence they could exert over their physiological reactions (p. 243).

The second study that Quattrone & Tversky (1984) conducted to test the confusion of causal and diagnostic contingencies is that concerned with the *voter’s illusion*, or that subjects do think that their voting behavior would influence other voters, e.g. induce them to vote too. Clearly, I would not consider this illusion as a case of self-deception, for there is no handling of contradictory evidence and there would presumably be no justification by the subjects upon questioning. Further, the authors (1984) mention that the case of Calvinists who, despite the belief that it is known before their birth whether they will go to heaven or hell, still try to live a virtuous life, because they believe that this is an indication that they are those who will come to heaven, although as such virtuous life will not change the fact of going or not going to heaven. Is this also a case of self-deception? Certainly, the concept of confusing contingencies is too broad to be included in the concept of self-deception. But even if we concentrate on the cold water study, it is not clear whether there is an inconsistency, let alone a contradiction, in trying their hardest upon the motivation given, especially after being said that this trying will not change the underlying fact, namely the health condition. Deflationary accounts of self-deception do not require an inconsistency to be present though. Yet, in any case, pain endurance paradigm would allow for testing a weaker kind of self-deception, if the amount of personal level recognition is taken as a measure of goodness.

1.3.2.3 Sloman: Dot-tracking study

Sloman et al (2010) hold that a condition for the possibility of self-deception is the *ambiguity* in the way that people represent actions (p. 268). According to the authors, there are two possibilities: representing an action as an intervention (= “deliberate choice”) or as an observation (= “result of external and internal forces impinging on the individual”) (pp. 268-269). Subsequently, misrepresentation either in the one or in the other direction is a kind of self-deception according to them (2010): e.g. addiction is a case of *interventional self-deception*, because addicts think that they “freely chose” the consumption (p. 269) and Quattrone & Tversky’s cold-water experiment is an example for a *diagnostic self-deception* (p. 270). Fernbach et al. (2014) add that self-deception involves a contradiction between belief and behavior, but not between two beliefs (p. 6).
Sloman et al (2010) have changed the procedure of Quattrone & Tversky’s study to incorporate the variation in the degree of vagueness of information to prove that (at least) diagnostic self-deception requires vagueness. The requirements that they pose on self-deceivers are

> (i) understanding the desirability of manipulating the behavior (this requires causal knowledge) and (ii) the absence of clear feedback that the behavior is manipulated. In other words, one must *know at some level* the diagnostic value of the variable manipulated (to see the value in manipulating it) while remaining unaware that the variable is being manipulated (to maintain an observational frame). (Sloman et al., 2010, p. 270; my emphasis)

In other words, the authors argue that if subjects are quicker to reach a dot after being told that this correlates with their intelligence, then this is an indication of self-deception, or that the subject knew of the desirable correlation, speeded up, but denied doing this on purpose. The procedure that according to the authors allowed testing diagnostic self-deception was the following:

1. **Tracking task** (without feedback): Participants had to move the mouse of a computer from the starting point to a dot. Location of the dot varied, but average distance was equal.
2. **Background information** and cover story is given to participants: The tracking task was to measure **spatio-motor speed**, which is a kind of processing speed. The experiment was said to measure the connection between spatio-motor and processing speed and the information was given that spatio-motor speed is either positively or negatively correlated with the general intelligence.
3. Participants then filled the **questionnaire** answering how likely it is that they possess an above-average spatio-motor speed and how happy they would be if it were true.

4. Then a **control task** was administered: "40 simple estimation questions," supposedly to return to baseline level of computational speed which included spatio-motor and processing speed task.
5. After that a **processing speed** task had to be accomplished: Math questions, like generating Fibonacci sequences, had to be answered, as well as word listing tasks.
6. Again, a **tracking task** had to be accomplished. This time, various sorts of feedback were given. **Vague feedback** included showing the participant on a horizontal bar after each trial whether they were 'fast,' 'slow' or 'average' with respect to the first time that they performed the tracking task. **Precise feedback** condition included

a. showing the participants on a horizontal bar dots on a scale that indicated their exact time in milliseconds for every trial during the first tracking task and

b. showing on top of the dots their time for each trial during the second tracking task.

7. Again, a **questionnaire** was administered that asked whether the background information according to participants affected their performance and whether they cheated.

The results to which Sloman et al (2010) came were that diagnostic self-deceivers speeded up given the cover story that speeding up indicated high general intelligence, but only if the feedback was vague. The authors also conducted a similarly structured experiment in which they compared whether the information about the "reference class" mattered: participants were either told that the feedback they will get is that about themselves, or about "a group of other Ivy League students" (p. 279). Self-deception is said to have occurred in both conditions. Sloman et al (2010) also varied the procedure in omitting point 3 – asking the participants about the computational speed prior to the test phase, because no correlation between the happiness to find out that one was above-average and the speed up could be established, which is actually counter-intuitive. Yet, this did not help to establish the correlation and led the authors to the following conclusion:

> It is not clear to us why a more positive inference was not drawn in Experiments 3 or 4. Perhaps it was too obvious and thus would have revealed the deception. That is, the failure to draw the inference explicitly may itself have been a form of self-deception. Another possibility concerns the nature of the attribute people were deceiving themselves about. Unlike heart type, participants already know something about their intelligence. It may be that self-deception in our experiments merely served to confirm what they already believed. (Sloman et al., 2010, p. 281)

In one follow-up experiment the authors also asked the participants to indicate on a (7-point) graded scale how uncertain they were about whether they manipulated their performance and got negative answers that participants did not manipulate the performance which suggests that they were not aware of the manipulations. This is according to the authors supposed to provide better evidence whether the participants possessed some awareness of their manipulation than Quattrone & Tversky's dichotomous yes/no response to determine awareness of influence" or Gur & Sackeim's "galvanic skin response measure" (Sloman et al, 2010, p. 277). Fernbach et al. (2014) have further linked diagnostic self-deception to *effort*, according to the rational that exerting effort should preclude participants from drawing the conclusion that their behavior is diagnostic of some beneficial outcome, given that efforts hints at participants' intervention at achieving this beneficial outcome. The cover stories that Fernbach et al. (2014) used were that pain endurance is indicative of the presence of certain chemical in the skin that enhances skin quality, as well as that finding hidden objects in pictures indicates good/bad self-control. The results indicated that participants that changed their behavior in line with the positive expected outcome reported less effort and, thus, supposedly engaged more in *effort denial.*

Summing up, subjects speed up after being told that this would show that they are more intelligent, but do so only if they are not shown on the screen how fast they were on a test trial, so that they could judge that they have speeded up and this speeding up does not correlates with measures of happiness. These subjects are also unaware of the manipulation, if they are asked about it. Is this a case of self-deception? Clearly, *motor* responses do not require justification. The fact that one was slower on the test trial could be explained by the subjects by insufficient motivation to excel on the task. Further, a positive emotional reaction – happiness – does not correlate with speeding up. The cover story may have *caused* the participants to change their behavior, but it was not present *as a reason* for such a change. As the original pain endurance paradigm, this study also makes an invalid inference from physiological (subpersonal) measures to psychological (personal) states of participants (see section 1.3.4).

1.3.2.4 Paulhus: Cheating on practice tests study

In Paulhus & Buckels (2012), the authors summarize two experiments their laboratory had done following the Quattrone & Tversky's paradigm which shows that people confuse "diagnostic contingencies with causal contingencies" (p. 370). Paulhus & Buckels' (2012) experiments attempt to show that self-deception "requires not only motivated cognition but also the additional feature of discrepant representations" (p. 367). Participants had to take an achievement test and were given the freedom to terminate the test even after the fixed amount of time that was given. In the first study there were three conditions (self-enhancement, reward and control). In the self-enhancement condition participants worked the longest on the task and their self-report indicated that they believed the cheated score. In the second study, instead of the reward condition, a handicapping condition has been added to confirm the presence of motivation.[138] The handicapping cover story was that high test scores are associated with schizophrenia. The results were consistent with the first study. The context the experimenters provided about getting high scores on the test (being successful or being prone to schizophrenia) affected the amount of time participants worked on the task, leading the authors to conclude that "[s]ubjects can be motivated to excel or fail on a test that is supposed to inform them about their current ability level" (p. 371). The authors present this as a case of self-deception, because the manipulation of the time spent working on the test should have precluded participants from concluding that the results of the test indicated their performance level (p. 372). Participants, however, justified the accuracy of the scores achieved by such cheating (p. 371). It is in question though, whether there is a contradiction for participants between the time they took to accomplish the tasks and the conclusions they drew about their performance, given that no information about physiological or some other measurements of tension was provided. Participants seem not to *experience* the contradiction between the time they took to accomplish the tasks and the conclusions they drew about their performance (and there is no information about physiological or some other measurements of tension).

Interim conclusion: I have review three studies – that of Quattrone & Tversky (1984) in which participants held their hands longer in cold water after being said that it is a sign of health, that of Sloman et al. (2010) in which participants reached the dot on the screen with a mouse more quickly after being said that this correlates with intelligence and that of Paulhus & Buckels (2012) in which participants worked longer on a task, but still held this

[138] A criterion for *motivated* behavior has been argued to be the possibility to induce a *reversal* of behavior: "To confirm that a behavior is motivated, the researcher must show that its direction can be reversed" (Paulhus & Buckels, 2012, p. 371).

task to be indicative of their ability. These studies clearly demonstrate that participants behaved in a motivated manner, but it is not the case that there need be an inconsistency in reaching the conclusions that the participants did, and thus, no contradictory evidence either. Especially in the first two studies, participants' motor responses were tested. Hypothesizing that they just try their best, without there being any inconsistent or contradictory evidence whatsoever is the fact that in Sloman's et al (2010) study the group being told that speeding up correlated with *low* intelligence did not slow down (p. 277). These studies are similar to Ditto's (2009) study in which participants who had been told that if they put their saliva on a piece of paper, it will change color in case of a positive medical diagnosis, waited longer for that piece of paper to change color. The difference is only that in Ditto's study it was not the case that the *same* participant had to put their saliva and wait twice. Thus, in Ditto's study it is clear that there is no inconsistency whatsoever and at least *any* kind of inconsistency should be there for something to be labelled 'self-deception.' In Quattrone & Tversky's and Sloman's et al study the same participant underwent the same procedure twice, one with and once without the motivation being given. But still, having a certain motivation to try one's best is not a case of self-deception. Paulhus' study comes near to showing an instance of self-deception, but also there, participants' criteria for measuring adequacy of the achievement test may have deviated from those assumed by the experimenters, namely the time of completion of the test. This may have influenced how participants fix the content of their attitudes and enforced a difference in content fixation between participants and experimenters.

1.3.3 Does motivation influence quantity or quality of processing?

In the previous section I argued that it is questionable whether Gur & Sackeim's, as well as Quattrone & Tversky's paradigm demonstrates a personal level inconsistency that self-deceivers have been accused of. The aim of this section will be to show that studies made according to the Quattrone & Tversky's paradigm also do not demonstrate any kind of *qualitative* influence on the belief forming processes of the self-deceiver. What the quantity view demonstrates is that the pain endurance paradigm allows for a weak, if ever, kind of discrepancy on the personal level.

I will proceed by means of a comparison. First, I will shortly present Kunda's (1990) popular theory of motivated cognition according to which motivation influences the *quality* of processing. Helzer & Dunning (2012) even posit motivated cognition as a paradigmatic case of self-deception. Then, I will briefly review Ditto's (1998) alternative *quantity* of processing view that has a very similar paradigm to that of Quattrone & Tversky. Ditto's color change experiment has also been mentioned by von Hippel & Trivers (2011b) as one of examples of self-deception.

Kunda (1990) promotes the distinction between two kinds of goals that affect reasoning – accuracy goals and directional goals. Both fall into the broad category of motivation which is defined as "any wish, desire, or preference that concerns the outcome of a given reasoning task" (p. 480). Kunda describes two possibilities by which motivation could influence the reasoning process: the one is by *confirmation bias* that serves as an intermediate variable between motivation and reasoning and the other one which consists of motivation influencing the *knowledge structures* taking part in the reasoning process (pp. 494-495). The choice of the hypothesis and the subsequent biasing of hypothesis testing has been taken to be the mechanisms of self-deception by Mele (2001).

Indirect influence of motivation	Direct influence of motivation
"One intriguing possibility is that the motive, or goal, merely leads people to ask themselves whether the conclusion that they desire is true; they ask themselves directional questions: "Do I support police intervention on campus?" "Am I extraverted?" "Is my date nice?" Standard hypothesis-testing processes, which have little to do with motivation, then take over and lead to the accessing of hypothesis-confirming information and thereby to the arrival at conclusions that are biased toward *hypothesis confirmation* and, inadvertently, toward goal satisfaction." (p. 494; my emphasis)	"These effects of directional goals on memory listing, on reaction time, and on rule use provide converging evidence for the notion that goals enhance the accessibility of those *knowledge structures – memories, beliefs, and rules* – that are consistent with desired conclusions. Such selective enhanced accessibility reflects a biased search through memory for relevant knowledge." (p. 494; my emphasis)

Table 16. Kunda: influence of motivation on reasoning.
Distinctions from Kunda (1990).

Though Kunda (1990) points out that evidence cannot rule out neither of these hypotheses (p. 495), in the given article the author argues for the *direct* kind of influence. Kunda writes that it is important to distinguish two kind of goals (accuracy and directed ones), because "there is no reason to believe that both involve the same kinds of mechanism" (p. 481). Those two kinds of goals have also been differentiated in the literature on self-deception. Hypothesis-testing whose motivation is directional may be biased and is the opposite of "objective processing" (Kunda, 1990, p. 491).

Thus, Kunda (1990) holds that directional goals influence the *selection* of

- declarative ("beliefs about the self, other people, and the world", p. 488) and
- procedural (inferential rules) knowledge structures
- with memory playing a constraining role, because people maintain the appearance of objectivity (p. 483).

Directionally motivated phenomena are among others self-serving biases, failures of probability estimates and base rate estimates (Kunda, 1990). More extensive and thorough processing – the characteristic of reasoning guided by the accuracy motivation – may still be biased (Kunda, 1990, p. 491). As an example of such, Kunda (1990) states that subjects of one study that were concerned to a high degree with the issue under consideration processed the information more thoroughly, but that they "listed more thoughts about their own position and had a higher percentage of thoughts that were unfavorable to the counterattitudinal arguments" (p. 491). Has here motivation changed only the duration of the process, or also the kind of processing itself?

Against Kunda Ditto (2009) argues that motivation changes the quantity and not the quality of information processing (p. 30). Ditto (2009) accuses the qualitative view in that psychological experiments induce motivated processing, demonstrate the results of motivated processing and based on the results make inferences about the reasoning process on the assumption "if motivational factors are to affect cognitive *outcomes*, they must do so by affecting some aspect of cognitive *process*" (p. 29). Ditto's (2009) view, on the contrary, is that preference-inconsistent information is coupled with negative effect, analyzed more thoroughly. An effortful cognitive analysis of information leads to the generation of a bigger amount of alternative hypotheses and to a stricter decision criterion on the acceptance of this information (p. 32). The assumptions on which this view bases are:

1. *Expectations and preferences* independently, but analogously affect the intensity of the reasoning process, e.g. if a medical test suggest I am sick when I have little reason to believe it given my prior health diagnoses, then I will be more skeptical, but even given two

individuals having equal prior health expectancies, the one who gets the unfavorable diagnosis would be more skeptical (p. 33).

2. The *amount of alternative hypotheses* that explain the unwelcome information is correlated with the duration of the reasoning process: the longer the reasoning process, the more alternative hypotheses (p. 33).
3. The motivation leads people to be *uncertain*, but does not influence the judgment of whether the unwelcome information is accurate (p. 34).

As evidence for this view, Ditto (2009) brings several studies he made with colleagues that demonstrated *motivated skepticism*. The design of one study was that participants had to put saliva on a piece of paper, wait for it to change color (within 20 seconds) having been told that the color change would indicate either a favorable or a unfavorable health condition and seal the envelope. The results were that those participants for whom lack of color change was associated with an unfavorable diagnosis waited longer and exhibited such skeptical behavior as e.g. reopening the envelope to check again (pp. 37-40). It is to stress that similarity to Quattrone & Tversky's pain endurance study that also had health prediction as motivation: duration of holding a hand in the cold water indicated a healthy type.

Along similar lines Ditto with colleagues demonstrated *motivated sensitivity* to unfavorable information that is argued to initiate more effortful processing for unfavorable information. Ditto (2009) argues that given that Kunda accepts the "illusion of objectivity" or constructing an explanation for desired conclusions, Kunda's model would predict symmetrical consideration of preference-consistent and inconsistent information (p. 41), whereas Ditto's model predicts asymmetrical sensitivity. I do not think that this needs to be the case, because from an *appearance* of objectivity it does not follow that favorable and unfavorable information should be considered in an equal way, but let me explain the experiment. The assumption of Ditto et al. (1998) has been tested using fundamental attribution error (tendency to underestimate situational factors). Male participants received flattering or unflattering evaluations by female confederate and they had to state how much the confederate liked them given the information that the evaluations she gave were made freely or were constrained. The results showed that participants were sensitive to the information about whether the confederate could choose which evaluation to give when they received unfavorable evaluation, but attributed the favorable evaluation to choice in both cases. Ditto et al. (1998) similarly changed the disease prognosis paradigm described above: Given statistical information about the accuracy of the test (1 out of 200 chance vs. 1 out of 10 chance), the participants accepted the negative results of the test when the test was perceived as highly accurate. These motivated sensitivity experiments differ from the motivated skepticism ones in that here it is not only *duration*, but there is also some sort of a cognitive mistake that participants can be accused to be guilty of. Thus, motivated sensitivity, in difference to motivated skepticism, might have changed the processing of information in a certain way.

Interim conclusion: In this section I presented two views: Kunda's *quality* of processing view, according to which motivated reasoning changes knowledge structures (memories, beliefs, rules) and Ditto's *quantity* of processing view, according to which motivated reasoning only changes the *duration* of processing. Which is the case in self-deception? And if both may be at work, can one of them be seen as a paradigmatic case of self-deception? Von Hippel & Trivers (2011b) consider waiting longer as self-deceptive in the context of Ditto's color change experiment[139] (*motivated skepticism*). Quattrone &

[139] Interestingly, among the advantages of his view Ditto (2009) names that the QOP avoids self-deception, because, instead of intentional "cherry-picking of the available evidence," there is only the goal of accuracy that is pursued (p. 44).

Tversky's pain endurance paradigm is of the same kind as Ditto's color change experiment – it demonstrates that participants wait longer for favorable results, but does not demonstrate changes in the quality of processing. According to the quantity of processing view, more *effortful* processing (*motivated sensitivity*) also does not need to evoke changes of knowledge structures. This is an empirical question, which view is correct – quantity or quality of information processing. The implication of the quantity view would be that self-deceiver's bias their belief-forming processes in a very weak sense. An interesting question is how much *control* over belief-forming the quality and the quantity view allows. I will come back to this question in section 2.2.1, because the question about control is connected to the kind of selectivity that is at work in both cases. So far, the quality view is preferred in the philosophical literature. Interestingly, recent predictive coding theory suggest that precision is correlated with the speed of decision making in a perceptual decision task (FitzGerald et al., 2015). Taken together with the claim that aberrant precision is responsible for psychopathology, predictive coding seems to be able to incorporate the quantitative view discussed in this section.

1.3.4 Inference from the subpersonal to the personal level

In section 1.3.1 I discussed questionnaires designed to test self-deception and argued that intentional objects of these measurements are unclear. I, then, introduced in section 1.3.2 two paradigms for testing self-deception: Gur & Sackeim's voice recognition experiment, as well as Quattrone & Tversky's pain endurance study, as well as two follow up studies. In section 1.3.3 I, thereafter, argued with respect to the second paradigm that it satisfies the disunity in unity constraint in only a very weak form. In this section I want to voice a point of critique that concerns both paradigms – inferences made from the subpersonal to the personal level.

In Gur & Sackeim's experiment skin conductance response was taken as an indication of a contradictory beliefs in the self-deceiver. In the Quattrone & Tversky's study, as well as Sloman's et al study participants are argued to cause changes in the two kinds of responses that are assumed to be uncontrollable – the pain response and the spatio-motor speed. Due to performing the task several times (with and without motivation) they are further argued to *notice* the change of responses that should have led them to the hypothesis that they themselves manipulated these responses. Yet, in Quattrone & Tversky's study it is unclear whether the phenomenology of pain perception really changed and in Sloman's study precise feedback is argued to preclude changes in motor speed. In Paulhus' study finally participants work longer on a task indicating their intellectual ability than they are supposed to. This study differs from the first three in that only there participants can be said to give *reasons* for their behavior, while in the first three studies motivation *causes* changes in participants' responses. Yet, Paulhus' study allows, particularly in virtue of participants justifying their attitudes, for the weakest amount of personal level discrepancy. [140]

My critique of the experiments will proceed as follows: On the basis of the distinction between personal level *explanations* and personal level *states* and subsequently the one

[140] Paulhus' study is similar to Greve & Wentura's studies on self-immunization in participants being able to justify their acquired attitudes. Self-immunization is the phenomenon, which consists in participants changing the description of a certain psychological trait that is important to them, in order to be able to further ascribe this desirable trait to themselves in cases in which their prior description of the trait does not fit anymore (Greve & Wentura, 2010).

between *doxastic* and subdoxastic states, I will argue that it is in question whether skin conductance response (or other physiological measure that might be made use of in psychological experiments testing self-deception) is a doxastic kind of state and even if it is, which kind of *content* should be ascribed to it. Particularly, content fixation might differ between experimenters and participants that were subject to certain testing procedures.

In previous sections I have already mentioned that one important question for self-deception is about the relation between the personal (reasons) and the subpersonal levels (causes) – the interface problem, as Colombo calls it by referring to Bermúdez (Colombo, 2013, p. 548; see table 17 for more on the distinction between different kinds of explanations). Self-deception is often described in folk-psychological terms (beliefs, desires etc.), yet explanation of this kind leave the question about the *mapping* from folk-psychological states to those posited by cognitive science open (Waskan, 2006, p. 37). Jonathan Waskan (2006) defends folk psychology on the premise that though it has difficulty predicting particular behavior, it can be *progressively refined* to gain knowledge about the *mechanisms* generating that behavior (p. 58). Waskan (2006) argues further that one can map beliefs and desires onto units of research in cognitive science, e.g. memories (p. 60). Yet, the question of interest here is not about folk psychological explanations in general, but those of self-deception in particular. And in explaining self-deception, folk-psychological explanation is stretched to its limits.

Zoe Drayson's (2012) has argued that an inference from persona/subpersonal *explanations* to respective *states* requires the presupposition of certain additional (metaphysical) claims[141] (p. 8-10). Subpersonal explanations can be treated as heuristics (p. 9). On the level of personal explanations, the inference from beliefs/desires to the postulation of internal representations and further to their nature requires additional argumentation.[142] When the distinction between explanations and states is not carefully handled, then the distinction between personal/subpersonal explanations cannot be drawn clearly.[143] One has also to be aware that there is, on the one hand, an explanatory gap in explaining psychological features in terms of physiological ones and, on the other, a mereological fallacy in explaining physiological features by psychological ones[144] (Drayson, 2012, p. 3). In the

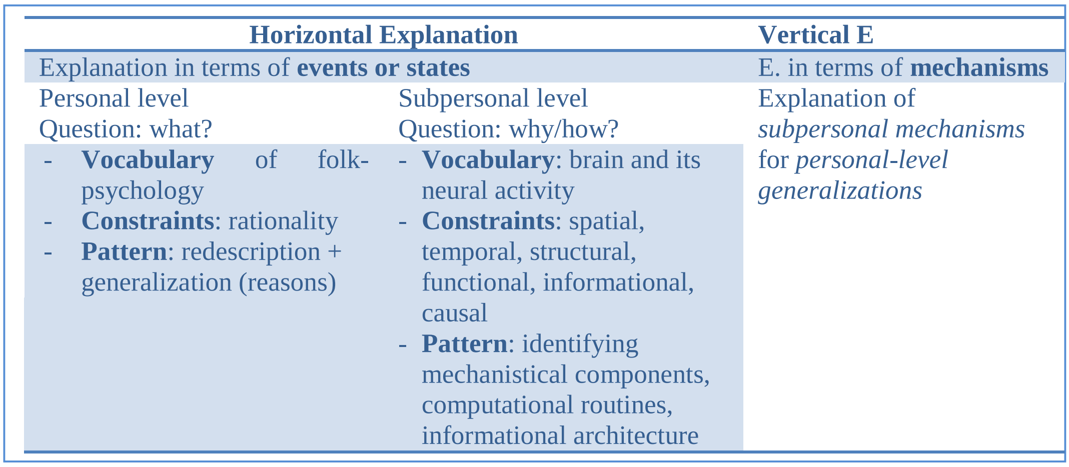
[141] The inference from explanations to states needs additional metaphysical assumptions: "The existence of personal and subpersonal psychological explanations does not directly result in a commitment to personal and subpersonal psychological states: for that, one needs to supplement the claim about our explanatory accounts with some metaphysical claims" (Drayson, 2012, p. 10).

[142] *Presence* of internal representations is not to be confused with their *nature*: "On an alternative construal of folk psychology, one might think that we instantiate a relation to a proposition in virtue of having an internal representation: an internal state that bears the content of the proposition. This representational theory of mind involves a three-place relation rather than a twoplace relation: the relation between people and the content of their thoughts is mediated by internal representations. But even on this view, we are not committed to any particular view of the nature of mental representations" (Drayson, 2012, p. 9).

[143] It is not a given that since there a personal level belief, they have a subpersonal level analogue, or even that those same beliefs can play an analogue explanatory role in personal, as well as in subpersonal explanations:
"Notice that when we identify the posits of personal and subpersonal explanations, whether or not computational theory is involved, we lose any notion of a distinction between personal and subpersonal *states*: the terms of personal and subpersonal explanations refer to the same entities. So not only does the original personal/subpersonal distinction fail to licence *any clear distinction* between personal and subpersonal states, personal and subpersonal states become *indistinguishable* when common metaphysical claims are combined with the explanatory form of the personal/subpersonal distinction." (Drayson, 2012, p. 11)

[144] Mereological fallacy: "The alternative approach would be to ascribe psychological instead of physiological predicates to a person's components: this would seem to count as genuinely

four experiments discussed personal level states are assumed to be present on the basis of certain physiological responses.



Horizontal Explanation		Vertical E
Explanation in terms of **events or states**		E. in terms of **mechanisms**
Personal level Question: what?	Subpersonal level Question: why/how?	Explanation of *subpersonal mechanisms* for *personal-level generalizations*
- **Vocabulary** of folk-psychology - **Constraints**: rationality - **Pattern**: redescription + generalization (reasons)	- **Vocabulary**: brain and its neural activity - **Constraints**: spatial, temporal, structural, functional, informational, causal - **Pattern**: identifying mechanistical components, computational routines, informational architecture	

Table 17. Colombo: horizontal/vertical, personal/subpersonal distinction
Distinctions from Colombo (2013, pp. 549-552).

Horizontal explanations may be personal or subpersonal, each characterized by its own vocabulary, constraints and patterns. Personal level explanations are formulated in folk-psychological terms, constrained by rationality and provide redescriptions or generalizations of the explanandum. Subpersonal kinds are governed by constraints on the brain architecture, representation and function. Note that Drayson (2012), in distinction to Colombo, seems to equate the personal/subpersonal with the horizontal/vertical distinction. Drayson (2012) states that personal level explanations are the folk-psychological ones and thus, horizontal, while subpersonal level explanations are vertical: personal level explanations are given in terms of sequences of mental events, while subpersonal ones – in terms of certain features of physiological components underlying psychological features (p. 3): "The point of the personal/subpersonal distinction is to emphasise that there is a type of psychological explanation which is not folk-psychological: it is not horizontal, and it does not consist in ascribing psychological predicates to whole persons." (Drayson, 2012, p. 7)

Even if personal level states are dependent on the right kind of neurocomputations for their occurrence,[145] some subpersonal states, e.g. skin conductance, may not be available at the personal level at all. This point is directed mainly at Gur & Sackeim's experiment. This is where the distinction between *doxastic and subdoxastic states* comes into play or the distinction between states that correspond to personal level mental states and those that do not[146] (Drayson, 2012, p. 13). It is different from the personal/subpersonal distinction. If the metaphysical claim is accepted that equates propositional attitudes with functional/computational states, then no distinction between *personal and subpersonal states* can be drawn, but one still can meaningfully distinguish doxastic from subdoxastic

psychological vertical explanation. Here, however, vertical psychological explanation can be accused of committing the mereological fallacy: ascribing to a part of something a predicate that can only correctly be ascribed to the whole thing" (Drayson, 2012, p. 3).

[145] A claim defended by Colombo (2013): "My argument is about etiology insofar as personal-level phenomena and behavior are causally generated by the *right kind of neurocomputations*. My claim here is twofold: first, at least some personal-level phenomena and behaviors are conceptually bound up with their being grounded in the right kind of neurocomputations—or internal causes; second, constitutive relevance can be relative to the state of empirical knowledge at a given point in time" (p. 566; my emphasis).

[146] The distinction between dochastic and subdochastic states: "This talk of cognitive subsystems makes it clear that Stich is drawing a distinction between two kinds of functional component that appear in subpersonal psychological explanations. His distinction is between those components that map onto everyday mental states, and those components that don't" (Drayson, 2012, p. 13).

states, because though every propositional attitude can be reduced to a certain subpersonal state, the reverse need not be true[147] (Drayson, 2012, p. 12; see figure 7).

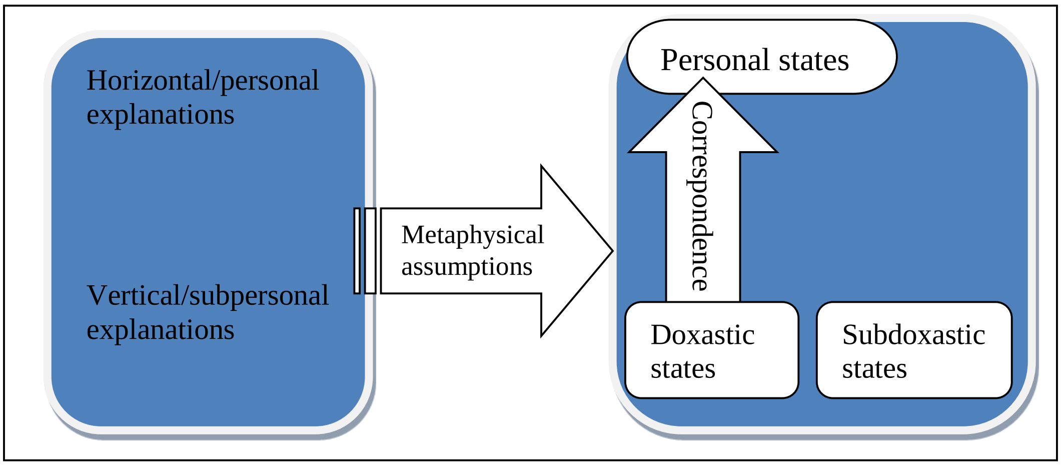


Figure 7. Drayson: kinds of explanations
Distinctions from Drayson (2012).

Is skin conductance response a doxastic kind of state and if yes, with which kind of content? In section 1.3.2.1 I mentioned at least two alternative kinds of contents that a doxastic state, originating from a skin conductance response, could have: a propositional kind of content, or an affective kind of content. If it is an affective kind of phenomenal content, then it may influence certain kinds of processes subpersonally without inducing certain kinds of propositional contents. Yoshie & Haggard's (2013) study shows that changes in the desirability of goal-representations may have a low-level sensorimotor basis. Yoshie & Haggard (2013) manipulated the emotional valence of action outcomes by following participants' key-press actions with negative or positive emotional vocalizations, demonstrating that the sense of agency can be modulated by the emotional effects of an action. The results of their study suggest that the sense of agency is attenuated for negatively valenced events and that this effect may possess a low-level sensorimotor basis,[148] indicating that the process of desirability evaluation is a subpersonal one. The further value of Yoshie & Haggard's study is that the desirability has been dissociated from context, because the negatively valenced event is a negatively perceived sound stripped of any informational context. Peetz et al. (2014), on the contrary, have found that *implicit* (associated with positive/negative valence, measured by IAT) future self-esteem correlated with future-oriented motivation in contrast to *explicit* (deliberate) future self-esteem (measured by Rosenberg Self-Esteem Scale) that correlated with positive fantasies and general optimism. These results question the lower-level basis of optimism, yet the validity of IAT may also be questioned. Peetz et al.'s (2014) hypothesized explanation is that the concept of 'explicit self' might be more influences by self-enhancement due to selective

[147] Moreover, not only is the personal/subpersonal distinction different from the doxastic/subdoxastic distinction, but also from the one between *conscious/unconscious states*. In the case of cognitive unconscious the distinction between doxastic/subdoxastic states may suit better, while in the case of Freudian unconscious no subpersonal states figure in the explanation, due to those explanations being only personal level horizontal ones (Drayson, 2012, p. 15).

[148] Noteworthy is that exaggerated belief in one's personal control has been suggested by Taylor (1989) to be one of the three self-deceptive illusions, among self-enhancement and unrealistic optimism.

attentional processes (p. 102). Summing up, how skin conductance response is to be interpreted at the personal level is in question.

Apart from the question about the presence of a doxastic counterpart of a subpersonal state per se and the question which kind of content such a doxastic state possesses, what has to be taken into account is the possibility that self-deceivers *fix* content in a different way. In this case, experimenters would ascribe different kinds of content to self-deceivers, e.g. in the pain endurance paradigm experimenters would ascribe to participants a personal level state with the content that one's actions are controllable. Participants themselves, on the other hand, would ascribe to themselves a personal level state that their actions are not controllable. More generally, in the described experiments certain beliefs are ascribed to self-deceivers on the basis of the causal connection between the motivation and the reactions of the participants of which participants are not aware. Only in Paulhus' experiment they can be plausibly argued to be aware of their manipulation of the duration of the test, but this makes the results dangerously close at falling prey to Davidson's paradox of irrationality: Depending on how plausible their reasons for manipulating duration are, it is not a case of self-deception anymore, because irrationality has been explained away.

That an explanation of self-deception requires a relational theory of consciousness has been argued by Clegg & Moissinac (2005) who argue that different observers would make different evaluations because of the ambiguity of objects of experience (p. 101). They do not mention content fixation though. Waskan (2006), from the other point of view, argues that we folk-psychologically, from an *observer* perspective,[149] ascribe content in an externalist way such that factors external to individuals are considered to influence this content (e.g., Twin-Earth experiment; p. 83). According to Waskan (2006), it is not a problem of folk psychology that the content of beliefs is codetermined by "doxastic surrounding," or a person's particular belief network (Waskan, 2006, p. 47), because the mechanisms of belief formation are shared among individuals (p. 68). In case of self-deception though, certain beliefs do not seem to be shared and, thus, there is a discrepancy in the content fixation. This has to be accounted for in the experiments testing for the presence of self-deception.

So far I have used the concept of 'belief' and 'representation' without further clarifications. I have noted in section 1.3.1 that beliefs are personal level entities, while representations are not necessarily personal level entities. Here, I want to note the difference in their content. Subpersonal biases process *information*, personal level reasoning processes operate with *beliefs* as units where beliefs are relations between persons and propositions. Both information and beliefs may be represented in a certain manner. Representations are certain information-bearing structures (Pitt, 2012). Information itself is different from representation, as well as from belief. Dretske (1982) holds that semantic content is the "outermost informational shell" (p. 178) in the sense in which, on the one hand, in the information, that *t* is a square, the information, that *t* is a quadrilateral, is contained (or nested), but, on the other hand, the belief, which has as its content that *t* is a square, does not have some other content that *t* is also a quadrilateral nested in it:

> A belief, therefore, exhibits a higher-grade intentionality than does a structure with respect to its informational content. Both may be said to have a propositional content,

[149] Waskan (2006) further argues that from an *actor* perspective, it is the isomorphy between our representations of the world how it is and the isomorphy between our representations of the world how we wish it to be that enables effective behavior (p. 96). The relational content of such representations are the conditions of their veridicality (p. 105).

> the content, namely, that *t* is *F*, but the belief has this as its *exclusive content* (to the exclusion, at least, of nomically and analytically related pieces of information) while the information structure does not. A physical structure cannot have the fact that *t* is *F* as its informational content without having *all* the information nested in *t*'s being *F* as part of its informational content. (Dretske, 1982, p. 174)

Though conceptual distinctions between information, representations and beliefs have never been an issue in self-deception, it would be beneficial for the reader to keep this distinction in mind. As suggested by Thomas Metzinger, *facticity* of beliefs or that something *is* the case is important to explain, particularly because in section 2.1.3 and later in chapter 4 I will show how *counterfactuals* (that something could be (have been) the case) plays a role in self-deception. Here, my point is only that observers' and self-deceivers' belief content might differ in that the former would not contain facticity. Something of which self-deceivers think that it really is the case might be a mere possibility for an observer. I will come back to this point in section 2.2.3.

Interim conclusion: The garden-variety cases of self-deception underlying its explanation, as well as the explanations themselves, have mostly been formulated in folk-psychological terms. Self-deception being the case of a violation of rationality, subpersonal explanations need to be taken into account. Yet, there is no direct correspondence between the folk-psychological terminology and internal representations of the self-deceiver's system. One has also to be sensitized to the assumptions one makes in switching from the personal level *explanations* to personal level *states*. One has to consider the possibility that there are subdoxastic states that do not feature in personal explanations which may improve the quality of the explanation, as well as that doxastic states might differ in the kind of personal level content or fixation of this kind of content between self-deceivers and observers, e.g. in controllability or facticity.

2 Self-deception – goal-directed subpersonal hypothesis testing

> Following this line of thinking, one might argue that resistance to scientific progress is often rooted in self-deception, a desperate clinging to outmoded (but familiar) ideas in lieu of risky new concepts.
> (Bornstein, 1997, p. 109)

In the first chapter I set four constraints on explanations of self-deception: disunity in unity, phenomenological congruency, parsimony and demarcation. The first two will be in the focus of this chapter. Namely, I will, first, argue that intuitions have played a great role in explanations of self-deception (2.1.1). I will, then, hold that it is more parsimonious to set out as an explanandum behavioral and phenomenological characteristics, instead of hypothesized kinds of internal states that might underlie self-deception. Self-deceivers exhibit inconsistency, which they justify, on behavioral level, as well as tension, insight and a counterfactual goal-directed pull on the phenomenal one (2.1.2 and 2.2.1). In section 2.1.3 I will argue that tension is a kind of a metacognitive feeling. In the subsequent sections I will be concerned with the choice of motivation, process and self-deceptive attitude[150] by means of phenomenological arguments (and parsimony). Subpersonal goal representations build a motivation for self-deception (2.2.1). Dolphin model of cognition, which I will compare to linear and dual process models, best explains the epistemic agent models that self-deceivers might possess (2.2.2). Finally, self-deceptive attitudes are characterized by degree of consciousness, uncertainty and transparency and it is a latter characteristic that actually distinguishes between the two main kinds of selection that underlie self-deception (see section 2.2.1 on selection types). Before I start, a tentative definition of self-deception:

SD_{Def}: Self-deception (SD) is
(1) a motivated kind of
(2) hypothesis-testing, that
(3) results in an evidence-incompatible mental representation of reality, which
(4) fulfills a belief-like role.

The first, as well as the third and fourth points stem from the analysis given in the first chapter. The second point will be, first, introduced in section 2.2.2 and defended in chapter 4.

2.1 Simplifying an explanandum for a theory of self-deception

In this section I will pursue three aims: First, I will argue that the kinds of approaches to self-deception offered in the literature so far depend on intuitions (2.1.1). Then, I will propose to restrict an explanandum of self-deception to behavior and phenomenology. It is certain to be a restriction, because no additional internal states will need to be postulated in this way. Afterwards, I will offer a characterization of the behavior (inconsistency + justification) and phenomenology (tension + insight). Last, I will propose a description of tension as an epistemic feeling based on Proust's (2013) theory of metacognition. It will serve as a bridge to the following section 2.2.1 in which I will argue that three different kinds of goal-directed object selection might be responsible for self-deception. This is

[150] The dolphin model of cognition (see section 2.2.2.3), as well as the properties of self-deceptive misrepresentations (see section 2.2.3) are modified versions from selected parts of Pliushch & Metzinger (2015).

because a kind of feelings arising from self-deception constrains the kind of selection that might be responsible for it.

2.1.1 Role of intuitions for explanations of self-deception

The aim of this section is, first, to argue that intuitions play a large role in explanations of self-deception and, second, to argue that self-deceptive attitudes and intuitions might share at least one feature, namely that the phenomenal signature of knowing has become transparent.

Mele (2001) distinguishes three kinds of approaches to definitions of self-deception: *lexical* which centers on definitions of "deceive", *example-based* that analyses representative examples of self-deception and *theory-guided* that is oriented towards an explanation of a commonsense theory about self-deception.[151] Mele (2001) states that the lexical approach is ambiguous in that there exists at least two definitions of "deceive:"

- the stereotypical intentional one (A knowing/believing not-p causes B to believe p) that elicits the well-known static and dynamic paradoxes of self-deception[152] and
- the non-intentional one ("deceive" as "to cause to believe what is false"; p. 8).

This further need of specification for the lexical approach lets me assume that a lexical definition itself depends on something, namely either on representative examples of self-deception or on commonsense theories.[153] Mele himself does not specify which kind of approach to self-deception he has, though he emphasizes that self-deception is an explanatory concept for him: the function of the concept of self-deception is to explain data (Mele, 2001, p. 10). Data from psychological experiments are also interpreted according to a certain theory. Examples may depend on a folk-psychological theory of self-deception as well. Thus, the given three kinds of approaches are interdependent. Moreover, Mele has left one crucial variable out: intuitions.

It seems as if a lot of authors would agree that intuitions are at least a necessary part of giving a definition to self-deception, so, for example, Kent Bach (1998) who speaks about *intuitively* acceptable examples of self-deception as a basis of an analysis of self-deception.[154] Scott-Kakures (2002) argues that an account that can satisfy intuitions about

[151] The approaches to defining self-deception have been described as follows: "In defining self-deception, three common approaches may be distinguished: *lexical*, in which a theorist starts with a definition of "deceive" or "deception", using the dictionary or common usage as a guide, and then employs it as a model for defining self-deception; *example-based*, in which one scrutinizes representative examples of self-deception and attempts to identify their essential common features; and *theory-guided*, in which the search for a definition is guided by commonsense theory about the etiology and nature of self-deception. Hybrids of these approaches are also common" (Mele, 2001, p. 5).

[152] The static paradox consists of the fact that an impossible state of mind is needed for self-deception: believing and not-believing p at the same time, while the dynamic paradox states that an impossible deceptive strategy is present in self-deception: a deceptive strategy aimed at oneself (Mele, 2001, pp. 7-8).

[153] Mele speaks only about a commonsense theory of self-deception, but if one takes into account the mutual influence of intuitions and theory then it becomes worth considering that as members of the scientific community develop a theory of self-deception, their intuitions about it change too, as well as some elements of the commonsense theory.

[154] An example of self-deception has been argued to need to be intuitive: "Any analysis should encompass a range of *intuitively* acceptable examples of self-deception (the analyst should take care that his choice of examples not be influenced by any self-deception about self-deception itself) and exclude such distinct phenomena as wishful thinking, denial, repression, and fanaticism. […] Also, insofar as the meaning of 'self-deception' is a function of the meanings of 'self' and of 'deception', an analysis cannot be oblivious to the meanings of the words making up the phrase" (Bach, 1998, p. 164; my emphasis).

self-deception should be favored (p. 591). Tamar Szabó Gendler (2007) speaks about taking into account a *natural* description of self-deception which, as I think, could be interpreted as intuitions (p. 240). Funkhouser (2009) also speaks about the dependence of the explanation of self-deception on our initial understanding[155] which is, according to me, nothing else than our intuitions, influenced by our folk-psychological theory (or maybe vice versa - our intuitions *influencing* our folk-psychological theory, or both). Nelkin (2002) argues that the analysis of self-deception starts with an example which is intuitively a case of self-deception and if one is not sure about the intuitions, one consults the account that one is developing. Thus, intuitions and the theory work together to explain the examples of self-deception:

> Second, there may be some variation among people's intuitions as to which cases count as self-deception. For example, although many on both sides of the debate about self-deception agree that the case of the jealous husband is really one of self-deception, some might not be as confident that it is. *Differences in intuitions* are often resolved when more detail is added to the cases in question and I will fill out the jealous husband case in various ways in what follows. But some doubts might remain even when the case is described in more detail, and then it is not unreasonable for one's ultimate judgment to be influenced by the *general account of self-deception one adopts*. In cases of this kind, I believe the proposal I ultimately defend will help allay any lingering doubts about whether they count as self-deception. (Nelkin, 2002, p. 388; my emphasis)

Mele (2010) also uses the notion of folk-intuitions in an attempt to reject Audi's claim about the superiority of his theory. Mele has decided to resolve his dispute with Audi[156] about the folk-notion of self-deception with the help of questionnaires that have been handed down among the students unfamiliar with the topic (Mele, 2010, p. 746). In these questionnaires different examples of self-deception were embedded which participants had to evaluate with respect to the question, whether the given example is an example of self-deception. The result was that, according to Mele's evaluation of students' answers, pre-theoretical notion of self-deception does not preclude cases in which the product of self-deception is a belief (Mele, 2010, p. 747). This is an example of an attempt to settle the question about what self-deception is by using the paradigm of an intuition-based X-Phi (see table 18). Apart from informing us about the intuitions of laypeople, it is questionable in how far Mele's experiment is useful for the scientific research of self-deception.

[155] Since analyzing a phenomenon depends on the idiosyncratic understanding of a phenomenon, differences in the analysis might stem from the difference of the analysandum: "After examining many of the theories of self-deception and resolutions to particular problems that have been offered, it is clear that *we are not all offering accounts of the same phenomenon*. This point holds for many other philosophically interesting concepts as well, such as knowledge, free will, belief, rationality, etc. What we take to be problems and solutions in these areas, as well as convincing theories, depends on our *initial understanding* of the phenomenon." (Funkhouser, 2009, p. 1; my emphasis)

[156] Mele (2010) takes Audi to think that one can do justice to the folk notion of self-deception (p. 747) in holding that self-deception does not need to involve a false belief, but only a false avowal (p. 746).

Experimental philosophy		Empirically informed philosophy
Producing empirical results by doing research		*Using* results of empirical research
Intuition-based X-Phi	Interdisciplinary Phi	
Method: Statistical estimation of phenomenological reports of others Role: Searching the landscape of possible phenomenal worlds in different populations and comparing the intuitions of lay persons to the academic ones	Method: Shaping the epistemic aim and the process of experiments Role: Methodological experiment of an intuition-free and actively interdisciplinary oriented philosophy of mind	

Table 18. Metzinger & Windt: Experimental vs. empirically informed philosophy
Distinctions from Metzinger & Windt (2014).

Interdisciplinary Phi: I understand the term ‘intuition-free X-Phi’ (Metzinger & Windt, 2014, p. 318) to mean philosophy that produces and uses experimental results (Knobe, 2015) in a manner that avoids using intuitions as arguments. Yet, there might be different approaches about the *neutral initial interpretation* of the data that one gets from empirical experiments (Weisberg, 2005, p. 2). One of them is *interdisciplinary constraint satisfaction* endorsed in Metzinger (2003).

Basing a theory of self-deception on intuitions has been argued to be problematic for at least two reasons. It is in question if there exists shared intuitions on self-deception and if these intuitions are then precise enough to give a fruitful definition of self-deception on which empirical studies can build upon. A negative answer to either of the two questions could be a reason why no agreement on what self-deception is has been reached until now in the literature. Neil Van Leeuwen (2008), for example, holds that intuitions might diverge about whether a person “which believes contrary to norms that she does hold but that are not rational in the first place” might be considered a case of self-deception (p. 195, footnote 10). According to Michel & Newen (2010, p. 732) too, folk intuitions on self-deception are heavily under-determined:

> Examples like this illustrate that self-deception is a familiar phenomenon. But the analysis of self-deception has proven to be a tricky endeavor. There is little consensus on the proper analysis of the cognitive state and the cognitive dynamics of self-deception. The two basic reasons for the high diversity in theory are, firstly, that there are various ways in which motivation can influence acceptance and, secondly, that the *pre-analytic folk-intuition* of self-deception is heavily *under-determined*. (Michel & Newen, 2010, p. 731; my emphasis)

Hirstein’s (2005) conclusion with respect to the flexibility of folk mental concepts to contain inconsistencies (p. 217) could also be interpreted in the way that intuitions about self-deception are under-determined and, as a consequence, tolerate inconsistencies.

It is also the case that intuitions do change. Nicholson (2007), for example, argues that folk-psychological and philosophical concepts of self-deception are based on other-deception and, thus, contain both the element of intentionality and the dual belief requirement (believing that *p* and believing that not-*p* at the same time; p. 53). Could we still say that *these* are the folk-psychological and philosophical intuitions? Nicholson (2007) say that “[t]here is something *intuitively* dissatisfying about removing both of these conditions from the formula, and continuing to call its product self-deception” (my emphasis p. 53). At this stage of the debate about self-deception, however, deflationary accounts prevail though. Nicholson (2007) further argues that garden-variety cases of self-deception are those about the self,[157] which may be a consequence of different evidential standards for the acquisition of beliefs about the self (p. 56). Given that garden-variety cases reflect our intuitions, what conclusions should we draw on the basis of a sheer amount of examples pointing to a certain interpretation? Importantly, whether intuitions *do* play a role in self-deception, and whether this *should* be the case, are worlds apart. Neil Levy (2009) denies the argumentative power of intuitions by stating that only *empirical evidence* could constitute a decisive argument in favor of one or the other position (p. 227).

So far I have discussed intuitivity as a property of a self-deceptive theory which means that a given theory is supported by our background beliefs, conscious/unconscious model of reality and metabolic energy (see table 19). I have also surveyed accounts arguing that psychological uncertainty is a property of self-deceptive beliefs (1.2.2), argued for a straightforward connection between anxiety and doubt (psychological uncertainty), as well as for a causal connection between epistemic and psychological uncertainty in self-deception (1.2.5). If self-deceivers' psychological certainty varies, why do they not relinquish self-deceptive beliefs? One possible answer is that a dissociation between psychological and epistemic certainty has taken place, namely that the phenomenal signature of knowing has become transparent, like it is the case for intuitions. That the phenomenal signature of knowing has become transparent means that an agent has a feeling of knowing without knowing why. It is to emphasize that an implication thereof is that each possibility of justifying such an attitude will seem post-hoc and artificial. Self-deceivers might possess *intuitions* of certainty or probability with respect to their self-deceptive attitudes. Such intuitions would explain resilience of self-deceptive attitudes to justification.[158] This would also mean that self-deceivers are guilty of the E-fallacy (= ascribing epistemic status to phenomenal states because of the phenomenal signature of knowing; see table 19) in a similar fashion as the researchers that study self-deception might be.

Intuitions - phenomenal signature of knowing - questionable epistemic status - embedded into PSM (content: one's own epistemic state) - form of embodied knowledge - Bayes-optimality of certain subdoxastic states		**Intuitivity** - property of theoretical hypotheses or arguments , relative to a class of representational systems with a specific functional architecture - phenomenology of an automatically constructed mental model of a theory or an argument
Intuitions of certainty	**Intuitions of probability**	**Causes of intuitivity:** 1a. good fit with respect to the network of our explicit background beliefs 1b. good fit with respect to our conscious and unconscious model of reality 2. metabolic price (energy)
Phenomenal signature of knowing has become transparent	Context, degrees of probability, reliability come into play	

Table 19. Metzinger & Windt: role and phenomenology of intuitions. Distinctions from Metzinger & Windt (2014).

The **phenomenal signature of knowing** is characterized by the phenomenology of direct accessibility of knowledge (which may be preceded by the initial phase of the phenomenology of ambiguity).
Epistemological fallacy or E-fallacy: ascribing epistemic status to phenomenal states because of the phenomenal signature of knowing.

[157] See also Holton (2001) for the view that self-deception is about the self.

[158] It is to be considered that there are also those who deny that phenomenology is a necessary characteristic of intuitions and who characterizes intuitions as *strong*, but *fragile* – strong in being supported by the evidence and fragile in being responsive and not resilient to counter evidence (Weatherson, 2014, pp. 526-527).

Phenomenally possible world: each world that can be simulated on the phenomenal level from a certain class of systems; the possibility of its simulation depends on the functional architecture.
PSM or Phenomenal Self-Model: A phenomenally available model of the construct we name 'self'

Interim conclusion: I have argued that three kinds of approaches to self-deception mentioned by Mele (lexical, example-based, theory-guided) are based on intuitions arising out of folk-psychological theories, as well as that one possibility to explain, why self-deceptive attitudes are upheld in the face of doubt, is that psychological and epistemic uncertainty becomes dissociated so that psychological uncertainty is accompanied by a transparent phenomenology of knowing. The question about which kind of approach is appropriate to the explanation to self-deception can be rephrased as to what is the starting point, the anchor, with which we start? Folk-psychological theory? Examples? Lexical definition? Intuitions guiding or folk-psychological theory? What is it that we want to know when we want to know an explanandum of self-deception? Levy talked about the balance of evidence we should change. What is then our evidence? I will argue in the following section that an explanandum for a theory of self-deception is self-deceptive behavior (inconsistency + justification), as well as self-deceptive phenomenology (tension + insight).

2.1.2 Explanandum: behavior and phenomenology

> All we know is that we sometimes engage in, for us, retrospectively, strange behavior and odd ways of forming beliefs, and so far we have reconstructed that as being self-deceived.
> (Borge, 2003, p. 12)

In this section I will first argue why taking behavior and phenomenology as an explanandum of self-deception is a simplification. I will then provide my characterization of the behavior (inconsistency + justification) and phenomenology (tension + insight) of self-deceivers. In the following section I will provide my own take on how to explain the self-deceptive explanandum, defined in such a way.

In the previous section I argued that theories of self-deception are heavily determined by intuitions. Restricting an explanandum of self-deception to behavior and phenomenology avoids such a dependence, because those can be empirically tested. An alternative would be to extend a self-deceptive explanandum to encompass certain kinds of internal states. I have already argued against internal states as a demarcation criterion for self-deception (see sections 1.1.2.5 and 1.1.3). The main argument was that the difficulty of empirical testing. During the consideration of empirical paradigms of testing self-deception I argued, instead, that what might distinguish self-deceivers from non-self-deceivers is the mechanism of content fixation (1.3.4). Here, I will revisit my previous ideas briefly for two reasons: first, to remind the reader why postulating certain kinds of internal states as part of an explanandum is not parsimonious; second, to connect a mechanisms of content fixation to goal-directed selective processes that will be the topic of section 2.2.1.

Representations, postulated to be inherent in the self-deceiver, are the consequence of the acceptance of a certain (intuitive) description (that the postulation of such representations is dependent on some metaphysical assumptions, see section 1.3.4). A similar claim, namely that both the intentionalist and deflationary accounts postulate *extra-experiential constructs* has been defended by Clegg & Moissinac (2005). The consequence that they draw is that "a relational conception of consciousness produces a more parsimonious and phenomenologically faithful account of those cases termed self-deceptive" (p. 97). Their solution for self-deception, interpreted as the problem that observer and self-deceiver would come to different conclusions given the same evidence is to see experience as idiosyncratic.[159] This explanation as such does not do justice to the phenomenology of the self-deceiver, or to the contradictory behavior of the self-deceiver, but the point is that it *is* true that one should be cautious of postulating extra-experiential constructs and that one should be aware of reasons for doing it:

> Under these accounts, self-deception does not involve multiple clandestine entities within the same individual but, rather, a kind of biased, or motivated, information processing that, over time, leads to biased evaluations of a situation. This approach transforms the central self-deception question from 'how can X simultaneously hold the beliefs p and not-p?' to 'how is it that X can come to believe p when the available evidence implies not-p?' Though this formulation of self-deception eschews the difficulties of a fragmented belief system, it still maintains the essentially fragmentary model of consciousness endemic to the self-deception tradition. *The means whereby self-deception occurs remain the province of extra-experiential constructs, though these are now cognitive or emotional processing systems rather than anthropomorphic entities.* It is nevertheless still the case that conscious judgment is manipulated, or perhaps produced, in a way unknown and unexperienced by the conscious self. Self remains divided against self in precisely the same way as in earlier accounts of self-deception—namely, by means of the unconscious, albeit an unconscious now stripped of its homunculi and repopulated with impersonal processes. (Clegg & Moissinac, 2005, p. 99; my emphasis)

I think that Clegg & Moissinac's (2005) rather radical position that no extra-experiential constructs (cognitive and emotional processing systems) should be used in the explanation of self-deception might be better abandoned in favor of a more moderate positions that self-deceiver's behavior and phenomenology as experiential constructs should be the *starting* point for every explanation. Nevertheless, they may have a point in that if a relationist conception of content is accepted, which would suggest that content supervenes on relational properties of mental states (Boghossian, 2000, p. 486), this warrants the possibility that the self-deceiver's and the observer's mechanisms of *content fixation* differ. At least two kinds of differences might be possible: 1. Difference in content fixation intra-individually, dependent on whether it is one's own or others' beliefs; 2. Difference in content fixation inter-individually. Concerning the first point, Hartwright et al. (2015), for example, argue that activation of brain regions in the case of reasoning about beliefs of others is modulated by the salience of self-perspective that contains competing knowledge (p. 189). Concerning the second point, according to Miłkowski (2015), content is constituted by *evaluative control*: for epistemic evaluation, errors are detected by the system (e.g., via predictive coding) and misrepresentation can occur due to referential opacity (truth or falsity changes if a term is substituted with another one that has the same referent). The errors are evaluated with respect to the goals of the system such that the success of the action provides success conditions for content being true or false. If Miłkowski (2015) is correct, then the most obvious way to explain the different

[159] Difference in attitudes has been described as difference in the points of view: "Further, it implies that the discrepancies evident in the paradigmatic cases of self-deception are not the result of a fragmented consciousness blinding itself to the inevitable facts of 'objective' reality but, rather, the result of different observers taking up different idiosyncratic objects, different phenomenological worlds, and thus making radically different evaluations" (Clegg & Moissinac, 2005, p. 101).

mechanisms of content fixation in the self-deceiver and observer, is to claim that they have different goals and, thus, different satisfaction conditions for content.

The positive claim that I argue for is that the explanandum of self-deception is the behavior and the phenomenology of the self-deceiver. The behavioral description of the self-deceiver on the folk-psychological level[160] is that

1. The self-deceiver behaves in an *inconsisten*t way (either as a case of inconsistent self-ascriptions, or a case of an inconsistent self-ascription and behavior).
2. The self-deceiver *justifies* his self-deceptive belief, or the belief that is inconsistent with the evidence, or as Michel & Newen (2010) put it, "classically conceived, a self-deceiver seems to be someone who actually gets things wrong and who will defend his doxastic *p*-commitment, if challenged." (p. 736)

Those two stand in conflict with each other with respect to the personal level implications, because the first one leads to the assumption that self-deceivers are irrational in virtue of possessing contradictory beliefs (inference from behavior to mental representations) and the second one leads to Davidson's paradox of irrationality: once the self-deceiver has justified his behavior, he is not irrational (from his own perspective anymore). Such contradictory beliefs assumptions have to be weakened though, if an assumption is to be maintained that self-deceivers recognize contradictory evidence on the personal level. Self-deception shares with delusions[161] both inconsistency and personal-level justification (Gerrans, 2013; see also sections 1.2.6 and 1.2.7). Thus, they are not distinguishing features of self-deception. Yet, the presence of justification may still distinguish self-deception from *some* phenomena, as Michel & Newen (2010) argue that justification distinguishes *imagination* from self-deception:[162]

> A probable reason [against self-deception being pretense] is that no one would commonly consider a normal *p*-pretender to be self-deceived, and therefore, the general claim that all self-deceivers are only engaging in a sort of pretense, imagining or fantasizing is counterintuitive rather than natural. Most, to the contrary, share the intuition that what self-deceivers do, is more like believing *p* than pretending *p*. A self-deceptive 'pretender' or 'fantasizer' is someone who seeks ways of avoiding acknowledging and facing what he knows. He engages in a kind of wishful thinking whereas *classically conceived, a self-deceiver seems to be someone who actually gets things wrong and who will defend his doxastic p-commitment, if challenged*. (Michel & Newen, 2010, p. 736; my emphasis)

In the following I will first show that my characterization of an explanandum is consistent with garden-variety cases of self-deception, second, argue that personal-level 'overhead' like irrationality or dual-belief requirement actually follows from self-deceptive behavior

[160] To remind the reader of the problem of a folk-psychological description of self-deception: "The problem is that even though the notion of 'self-deception' is part of our folk psychology, the folk-psychological notions of 'belief,' 'deception,' and 'self' seem, on closer inspection, to rule out the very possibility of self-deception" (Borge, 2003, p. 4).

[161] The similarity between mechanisms employed in delusion and self-deception has been described as follows: "Tacit ambivalence is manifest in behaviour. It can take the form of defensiveness, evasiveness or confabulatory rationalisation provoked by requests for justification or, in some cases, a compartmentalising or partitioning of delusion from disconfirming evidence and background knowledge. In these respects delusions resemble other attitudes such as motivated irrational beliefs and self-deception that are problematic for doxastic theories" (Gerrans, 2013, p. 85).

[162] Belief-ascription depends on the extent of irrationality though: "Equally however if an attitude is entirely resistant to evidence and argument one might start to wonder if belief is really the right word for it. These problem cases are somewhere on a continuum between rationally anchored beliefs sensitive to evidence and inferential pressure, and states completely insulated from processes of rational belief fixation" (Gerrans, 2013, p. 85).

with additional assumptions and, third, propose an enactment criterion for self-deception ascription that follows from self-deceiver's justification. Finally, I will turn my attention to the characterization of self-deceptive phenomenology. In this context, it is the behavior and phenomenology and not the postulated mental states that are in the focus of garden-variety cases 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. (Rorty, 1988, p. 11)

> For example, suppose Bill desires to feel joy at attending his brother's wedding, despite not really liking his brother's bride. This desire can cause Bill to selectively pay attention to particular aspects of the wedding. This selective attention can influence Bill's emotion experience, which can be further influenced by Bill's paying attention only to those aspects of the emotion experience that support his desire to feel joy. After the wedding, Bill believes both that he felt joy at the wedding and that this joy represents his evaluative response to the wedding. Yet Bill is self-deceived because had he not engaged in selective data gathering and selective attention in order to support his desire to feel joy at the wedding, Bill would have had a negatively valenced experience at the wedding that would have reflected his true evaluative response to the event. So it might, in fact, be to Bill's advantage to employ this strategy of being self-deceived, but it remains true that he is nonetheless self-deceived. (Damm, 2011, p. 263)

The difference between the first and the second example is that in the second the description of behavior and phenomenology is intermingled with an interpretation of what is responsible for it – selective processing of information. Nonetheless, I think that self-deceivers' behavior and phenomenology as experiential constructs that hopefully are least susceptible to intuitions and more susceptible to direct testing should be in the focus of the explanation, while the supposed mechanisms influencing the processing of information, as well as representations postulated to be possessed by the self-deceiver should be seen as dependent on the descriptions of behavior and phenomenology. This is also what is found as an implicit assumption in the literature. Namely, it is the case that the behavior of the self-deceiver is often taken as an argument in favor of one or the other explanation of self-deception (see section 1.1.2.5). Against the assumption that self-deceivers believe the truth at some level Nelkin argues that "[t]he *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" (Nelkin, 2012, p. 124; my emphasis). What Nelkin here does is taking the behavior of the self-deceiver as evidence in the argumentation on the nature of the motivation of the self-deceiver. Thus, it is her *interpretation* of the behavior of the self-deceiver that she considers as evidence for the nature of motivation in self-deception. Let me consider another example. Porcher (2012) argues that inconsistency and instability can be derived as characteristics from the behavior of the self-deceiver. While the first characteristic can be observed directly and is described at the folk-psychological level, at least when it is conceptualized as inconsistency between the behavior and the reports about mental states of the self-deceiver, the second is an inference about the mental state and the phenomenology of the self-deceiver, when it is conceptualized as uneasiness, tension (for explanations of notion of tension see Lynch, 2012, Funkhouser, 2005):

> 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. (Porcher, 2012, p. 81; my emphasis)

Porcher (2012) holds self-deceptive belief to be an instance of in-between-belief in Schwitzgebel's sense, from which he draws the conclusion that the aim for self-deceptive accounts is to *refine the description*, instead of clarifying the question of what the self-deceiver believes (p. 80). What Porcher (2012) is correct about, though, is that the description of the self-deceiver's behavior and phenomenology should be as refined as possible, before an attempt is taken to explain it at other levels of description, including also the representationalist (for different levels of description see Metzinger, 2003, p. 110). The requirement that self-deception is irrational stems, at least partly, from self-deceptive behavior (the other part is phenomenological tension). The argumentative chain from self-deceptive behavior to irrationality is as follows: The behavior of the self-deceiver is ambiguous and inconsistent, leading to a dual-belief assumption that the self-deceiver possesses contradictory beliefs:

> A familiar defense of the claim that self-deceived people satisfy the dual-belief conditions proceeds from the premise that they *behave in conflicting ways*. (Mele, 2001, p. 79; my emphasis)

> We typically say of an agent we describe as self-deceived that on the *basis of certain behavior* he seems to believe that *p*, while also seeming to believe that not-*p*, since, if asked, he would sincerely assent to not-*p.* (Borge, 2003, p. 2; my emphasis)

A change from the description of behavior to assumptions about the mental representations, that one possesses, is to be noted here. The dual belief-requirement is irrational, on the assumption that beliefs are internally governed by rules of rationality in the self-deceiver. It is to be emphasized that the personal level description has led to the acceptance of the rules by which such personal level ascription is governed, namely by those of rationality (Bermúdez, 2000a). From the folk-psychological assumption that there are dual beliefs, it does not follow that the self-deceiver possesses contradictory representations. The folk-psychological and the representationalist levels of description are to be distinguished. Often, the defense of an intentionalist position (see section 1.1.1), is based on the behavior of self-deceiver, e.g. that it is *too skillful*, for the truth not to be known too:

> Interestingly, one of the reasons for concluding that the patient *unconsciously* knows of the incurable malignancy is the very success of the defense. How could that defense be maintained so skillfully *without* using knowledge of the unwelcome fact to anticipate the forms in which it might try to intrude itself on consciousness? (Greenwald, 1988, p. 113)

Apart from behavior, other arguments in favor of the dual belief requirement have been those of rationality, motivation and phenomenology[163] (Michel & Newen, 2010). I argue that the support of the belief requirement by the rationality of the self-deceiver (e.g. since self-deceivers are usually rational, when they are not, they possess contradictory representations) occurs as an inference from behavior to the rules by which representations in the system of the self-deceiver are governed, namely those of rationality. The assumption, that the unfavorable belief has to be the motivation for self-deception, has been defended by early intentionalist positions[164] (e.g. Davidson, Pears), and it has been relinquished in favor of a desire-driven approach.[165] To sum up, the dual-belief requirement is justified primary by self-deceiver's behavior, coupled with the assumption that personal level processes are governed by the rules of rationality and that it is *beliefs* that paradigmatically guide behavior (see Van Leeuwen, 2007a in 1.2.1).

Third, that fact that self-deceivers justify their beliefs, brings in a *social* component into self-deception. I want to discuss this question shortly from the perspective that comes short in self-deception literature, namely how others understand self-deceivers (and not how self-deceivers appear or want to appear to others). Understanding self-deception from *inside* has been popular in the literature (focus on the perspective of the self-deceiver, how she acquires and maintains self-deception). Not least, I have shown that it is difficult to disentangle self-deception from other related motivated phenomena in this way (see section 1.2.7). I think that it may be more fruitful to offer an *external* criterion that distinguishes self-deception from other phenomena, namely a criterion that is based on how others understand self-deceivers. I think that it is plausible in virtue of the fact that self-deceivers cannot judge whether they are self-deceived or not at the time of being self-deceived anyway. Newen (2015) has recently proposed a multiplicity view of understanding others. According to him, we form *person models* of others that combine at least four strategies via which we accomplish the understanding: intuitive (direct perception + interaction) and inference-based (high-level simulation and theory-based understanding; p. 8). I argue that one of the ascription criteria of distinguishing self-deception from other phenomena is the *enactment criterion*, introduced by Newen (2015, p. 3): In certain cases, e.g. delusion of persecution, we are not able to use enactment imagination for understanding, due to the (radical?) difference in phenomenology. I agree with Newen (2015, p. 3) that understanding others requires the "necessity of 'quarantining' my idiosyncratic background beliefs," but

[163] Arguments for the dual belief requirement have been summarized by Michel & Newen (2010) as follows: "So, it appears that there could be at least four good reasons to suppose that a person who is self-deceived in accepting that p will also believe that not-p:

- *Rationality*: Since **S** is sufficiently rational and evidence-sensitive, **S** simply recognizes that not-*p*.
- *Behavior*: The assumption that not-*p* is believed by **S** explains typical self-deceptive behavior.
- *Motivation and activation*: Only a belief that not-*p* explains why a process of self-deception is initiated.
- *Phenomenology*: **S**'s experience of a characteristic psychological tension is caused by his believing that actually not-*p*." (p. 734)

[164] Lazar (1999) states as the difference between self-deception and bias: "Cognitive biases are persistent and highly prevalent patterns of biased reasoning. They are exhibited regardless of subject matter. In contrast, self-deception is *thematic:* the content of the irrational belief is relevant to the explanation of its formation" (p. 267).

[165] Scott-Kakures (2002), for example, argues that "reflective reasoning" can be influenced by motivation (p. 592): "Reflective reasoning paired with a certain error of self-knowledge makes for self-deception. We are at permanent risk of self-deception because, in testing hypotheses as we typically do, and in our capacity as reflective cognizers, we are *unwittingly moved by desire*" (p. 599; my emphasis).

that it is difficult to decide, which ones are those in need of such quarantining. I argue that in case of self-deception, others hold enactment of the self-deceiver to be possible and the need to quarantine idiosyncratic beliefs minimal, if at all. This also follows from Mele's (2001) impartial peers test of self-deception: if it is not for the motivation, self-deceivers would come to believe the same as their impartial peers. So, as for peers, they ascribe self-deception exactly in cases in which they believe in the possibility to simulate the decision process of the self-deceiver, but fail to come to the same conclusion as the self-deceiver and, more importantly, fail to persuade the self-deceiver in the falsehood of this acquired belief-like representation. It is further to be emphasized that one of the arguments for Newen's (2015) multiplicity view can be used as an argument as to why folk-psychological or narrative theories of self-deception are insufficient. Newen (2015) argues that (propositional) knowledge, organized as a folk-psychological theory or as a narrative, does not cover basic intuitive understanding of others. The transfer of this idea to self-deception might look as follows: understanding self-deception as belief-formation or narrative construction is insufficient, because self-deceivers, just like others, possess a broader repertoire of means of self-deceiving, than only those two. Maybe an 'interactive turn' is needed not only in social cognition (about the latter see Quadt, 2015).

Interim conclusion: Self-deceivers are either synchronically or diachronically inconsistent, yet justify their inconsistent behavior. That they are also irrational or possess contradictory beliefs can be inferred from self-deceptive behavior if the disunity in unity constraint is solved by assuming a *personal* level unity governed by logic and rationality and the kind of disunity present in self-deceivers to be formulated in folk-psychological terms. The fact that we, as observers, are so puzzled by self-deceiver's behavior is because we try to enact them and fail. The enactment criterion is an *external* criterion for self-deception ascription that can be seen as an extension of Mele's impartial peers test.

In the remaining part of the section I will offer two phenomenological characteristics of self-deception – tension and insight. Self-deception can't possess a phenomenology as self-deception, because at the time of self-deception the self-deceived does not realize in what a condition one is (Borge, 2003, p. 12; see figure 8). Thus, the phenomenological description of the self-deceiver[166] is that

1. The self-deceiver experiences *tension* or a feeling of discomfort to a different degree (from uneasiness to anxiety).
2. The self-deceiver experiences feeling of *insight* at the time when self-deception is relinquished.

[166] Interestingly, Gendler (2007) has proposed that self-deceiver's phenomenology is also characterized by a feeling of being in a realm of make-believe, she is to my knowledge the only one though to suppose such kinds of feelings in self-deception: "Moreover, what is distinctly characteristic of paradigmatically perplexing cases of self-deception – in contrast to cases of mere motivated false belief – is the *phenomenology* associated with such overriding. For the feeling associated with ceasing to guide one's actions on the basis of not-P and of starting to guide them on the basis of P is not the feeling of replacing one reality-reflective attitude by another with contrary content; rather, it is the feeling of leaving a realm of make-believe where one has allowed one's thoughts and actions to be governed by some sort of fantasy, and returning the realm of reality-receptiveness where one's thoughts and actions are responsive to things as they actually are" (p. 245; my emphasis).

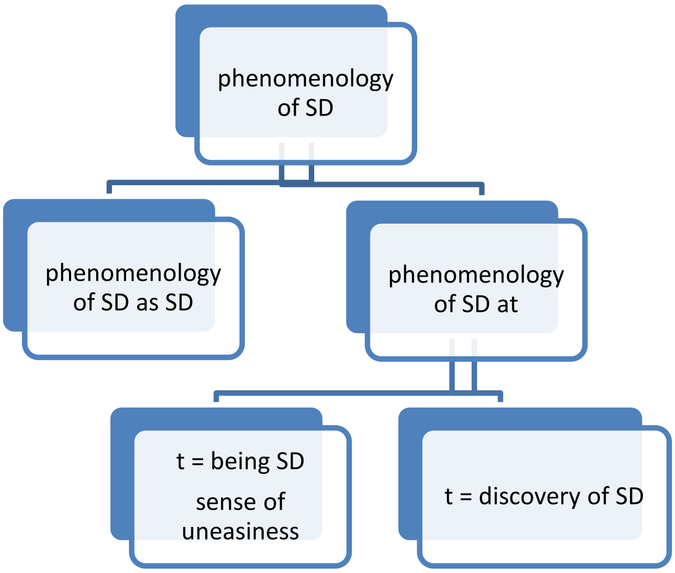


Figure 8. Phenomenology of SD.
Distinctions from Borge (2003) and Gendler (2007).

That tension is a characteristic of self-deception is a phenomenal counterpart to the dual-belief requirement or any other inconsistency between two personal level mental states (in virtue of the unity of the personal level by rules of logic and rationality), hence also a distinction between cognitive and behavioral tension:

> Funkhouser divides it up into ''cognitive'' and ''behavioral'' tension (2005, p. 296). ''Behavioral tension'' may refer to the alleged fact that the self-deceiver displays some behavior that seems more consistent with believing that p and other behavior more consistent with believing that not-p. Besides this, self-deceivers supposedly also experience mental/cognitive ''conflict,'' ''tension,'' or ''discomfort'' (Funkhouser, 2005, p. 299; Graham, 1986, p. 226; Losonsky, 1997, p. 122). Graham elaborates this as the experience of being afflicted with ''doubts, qualms, suspicions, misgivings, and the like'' (1986, p. 226) concerning the belief we are self-deceived in holding, or in Losonsky's words, ''recurring and nagging doubt'' (1997, p. 122; see also Funkhouser, 2005, p. 299). Noordhof (2009) similarly speaks of an essential ''instability'' present in the self-deceived state. I would assume that these philosophers think that such mental tension is the experiential accompaniment for those cases in which behavioral tension is present or liable to occur. (Lynch, 2012, p. 435)

As existence of contradictory beliefs in self-deception, presence of *phenomenological* tension can also be questioned. Both provide a certain kinds of answer to the disunity in unity constraint, namely that the personal level governed by logic and rationality is the unity that is disunified either in virtue of contradictory cognitive attitudes, or a certain kind of affective attitudes that, again in virtue of logic and rationality, should have evoked doubts about the cognitive attitudes we possess. It is not inconsistency per se, that humans supposedly avoid, but aversion that stems from the violation of norms of rationality (see Van Leeuwen, 2008).

> [T]hey do not account for the strong emotions generated *when self-deceptive beliefs are challenged*. What prevents the self-deceived from enjoying their false beliefs with smug complacency? There is no explanation for the brittle quality of self-deception (Audi 1985; Bach 1997), and the defensiveness associated with a self-deceptive personality.
> (Mijovic-Prelec & Prelec, 2009, pp. 228–229; my emphasis)

> When people are motivated to cling to a belief, they *do not feel comfortable* with blithely ignoring adverse evidence or simply shutting their ears to anyone who opposes their view. Instead, they engage in more subtle forms of ad hoc reasoning, rationalization, and special pleading to arrive at their desired conclusions and to justify their beliefs to others […]. (Boudry & Braeckman, 2012, p. 346; my emphasis)

The *intuition* that "reality always threatens to break in" (Noordhof, 2009, p. 71) in self-deception is actually a premise about connections between personal level attitudes. Granted that self-deceivers do, in at least some cases of self-deception, experience phenomenological tension, why don't they recognize their self-deception in that instant? Borge (2003) argues that emotions might be sources of self-knowledge (p. 15), which might be extended to affective states in general and feelings in particular. There are at least three possibilities why self-deceptive affective states do not serve as sources of self-knowledge concerning the presence of self-deception: first, phenomenological tension is *vague* (in analogy to dot-tracking study where the argument was that self-deception requires vague feedback/evidence, see section 1.3.2.3). If this were the case, self-deceptive affective states might be *misattributed* (see also the cognitive dissonance paradigm on this, section 3.1.1). Second, tension elicits doubts with respect to the contents of self-deception, but psychological uncertainty is dissociated from the epistemic one so that self-deceptive attitude acquires the phenomenology of knowledge like intuitions do (2.1.1). Third, tension might elicit further hypothesis testing cycles so that it disappears in the next round. Here, tension-full self-deception might become tension-free (and more robust by that).

Tension will also be the topic of the following section, so I will leave it here and conclude with a phenomenological characteristic of self-deception that is, more often than not, overlooked, namely insight. Demos (1960) has argued that self-deception is characterized by the feeling "I knew it all along." Müller & Stahlberg (2007) argue that the hindsight-bias depends on surprise in a certain way. Surprise is conceptualized as the subjective experience of the difference between the expectation and outcome (Müllber & Stahlberg, 2007, p. 170). Their basis is Pezzo's sense-making model which states that *initial surprise* (= elicited after an unpredicted outcome) triggers the process of sense-making which, if it is successful, results in a *hindsight bias* and if it is not successful – in *resultant surprise* that evokes the reverse *hindsight bias* (Müller & Stahlberg, 2007). In reliance on dual theories, Müller & Stahlberg (2007) extend Pezzo's model by arguing that surprise can serve as a direct metacognitive heuristic cue (under insufficient motivation or time constraints) or can elicit biased deliberative and conscious sense-making (p. 172).

Since self-deceivers stick to their self-deception *before*, but have the feeling of having known the truth *after* relinquishing self-deception, then this means that something must have changed in between – they have gained *insight*. I think that, instead of hindsight bias, this peculiar experience of retrospective recognition in self-deceivers which has been described as a feeling to have known that one was self-deceived *during* self-deception already, even if one did not want to recognize it, fits the description of *insight*. Recently, Cosmelli & Preiss (2014) described insight as a dynamic past-future interplay which means that it has a gap-filling effect in virtue of certain argumentative context having been understood in a different way.[167] Such gap-filling takes place upon recognition of self-deception too, if self-deceivers are ever to experience hindsight bias.

[167] Cosmelli & Preiss (2014) actually use the distinction opaque/transparent to denote the context change. In philosophical tradition though, the distinction transparent/opaque has been applied to belief-contexts Jackendoff (1975). *Referential* transparency/opacity is here in question: if a term can be substituted by a co-referential term in a linguistic context, then this is a case of referential transparency, if not – referential opacity (for a relation between the distinction

If self-deception is characterized by *lack* of insight (before recognition), then the next question concerns the *kind* of insight that is lacking. David et al. (2012) differentiate between at least two forms of insight, *clinical* (awareness of the illness in question, e.g. anosognosia) and *cognitive* (metacognitive flexibility in the attribution of mental states – beliefs, judgments, experiences; p. 1383). The latter is said to encompass *self-certainty* as overconfidence about the correctness of one's mental states and *self-reflectiveness* as acceptance of correction from others about one's own mental states. What appears to be certain is that self-deception is characterized by lack of clinical insight (one does not acknowledge that one is self-deceiving). It might be too strong, though, to claim that self-deceiver's *personality* is characterized by general self-certainty and self-reflectiveness, apart from the mental states that depend on the issue in question, namely, the one that the person is self-deceiving about. Yet, self-deceivers are self-certain and not self-reflective with respect to the matter they are self-deceiving about. The mechanism by which self-deceivers gain insight seems to be the opposite of the one by which tension-loaded self-deception becomes tension-free.

Interim conclusion: Self-deceivers are either synchronically or diachronically inconsistent, yet, justify their inconsistent behavior. That fact that they are also irrational or possess contradictory beliefs or experience tension, can be traced from self-deceptive behavior if the disunity in unity constraint is solved by assuming a *personal* level unity governed by logic and rationality and the kind of disunity present in self-deceivers to be formulated in folk-psychological terms. The fact that we as observers are so puzzled by self-deceiver's behavior is because we try to enact them and fail. The enactment criterion is an *external* criterion for self-deception ascription that follows from the justification characteristic of self-deceivers. As inconsistency-tension builds one dependency pair, justification-insight builds another: inconsistency may trigger tension, while justification should lead to insight into self-deception, at least this is the aim of the observer. Yet, in neither case, self-deception is abolished and the crucial (empirical) question is why. I will revisit these ideas in chapter 4.

2.1.3 Tension as an epistemic feeling with an indicator function

> What we call rational thinking simply can't take place without a huge complex background of intuitive thinking that's registering in consciousness only as 'Uh-huh, that follows' or 'Nope, it doesn't follow.'
> (Jackendoff, 2012, p. 214)

To sum up, the following results have been crystalized, that will be relevant for this section: self-deceptive phenomenological tension may either cause a self-deceptive process as such, be triggered by a self-deceptive attitude in virtue of its inconsistency with some other attitude, be part of the self-deceptive process itself, or indicate the rightfulness of the self-deceptive process (see section 1.2.5). As for the process generating self-deception, I distinguish between personal (agentive) and subpersonal selection (section 1.1.2.4) on the one hand, and belief-formation and narrative construction, on the other (section 1.2.7). In combination, subpersonal selection as narrative construction might be seen as a kind of selection responsible for our world- and self-model, instead of an epistemic agent model construction that is embedded into the former one. How different kinds of selection depend

between referential opacity/transparency and Frege's distinction between extension/intension see also Jackendoff, 1975).

on goal representations will be the topic of following section 2.2.1. Here, I want to offer an explanation for tension as a combination of personal/subpersonal selection with the indication that certain criteria of goodness of the self-deceptive process have been fulfilled. For this, I will first review Joëlle Proust's (2013) theory of metacognitive feelings. I will, then, argue that tension is a certain kind of a metacognitive feeling and compare it to other related kinds of metacognitive feelings, as well as, last, extend Dokic's (2012) idea that metacognitive feelings provide *modal* knowledge to the argument, defining which role *possible worlds* play in self-deception. This last point will be important for at least one type of selection, namely counterfactual selection that I will introduce in section 2.2.1.

Here comes a concise list of results. Tension is a kind of metacognitive feeling because, first, the indicator-function, ascribed to metacognitive feelings by Proust, fulfills the functional role ascribed to tension – to indicate on the personal level to the agent that something has gone wrong, e.g. inconsistent attitudes are present or the acquisition process of the attitude has not fulfilled the evidence handling requirement. And second, the phenomenological description of tension as a feeling of *uneasiness*, *distress* and *anxiety* fits the *phenomenological* profile of an indicative metacognitive feeling about a faultiness of some mechanisms. Dokic's *modal* description of metacognitive feelings is that those feelings indicate the extent to which a certain process would be successful in nearby possible worlds. Such 'possible worlds' might be understood either as personal or subpersonal level counterfactuals, as 'selection' or 'inference'. On the personal level of description, one might say that tension arises when almost all constructed mental models (see Khemlani & Johnson-Laird, 2013; Khemlani & Johnson-Laird, 2012), that vary in the kinds of evidence, taken as premises for the belief-forming process, lead to the result that a certain accepted attitude is to be rejected. On the subpersonal level, one could ask how counterfactuals constructed during self-deception change the experience (tension) itself (see Seth, 2014). If there is some sort of cycle by which tension-loaded self-deception might become tension-free and vice versa, as suggested in section 2.1.2, then tension might lead to its disappearance if a certain interplay between counterfactuals is given. An open question, that begs to be answered, is whether in these cases tension also preserves its indicative function such that it is agentive control that triggers the subpersonal counterfactual changes. Alternatively, tension could also be a mere epiphenomenon, as Dokic (2012) argues, that when no *deliberate* reasoning takes place, noetic feelings might be present, but they would in this case remain *epiphenomenal* (pp. 312-313). The answer to the latter question is that, if agentive control were the criterion for determining whether self-deceptive tension were an epiphenomenon or not, then in at least some cases this tension would be an epiphenomenon (see section 2.2.1 for self-deception being motivated by subpersonal goal-representations, section 2.2.2 for the description of self-deception in terms of dolphin model of cognition and section 4.5 for an argument in favor of a post-hoc sense of control in self-deception).

Now, that my goals for this section are set out, I will proceed with Proust's (2013) theory of metacognitive feelings. Since I have been honored to have written a commentary (Pliushch, 2015) on her article about the nature of feelings in general (Proust, 2015b), I will reply to her criticism on my idea about tension as a metacognitive feeling at the end of this section. First, I would like to write a few words about a distinction between feelings and emotions and their relations to cognition that was the topic of Proust (2015b). De Sousa (2014) argues that in feelings the bodily component is central and that those are similar to sensations, while emotions possess an intentional object (moods being emotions without an intentional object). This opens up the question which role cognitive and conative attitudes do play in emotions. Appraisal theories of emotion, for example, would claim that it is the associated cognitions that give an emotion its distinct color (de Sousa, 2014). The appraisal component can be understood through the prism of Schachter & Singer's misattribution of arousal paradigm: given different context, arousal can acquire a positive or a negative valence (=interpretation of bodily changes). Which role beliefs and desires do play in emotion, is also subject to debate. De Sousa (2014) argues on the basis of the Othello example that emotions *weigh* arguments in our experience. It is to be noted that the Othello example has been stated as the one of self-deception, that directing one's attention is part of every explanation of the mechanism of self-deception and that such an interpretation presupposes that they do change our experience of reality. Proust (2015b), like de Sousa, sets out the presence of an intentional object as one of the distinctions between feelings and emotions, another important one being the expressive non-conceptual mode of feelings that makes them cognitively impenetrable. In section 1.2.5 I have stated that in the self-deceptive literature no distinction between feelings and emotions has been set and used the term 'affective states' to generalize upon both and to distinguish different roles that affective states have played in self-deception (for the difficulties to distinguish between feelings and emotions see also Pliushch, 2015).

After this brief reminder, let me get back to the role of noetic feelings (feelings of knowing) in procedural (non-conceptual) metacognition.[168] Proust (2013) defines *metacognition* as "the set of capacities through which an operating cognitive subsystem is evaluated or represented by another subsystem in a context-sensitive way" (p. 4). An example for such feelings might be "Yes, I feel I know the answer!" as the result of a query to remember something (Proust, 2013, p. 56). Proust (2013) argues that metacognitive feelings are involved in reasoning. Reasoning is defined by her as a *mental act* or "the process of intentionally activating a mental disposition in order to acquire a desired mental property" (Proust, 2013, p. 149). A refined definition would be "[b]eing motivated to have goal G realized → (=causes) trying to bring about H in order to see G realized by taking advantage of one's cognitive dispositions and norm-sensitivity for H reliably producing G" (p. 162). Mental acts rely on two types of norms according to her: *instrumental* (means-to-end) and *constitutive* (enforcing the correct outcome) (p. 152). Thus, one can fail in a mental act either because one has chosen a wrong strategy, or because one has failed to fulfill the normative requirements for that strategy (p. 155). According to Proust (2013), a mental act is preceded and followed by *self-evaluation* (p. 162) which is an initial and last step of an action, but not a kind of mental action itself (p. 165). The questions that describe this kind of self-evaluation might be: can my cognitive dispositions be reliably activated? Does the outcome match the expected goal? (p. 165) Self-evaluation itself is not a kind of mental action, because it is dependent on the mental action that it affects, or as Proust (2013) formulates it, "the scope of a single mental action supervenes on a functional, normative, and motivational continuity between the metacognitive phases and the mental action core" (p. 191). Here comes an example of an instance of remembering that includes all the components just mentioned:

> The error-signal often consists in a temporal lag affecting the onset of a sequence of action. For example, in a conversation, a name fails to be quickly available. The error-signal makes this manifest to the agent. How, from that error-signal, is a mental act selected? In some favourable cases, an instrumental routine will save the trouble of

[168] Proust (2013) argues that "[f]eelings, that is, states of a comparator, are indexing neither events, nor objects, but possibilities of cognitive success or failure. They do not properly 'refer', because they do not engage propositional thinking" (p. 77). Thus, they do not refer neither de re, nor de dicto.

> resorting to a specific mental act: 'Just read the name tag of the person you are speaking to'. When, however, no such routine is available, the speaker must either cause herself to retrieve the missing name, or else modify the sentence she plans to utter. In order to decide whether to search her memory, she needs to consider both the uncertainty of her retrieving the name she needs to utter, and the cost-benefit ratio, or utility, of the final decision. *Dedicated noetic, or epistemic, feelings help the agent evaluate her uncertainty*. These feelings are functionally distinct from the error-signals that trigger mental acts. Nonetheless, *the emotional experience of the agent may develop seamlessly from error signal to noetic feeling*. (Proust, 2013, p. 166; my emphasis)

To sum up, Proust focuses on the role of epistemic feelings for *mental* actions that are characterized by the presence of *intentions*, as well as by *goals* and *rules* for conducting them. Is self-deception a mental action (1) and by which goals (2) and rules (3) it is guided? These questions have to be answered for the transfer of the indicator function of metacognitive feelings to tension. Regarding the first, if self-deception involved reasoning (= *conscious* belief-formation) or conscious narrative formation as epistemic model construction, then it could be called a mental action. If self-deception changes the transparent world- and/or self-model, then the process by which such a change is being made would not be available in virtue of the definition of a *transparency* as unavailability of earlier processing stages. I am inclined to think that in at least some cases of self-deception the latter is the case. The reason for this is the same as the one for arguing that tension-loaded self-deception might become tension-less: While for *researchers* of self-deception the challenge for explaining the phenomenon lies in adding as much as possible personal level recognition, for *self-deceivers* the less personal level availability, the better and robuster the self-deception. If metacognitive feelings arise on a process that is also consciously available, e.g. reasoning, then the origin of such a metacognitive feeling is easier to track to that process in order to *control* it, e.g make changes and relaunch it (see section 2.2.2.2 on dual processes). In this respect, one has to be cautious not to confuse two different kinds of references: what do feelings refer to and what we take them to refer to may differ[169] (Nagel, 2014). In addition to an indication function of metacognitive feelings, Proust (2013) argues that metacognitive feelings are responsible for the *sense of agency of one's thoughts*, because metacognitive feelings provide a non-conceptual equivalent to conceptual entitlement[170] (p. 226). Entitlement is, according to her, one of the two alternatives of justifying one's beliefs, on the par with explicit justification (p. 209):

- *Explicit justification* is based on one's knowing the reason for believing a proposition;
- *Implicit entitlement* is based on one's trust into an experience one has that the given proposition is compelling.

Such implicit entitlement might describe a phenomenology similar to that when a phenomenal signature of knowing has become transparent (see section 2.1.1). Self-deceivers justify their self-deception, but this does not mean that the acquisition process was consciously available. On the one hand, Proust (2013) states that it is "an a priori

[169] The kind of feeling one is having might be underdetermined with respect to the indicator function that it possesses: "However, even if the feeling itself does not have a dedicated internal meaning, it could still be true that fluency only ever serves as a guide for thinking when we take it to be about something, and more specifically, about a mental state of ours, a state seen (perhaps wrongly) as a state of knowledge, perception, recall or imagining." (Nagel, 2014, p. 715)

[170] Metacognition and the sense of agency have been argued to be related so that having the former is an indication of the latter: "On the view presented here, a subject is entitled to a sense of agency for a particular mental action if the metacognitive feedback relevant to that action both causes and justifies her performing that action (or, in the retrospective case, if it allows her to evaluate the success of her action" (Proust, 2013, p. 226).

necessity that a *mental agent* permanently monitors and changes her own knowledge state, her emotions, or her present conduct" (Proust, 2013, p. 233; my emphasis). On the other hand, Proust (2013) not only argues that metacognition operates at the *subpersonal level* (p. 298) and that "unconscious heuristics generate conscious noetic feelings that will allow a thinker to guide her epistemic decisions" (p. 72), she goes further to claim that "[a]ll our flexible thoughts are also generated subpersonally" (p. 299). The decision about which kind of selection and control self-deceivers employ, would need to depend on their phenomenology and memory: When self-deceivers justify their self-deceptive attitudes, do they remember acquiring them? If they post-hoc rationalized, why do they not doubt and relinquish these attitudes? As I noted in chapter 1, since there may be different kinds of self-deception, there may be different kinds of selection responsible for it too. Thus, I cannot a priori rule out the possibility that self-deception is a mental act, though in virtue of the intentionality feature the latter would bring back paradoxes discussed previously (see section 1.1.3 for summary). That self-deception involves unconscious/subpersonal belief-formation such that only the resulting attitudes pops out in the head of the self-deceiver is a more parsimonious option. That the feelings are consciously, present though the process, whose rightfulness they should indicate, is not, is the extension of Proust's theory to self-deception that I propose. Whatever phenomenological accompaniment the process responsible for the existence of metacognitive feelings possesses, the subpersonal level might be described similarly in both cases, e.g. as a set of comparators, as argued by Proust (2013, p. 253-257). "[P]henomenology tracks epistemic adequacy" during such monitoring (p. 256). In Proust's (2013) words, metacognitive feelings provide "immediate *subjective, phenomenological* access to the comparator's verdict" (p. 223). Comparators are, according to her, structures that compare observed properties of a mental action with those necessitated by the chosen norm for this mental action[171] (p. 256). The basic schema of comparisons taking place and errors, being transmitted to higher levels, is also present in predictive coding – an uprising (subpersonal) theory for explaining perception, cognition and action (see chapter 4).

To recapitulate, I started with the question of what kind of a process, and with which goals and rules, might lead to tension in self-deception? I left it open whether it is a kinds of mental action, e.g. reasoning (conscious belief-formation) or narrative construction, or some other possibly subpersonal process. The purpose of the elaboration was to make the reader aware of different possibilities of such processes, not to narrow them down. The enlistment of rules, that might be broken in mental actions, shows that those are also potentially the rules that might have been broken in self-deception. The three kinds of errors that Proust (2013) argues to be found in acceptances (p. 180), where *acceptances* are decisions to regard some proposition as true (p. 172):

- *instrumental* – selection of an epistemic norm that is inappropriate to the given context;
- *epistemic* – misapplying the norm or misjudging confidence in the performance; epistemic norms might be those of truth, exhaustivity, consensus, coherence, intelligibility, relevance

[171] Interestingly, Proust (2013) argues that her comparator-view is compatible with Hierarchical Bayesian Network explanations (p. 261, 263): "This proposal is similar in many respects to Koechlin's Cascade model and to Fletcher and Frith's recent proposal in favour of a Hierarchical Bayesian Network. The main difference is that our proposal offers an alternative interpretation of what other authors take the higher level of the action system to consist in: an organizational or attentional-supervisory system. On our view, it should be construed as emerging from distributed metacognitive abilities, thanks to which mental actions can be selected, monitored, and controlled. *Our model thus enriches the understanding of executive impairments by specifying the metacognitive nature of the mechanisms upon which a rational execution of action depends.*" (Proust, 2013, p. 263; my emphasis)

(pp. 174-176); their multiplicity "is a consequence of bounded rationality" (p. 184); See e.g. Van Leeuwen (2008) and Sturm (2009) for the role of bounded rationality in self-deception.

- *strategic* – "incorrectly setting the decision criterion given the stakes"; strategic acceptance, like the epistemic one, is associated with "dedicated emotions (with their associated somatic markers)"(p. 182); notice the similarity of the latter to Mele's (2012), Lynch's (2012) and Talbott's (1995) emphasis of the role of a decision threshold in self-deception.

Interestingly, Proust (2013) also argues that metacognitive feelings provide a defense against the kind of self-deception that consists in subjects wrongly supposing that they have fulfilled the normative requirement on the given action, e.g. "believe wrongly that she succeeded in comparing two plans of action, when she in fact only considered one" (p. 219). As I have stated in section 2.1.2, though the functional role ascribed to tension is exactly that, by misattribution, as well as other means, tension does not fulfill this role in self-deception, even if metacognitive feelings fulfill their function *reliably*. Proust's (2013) thought experiment to demonstrate that the reliability of metacognition stems from it being formed "on the basis of an extended sequence of *dynamic couplings* effected in prior interactions between mental actions and monitored outcomes" (pp. 203-204), resembles those given as theoretical examples of self-deception:

> Let us illustrate the incompleteness of an account that ignores the distal source on which epistemic feelings are grounded through the following 'brain in the lab' experiment. Suppose that a mad scientist provides Hillary with regular spurious feedback on how she performs in a type of mental task. Whenever she performs a given *type* of mental action (such as retrieving a name, performing an arithmetic calculation, controlling her perception for accuracy, checking the soundness of her reasoning), she will receive *consistently biased feedback*; she will be led to believe that her mental actions of that type are always correct. For the sake of the argument, let us assume that Hillary has *no way of figuring out that the feedback that she receives is systematically biased*. There are several ways of exposing Hillary to biased feedback. The mad scientist can *explicitly misinform* her, by systematically telling her – after a block of trials – that she is performing well above average, even when it is not the case. Or, still more cunningly, the scientist can *use implicit forms of spurious feedback*, extracted by Hillary in ways that she cannot consciously identify. For example, self-probing can be *manipulated by the perceptual aspect* of the tasks: using familiar items for new tasks misleads her into believing that these tasks are easier that they are. Another trick is to *manipulate the order* in which the tasks of a given difficulty are presented. When the tasks are ordered from more to less difficult, Hillary might have a misleading feeling of growing self-confidence. *Priming* can also be used to prompt Hillary to the correct solution, which she will believe to have found herself in. The mad scientist can combine these various ways of manipulating self-confidence. To prevent Hillary from having the sense that the is being manipulated, there are several strategies that the mad scientist can use, he can, for example, organize the temporal pattern of the responses in a way that prevents Hillary from performing a careful post-evaluation. Alternatively, he can erase her own post-evaluations from her memory by applying, for example, well-targeted transcranial magnetic stimulations each time she performs one. (Proust, 2013, p. 199; my emphasis)

This description of the interaction between the evil scientist and the reasoner resembles a folk-psychological description of the interaction between the deceiver and the deceived in self-deception. In the case of self-deception though, the self-deceiver does not know how to fix the problem indicated by epistemic feelings,[172] because of the process not being

[172] Some further characteristics of noetic feelings:

- Norms, to whom noetic feelings are sensitive, are not those of *utility*, as studies on monkeys show (p. 65). *Utility* is to be distinguished from *epistemic correctness* (p. 125). *Utility* determines the context or the choice of the relevant epistemic norm, e.g. exhaustivity or available to revision. The above enlistment of different kinds of ways in which a process oriented at discovering the truth on a certain matter might be skewed shows, on the one hand, how everything that fits the skewing role might be connected to self-deception and, on the other hand, poses one question that I will pursue in the next section: What is the explanatory value for self-deception of enlisting and categorizing different skewers for self-deception, if they will be neither unique nor exhaustive (demarcation problem)?

Let me entertain the third point after the discussion of the kind of process and kind of skewers of such process, namely the goal representations involved. My assumption is that the rules that are violated depend on the *goal* that is pursued. The goal of a belief-forming process is to acquire a truthful attitude. The goal of a self-deceptive process might vary in this respect, e.g. either to acquire a *certain* attitude (on different kinds of intentions from which one could make a transfer to self-deceptive goals, see Bermúdez view in section 1.1.1.5) or to bias the belief-forming process such that a certain attitude is acquired, or a certain goal reached. Multiple goal representations can, thus, play different roles in self-deception. One goal representation might trigger self-deception, another – be the desired outcome of a self-deceptive process. As for the former, Baumeister & Newman (1994) have, similarly to Proust, proposed that affect can "serve as a cue for the correctness and acceptability of an inference or decision" (p. 15). What they have in mind is that negative affect indicates that the *goal to achieve a favorable conclusion* has been violated and that as a consequence negative affect leads to a *personal level* biasing, e.g. reassessing of the evidence, and, as such, that negative affect *causes* self-deception[173] (p. 15). It may be the case that, apart from the failure to acquire a truth-conducive representations, the failure to achieve other goals might enhance the intensity of the generated metacognitive feeling. This distinction between the goal to find out the truth and the other one might find its reflectance in the kinds of metacognitive feeling that one has. If belief-formation is taken, three kinds of metacognitive feelings might arise from it. Those are *intuitivity*, *counter-intuitivity*, and *anxiety*, if one classifies them according to the phenomenology and not according to the norm that they control. Intuitivity indicates the appropriateness of a given belief-forming process whose aim is to find out the truth per definition. The reasons for the ascription of the given functional role to intuitivity is that intuivitity signals 1) a good fit with respect to the network of our explicit background beliefs and 2) a good fit with respect to our conscious and unconscious model of reality (Metzinger & Windt, 2014). An appropriate belief-forming process provides a good fit with respect to 1) and highly likely also with respect to 2). I further argue that counter-intuitivity represents that a certain cognitive process violates the chosen criterion of appropriateness (truth-finding), but is neutral with respect to the system's other goal representations, while tension or anxiety represents that the cognitive process violates at least some important goal representations.

accuracy (p. 174): "utility drives the selection of a specific norm; epistemic content, however, is in itself indifferent to utility" (p. 177).

- Proust (2013) further contrasts *engagement* in metarepresentation and metacognition: the former one is a simulation, the latter one "requires actual performance, the intention to perform a primary task and a motivation for being successful in it." (p. 55)
- Proust (2013) calls the question whether noetic feelings can occur unconsciously a "thorny question" (p. 58) and argues that they have "*embodied* counterparts," e.g. "activity in the corrugators muscles expresses effort of processing" (p. 282; my emphasis).
- Metacognition possesses a *social* component – it serves communication and can be misrepresented to others, but to a limited degree, to make communication possible (pp. 287-288).

[173] There is also a dual processing view that reasoning biases are fluent (quickly generated) and lead to the feeling of rightness (FOR) and constrained analytical processing (Thompson, 2009).

In the case of self-deception, the clash between the goal to think well/healthy of oneself/in-group, on the one hand, and the evidence, on the other, may lead to tension and a biased belief-forming process that will again lead to tension, now in virtue of not being truth-conducive, which at some point may be *compensated* by the absence of inconsistency with other goals such that tension disappears. My point is that, if different goal representations cause tension in self-deception, then, at certain points in time, the intensity of tension might rise or fall proportionally to the amount of violated goal representations, e.g. might lead to a feeling of uneasiness or anxiety.

A further distinction, that I want to make, between counter-intuitivity and anxiety, is best explained in terms of possible worlds. Dokic (2012) argues that the kind of knowledge noetic feelings[174] provide is a *modal* one, e.g. how easy one could accomplish a mental task, which can be modeled as "the degree of proximity of the worlds in which their performance would succeed or fail" (p. 316). Here is one example:

> Something might easily happen if it is the case in nearby possible worlds (where the notion of modal proximity is context-dependent). For instance, the feeling of knowing is the feeling that one's performance is or will be successful in possible worlds close to the actual world. Now these worlds can be more or less close to the actual world, depending on the robustness of one's competence. The more robust one's competence is, the less easily one's performance might fail. If one's competence is fragile, one's performance might fail in possible worlds not too distant from the actual one. One might suggest that *degrees* of noetic feelings can then be modelled in terms of the modal extent to which one's performance is successful. A strong feeling of knowing indicates that one should not expect one's performance to fail too easily. In contrast, a weak feeling of knowing indicates that while one can still do the task, one's performance might more easily fail. (Dokic, 2012, p. 316)

On the premise that on the computational level metacognitive feelings arise as the result of modelling of phenomenally possible worlds in which the metacognitive feelings were representing correctly, there are at least two ways in which one could distinguish between intuitivity and self-deceptive tension (anxiety):

1. In case of counter-intuitivity no phenomenally possible worlds are available, while in case of anxiety the phenomenally possible worlds are goal-impeding (or threatening the integrity of the self-model). Yet if this is the case, no (at least garden-variety) self-deception could

[174] In general, Dokic (2012) agrees with Proust that empirical evidence speaks in favor of the view that metacognitive feelings are

- neither metarepresentational;
- nor special kinds of introspection that provide partial access (e.g. "The capital of Peru is called ___", p. 306);
- instead they are bodily experiences that indicate mental conditions, e.g. ability to remember something or to be able to successfully execute a mental act.

Important characteristics of metacognitive feelings mentioned by Dokic (2012) are the following:

- the indicator-function is *derived*, e.g. subjects have to *learn* what the feelings mean to be able to exploit them in reasoning and judgment (p. 309). The derived nature differentiates them from emotions whose content is supposedly robust ("It is difficult to imagine fear that does not have the function of detecting danger," p. 308);
- noetic feelings are *motivational* "in the sense that they reflect behavioural inclinations that are already in place" (p. 313). The idea here is that *deliberate* metacognition is always based on *procedural* metacognition where the latter has the function of monitoring first-order processes;
- noetic feelings provide *premises* in theoretical and practical reasoning (p. 314) due to them providing *knowledge* (p. 303).

possess a phenomenology of counter-intuitivity, given that it is motivated (goal-directed) and counter-intuitivity is neutral with respect to goals (because phenomenally impossible).

2. The degree of availability of phenomenally possible worlds might be the distinguishing feature between the phenomenology of counter-intuitivity and anxiety.

Anil Seth (2014) further suggests that counterfactual richness is responsible for the experience of realness. I will argue that the self-deceptive representation oscillates in degree of realness (in section 2.2.3). Noetic feelings might constitute the force stripping the self-deceptive representation of its phenomenal realness and enforcing at least some congruence with the evidence.

In Pliushch (2015) I have presented the idea, summarized in this section, that tension is a kind of metacognitive feeling and I would like to counter Joëlle Proust's (2015a) critique of my rationale.

- Object of appraisal: I argued that self-deceptive tension is a kind of metacognitive feeling. Proust asks what this kind of feeling appraises then. The three possibilities of appraisal that Proust (2015a, p. 5) names are: appraising the content of belief, appraising the kind of process that generated it and, finally, appraising *dynamic properties* of the process. The latter possibility is argued to be prevalent in metacognition research and applied by me to tension in self-deception.
- Kind of metacognitive feeling: With respect to my distinction between intuitivity, counter-intuitivity and anxiety as types of metacognitive feelings during belief-forming processes, Proust (2015a) notes that intuitivity is the feelings of fluency and that it is not goal representations, that might be violated in self-deception, but heuristics of *self-consistency*, as well as that it is the amount of *effort* needed to accomplish the task that generates tension (p. 6). As for self-consistency, I agree that it is *one possibility* (of those listed at the beginning of the section) for tension to be generated. As for the amount of effort, it fits best into the quantitative picture on self-deception – it being generated not by certain biasing, but by an *extended* process (see section 1.3.3). It is true that I enlisted different epistemic norms that might be violated above, but did not describe dynamic properties that might indicate the violation of those norms, for which one would need to consider a lower level of analysis. At this point I do not see how one could theoretically encircle the dynamic properties to be violated. I will provide a tentative sketch of a computational analysis of self-deception in chapter 4.
- Extension: In how far metacognitive feelings extend to unconscious belief-forming processes is questioned by Proust. The possibilities named by her here are: dynamic properties leading to feelings are unconscious and feelings themselves are unconscious (Proust, 2015a, p. 6). The idea that I have is rather that the *process* that the feelings evaluate is unconscious, as I described above.
- Cognitive impenetrability as substitution: Proust (2015b) argued that feelings are cognitively impenetrable. I pointed to different kinds of dependencies of feelings on concepts and Proust (2015a, p. 5) made the point that *modulation* of feelings necessitates their *substitution* by other feelings, but that feelings themselves are cognitively impenetrable. I presume that cognitive impenetrability follows for Proust not least from them being a kind of affordance sensing which is different from inference (see next point). Whether feelings are penetrable or not, does not endanger my argumentation about tension as a metacognitive feeling, but an answer would make the relations between affective and cognitive states much more articulated, whose mutually dependent workings bring about self-deception. Cognitive impenetrability might also follow on the assumption of *informational encapsulation*. Proust (2013) names informational encapsulation as a characteristic of procedural metacognition so that the latter "[i]t guides action on the mere basis of the information generated by the on-going cognitive activity and the associated reward" (Proust, 2013, p. 62). Lastly, cognitive impenetrability would also follow from noetic feelings being *immune to revision*, even in the light of the recognition that they "have been produced by some biasing factor" (Proust, 2013, p. 106). In the light of predictive coding (chapter 4), I think that the standard case to assume

would be that of cognitive *penetrability* and hence of various different connections between affective and cognitive states. See for example Anil Seth, 2015b, p. 11): "Instead of distinguishing 'physiological' and 'cognitive' levels of description, interoceptive inference sees emotional content as resulting from the multi-layered prediction of interoceptive input spanning many levels of abstraction. Thus, interoceptive inference integrates cognition and emotion within the powerful setting of PP [predictive processing]."

- Distinctness and directness: I argued that feelings might be as indirect as perceptions, since both are beyond the evidentiary boundary (Hohwy, 2014). According to her, affordances are taken in by *sensing*, which is an ability to be distinguished from *perception* or *inference* (p. 3) and feelings as results of affordance sensing possess their own representational format, which is different from the conceptual one: feelings stem from the "dynamic characteristics of the processes underlying concept use" (p. 4). As noted above, in predictive coding affective states are also results of a certain kind of inference, as perception. According to Quadt (2015), the enactivist/relational and the cognitivist views on the mind can be argued to stand in tension, on the assumption that the former holds perception direct while the latter is not (p. 4). She also distinguishes between two different kinds of directness: directness as a phenomenal quality of some *mental states* and as the quality of an *epistemological mechanism* (p. 6). I think that postulation of an additional kind of process, apart from inference, is not parsimonious and should be avoided, if possible. Particularly, Friston, Shiner et al. (2012) argue that affordance as an attribute of a cue also has to be inferred. Further, the claim that feelings and cognitive attitudes differ in their directness stands and falls with the assumption that both differ in their cognitive penetrability, which I questioned above.
- Predictive coding explanations of feelings: Feelings as violations of transition probabilities (which was the predictive coding description that I provided) is argued to be not new, e.g. to be found in reinforcement learning, and "[t]he concept of free energy, however, is no more equipped to provide any mechanistic account of brain function as any other evolutionary theory" (Proust, 2015a, p. 6). To recapitulate, there is a theory according to which emotional valence can be modeled as a rate of change of free energy (Joffily & Coricelli, 2013) so that minimization of free energy elicits positive valence. Systems tend to minimize free energy, because negative free energy stands for the evidence in favor of a certain model (Friston, Adams, Perrinet, & Breakspear, 2012) that agents want to maximize. There have been argued to be two kinds of transitions in predictive coding: between hidden states (= modeled states of the world) and control states (= beliefs about future actions of the agent) so that they mutually influence each other and lead to a perception-action cycle (Friston et al., 2013). In other words, control states replace goal representations in this framework and transition probabilities are the connections between them. Transitions – that take place because there is self-organized instability - have been argued to relate to consciousness (Friston, Breakspear, & Deco, 2012). Proust (2015b) herself argues that feelings express the rate of discrepancy reduction towards a certain goal. To sum up, *generative models of sensory states* possess free energy, *control states* or *hidden states* are connected by transition probabilities. The rate of change towards a certain goal would in predictive coding terms be the transition from one control state to the next. My aim by arguing that feelings express violations of transition probabilities (without specifying which – between control states or hidden states), instead of more general claim that they express the rate of change of free energy, was chosen to draw attention to the ambiguity and multifunctionality of transitions in predictive coding that I will entertain in great detail in chapter 4. As for the concept of free energy being superfluous, this depends on whether one accepts the claim that predictive coding can also incorporate other frameworks, e.g. reinforcement learning (Friston, 2010; Friston et al., 2013).

Interim conclusion: As seen in chapter 1, characterizations of tension in self-deceptive literature have been quite general and incomplete so far. Thus, in this section I apply Proust's theory of metacognition to characterize tension in self-deception as a certain kind of a metacognitive feeling. I left open which kind of process generates self-deceptive tension (and used the standard folk-psychological term 'belief-formation'), but noted that this process itself may be subpersonal and proposed to categorize the nature of metacognitive feelings arising out of such belief-formation as intuitive, counter-intuitive or anxious, depending on goal representations that are violated in each case. Dokic's modal characterization of metacognitive feelings as successful execution of the process in close possible worlds has tied metacognition to counterfactuals and through that to the question whether metacognitive feelings in self-deception contribute to the disappearance of tension due to the changes in experience evoked by subpersonal counterfactuals. In other words, though tension as a metacognitive feeling should indicate the inappropriateness of the self-deceptive process with respect to the goal of truth-acquisition, self-deception is not abolished, which means that either tension was misattributed, or it disappeared. Worry has been argued to *improve* metacognition (Massoni, 2014), but in self-deception, at least with respect to the self-deceptive attitude, this does not take place.

Last, I noted that in predictive coding terms feelings might express changes in transition probabilities (between control, or hidden states, or both due to their mutual dependency) which is a description congruent with that provided by Proust. Transitions require instability (dynamics) to explore alternative hypothesis, for which the certainty in the current hypothesis about the state of the world has to be lowered (Friston, Breakspear et al., 2012). Friston, Stephan et al. (2014) argue that metacognition is directly linked to *subjective* precision (measure of uncertainty; p. 152), which is also the case for *psychopathology* (about similarity between delusion and self-deception see section 1.2.7). I will speak at length about precision in chapter 4, but here it can already be said that it shares with phenomenological descriptions of tension at least one characteristic, namely its unspecificity:

> However, one might ask if the notion of aberrant precision (ie, neuromodulation) is so inclusive that it is nonspecific? Clearly, to understand the specific ways that aberrant precision is expressed, one has to understand the myriad of neuromodulatory mechanisms (e.g, different neurotransmitters acting on different receptor subtypes in different parts of the brain) in relation to functional anatomy and neurodevelopment. This understanding might be necessary to understand the diversity of psychiatric disorders and the mechanistic basis of their classification. (Friston, Stephan et al., 2014, p. 155)

To sum up, in section 2.1 I first argue that intuitions have so far played an important role in categorizations of self-deception (section 2.1.1). There, I also propose that transparency of the phenomenal signature of knowing that characterizes intuitions might also characterize at least some self-deceptive attitudes which would, first, lead to the dissociation between psychological and epistemic uncertainty, that so far has not been considered in the self-deception literature and second, which would explain why doubt (= psychological uncertainty) does not lead to the abolishment of self-deception. In order to avoid intuition loaded characterizations of self-deception, I proposed (in section 2.1.2) to restrict the explanandum of self-deception to empirically testable third- and first-person characteristics: behavior (inconsistency+justification) and phenomenology (tension+insight). Last, section 2.1.3 was devoted to a more precise characterization of tension as influence of affective states on selection employed in self-deception. In the following section I will say more on goal-directedness of selection (section 2.2.1), the characterization of the process of self-deception (2.2.2) and transparency as one of the properties of self-deceptive attitudes (2.2.3).

2.2 The building blocks: Motivation? Process? Product?

In section 2.1 I have argued that the explanandum of self-deception is its behavior (inconsistency + justification) and phenomenology (tension + insight). As constraints on an explanation of self-deception I have set parsimony, disunity in unity, phenomenological congruency and demarcation (chapter 1). Disunity in unity constraint requires an explanation of behavioral explananda of self-deception and phenomenological congruency – of phenomenological explananda. I justified the choice of an explanandum by parsimony, even though it is in such a form too general to enable a demarcation of self-deception from other phenomena. Further behavioral or phenomenological criteria can be added or the existing refined, though. As such, in this section I will be concerned with the choice of motivation, process and self-deceptive attitude by means of phenomenological arguments (and parsimony).

First, why should one bother about the triad motivation, process and resulting attitude? Those have served so far as building blocks for every theory of self-deception. Funkhouser (2009) argues that it is the nature of motivation and the resulting doxastic state that have to be specified.[175] Van Leeuwen (2007a) also claims that it is the deceptive element and the cognitive attitude that have to be specified.[176] Nelkin (2012) identifies some additional questions that an account of self-deception has to answer:

> 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 pretense 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)
> […] 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. (Nelkin, 2012, p. 119; my emphasis)

The first three questions describe the major ingredients of an explanation of self-deception: the starting point (motivation), the process and the result. The answer to the 6th question constrains the product of self-deception. The 5th question specifies the process question

[175] Nature of motivation and the resulting doxastic state are two building blocks: “This much, however, is widely accepted: self-deception is some kind of motivated irrationality, in which the self-deceiver fails to handle the evidence available to her appropriately. Controversy arises when we try to specify the *nature of this motivation* and the *resulting doxastic state* of successful self-deception” (Funkhouser, 2009, p. 2; my emphasis).

[176] Also in the following account, deceptive element (motivation) and cognitive attitude are two building blocks: “Although the literature is a mess, there are strands of consensus. First, it is widely agreed that some motivational attitude is constituitively involved in causing self-deception; I shall call this the *deceptive element*. Second, it’s uncontroversial that the self-deceiver has to have some sort of access to information that would justify believing the doxastic alternative. Third, it seems agreed that the product of self-deception is some cognitive attitude, where a *cognitive* attitude is one that can be evaluated as true or false – as opposed to conative attitudes, like desires” (Van Leeuwen, 2007a, p. 422).

more precisely, suggesting that the answer to the process question is a certain theory of human reasoning that can accommodate the kind of dissociations to be specified that allow the phenomenon of self-deception to occur (my disunity in unity constraint). The theory of human reasoning is dependent on the theory of the self one embraces. The 4th question also relates to the process question: is our reasoning process generally rational?

The last point to notice is that the *weight* one ascribes to different constituents might change the explanation of self-deception. An example that supports this claim is the debate between Mele and Audi about whether self-deception is an act or a state and if both, whether self-deception being an act or a state should be weighted more in the explanation. Donald Davidson (1986) holds self-deception to be both a state and an act: "self-deception must be arrived at by a process, but then can be a continuing and clearly irrational state" (p. 90). The existence of both static and dynamic paradox of self-deception is also an argument pointing into that direction. A while ago though, Audi has accused Mele's theory of self-deception of being an act-model of self-deception, instead of a state-model which would be able to accommodate the notion of tension necessary for self-deception. Audi's critique is that self-deception is not a historical concept:

> One theoretical suggestion I am making, beyond the point that self-deception seems a kind of dissociational phenomenon, is that whether one enters it is determined more by the *kind* of state one enters than by the kind of *path* one takes in getting there. […] self-deception is not a *historical* concept. If I am self-deceived, so is my perfect replica at the very moment of his creation. (Audi, 1997, p. 104)

Mele (2012) answers that self-deception has an essential historical feature – the time consuming quality of deception (Mele, 2010, p. 748). Independently of the matter of who is on the right track, the "difference in orienting models" (Mele, 2010, p. 745) shows that Mele emphasized more the process of getting self-deceived, while Audi – the resulting state. The latter (self-deceptive attitude) McKay & Dennett (2009) characterize that *as* a kind of misbelief is ascribed to her person as a whole.[177] The relations between the building blocks are presented in figure 9. It emphasizes the connection between the phenomenal world/self-model one might possess and an epistemic agent model that one might construct, which is transparent conscious self-representation of executing epistemic actions.

[177] There might be *subpersonal* and *personal misrepresentations* and it is the latter case that seems to be more interesting: "the dynamics of actual belief generation and maintenance create a variety of phenomena that might be interpreted as evolved misbeliefs. In many cases these phenomena are better seen as prudent policies or subpersonal biases or quasi-beliefs (Gendler's "aliefs"). […] What is striking about these phenomena, from the point of view of the theorist of beliefs as representations, is that they highlight the implicit holism in any system of belief-attribution. To whom do the relevant functional states represent the unrealistic assessment? If only to the autonomic nervous system and the HPA, then theorists would have no reason to call the states misbeliefs at all, since the more parsimonious interpretation would be an adaptive but localized tuning of the error management systems within the modules that control these functions. But sometimes, the apparently benign and adaptive effect has been achieved by the maintenance of a more global state of falsehood (as revealed in the subjects' responses to questionnaires, etc.) and this phenomenon is itself, probably, an instance of evolution as a tinkerer: in order to achieve this effect, evolution has to misinform the whole organism." (McKay & Dennett 2009, p. 508; my emphasis)

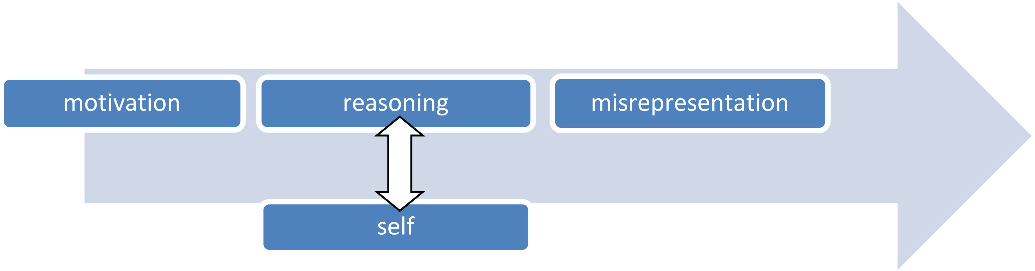


Figure 9. Ingredients of a personal level explanation of self-deception.

The blue arrow denotes the temporal aspect and the white arrow – the possibility of bidirectional relationships, e.g causal or of other sort (for more food for thought on this possibility see section 1.3.4).

The results will be as follows: In section 2.2.1 I will argue that the motivation for self-deception consists of multiple subpersonal goal-representations. Since goal representations fulfill the functional role of context-sensitive selection, I will sum up the kinds of selection according to phenomenology (personal and subpersonal), type of process (world/self-model construction vs. epistemic agent model construction), object of selection (information/evidence, internal knowledge activation, counterfactual model). I will then argue that since self-deceptive selection is subpersonal, the two main types of self-deceptive selection differ with respect to one phenomenological property - the transparency property of self-deceptive attitudes. If the resulting self-deceptive attitude is transparent, which means that it has acquired the property of realness for the self-deceiver (it is part of how the world is and not how one takes the world to be), then world-model selection has taken place and if, despite absent or gap-full epistemic agent model, the resulting attitude is opaque, which means that it is still experienced as an attitude one has acquired and has to defend, then epistemic agent model selection has taken place. How to conceive of such gap-full or inexistence epistemic agent models that produce self-deceptive opaque attitude on the phenomenal level will be then the topic of section 2.2.2, where I will first review the two alternative views: single strand of thought or linear model, dual strand of thought or dual process model and multiple strands of thought that bind into each other on the personal level as dolphin model of cognition. Last, in section 2.2.3 I will summarize three properties of self-deceptive representations – degree of consciousness, certainty and transparency – and focus on exploring the possibility of how self-deceptive attitudes might acquire the property of realness.

I would like to start with revisiting the general argument for why the subpersonal level matters for explanations of self-deception:

1. Personal level explanations employ the vocabulary of folks-psychology and are governed by rules of logic (see section 1.3.4).
2. Self-deception cannot be explained by embracing folk-psychological terminology in a paradox-free manner if a *dispositional* account of belief is accepted (section 1.2.6), because it assumes the possession of states with contradictory content and such possession would go against the rules of logic (or weaker, against coherence).
3. On the *constructionist* folk-psychological account of belief, self-deception could be understood as narrative construction (see section 1.2.7). Still, either one would need to answer the diachronic version of the contradictory content problem, because one would remember what one believed a week ago,[178] or one would delegate the explanatory weight to the (subpersonal) selective processes that not only choose the states upon which a belief at

[178] That this is not the case for the recognition of anosognosia via caloric vestibular stimulation see section 3.1.2.1.

each moment in time is constructed, but also those that one is conscious of. In the second case it would be mechanism underlying *global availability* that would explain self-deception.

4. Global availability is one of the constraints on the concept of consciousness (Metzinger, 2003) that has been criticized by Weisberg (2005) for the reason that plays a central role for any explanation of self-deception, namely selective guidance of behavior. Weisberg (2005) argues that in the face of "coherent folk-psychological explanations [of selective behavior] in terms of nonconscious states" (p. 10), global availability is not necessary for selective behavior, which makes the function of this constraint less clear. A case of self-deception is the one where one would like to ascribe selectivity to conscious, as well as unconscious states, but where those two sets of states would contain a contradiction and would violate the rules of logic. Thus, at least in the case of self-deception, (a) either one restricts personal (folk-psychological) explanations only to the domain of the conscious, or (b) one accepts that the rules of logic apply independently to conscious and unconscious states, which would mean that there is a boundary between the two. I think that the first option is more parsimonious.
5. Subpersonal states are unconscious, but these two distinctions (conscious-unconscious, personal-subpersonal) do not rule out that there are personal unconscious states. However, for parsimony reasons, I would like to avoid psychodynamic connotations.
6. Thus, the subpersonal level should play an important role in the explanation of these constituents of an explanation of self-deception.

Given the necessity of taking the subpersonal level into account in explanations of self-deception, one should be open to the possibility of the necessity to *revise* the folk-psychological concepts used to explain self-deception, as there need not be a mapping between the two (Anderson, 2015).

2.2.1 Motivation: goal representations

> When people are actively pursuing a goal, by definition they *want (desire)* those things that can help them achieve the goal, and similarly should not want those things that prevent them from reaching the goal.
> (Fishbach & Ferguson, 2007, p. 497; my emphasis)

This section possesses multiple aims: First, I want to propose the alternative to the folk-psychological description of the motivation for self-deception as intentions or desires, namely its description as *subpersonal* goal representations. The personal and the subpersonal level of description do not enforce the *same kind* of description, namely the one in folk-psychological terms. Where on the personal level we speak about intentions and desires, on the subpersonal one, we should speak about goal representations (Frankish, 2009).[179] Moreover, in self-deception more often than not there will be no conscious intentions or desires, but *only* interactions between subpersonal goal representations. The argument in favor of this proposal is that goal representations can exhibit context-sensitivity even when operating effortlessly, unconsciously and implicitly. Second, on the example of self-regulation that has been argued to be the process by which self-deception is accomplished, I will show the difference in describing the same process in agentive vs. non-agentive terms. Third, I will extend and characterize the kinds of selection that might lead to self-deception. To the distinction between personal/subpersonal selections,

[179] Saunders & Over (2009) define an instrumental view of *rationality* as the one "according to which people are rational to the extent that they reliably achieve their goals" (p. 318). Frankish (2009) notes that if one wanted to "assimilate" the concepts of belief and desire to some "theoretical concepts of subpersonal cognitive psychology" by their functional role, then they would correspond to "concepts of knowledge (or memory) and goal structure" (p. 91).

selection of world-/self- or epistemic agent model will be added the distinction according to the objects of selection: evidence (external), knowledge and counterfactual models (internal). I will then argue that the second distinction is the most important and can be phenomenologically distinguished on the basis of self-deceptive resulting attitudes. Fourth, I will enrich the description of tension by the counterfactual goal-directed pull condition.

To remind the reader, philosophical theories propose intentions and different kinds of desires as motivation for self-deception (for a summary see section 1.1.3). *Complexity* of self-deception has been used as an argument that self-deception cannot be automatic, but would need effort on the part of the agent (see section 1.1.1.2). Complexity of self-deception stems from its *selective* nature that both intentionalist[180] and deflationary accounts have trouble to explain (see demarcation constraint in section 1.1.3). There are three distinctions in the psychological literature that imply a complexity and agency difference. Those - unconscious-conscious, automatic-controlled, implicit-explicit, - have been argued to play a role in self-deception (von Hippel & Trivers, 2011b; see section 3.2 for their evolutionary theory). In particular, von Hippel & Trivers (2011b) argue that in self-deception there is a dissociation between socially desirable controlled goal representations and socially undesirable automatic goal representations[181] (p. 7).

In the following I will present empirical support for the idea that goal representations are *context-sensitive, creative and selective* (see table 20) despite operating *unconsciously*, *automatically* and *implicitly*.[182] Goal representations are selective in that they impact the activation of internal knowledge structures, but also consideration of external information. In virtue of their selectivity they are creative, since differential activation of affective and cognitive states will influence one's experience of the world and one's argumentation. Goal representations are also argued to be flexible and *context-sensitive*, and particularly to be so in an *unconscious* manner due to the fact that they operate in a *complex* environment[183] (Aarts & Custers, 2012). Goals pursuit is argued not only to be *context dependent and situation specific* in the sense that the context provides a departure point, but it is also argued to alter the applicability of activated knowledge, as well as to be able to also alter the memory structure during the usage of concept.[184] The strength of association between

[180] Model of intentional action can be misapplied: "Advocates of this thesis [without awareness no reason to bias data] tend to understand self-deception on the model of *intentional action*: the agent has a goal, sees how to promote it, and seeks to promote it in that way" (Mele, 2001, p. 53).

[181] This is the example given in the automatic-controlled section that demonstrates that the boundaries between the difference between automatic and unconscious are not always clear: "For example, a student whose parents want her to be a physician but who wants to be an artist herself might follow her conscious goal to major in biology and attend medical school, but her unconscious goal might lead her not to study sufficiently" (von Hippel & Trivers, 2011b, p. 7).

[182] Note that an *automatic* process can also be *intentional,* as "[p]eople who have a goal that they want to accomplish over a period of time can automate that goal pursuit" (Andersen et al., 2007, p. 144). Andersen et al. (2007) have "intentional, controlled, effortful and aware processing" in mind (p. 143). This is not a kind of intentionality though that one would suppose is at work in self-deception, because self-deception is not a kind of habit formation.

[183] Dynamic nature of the world has been brought about as a reason why unconscious goal pursuit should be possible (Aarts & Custers, 2012, p. 238).

[184] Context has been argued to influence the accessibility of knowledge structures and the connections between those structures: "Indeed, regardless of the underlying structure and operation of the representational system, two basic principles emerge from this discussion. First, context provides a starting point for the rumblings of the representational system, determining the *departure point* from which some subset of available knowledge will be made relatively accessible. Second, because of the malleability of the process, the triggering and use of a concept *alters the memory structure* in a way consistent with the situational specificity. Thus, the construct triggered by and used in a specific context is altered accordingly in the

stimulus and knowledge also determines which concepts are activated if there are competing cues.

Definition	
Fishbach & Ferguson (2007) “define a goal as *a cognitive representation of a desired endpoint that impacts evaluations, emotions and behaviors*” (p. 491).	
Selectivity of goal representations	**Content of goal representations**
1. Goal representations fluctuate in accessibility, as **memory constructs** do. 2. Multiple interconnected memories relate to a certain goal representation. 3. Classical *knowledge activation processes* are responsible for the activation of these memories. In particular, these memories possess **excitatory and/or inhibitory links** to the others, which may be bidirectional. Moreover, the activation of a goal representations “only dissipates when the goal has been reached” (p. 492).	- **end state** which has to be desirable (= associated with positive affect). This can be achieved in “a more implicit, nonconscious fashion” or conditioning - **means** to reach that end state: behavior, objects, plans, other end states etc.
Interaction types between goal representations	**Context creation** by influencing “knowledge accessibility, evaluations, and emotions” (p. 496)
• **Competition** for “limited motivational resources” if the goals are competing ones (p. 501); • **Multiple goal attainment** in case of multifinal (conducive to more than one goal) means (p. 502).	• Accessibility of *knowledge* that could lead to goal attainment is increased (pp. 496-497); • *Stimuli* consistent with a given goal are evaluated more positively when that goal is activated, while inconsistent stimuli are evaluated more negatively (p. 498). • Goals can be associated with different *moods/emotions* and the latter can influence the behavior, thereby regulating the goal attainment process (p. 499).

Table 20. Fishbach & Ferguson: characteristic of goal representations. Distinctions from Fishbach & Ferguson (2007).

As for the last triad – unconscious, automatic, implicit - activation and inhibition of goal representations are assumed to be able to occur *implicitly* (Andersen et al., 2007, p. 145). Goal representations can influence the information processing in an *automatic (unconsciuos)* way so that the behavior of subjects appears inconsistent over time (Huang & Bargh, 2014b, p. 127) and one’s experience of the world gets transformed in the process without the subject noticing it (p. 123-124). That automatic processes are also argued to be (often) unconscious, can be seen on two following definitions of automaticity:

> If initial stimuli are perceived consciously but a provoked process occurs outside awareness, or even despite attempts to prevent it, this is also automatic. Accordingly, automaticity is granted if the perceiver *lacks awareness* of the process, *does it with efficiency* (i.e., with minimal use of cognitive resources), has *no intention* to do it, or *cannot control* it. This approach frames automaticity as a continuous variable rather than as a discrete class of processes and as a matter of degree rather than kind. (Andersen et al 2007, p. 139)

> The defining features of automatic processes are that (1) they do not involve conscious awareness; (2) they do not require a person’s intention to be started; (3) they operate even under limited cognitive resources; and (4) they cannot be stopped or altered voluntarily. Conversely, controlled processes (1) operate under conscious awareness; (2) require a person’s intention to be started; (3) fail to operate when cognitive resources are limited; and (4) can be stopped or altered voluntarily […]. (Gawronski & Bodenhausen, 2012, p. 23)

process, such that whatever is stored is stored in modified form” (Andersen 2007, p. 143; my emphasis).

Summing up, recent empirical literature points into the direction of goal representations operating unconsciously[185] (Fishbach & Ferguson, 2007), implicitly, automatically (Andersen et al., 2007), but despite that providing context-sensitivity that might be enough for self-deception. Interestingly, Huang & Bargh (2014b) compare goals to genes in order to emphasize that both goal representations and genes might lead to actions detrimental to the individual making them. I will talk more about the role of genes for self-deception in section 3.2.2 when I discuss Trivers' evolutionary theory of self-deception. At this point, I want to provide an example what difference it makes if a certain goal-directed phenomenon – self-regulation – is described in agentive or non-agentive terms. Self-regulation is the last stage at which influence can be wielded upon the degree of activation of a knowledge structure (see figure 10). At the same time, self-regulation has been argued to be the process by which self-deception occurs (Baumeister, 1996).

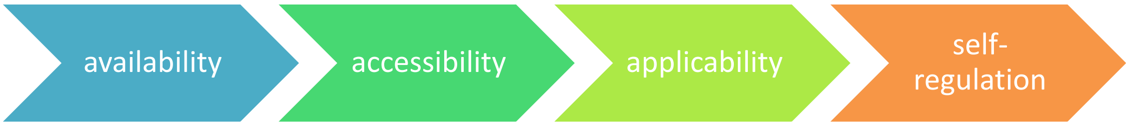


Figure 10. Andersen: processes determining knowledge use.
Distinctions from Andersen et al. (2007).

Description of the processes by the example of social knowledge:
"First, social knowledge must be *available* in memory if it is to guide subsequent processing, and thus *availability* of social knowledge is a precursor to its use and understanding knowledge acquisition is thus relevant. Second, social knowledge must be accessible. *Accessibility* refers to a construct's readiness to be used, commonly defined as the degree to which it can be automatically *activated.* Third, accessible social knowledge may or may not be applied. *Applicability* refers to how well social knowledge matches attended-to features of a stimulus or situation. This match, or the usability of accessible social knowledge, determines in part whether or not the knowledge is applied. *Self-regulation* covers a heterogeneous set of processes that affect the accessibility and application of social knowledge such as inhibition, suppression, adjustment, or enhancement. Self-regulation can thus short-circuit activation at the outset, or redirect it once it has occurred, and can also prevent application or introduce a postapplication correction."
(Andersen et al., 2007, p. 139; my underscore emphasis)

Baumeister's description of self-regulation, by which he also explains self-deception, possesses a very strong agentive flavor, while Andersen's et al (2007) description is more neutral in this respect. The reader is encouraged to estimate, how big the changes in the descriptions are. Baumeister (1996) explains self-deception in the self-regulation framework. Self-deception is self-regulation that is guided by goal representations other than acquisition of truth[186] (Baumeister & Newman, 1994) and a defensive response is a major form of self-deception: "[r]efusing to draw obvious but undesirable conclusions in order to cling to preferred beliefs is one major form of self-deception" (Baumeister, 1996, p. 29). High self-esteem is argued also to be defensive, if it correlates with high scores on Paulhus' self-deception questionnaire (Baumeister et al., 2003, p. 5; for Paulhus' BIDR and criticism see section 1.3). Trait repressiveness is further said to be indicative of self-deception (Baumeister & Cairns, 1992, p. 853). Baumeister's self-regulation theory, thus, *assumes* a defensive function of self-deception to begin with. This is due to the central idea

[185] Withstanding temptation can be achieved nonconsciously by implicit goal representations being activated by temptation cues (Fishbach & Ferguson 2007, p. 504; my emphasis).

[186] Self-regulation as a means to self-deception: "To the extent that the preferred conclusion is not the same as the accurate or optimal conclusion, self-regulation may serve ends of self-deception - that is, convincing oneself of something that is not correct or not consistent with all the evidence" (Baumeister & Newman, 1994, p. 4).

of the theory that an undesirable conclusion is an “ego threat” to self-esteem which self-regulation is bound to ward off (Baumeister, 1996, p. 29). In a study Baumeister & Cairns (1992) tested that upon being presented with an unfavorable personal profile of which one knows that it will be shown to others, repressors listed more thoughts related to it (p. 858). Baumeister & Cairns (1992) interpreted this as that repressors were having more *ruminative* thoughts due to an attempt to generate refutations (p. 858). Consistent with it, repressors remembered negative feedback better (p. 861). As one can see, if self-deception is equated with repression, then a profile of the self-deceiver is no more that of somebody who avoids unfavorable evidence, but who remembers it better. Thus, it is a case in which all evidence is processed at the personal level and then refuted, so that a desired conclusion can be reached, as is evident from this quotation where the self-deceiver is the ‘intuitive lawyer’ and the neutral reasoner is the ‘intuitive scientist’:

> But the intuitive lawyer's goal of reaching a specific conclusion can be well served by regulating the assessment of implications, because here one already has evidence and is evaluating its quality and clarity. *The individual already knows at this point whether each bit of evidence is favorable or unfavorable to the preferred conclusion, because those implications were automatically recognized as soon as the evidence was found, and so one can manage the assessment process so as to slant it toward discrediting the disagreeable information while allowing the favorable evidence to stand.* [...] In short, there may be some tendency for the intuitive scientist to concentrate on regulating the collection of evidence, whereas the intuitive lawyer concentrates on the assessment of implications. (Baumeister & Newman, 1994, p. 8; my emphasis)

So, here, it would seem that self-deception would occur when an *agent* would misinterpret the evidence in a favorable way, that it is an effortful activity with its own phenomenology. Andersen et al. (2007), from a different point of view, characterizes self-regulation as an *automatic* phenomenon. Among the goal representations that Andersen et al. (2007) mention as “core human motives” underlying self-regulation are those that are mentioned in the self-deception literature to underlie self-deception, such as the goal to enhance self-esteem, as well as the goal to be accurate and to avoid bias[187] (p. 155). The authors speak of the emerging evidence in favor of the hypothesis that goal representations can shape self-regulatory processes “relatively unconsciously and effortlessly” (p. 155), stereotype inhibition is further argued to occur *automatically* (Andersen et al., 2007, p. 156), not least by automatic processes of *selective attention* (pp. 156-157). Andersen et al. (2007) elaborate further that goal representations are activated automatically by contextual cues, that goal satisfaction leads to positive affect and to tension if not,[188] goal representations could become chronic (more readily accessible than others) and influence the way information is processed (pp. 148-149). Unconscious goal representations that influence information processing and lead to negative affect, if not satisfied, is a description that should be familiar to the reader by now from descriptions of self-deceptive processes.

Before I evaluate the difference in description between self-regulation by Baumeister and Andersen et al., I want to shortly note how latter authors connect self-enhancement, in-group/out-group biases and cognitive dissonance to self-regulation. This is because those have been taken to explain self-deception by different researchers (see section 3.1) and

[187] Two other goal representations that I find important are the goal to understand or comprehend, as well as the goal to get along with others and making a good impression (Andersen et al, 2007).

[188] Tension as accompaniment of goal pursuit: “Rather, desiring an end state involves experiencing associated tension prior to goal satisfaction and is a precondition for nonconsciuos goal priming” (Andersen et al., 2007, p. 149).

Andersen et al. suggest that those also operate non-agentively. In self-regulation, the *goal representation of self-enhancement* becomes active when the self-concept is threatened by triggering negative stereotypes that are relevant to the latter. In these cases automatic self-protective biases, such as "ingroup favoritism and outgroup denigration" can occur (Andersen et al., 2007, p. 158). The latter are mentioned by (Trivers, 2011) to be self-deceptive. Andersen et al. (2007) also mention that *cognitive dissonance* plays a role in self-regulation. Scott-Kakures (2009) takes cognitive dissonance to explain self-deception. Cognitive dissonance is argued by Andersen et al. (2007) to dependent on the goal context: "However, cognitions seen as dissonant in the context of one goal or standard may not seem dissonant in the context of another" (p. 161). Tension is evoked not by discrepant cognitions per se, but by unfulfilled goals: "Unfulfilled goals evoke tension, which then evokes a drive to reduce it" (p. 161). If dissonance reduction can be primed in cases when certain goals/standards are activated, then it can be taken as "evidence for automaticity in dissonance"[189] (p. 161). Cognitive dissonance may also lead to selective attention to goal-relevant stimuli, if this dissonance is caused by goal representations that could not be fulfilled and are, thus, incomplete and prone to trigger compensatory responses (Andersen et al., 2007, p. 161; my emphasis).

I see the core distinction between Baumeister's and Andersen's et al. descriptions in that the former assumes the existence of the *distortion* of the evidence in certain self-regulatory processes, e.g. self-deception, while the latter emphasizes its *selection*. It is to note that 'distortion' is a verb with a strong agentive connotation, while selection is not. This is not to say that selection cannot be agentive. According to Wayne Wu (2013), mental action is then characterized by finding a selected[190] (intended) path in the behavioral space with multiple inputs (memory contents) and outputs (possible kinds of behavior). *Mental action* for him is intentional selection such that intention may be unconscious (Wu, 2013). I would refrain from using the term 'intention' to denote subpersonal goal representations, but for me, selection per se does not have agentive connotations. As seen above, goal representations fulfill a selective function, yet whether there is an accompanying phenomenology to that selection and which *kind* of phenomenology that is, is in question. Particularly, a useful distinction is one made by Irving (2015) between *motivation* (goal-caused) and *guidance* (dynamic phenomenal feature). Guidance can be *habitual* (passive phenomenology of contents of experience unfolding over time) or *goal-directed*

[189] Motivation, e.g. for self-consistency or self-affirmation, has been argued to be susceptible to priming and, as such, provides evidence for the *automaticity* of dissonance reduction (Andersen et al., 2007, p. 161). This is interesting, because it is indirectly a piece of criticism for explaining self-deception by self-affirmation (see von Hippel and Trivers, 2011b for such an explanation) for those who accept an intentionalist position on self-deception, because intentionalists (see section 1.1.1) argue that self-deception cannot be explained by mechanisms that operate automatically.

[190] Jennings & Nanay (2014) argue against Wu in that *attention* is not necessary for action, since there may be cases where one does not have to select an appropriate behavior given the input, but where it is *mental preparation* that distinguishes action from a mere reflex:

"In the case of the café regular, for example, the ordering of a favoured dish counts as action because it requires the café regular to perceive it as his or her favoured dish. If the café regular ordered the dish 'just because,' it might count as reflex, but the café regular orders the dish because it is his or her favourite, which is a process that includes specific mental preparation." (p. 3)

Their argument here is that in the cases, in which an agent would behave in the samemanner (no variation), there is mental preparation, but not selection and thus, no role for attention, e.g. jumping back in order to avoid a ball although there is a glass wall between the person and the ball (p. 4). Jennings & Nanay may be right in the case of action in general, but I agree with Wu that in the case of mental action attention is essential.

(phenomenology of agency over contents of experience in accordance with one's reflected goals), but both share the counterfactual condition for guidance: if focus on goal-related information were to cease, discomfort would follow. Rumination and absorption are according to Irving guided, while mind wandering is not. So, personal level selection might be either habitual or goal-directed. I would not use the term 'goal-directed' for the second type of guidance, since I think that if the authors I cited above are correct about goal representations operating unconsciously and automatically, motivation can allow for goal-*direction* already (as phenomenally unavailable goal-pursuit). But I agree that if selection is phenomenally available, it might either come with a sense of agency (= sense of control) or not.

Which kind of selection is responsible for self-deception? Several might be, depending on which kind of phenomenology and behavior description one provides for self-deception, particularly for the self-deceptive process. Before I enlist all the selection options, let me answer one question: can self-deception be habitually guided? Is this intermediate option between subpersonal and personal (agentive) selection a possibility? I ask this question now, because Irving (2015) argued rumination to be motivated and habitually guided and Baumeister described self-deceivers as ruminative repressors. Further, allocation of attention has been offered as a means of deceiving oneself (see Lynch, 2014 for critique, but see section 1.1.2.4 for defense). Answering this question shows how a choice of a certain phenomenology of self-deceptive process leads to a choice of a certain kind of selection responsible for it. Lynch (2014) argues that two roles have been ascribed to attention in self-deception: forgetting the unfavorable evidence by shifting attention away from it or, on the contrary, shifting attention to it and rationalizing it (p. 66). Lynch's (2014) argument against attention as means of deceiving oneself is that shifting attention away from evidence does not lead to its forgetting, especially in the light of the *thought-suppression* literature pointing to difficulties of successfully suppressing thoughts. As the author mentions, greater vividness of anxiety provoking thoughts makes those even more difficult to suppress (p. 71). Anderson & Hanslmayr's (2014) review of the literature not only suggests that motivated forgetting is possible, but also that it happens due to inhibitory control processes which can affect different information-processing stages and not that it is achieved in a passive way. More specifically, the authors argue that inhibition might evoke a *context shift* along selected dimensions (and not only the temporal one), given that participants could selectively forget items based on the gender of the speaker that mentioned those items (Anderson & Hanslmayr, 2014, p. 283). Suppression is seen by Anderson & Hanslmayr (2014) as only one of the approaches that limit awareness, along with self-distraction (p. 286). The authors even argue that experimental results of the suppression literature might underestimate the effect of suppression due to the absence of the powerful motivation for participants to exhibit suppression upon experimental items, e.g. random words or lists (Anderson & Hanslmayr, 2014, p. 289). Further, suppression-induced forgetting is 'cue-independent' – it generalizes to novel, contextually related cues (p. 284). I think that a generalization from thought to emotion suppression is possible. Emotional saliency of the given contents would be then regulated via subpersonal selection mechanisms and the difference between normal individuals and pathological cases would consist in the success of these subpersonal mechanisms in the former, but not necessary in the latter case. An example would be an individual with ruminative thoughts that cannot disengage from them and remains captive inside the ruminative context. In extreme cases such captivation might hinder goal-directed actions and reduce life quality in general. Normal individuals might trigger these subpersonal processes by engaging in mind wandering or just communicating with other people.

So, let us go back to the question about the phenomenology of rumination and self-deception: Successful thought suppression is the end for rumination. Irving has argued that during rumination there is the phenomenology of being pulled to the contents, but without the sense of agency for choosing those contents. I do not think that thought suppression possesses a sense of agency, since this would be counter-productive: If I want something to disappear from phenomenology, then just disappearance of certain contents would be better than if it were accompanied by sense of agency about letting those contents disappear. But in the end, it is up to the reader to check his memory about his experiences of rumination and thought suppression. If unfavorable evidence triggers rumination that is eliminated by thought suppression in self-deception, then self-deceivers would be passive observers of their self-deception that they could not justify, because when questioned by others, the worst might happen - ruminative thoughts might return. For this not to happen something must change – some kinds of contexts need to permanently change and this change will not be phenomenally available for the rumination-thought suppression pair. Thus, habitual guidance might precede self-deception, but it will not describe the self-deceptive process itself, because the crucial context changing selection will not be part of this guidance. Yet interestingly, Irving's *counterfactual* condition (pull towards certain kinds of information and discomfort if it were not possible), has been also introduced for self-deception by Van Leeuwen:

> Synchronically, they are all characterized, as noted, by the holding of a cognitive attitude contrary to the agent's evidence and epistemic norms, where this holding is under the influence of a desire or other motivational component. Diachronically, each kind of self-deception involves the influence of that motivational component on *attention* in a way that produces and supports the cognitive attitude or product of self-deception. The attention is pushed by aspects of the motivational element *away* from the greater evidence that supports belief in the doxastic alternative and onto the scantier evidence that supports the product of the self-deception. What makes this a process of *self-deception* is that the agent feels a rational pull to attend to evidence she's not attending to (which would dictate believing the doxastic alternative if properly considered), but she motivatedly ignores that evidence nonetheless. (Van Leeuwen, 2007a, p. 425; my underscore emphasis)

Van Leeuwen argues for a *rational* pull. But in virtue of the fact that if goal representations are not fulfilled, tension might also occur (section 2.1.3), I am inclined to argue for a goal directed pull also in case of self-deception, like Irving postulates. What one might have then is two different pulls – rational and non-rational – leading to similarly and, maybe, even undistinguishable phenomenology blending together into that feeling that self-deceivers then in the end experience as tension. Summing up, guidance is not a property of a self-deceptive process per se, since a cardinal change of the context would be needed for self-deception, but *counterfactual goal-directed pull* can be taken as an enrichment of the description of phenomenological tension. Let me now summarize different kinds of selection that might lead to self-deception (see table 21).

Phenomenology of selection	**Process involved**	**Object selected**
• present (personal selection) • absent (subpersonal)	• world/self-model • epistemic agent model ○ belief-formation ○ narrative construction	• information/evidence from environment • knowledge structures • counterfactual models

Table 21. Kinds of selection

In chapter 1, I have already spoken about personal/subpersonal selection, as well as selection as construction of the world/self-model, in difference to selection as construction of an epistemic agent model. Self-deceptive representations have been argued to be acquired on the basis of *selective* information (e.g., Mele, 2001), goal-conducive information needing to be evaluated more positively (e.g., Van Leeuwen, 2008; see section 3.2.3.1), negative affect being a motivating force leading to biased personal level belief forming process (see comments on Baumeister & Newman (1994) in the tension section). The review of the goal literature given in this section has added the *knowledge* selection to the list. Wu (2013) also distinguishes between *two kinds of attention – perceptual* and *cognitive*. The difference between perceptual and cognitive attention is according to Wu (2013) that in the first case input is simultaneously present for selection, but in the case of cognitive attention it must first be somehow activated. It should be also noted that this view about how goal representations influence information processing (by activating certain knowledge structures) is based on certain assumptions about how knowledge is stored (in a connectionist network), activated and by means of being activated influences other psychological processes. The assumption is that goal representations activate other representations and only those representations that have been activated have the potential to subsequently influence the belief-forming process. Different degrees of activation allow for different kinds of influences, which those representations can exert over the belief forming process. The use of knowledge structures has been argued to be determined by four basic processes, each of which describes a stronger influence of the knowledge structure (Andersen et al., 2007).

Last but not least, apart from evidence and knowledge selection, selection can also exhibit goal-directed influence on our reasoning process. Counterfactual thinking could directly contribute to the optimism bias, as Brigard & Giovanello (2012) have found that "participants were more likely to think that it is more plausible that a past event with a positive outcome could reoccur in the future, than to think that a past event with a positive outcome could have occurred with a negative outcome" (p. 1093). Gerlach et al. (2014) have found episodic counterfactual simulations (imagining events with different outcomes) to distort the memory of an event, more so in older than in younger participants. For this reason, Gerlach et al. (2014) call episodic counterfactual simulations a "kind of internally generated misinformation" (p. 158). The parallel to von Hippel & Trivers' (2011) claim, that older people are more optimistic than younger ones, is to be emphasized. Girotto et al. (2007) also found that actors are more prone than readers to generate counterfactuals that modify the problem, rather than change the outcome. For example, upon given choice to select an envelope with either a simple or difficult math problem, actors (those who had to actually solve the problem) more eagerly constructed a counterfactual situation in which they had a calculator to solve a difficult math problem, instead of just changing the outcome by choosing another envelope (p. 514). Girotto et al. (2007) argue that this behavior depends on the *saliency* of information which is, according to me, a matter of selection. McCulloch & Smallman (2014) explore the hypothesis that different counterfactuals elicit different processing styles, or that additive counterfactuals (imagining new actions that could change the result of an event) elicit an expansive processing style (fosters creativity) and subtractive counterfactuals (imagining undoing certain actions) elicits a relational processing style (concentrates on the existing relations among entities in the event).

Counterfactuals not only bias, but also evoke certain kinds of affect. Myers' et al. (2014) study indicates that counterfactuals (thoughts of the form "If only I had done x in situation y") influence goal-directed behavior, e.g. solving anagram tasks (rearranging letters of a word to produce another word), but only if participants were in the negative mood. On the premise that negative affect triggers the motivation to improve, the authors affect such personal level counterfactuals to goal-directed behavior: when motivated, subjects would use counterfactuals for planning and changing the strategy of solving a certain problem. Gleicher et al. (1990) argue that, especially in the case of a negative outcome of an event, counterfactuals are generated so that they lead to negative affect. They argue more precisely that "an extremity of one's affect toward an outcome will be directly proportional to the judged likelihood of occurrence (and judged valence) of a counterfactual outcome that has a valence different from reality. Thus, a person should feel extremely bad if a good alternative outcome is judged as having been quite likely," (p. 293) but did not occur.

So far, I have enlisted different selection possibilities, opted for subpersonal selection in the case of self-deception, but described literature on the influence that personal level counterfactual thinking may exhibit on our cognition. In the previous section on tension, though, I have also described Dokic's model view on metacognitive feelings that represents those feelings dependent on proximal possible worlds in which an action would be successful or not. So, why should one bother with personal level counterfactuals and in how far do they help in self-deception? Since a folk-psychological description of a self-deceiver is easier to accomplish than a construction of a computational model, the empirical literature, cited above, serves the aim to activate the available folk-psychological intuitions on self-deception that may make it easier to make afterwards a transfer to the subpersonal level. Subpersonal counterfactuals influence experience, but are not part of that experience. I will talk about them at length in chapter 4. If self-deceptive selection possibilities are narrowed down to inward (knowledge, counterfactuals) and outward (environmental information) subpersonal selection, then, since no epistemic agent model of self-deception will be constructed, self-deception will be a case in which either the world/self-model is changed, or self-deceptive attitudes "pop-out" in the mind of the self-deceiver. Whether self-deceptive process is named 'belief-formation' or 'narrative construction' on the personal level does not matter, since on the subpersonal level both be a certain kind of hypothesis testing (4). Actually, construction of a world/self-model would also be a kind of hypothesis testing according to predictive coding, so the difference between the two main kinds of self-deceptive selection – world/self-model construction and some kind of epistemic agent model – would be a phenomenal one, namely whether the resulting attitude is transparent or not (see section 2.2.3). Why do self-deceivers do not notice that some attitudes just appeared without an epistemic agent model being constructed for their acquisition? Or with an epistemic agent model possessing gaps? This will be the topic of section 2.2.2.

I want to conclude by commenting on two things: first, the unit of complexity that determines behavior and second, the relationship between goal representations and affects, affordances, reward and feelings. First, there is no agreement on which unit of complexity ultimately determines the behavior: whether it is reward (Ainslie, 2014) or affect (affect = hedonicity + arousal; Bliss-Moreau & Williams, 2014) as components of goal representations, goal representations themselves as Huang & Bargh (2014b) propose, or overarching units, such as personality and self-concept (Hirsh, 2014) or a (motivational) self (Baumeister & Winegard, 2014; Fishbach, 2014). Also, there is no agreement on the relationship between affective states and goal representations, both of which have a selective quality. The first possibility is that it is a *reward value* of a goal state that leads to the pursuit of an unconscious goal ("unconscious goal pursuit may result when a preexisting desired goal is activated, which, because of its association with positive affect, sets off a positive reward signal," Aarts & Custers, 2012, p. 241). In this case goal representations would be independent from affective states. The second is that "the mere coactivation of a neutral goal concept and positive affect produces unconscious goal pursuit" (Aarts & Custers, 2012, p. 241). In the second case, the affective state of the system would, at every moment of time, influence goal pursuit. If Proust's theory of metacognitive feelings is accepted (that they express affordances), one might argue the other way around, namely that since there is a continuous interplay between goal representations and affordances (opportunities in the environment; Huang & Bargh, 2014b, p. 125), goal representations might change the perception of affordances and, hence, feelings. This interplay is important for self-deception, since it might help to elucidate the fact how tension-loaded self-deception becomes tension-free, namely by feelings influencing the desirability of subpersonal goal representations such that in the next hypothesis-testing cycles different attitudes are acquired.

Finally, there is also an argument against goal representations as determinants of behavior. If every representation possesses a motivational force (Huebner & Rupert, 2014) and there are multiple and hence, redundant representations for the same circumstance (Ibid.), it is the interaction of those partly overlapping and redundant representations that determines behavior (Ibid.). Here I will adopt a, probably, more conservative view that it is the goal representations that determine behavior, which is a matter of convenience, but I think that it is plausible to assume that not always only *one* goal representation guides behavior. Self-deception provides a paradigmatic example for the case in which *creative tension* between different goal representations (acquire a truthful attitude/model of reality vs. acquire another desirable attitude) may be the motivating force.

I would like to conclude by stating that one can defend a hypothesis that both *emotion* and *motivation* (goal representations) compete, but also interact in order to influence perception and executive control (Pessoa, 2015a), that the general form of influence – redirection of information flow – similar in both (p. 11). This hypothesis is different from the one that *cognition (attention)* and *motivation* are embedded (in the neuronal level) too (p. 13). Pezzulo et al. (2015) speak of integration of emotion and motivation, but also exteroceptive, proprioceptive and interoceptive signals, as well as cognition and emotion in predictive coding:

> All organisms must integrate cognition, emotion, and motivation to guide action toward valuable (goal) states, as described by active inference. Within this framework, cognition, emotion, and motivation interact through the (Bayesian) fusion of exteroceptive, proprioceptive, and interoceptive signals, the precision-weighting of prediction errors, and the "affective tuning"[191] of neuronal representations. (Pezzulo et al., 2015, p. 37)

On the one hand, there are several kinds of integrations that support goal satisfaction, what means, that the claim, that self-deception is motivated by goal-representations, is actually a very broad one that includes the possibility that several affective influences also play a role. On the other hand, because of the structural-functional entanglement of emotion and cognition, the claim, that those have to be distinguished on the phenomenal level (Pessoa, (2015b), can be questioned too.

[191] Affective tuning means in this context that expectations of valuable states are encoded in more detail – with higher resolution (Pezzulo et al,. 2015, p. 38).

2.2.2 Process: subpersonal hypothesis-testing

> If the skeptic is right, *normal subjects* would generally be *wrong* in attributing to themselves agency in thought. Thoughts are the moment-to-moment expression of an unending process of combination and retrieval; they exploit brain structures, inferential principles and motivations of the system in much the same way as viruses do; they don't engage any 'authorship' of a thinker.
> (Proust, 2013, p. 209)

In this section I will first review two kinds of accounts in terms of which self-deception is usually analyzed: linear simplified model (section 2.2.2.1) and the dual processing theory (section 2.2.2.2). Finally, I will argue that self-deception is best explainable in terms of a dolphin model of cognition (section 2.2.2.3).

2.2.2.1 Balcetis: linear simplified model of motivated cognition

> Each of these approaches [to the explanation of self-deception] is vulnerable to a kind of dilemma argument. Each takes something - some stages or type of processing σ - that is necessary for the knowledge that p, and situates its theory of self-deception in the withdrawal or otherwise thwarting of this processing.
> (Lockie, 2003, p. 136)

Balcetis (2008) gives an elaborate account of conditions that are conducive to self-deception, pieces of which can be found by other authors who writes on self-deception. For example, necessity of mixed evidence is explored by Van Leeuwen (2008, p. 199; see section 3.2.3.1), deliberative/implemental mindset distinction is approached by Taylor & Brown (Van Leeuwen, 2009, p. 116), demand of the plausibility constraint on self-deception is touched by von Hippel & Trivers (2011a, p. 43), self-schema and its relation to self-deception lies in the center of the self-immunization theory of Greve & Wentura (2010; see section 3.1.2.2). Her account, as a simplified *linear* biasing account, will also serve as a contrast to the *dual processing* account presented in the following section.
Emily Balcetis (2008), as well as Taylor (1989), acknowledges a connection between self-enhancement biases and self-deception. In distinction from Taylor, however, who holds that self-enhancement bias is one type of self-deception/illusion, Balcetis claims self-deception to be a mechanism for the maintenance of self-enhancing views:

> The recipe for such self-views calls for a dash of self-deception. To maintain overly and perhaps undeservedly positive impressions of oneself, it is imperative that people remain unaware of the distortions they place on their thinking. (Balcetis, 2008, p. 362)

The definition of self-deception given by Emily Balcetis seems unsatisfying, since it does not specify,

1. whether she holds self-deception to be a conscious or an unconscious phenomenon,
2. justification for why she omits self-deception being a state,
3. the relationship between self-deception and belief/behavior. Moreover, she uses the notion of rationalization, which possesses a lot of intuitive connotations, without defining it:

> Rather than simply being wrong about one's own motives, self-deception is the process of ignoring, rationalizing, or manipulating some thought or behavior to create consistency between that thought or behavior and one's sense of self. (Balcetis, 2008, p. 362)

As opposed to Taylor and von Hippel & Trivers, Balcetis (2008) holds that self-deception possesses a *defensive* and not an offensive function of maintaining a positive self-view[192] and more generally that motivated cognition has a *different route* from non-motivated processing, because of a short amount of time[193] needed by motivated cognition to occur. However, I would question that avoidance motivation could be activated by pushing a joystick away from one's own body which is the case in the empirical research that the author considers[194] (p. 373). Balcetis (2008) acknowledges that the linear four stages model of cognition[195] is a simplification, given that both sequential and parallel processing, as well as top-down and bottom-up constraints should be taken into account (pp. 374-375). Bur for clarity matters she still elaborates the connection between motivation and cognition by using the mentioned four stages model that will be presented below (Balcetis 2008, pp. 364-9). The benefit of using it is to summarize the psychological take on the ways that motivation might bias cognition (see figure 11), in addition to philosophical theories reviewed in the first chapter (1).

[192] Psychological immune system has been seen as the deceiver-element in self-deception: "As its charge is to defend against unhappiness, the psychological immune system may provide the impetus and the resources to transform, manipulate, and cajole self-relevant and social information in order to maintain a positive view of the self." (Balcetis, 2008, p. 363)

[193] Balcetis (2008) presents evidence that amygdala processes not only negative, but also positive and thus, more generally, *goal-directed information* (promotion and prevention focus, p. 372). She claims that basal ganglia (which process positive information) and amygdala (which process goal-directed information) are a complementing dynamic duo which is responsible for motivated cognition.

[194] In the joystick experiment, in which participants pushed the joystick away after seeing a white or a black face, it was the case that difference in activation has been measured to occur in a short amount of time, which leads Balcetis to conclude that "[u]sing electroencephalogram, research has demonstrated motivations bias information processing as quickly as 100 milliseconds after exposure to a threatening stimulus" (Balcetis, 2008, p. 373).

[195] Greenwald (1988, 1992) describes a sequential stage model for an explanation of self-deception by means of a *junk-mail* analogy. The formation of representations starts with the preattentive analysis (feature detection), which happens unconsciously, and is followed by focal attention, comprehension and elaboration reasoning (Greenwald, 1988, pp. 120-121) which are conscious. At the comprehension state, propositions are formed and at the elaboration reasoning stages, inferences are drawn from them (p. 120). A failure to proceed to the next level can happen between any levels, and not only between the third and the fourth, making knowledge avoidance an ordinary phenomenon (p. 123). Thus, comprehension, attention and exposure can also be avoided. Counter-intuitively, Greenwald (1988) states that *thought-stopping* is a viable strategy of executing knowledge avoidance: "the anxiety produced by thinking about high places should be controllable cognitively by not thinking such thoughts" (p. 125). Having applied the stages model to knowledge avoidance, the author concludes that having contradictory beliefs simultaneously is not required for cognitive defense (p. 126). Even if Greenwald's junk mail analogy model is accepted, it is unclear though *why* the processing of information is not taken to the next stage.

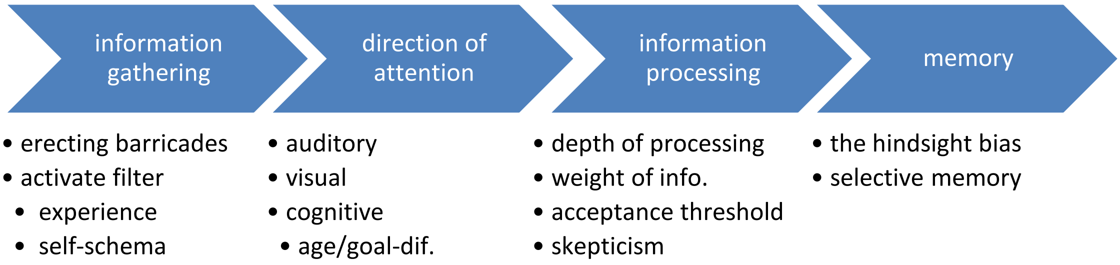


Figure 11. Balcetis: influence of motivation on cognition.
Distinctions from Balcetis (2008).

Humans use different methods for gathering information which can be potentially biasing. They may prevent themselves from gathering some information at all (erecting barricades) if they know that they will not be able to change the negative circumstance in question, or they may activate filters, such as experience or one's self-schema, to categorize the given information (modifying the information in that way):

> In short, abstract traits are defined egocentrically, using the self as a standard of comparison, and then become the metric by which others are judged. The inferences that are then drawn about others are rarely produced in a sterile environment but are instead the product of sifting social information through motivational filters. (Balcetis, 2008, p. 365)

Human beings can also actively control their attention: auditory (cocktail-party phenomenon), visual (the case of binocular rivalry) and cognitive. Cognitive selection of attention is most salient with respect to older people who, due to the goal of emotional well-being, sacrifice negative information in favor of the positive one (Balcetis, 2008, p. 365-366). On the stage of information processing, motivation can influence cognition by changing the depth of processing (whether one is fully attending to information or using only shallow criteria to judge it), the weight of information, acceptance threshold (minimum level of evidence in order to accept some information) and skepticism concerning the given information. Even memory can be susceptible to motivation, giving existence to such phenomena as the hindsight-bias ("I-knew-it-all-along" effect) and presence of selective, goal-dependent memory.

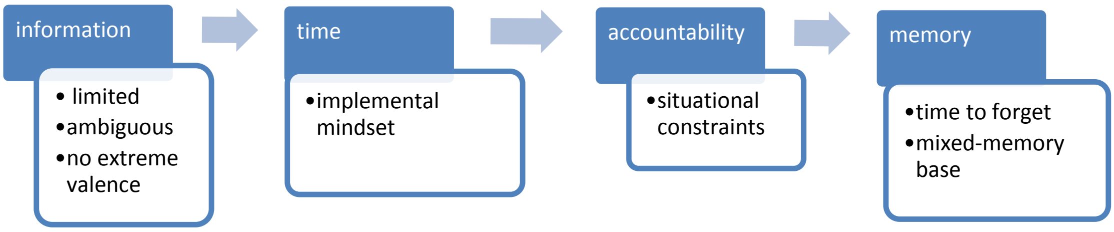


Figure 12. Balcetis: limitation criteria for SD.
Distinctions from Balcetis (2008).

Described effects of motivation on cognition have limitations of information, time, accountability and memory (Balcetis, 2008, p. 367-8; 369-70). For self-deception to occur, information has best to be limited (thus leaving less space for the evidence to the contrary), ambiguous (if both pro- and contra information is available, it is easier to change the weight of it, thus accepting the desirable belief) and without extreme valence. Implemental mindset is more conducive to self-deception than a deliberative one.[196] Before a decision is made one has to analyze different options, so different informational sets will be activated and it could hinder self-deception. After a decision is made and the focus is on the course of action, information is easier to get biased. Situational constraints could make some aspects difficult to ignore and, thus, force the gathering of evidence which could hinder self-deception. Memory, as information, should be best mixed[197] and it should be given some time for negative feedback to fade.[198] The linear model presented comes at the cost of an intuition that the more elaborate the process of negative information has been, the worse:

> Because motivated forgetting requires by definition that the less flattering information be first processed, perhaps it should be considered a backup plan and not a primary system of defense. (Balcetis, 2008, p. 370)

One more point about selective memory mentioned by Balcetis (2008) is that it is most often used with respect to oneself than with respect to others, thus narrowing the topic of self-deception (p. 369). It fits to Taylor's narrowing of the possibility of self-deception to the sphere of social attributes, as well as to Holton's (2001) view that self-deception is about oneself. In section 3.2.2, I will consider Robert Trivers' view that self-deception is about deceiving others, and not about maintaining a positive self-concept.

Interim conclusion: Balcetis' account highlights different *stages* of information processing at which certain biasing strategies may operate. In the following section, I will review dual processing theories that focus on different *levels*, each of which in possessing of a characteristic processing style.

2.2.2.2 Dual processing theories

> Although some conscious and unconscious processes may involve entirely different and separate neural substrates, creating a divided mind, some forms of self-deception might be related to the activation or inhibition of *alternate subroutines* in intimately entwined neural circuits.
> (Surbey, 2004, p. 134)

This section serves three aims: first, as a contrast to the simplified line model discussed in the previous section. Second, it is the basis for Proust's (2015b) theory of feelings, so it will offer more background to the idea of tension being a metacognitive feeling. Third, the shortcomings of the tools, that this theory provides for the explanation of a self-deceptive process, will be listed to be dealt with by dolphin model of cognition in the following section.

The assumption that in self-deception incompatible kinds of information have *both* to be processed suggests an explanation in terms of dual processing theories. Further, the distinctions between implicit and explicit memories, automatic and controlled processes,

[196] Balcetis does not use the distinction deliberative/implemental mindset. Van Leeuwen (2009, p. 116) cites in his article Taylor & Brown's (1989) quotation about this distinction. I hold that what Balcetis means with "before and after decision is made" reflects the deliberative/implemental distinction from Taylor and Brown (1994).

[197] Van Leeuwen also holds that the evidence should be mixed for self-deception to be possible (Van Leeuwen, 2008, p. 199).

[198] A metaphor for the working of motivation as marination: "In order to think well of oneself even in the face of negative feedback, memories need time to marinate in the motivational juices. It takes time to tenderize a tough memory" (Balcetis, 2008, p. 370).

as well as unconscious and conscious ones that have been used by von Hippel & Trivers (2011a,b) to demonstrate the empirical possibility of self-deception,[199] are incorporated by dual-system theories of reasoning as features of either the first or the second system of cognition: system 1 (evolutionary old, unconscious and automatic processes, implicit knowledge) and system 2 (evolutionary recent, conscious and controlled processes, explicit knowledge) (Frankish & Evans, 2009, p. 15). Further, Proust's (2013) theory of epistemic feelings that I applied to self-deceptive phenomenology (see section 2.1.3) is also based on dual processing theory (Proust, 2015). Thus, I will review dual processing theories in this section, but nevertheless argue in the following one that the process of self-deception is better to be explained by the dolphin model of cognition.

I would like to start with two conceptual distinctions: assumptions about the existence of dual *processes* are not to be equated with assumptions about the existence of dual *systems*[200] because of additional ontological assumptions one has to make.[201] One either has to postulate two separate neural systems (Frankish, 2009, p. 96), or one neural and one virtual/emergent system[202] (Frankish & Evan, 2009, Frankish 2009, p. 97). Systems have further been argued to possess their own goal representations, inferential control mechanisms and their own knowledge base[203] (Frankish & Evans, 2009; Frankish, 2009, p. 102). Evidence from neuroimaging studies, though, is argued not to support the distinction between two systems in terms of how old/new the brain regions are that each system employs (Evans, 2009, p. 41). The recently presented network perspective on brain function also speaks against such oversimplistic dichotomization (Pessoa, 2015a). According to it, it is a certain *network* and not *brain area* that underlies certain processes that are *emergent* properties of networks. Possibly *overlapping* networks should be seen as units of interest in discussing the interaction between emotion, motivation, perception and cognition and emphasis should be given to the *dynamics* of the relationships between these networks

[199] The reasons for the employment of these paradigms in explanations of self-deception is that they are taken to explain how contradictory behavior of the self-deceiver can be explained, namely by positing contradictory representations in his mind that, independent of its accessibility to the self-deceiver, *trigger* goal directed action: "as the conscious memories could be those that are consistent with the fiction that the person wishes to promulgate, whereas the unconscious memories could be the facts as originally encountered. By maintaining accurate information in unconscious memory, the individual would *retain the ability to behave in accordance with the truth*" (von Hippel & Trivers 2011b, p. 6; my emphasis).

[200] Frankish & Evans (2009) make the following distinction between dual-process and dual-system theories: "As already observed, dual-process theories distinguish fast, automatic (type 1) processes from slow, deliberative (type 2) processes. Dual-*system* theories attribute the origin of these processes to two distinct cognitive systems" (p. 15).

[201] Frankish & Evans (2009) argue that it is the *neuropsychological* evidence that offers a decisive argument in favor of the system view, while the processing view is itself supported by evolutionary arguments (p. 12).

[202] A notion that is used, according to the authors, by Dennett to characterize the conscious mind (Frankish & Evans, 2009, p. 21). Neural activity has been argued to provide evidence in favor of the emergent view: "If type 2 processing is an emergent property of type 1 systems, then we should not expect a switch to wholly distinct neural areas when this kind of thinking is activated" (p. 22).

[203] The *characteristics* of system 1 have been argued to be the following: "System 1 is assumed to have its own propriety knowledge base and *goal structure*, formed by routine belief-forming and desire-forming mechanisms, in response to perceptual information, bodily needs, and so on" (Frankish & Evans, 2009, p. 7; my emphasis). According to them, these goals may be "genetically determined," but no repression is involved (p. 7). System 1 is further said to possess "a set of inferential mechanisms for the control of various aspects of everyday behavior, and has a rich representational structure," but it is not a "reasoning system" (Frankish & Evans, 2009, p. 7). It has a direct and beneficial role insofar as "its outputs are directed outcomes of its goals" (Frankish & Evans, 2009, p. 7).

(Pessoa, 2015). Pessoa's arguments undermine Samuels' (2009) argumentation that the properties, by which each of the systems is described, build *clusters* (e.g., the properties of system 1 – associative, heuristic, unconscious and others belonging to this system – build a cluster, cf. 131) and that it is *inference to the best explanation* that if logically independent properties build a cluster that there exist underlying *mechanisms/systems* that explain the co-variation of these properties (Samuels, 2009, p. 131). Further, if the distinction between system 1 and system 2 is reducible to the one between *subpersonal and personal* states,[204] because "subpersonal reasoning is typically fast, automatic, effortless, and non-conscious" (Frankish, 2009, p. 96), then it is redundant and in the light of additional ontological assumptions one has to make and as such not parsimonious.

Generally, there is a trend of "backing off from definitions of the two systems in terms of the processing styles involved – heuristic and associative on the one hand, analytic and rule-governed on the other" and hypothesize that "it is likely that future two-*systems* theories will need to posit multiple kinds of cognitive *processing*" (Frankish & Evans, 2009, p. 23). The minimal definition of dual processing types[205] is similar to that of systems in terms of characteristics that are ascribed:

- type 1 processes are characterized as "fast, automatic, high processing capacity, low effort" and
- type 2 respectively as "slow, controlled, limited capacity, high effort" (Evans, 2009, p. 33).

Those two types of processes are argued to either occur in parallel, or in a sequence one after the other[206] (Evans, 2009, p. 43):

- *parallel-competitive*: Two kinds of processing are assumed to happen simultaneously and compete with each other, e.g. implicit (type 1) vs. explicit (type 2) knowledge activation (Evans 2009, pp. 43-45).
- *default-interventionist*: Two kinds of processing build a sequence, where the second kind of processing might correct the first, default one, e.g. heuristic (type 1) vs. analytic (type 2)

[204] Frankish (2009) holds that subpersonal/personal processes can be mapped onto the unconscious/conscious distinction: "In the case of mental *processes* at least, the distinction between personal and subpersonal corresponds roughly with that between conscious and non-conscious. We are typically conscious of our personal mental processes, but not of our subpersonal ones" (Frankish, 2009, p. 91).

[205] Nonetheless, Evans (2009) argues that an appropriate kind of description is also that in terms of a "two minds hypothesis:" "Here I define 'mind' as a high-level cognitive system capable of representing the external world and acting upon it in order to serve the goals of the organism" (p. 35).

[206] Similarly to Evans' (2009) distinction between parallel-competitive and default-interventionist interactions, De Neys et al. (2008) distinguish between two kinds of models of reasoning in which heuristic and probabilistic reasoning interact with each other:

- *bias-as-detection-failure* which states that reasoning process is of a *heuristic kind* by default and that conflict-monitoring processes fail, thus there is a failure to detect the inconsistency between the favored heuristic conclusion and the conclusion that would stem from probabilistic reasoning;
- *bias-as-inhibition-failure* which states that the reasoning process is divided into heuristic and probabilistic which occur in *parallel,* conflict-monitoring process is intact, but there is an inhibition failure or a failure to inhibit the heuristic response.

De Neys et al. (2008) argue on the neural basis of conflict detection (anterior cingulated cortext, ACC) and response inhibition (right lateral prefrontal cortex, RLPFC) in favor of the second model. Yet neural data does not tell us anything about the conscious level. De Neys, Vartanian & Goel (2008) use personal-level terminology and even suggest that there might be an affective component to inhibition response, e.g. regret:

"One might wonder whether the inhibition failure also has an affective component (e.g., do people 'regret' their stereotype-based response after an inhibition failure?). Our data do not speak to this issue, but as one reviewer noted, possible affective reactions might be linked to cases of 'weakness of will'." (De Neys et al., 2008, p. 488)

> processing (Evans 2009, pp. 45-46).[207] Here heuristic processing is understood as the "*implicit* processing of *explicit* knowledge" (Evans 2009, p. 45) and the analytic one is argued to require working memory[208] (p. 43) "[a]nd in so far as the fleeting contents of working memory are conscious, *something* about the working of a type 2 system will become conscious." (Evans, 2009, p. 38).

In the case that type 1 and type 2 processing conflict, it is a third type of processing that has to take over for "resource allocation, conflict resolution, and ultimate control of behavior" (p. 48), but which is again not necessarily conscious, e.g. switching attention to the road if during the process of driving a hazard is met (p. 48). The latter fact, namely that dual system theories do not clarify how the property of consciousness is distributed over the systems – an analog point can be made about processes - is criticized by Price & Norman (2008, p. 37). They argue that *intuitive feelings* are an "interface between non-conscious and conscious processes" (p. 37) such that upon focusing on these feelings, the non-conscious processes, that have produced them, might be able to be brought into consciousness (p. 33). They propose to situate mental phenomena in a multidimensional space (see figure 13). It should be noted that transparency as unavailability of earlier processing stages (Metzinger, 2003) is one possible characteristic of intuitive feelings, given the definition that "[a]n emotional feeling would be *intuitive* to the extent that there was a conscious summary signal of some kind, but limited access to the antecedents of that signal" (Price & Norman, 2008, p. 38).

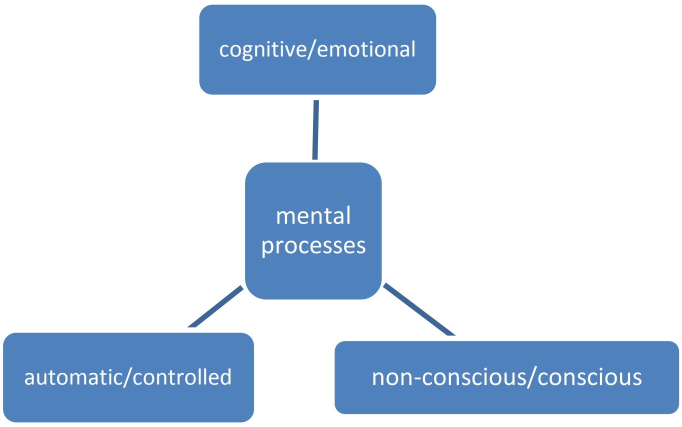


Figure 13. Price & Norman: multidimensional space of mental processes
Distinctions from Price & Norman (2008, p. 38).

As Price & Norman (2008), Thompson (2009) has also proposed that system 1 processes generate certain feelings. They generate the FOR (feeling of rightness) – an affective response about the heuristic output of system 1. A FOR is taken as input for the JOR – a

[207] I assume that it is this interpretation of dual processing theories that was criticized by Andersen et al. (2007). Andersen et al. (2007), in an overview article on the concept of automatic thought in psychology, characterize dual-process models as those in which automatic processes temporally precede controlled ones (p. 141) and criticize them insofar as "the simple temporal progression from automatic to controlled does not capture all that occurs" (p. 142).

[208] Contents of working memory are argued to be supplied by implicit cognitive systems: "But working memory does nothing on its own. It requires, at the very least, *content*. And this content is supplied by a whole host of *implicit* cognitive systems. For example, the contents of our consciousness include visual and other perceptual representations of the world, extracted meanings of linguistic discourse, episodic memories, and retrieved beliefs of relevance to the current context, and so on. So if there is a new mind, distinct from the old, it does not operate entirely or even mostly by type 2 processes. On the contrary, it functions mostly by type 1 processes." (Evans, 2009, p. 37)

meta-cognitive judgment about the FOR and either FOR or JOR determine whether system 2 intervention will take place. Interestingly, Thompson postulate that a minimal awareness of FOR as "a vague unease about the heuristic output" is on the personal level enough to trigger subsequent system 2 processing[209] (p. 183). The joint results of Frankish' (2009), Evans' (2009), Price & Norman's (2008) and Thompson's (2009) accounts are that (1) control is more often than not ascribed to processes that might become conscious, (2) feelings are the means that indicate the need for control in virtue of (3) certain (subpersonal) kinds of processes not being available. Curiously, Evans' (2009) arguments against the claim that feelings are *necessarily* the output of type 1 processes resemble the debate between intentionalists claiming that tension is necessary and deflationary positionists, arguing otherwise:

> This [type 1 processes are followed by an intuition and a metacognitive feeling] is what Sloman has in mind with his criterion S when he talks of people believing in two contradictory responses simultaneously and *experiencing a conflict* between them. It is true that we sometimes become aware of dual-process conflict, as when a man who wishes to visit his sick mother is unable to cope with his flying phobia, or when a compulsive gambler cannot overcome her habit by rational reasoning about chance and probability. In general, however, I think that Criterion S is too strong for many applications of parallel-form dual-process theories. For example, there is much evidence of automatic and nonconscious processing in social cognition (Bargh 2006; Forgas et al. 2003; Wilson 2002) where people do not generally seem to be aware of conflict. *If people respond according to an implicit attitude or implicit stereotype, they must generally lack awareness that these behaviors are in conflict with their explicit attitudes and beliefs in order to maintain their core beliefs about themselves.* (Evans 2009, p. 44; my emphasis)

Price & Norman's (2008) claim that feelings, generated by type 1 processes, are intuitive in virtue of their *transparency*, resembles the features of the second phenomenological characteristic of self-deception, namely *insight* (see section 2.1.2). I argued that during insight a self-deceptive context is suddenly understood differently. This may be the case particularly because the kinds of processes (I will not restrict myself here to suppose that there is a certain type 1 process) that generated it lose their quality of transparency.

Summing up the review, according to dual process theories, there is one process that is phenomenally available (one epistemic agent model has been constructed) and another one that is not phenomenally available, but such that certain feelings arise if the results of the two processes diverge. The purpose of those feelings is to provide better personal level *control*. I argue that this kind of simplification – two processes in parallel, or serially one after another controlling each other - doesn't satisfy to the phenomenology of the self-deceiver's process and probably not only it. Vlassova et al. (2014) conducted a series of perceptual experiments (different stimuli were presented to each eye such that one stimulus was visible and the other occluded and when the presentation of the occluder stopped participants had to make a perceptual choice about the target stimulus) indicating that an unconscious process has to be followed by a conscious one to increase accuracy, but that this increased accuracy is not reflected at the meta-level as confidence. They then assume

[209] In the cases where system 1 (a heuristic system) does not produce a solution accompanied by strong FOR and JOR, the Judgment of Solvability (JOS) determines according to Thompson (2009) the "type and extent of analytic engagement" (pp. 183-184). JOS is determined by the assessment of the difficulty, as well as goals and motivations (p. 184). The FOR and JOR proposal for reasoning is drawn in analogy to the phenomenal experiences that have been analyzed in memory research: Feeling of Familiarity (FOF), Feeling of Knowing (FOK), Judgment of Learning (JOL) (Thompson, 2009, p. 175).

that "it is possible that the unconscious information is processed but requires conscious information to *bind* to for us to be able to access and use it" (p. 16216; my emphasis). The formal reason for dual process theory not being enough for an explanation of self-deception is that since per definition a self-deceptive process cannot possess an epistemic agent model in which epistemic actions are directed at self-deceiving, the aims of an epistemic agent model that the self-deceiver currently has and other kinds of belief-forming processes without an epistemic agent model that she might at the same time possess will diverge. Further, the dual process model assumes that an epistemic agent model is restricted to *one* process, but if one thinks about such cases as mind wandering in which the boundary between it and epistemic agent models that might precede or follow is not phenomenally available (Metzinger, 2013a), then there is no reason to leave out an option that multiple processes might bind into each other so that those are experienced as *one* epistemic agent model. That attitudes might "pop-out" without an epistemic agent model for their acquisition having been available and without an agent's memory that it was not available is another reason for the need to extend the dual process model. As for the control function of feelings, I have already noted in section 2.1.3 that in case of self-deception the kind of control they provide fails to be that of leading to truth-acquisition and instead it might be a subpersonal kind of control that leads to self-termination (tension-loaded self-deception becoming tension-free in virtue of feelings changing the subsequent hypothesis testing cycles). In the latter case the phenomenal aspect of these feelings might be seen as an epiphenomenon, as argued by Dokic.

2.2.2.3 Dolphin model of cognition[210]

> Even when we're aware of our *thoughts*, we're not aware of the process of *thinking*.
> Jackendoff (2012, p. 212)

In previous two section two simplified models of cognition have been introduced: the linear and the dual process one. Those will serve as a contrasting background to the dolphin model to be introduced in this section. Particularly, I have argued that the dual processing model, which is quite popular in the psychological literature, does not do justice to the phenomenology of cognitive processes, particularly to the epistemic agent models we experience. In case of self-deception in particular and cognition in general, the possibility that different cognitive processes *bind* on the phenomenal surface has to be considered. The structure of this section will be as follows: First, I will introduce the concept of 'binding' and how it has been used in the literature. Second, I will argue that binding is plausible to assume as a binding of *processes*, as well as the binding of *representations* that the latter operate with. Third, I will describe the phenomenological binding that does justice to the phenomenology of self-deceivers. Fourth, I will indicate how binding might be understood in the predictive coding framework.

First, the term 'binding' has been applied to a variety of problems, e.g. consciousness-related binding as "a problem of finding the neural mechanisms which map the unified contents in phenomenal consciousness to corresponding neural entities in the brain" (Revonsuo, 1999, p. 175). Cognitive binding, modelled as strengthening the weights in a connectionist network, has also been proposed as a solution for the dependencies in how good participants solved tasks measuring explicit memory, e.g. recognition and recall (Metcalfe et al., 1992). Hippocampus has been proposed as a neural structure whose

[210] This section is a modified and extended version of section 4.2 in Pliushch & Metzinger (2015).

function is to “rapidly, continuously, and obligatorily form associations among disparate elements across space and time, and further, to enable the comparison of internal representations with current perceptual input” (Olsen et al., 2012, p. 2). Note that there is not only a distinction between *kinds* of binding problems (see Treisman, 1996 on that), but also *levels* of description for these problems, those levels being at least the phenomenal, cognitive (level of modules) and neural (Revonsuo, 1999, p. 176). Cognitive and neural levels are regarded as offering descriptions of the mechanisms (p. 177). There is also a term ‘cognitive unbinding’ that refers to “impaired synthesis of functionally specialized cognitive modules in the brain, which is posited to interrupt conscious representation” (Mashour, 2013, p. 2752). For this author, the relationship between cognitive binding and unbinding is such that the first is necessary for consciousness, while the second – sufficient for unconsciousness (Mashour, 2013, p. 2753). Cognitive unbinding is regarded as playing a role in the exclusion of certain *contents* from consciousness, enabling repression and dynamic unconscious (Mashour, 2013, 2757). Recently, a thesis has been defended that feature binding is an ill-defined problem in that there is no need for a separate feature-binding process, just a hierarchical comparison of a high-level perceptual hypothesis to lower-level activity is enough (Di Lollo, 2012). This description reminds that of predictive coding (apart of the critique of attention as a useful explanatory tool because of the missing exact description of the mechanism by which attention might ‘glue together’; Di Lollo, 2012, p. 318). Thus, according to Di Lollo (2012, p. 318), such a hierarchical comparison and binding are rival explanations. This idea may stem from the intuition that no top-down control is necessary for binding (for the acceptance of the latter thesis see Olsen et al., 2012, p. 10). From a different point of view, recently a hypothesis has been defended that cognitive control results from arousal-influenced binding processes (Verguts & Notebaert, 2009). The binding process is here understood again as changing weights in a connectionist network and arousal is argued to increase in volatile environments (Verguts & Notebaert, 2009, p. 255). Tschacher & Bergomi (2011), in difference to the rivalry hypothesis, argue that “optimal inferences about the causes of sensory input (ie, binding) may be impaired in schizophrenia by a failure to optimally modulate synaptic gain, resulting in the abnormal *intersensory binding* reported in this study” (p. S20; my emphasis). To sort the use of the term ‘cognitive binding’ in this thesis according to the distinctions just mentioned, cognitive binding problem posed here is a *phenomenal* unity of higher-order cognitive processes and its scope is not consciousness in general, or explicit memory, but processes underlying our experienced strands of thoughts. As for the relationship between binding and inference, I will come to this at the end of the section. It should be noted here already that control by arousal-induced binding may be one candidate for an explanation of how tension-loaded self-deception becomes tension-free.

Let me consider the second point: How exactly should one conceive of binding of different processes on the functional and representational level? One possibility is that there is a functional overlap. As elaborated in the previous section, postulation of different processes has been sometimes accompanied by the postulation of respective systems. Thus, a mapping has been drawn from a functional aspect of the existence of a process with certain attributes to the respective brain area that manages this kind of processing. A recent model of interactions between cognition and emotion (Pessoa, 2015) emphasizes that a certain process may involve the interaction of different overlapping networks (anatomical regions), each of which is characterized by structural and functional constraints. Pessoa’s emphasis is on the multi-functionality of *brain regions* which the author holds to possess a *functional fingerprint* (= family of tasks it is involved in). Pearson et al. (2014) also claim that neurons are functionally heterogeneous. If function is ascribed to *overlapping* networks, is it still valid to view a certain kind of processing as possessing certain kinds of knowledge base, computing only certain kinds of representations or having access only to certain kinds of representations (those are dual processing assumptions that are also *modularity* assumptions, see section 3.2.1), given that more than one function can be executed at a time?

The processes might bind into each other by means of changing its interaction with each other, [211] or by binding together the representations[212] that they manipulate. As Mahon & Caramazza (2008) put it, one should avoid the conflation of "claims about the dynamics of activation flow with claims about the structure and format of representations" (p. 67). The binding of representations has already been proposed in the context of explaining *spontaneous confabulation*:

> Perhaps the spontaneous confabulator sometimes has very patchy memory experiences which come to mind in a piecemeal kind of way and, while endorsed (as received delusional primitives), trigger both reflection and further memory search. Consider, for example, delusions of alien abduction. […] That the hypothesis successfully "binds together" the individual's memory fragments gives it an initial hold […] which – with time, further reflection, further encoding, and further retrieval – comes to be recalled as a fully-fleshed-out memory experience of having been abducted by aliens. (Langdon & Bayne, 2010, p. 338)

The question about the decomposition of a cognitive process is to be distinguished from questions about the format of mental representations or the modularity of the processes involved, though the former topic would benefit from answers to the following two questions. The question about the format encompasses more complicated subquestions: Is there a global representational format? If yes, what is the nature of this format – symbolic or subsymbolic? If not, how do different representational formats interact? These, as well as other questions are posed by Anderson (2003) in an article on embodied cognition. The question about the representational format and its binding has a straightforward application to self-deception. In the previous section on dual processing I have mentioned that the implicit-explicit distinction has been used in the psychological literature to solve the disunity in unity constraint for self-deception (see section 3.2.2.3). It is, namely, argued that the disunity is between the explicit and implicit cognitive processes. A critique of such a view might be that in this case there is no unity at all, since there are no, at least rational, constraints on the unification between these two processing levels: no personal level recognition and, hence, no self-deception in a challenging sense if personal level recognition is used as a measure of goodness (see chapter 1). But in any case, if the implicit-explicit distinction were to play a role in self-deception, the question about the binding of two different kinds of representations (an explicit and an implicit) would be important (Anderson, 2003, p. 121). More generally, it is not only two kinds of *cognitive* processes that might simultaneously access and change a representation, but also a cognitive and a

[211] A process might be distinguished from another one on the basis of the goal representation it attempts to satisfy, in reasoning this is truth-acquisition: "Here a subject cannot want to remember or perceive a given content because it is a condition of satisfaction of the corresponding mental act that it responds to truth or validity, rather than to the thinker's preferences" (Proust, 2013, p. 151).

[212] Binding of representations dependes on the representational format, because the representational format constraints the operations possible on these kinds of representations. Cognitive representations have been recently hypothesized to represent intrinsic, inter-dimensional contraints (Waskan, 2006, p. 187) or a region of a conceptual space where the latter is a set of quality dimensions (Gärdenfors, 2004).

motor/affective process. The latter, with respect to self-deception, is important not only for the epistemic agent models generated, but also for the clarification of tension and its role in upholding self-deception. Proponents of embodied cognition argue that cognition is, first and foremost, instantiated by means of *modal* representations on which sensory/motor systems operate (Barsalou, 2008). Opponents object that empirical evidence (e.g., activity of sensory/motor areas during conceptual tasks) is compatible with the claim that there is an *interface* between modal and amodal representations (Mahon & Caramazza, 2008, p. 60). A third option is that the modal representations belonging to the same concept[213] also need to be integrated for which task there might be a "conceptual hub" (Kiefer & Pulvermüller, 2012, p. 819). I will leave the question open in how far an interface and a conceptual hub are similar to each other (I assume that the former suggests a tighter form of interrelations, akin to connectionist ideas, in comparison to the latter), because my general point is that neither processes nor representations are to be seen as rigid. Some binding might take place in the former, as well as in the latter case and how exactly this happens on the functional level is an empirical question that has straightforward implications for an explanation of self-deception not only on the level of phenomenology, but also regarding the question of mechanisms of content fixation and their difference in self-deceivers. In section 1.3.4 I have criticized experiments testing self-deception that instead of self-deception what is found is the difference in content fixation between self-deceivers and experimenters. Mahon & Caramazza (2008) argue that the interactions between modal and amodal representations might be such that there is "a decision mechanism that promiscuously uses information that, on a logical analysis of the task requirements, would not be necessary" (p. 65). For experimenters testing self-deception it would be important to consider the possibility of such mechanisms changing content fixation to kick in. At the end of the following section I will summarize my suggestions for experiments.

The third and central point of this section is the description of a phenomenological kind of binding that I hold to take place in self-deception. Before that, let me comment on two conceptual tools that will be used: "mental autonomy" (M-autonomy) and the notion of an "epistemic agent model" (EAM; Metzinger, 2013a,b). M-autonomy is the ability to control the conscious contents of one's mind in a goal-directed manner and functionally presupposes the ability to suspend or terminate cognitive processes at will, by means of attentional or cognitive agency. Cognitive processing *becomes* a personal-level phenomenon by being functionally integrated into an EAM, which is a transparent conscious self-representation of possessing the capacity for epistemic agency and/or actually executing epistemic actions, where epistemic agency involves attentional agency (the ability to selectively focus one's attention) and cognitive agency (the ability to control goal-directed thought). Metzinger (2013a,b) applies these conceptual tools to the phenomena of dreaming and of mind wandering, and Pliushch & Metzinger (2015) apply them to self-deception. The involvement of belief-forming processes is a characteristic of self-deception that differentiates it from cases in which goal representations influence the behavior directly. As Donald Davidson (1986) puts it, self-deception is a 'weakness of the warrant,' to be differentiated from the weakness of the will.

[213] Apart from the feature of modal specificity/idependence, concepts have been characterized as flexible (as opposed to stable or context-independent), experience-dependent (as opposed to innate), distributed across sensory and motor systems (as opposed to local) (Kiefer & Pulvermüller, 2012, p. 817).

Mind wandering[214] is a phenomenally represented, but still subpersonal form of cognitive processing (Metzinger 2013b). Similarities between mind wandering and self-deception consist in its goal-directedness[215] (on the premise that mind wandering concerns goal representations other than the current ones, Schooler et al. 2011, pp. 323 – 324) disengagement of attention from perception, restricted meta-awareness and it being a "*subpersonal* process that functionally results from a cyclically recurring *loss of M-autonomy*" (Metzinger, 2013b, p. 15; my emphasis). The self-deceptive belief-forming process is subpersonal cognitive processing as well, yet it is an open question whether it is always phenomenally represented (see section 2.2.1). Moreover, this question is independent of the one of whether tension is a necessary phenomenological component in self-deception: "I feel discomfort and identify its origin in me possessing certain representations" does not imply "I am an epistemic agent engaged in cognitive processing and I feel tension *because* I feel that something is wrong with the latter."

Taking for granted that at least some self-deceptive subpersonal belief-forming processes are phenomenally represented, let me draw the analogy between these instances and mind-wandering (Pliushch & Metzinger, 2015). If subpersonal cognitive processing is like a group of dolphins traveling below the surface, then mind wandering is like an absent-minded tourist passively monitoring a single dolphin's path. This happens in the absence of an EAM, whereas normal, as well as self-deceived, reasoning is like tracking an individual animal's leaps plus an experience of being able to control its trajectory, perhaps even involving the phenomenology of identification. What this analogy does not capture, however, is that we do not experience reasoning as singled out splashes, but as a connected whole. Just as real-life dolphins are often still visible below the surface or even partly in the water and partly above it, and therefore effortlessly perceived as one and the same dolphin over time, cognitive processing on the level of the EAM is experienced as stringent and internally connected. What conscious reasoning shares with the blind spot in vision and inattentional blindness in perception is the common characteristic that though only parts come to consciousness, we experience the overall process as complete and without gaps ('cognitive binding', Pliushch & Metzinger, 2015). The continuous process of integrating cognitive processing into an EAM may result in a "glossing over" of temporal discontinuities. To understand how cognitive binding is possible, one has to precisize the format of thought on the phenomenal level. Siewert (1998), for example, argues that apart from image-like thoughts and verbalized thoughts, there might be *wordless thoughts* or those that we notice by just noticing their content.[216] Wordless thoughts lead to "variations

[214] Core processes of mind wandering: "Mind wandering (i.e. engaging in cognitions unrelated to the current demands of the external environment) reflects the cyclic activity of two core processes: the capacity to disengage attention from perception (known as *perceptual decoupling*) and the ability to take explicit note of the current contents of consciousness (known as *meta-awareness*)" (Schooler et al., 2011, p. 319; my emphasis).

[215] When subjects had to rate whether adjectives described themselves (about *one's personality*), this increased the subsequent frequency of self-caught episodes of mind wandering (Chin et al., 2012, p. 82). Further, this frequency can be manipulated by using the *thought suppression paradigm* (pp. 82-83; my emphasis).

[216] Example for wordless thoughts: "Cases of sudden wordless thoughts that, like the examples I have just chosen, involve a sudden surge of anxiety, may be relatively easy to notice and recall, but we can take note of the occurrence of such noniconic thoughts even where they involve no such marked change in one's emotions. Suppose you are riding a bicycle, approaching a green traffic light far ahead, and it suddenly occurs to you that the light will change soon, and you wonder whether to speed up to make the light – but this experience involves no remarkable change in your emotional state. You may have known this thought – that the light's about to

in the phenomenal character of thought, distinct from differences in the phenomenal character of imagery and sense-experience” (Siewert, 1998, p. 281).

Let me elaborate cognitive binding in more detail. Keeping in mind the three levels of the reasoning process (unconscious, conscious and uncontrolled, conscious and controlled), the mind wandering process is an uncontrolled one that is at certain points coupled with the phenomenology of control due to a transient integration into the EAM. The general point about conscious cognition, which, however, has to be accounted for in explanations of self-deception, is that even when we have the phenomenology of a controlled reasoning process, it can arise from the conscious experience of only some *parts* of the subpersonal reasoning process being currently integrated into an EAM, which themselves may be non-veridical or appear outright hallucinatory when detected from a third-person stance. We may experience ourselves as epistemic agents while we are not. In philosophical and psychological models of reasoning, introspectively accessible parts that are specified as conscious, and those containing biases that might or might not be conscious are often combined without specifying how exactly conscious and unconscious parts are dynamically integrated into a holistic PSM. This problem is different from Bermúdez’ (2000a) interface problem, insofar as this is not the problem of how to describe the *transition* from subpersonal to personal-level states, but how to explain their interaction in creating a unified experience of the EAM in general and a unified experience of a continuous reasoning process in particular.[217]

Cognitive binding is a different possibility from those so far considered in the literature about the relationship between conscious and unconscious belief-forming processes. For instance, Evans (2009), a proponent of dual-system theories of reasoning, considers the two possibilities to be either that an unconscious reasoning process is tracked, monitored, and corrected by the conscious one or that both independently run in parallel. Yet both models leave unaccounted the possibility that the reasoning process, as consciously experienced and controlled, is the result of parts of the unconscious reasoning process being sporadically integrated into the unified EAM. Cognitive binding should also be differentiated from the “illusion of objectivity” or the control exercised by the agent over the reasoning process to appear unbiased (e.g., Pyszczynski et al., 1999, p. 839).

There are two inferences to be drawn from cognitive binding. First, verbal reports about participant’s reasoning processes generated in the context of empirical studies may actually refer to the introspectively available results of cognitive binding. Second, the need to justify the mental representations derived from reasoning processes that involve cognitive binding will lead to attempts of justification that, by necessity, can at best be only partially correct, because they must rely on what is introspectively available. In principle it would be possible that one dolphin changes positions with another without “you,” the tourist, noticing this fact. The tourist is the EAM, and like any model it can always be false or incomplete.

Fourth, let me conclude this section by commenting on inference, binding and their importance for predictive coding as a model of perception and cognition that I want to use for self-deception. Predictive coding explains perception, cognition and action in terms of *subpersonal* inference. This kind of inference is to be distinguished from the one that we might experience when an epistemic agent model of a *reasoning* is formed, namely a

change – occurred to you and that it is conscious, though you did not utter to yourself the words, ‘The light is about to change,’ or visualize the light changing.” (Siewert, 1998, p. 277)

[217] Klinger (2013) in the discussion of mind-wandering hypothesizes that “mental content continually jumps from goal-related topic to goal-related topic in brief segments that may or may not return to the same topic as previous segments” (p. 4).

personal kind of inference by which self-deception is often tried to be explained (see chapter 1). So, explanations of higher-order cognition using predictive coding (see section 4.3) should avoid the equivocation of the concepts of "reasoning" and "inference." The hierarchical inference of causes using prediction errors is not to be confused with the personal-level notion of inference. Section 4.2.2 will be devoted to the discussion of this topic. To illustrate this point, Blokpoel et al. (2012) argue that it may be difficult to explain higher-order cognition with predictive coding, because of the complexity of causal models at higher levels required for Bayesian inference to take place. They make clear their point by citing the Dr. Jekyll and Mr. Hyde thought experiment discussed by Kilner et al. (2007) from Jacob & Jeannerod (2005): Imagine that you are observing the identical movements of a person that, in one case, is Dr. Jekyll having an intention to cure and in another case is Mr. Hyde having an intention to hurt. The difficulty is to distinguish whether this person performing the same movements has the intention to cure or to hurt. Kilner et al. (2007) argue that it is possible to distinguish these two intentions due to the dependency of prediction error on context. Yet it is not the case that a personal-level inference is responsible for the given discrimination. Thus, it does not follow from predictive coding that the conscious reasoning process, as well as the EAM, involves Bayesian reasoning on the level of its content, even if both arise out of subpersonal process approximating Bayes-optimality.

As for cognitive binding and inference, a natural conclusion from the fact that predictive coding employs the same explanatory schema – the one operating in terms of hierarchical subpersonal inference – is that cognitive binding, if there is such, is also to be explained by means of it. Yet as indicated above, some see binding and inference as rivaling explanations. I think that if predictive coding can explain the *temporal dynamics* of our cognitive and affective processes, then it can also explain cognitive binding as I defined it (see chapter 4). Self-deceptive belief-forming processes are highly dynamic and context-sensitive, so there will never be a static resulting mental representation with a fixed and determinate content. I have noted in section 2.1.3 that *transitions*, as well as the *instability* of the hypotheses about the world and ourselves is one of the key points in explaining such dynamics. Interestingly, itinerant fluctuations of ongoing brain activity that provide the "brain's internal context for processing external information and generating behavior" have been argued to aid in explaining mind-wandering[218] (Sadaghiani et al., 2010, p. 12).

2.2.3 Properties of self-deceptive misrepresentations[219]

This section concludes the building blocks for the self-deceptive explanation – motivation, process and resulting attitude. I have so far argued that multiple subpersonal goal representations provide the motivation for self-deception, that two central subpersonal kinds of selection responsible for self-deception can be distinguished phenomenally on the basis of them being experienced as attitudes, or as parts of the world/self-model and that the dolphin model of cognition best describes the epistemic agent models generated during self-deception: In virtue of cognitive binding, the self-deceiver does not notice how certain

[218] Sadaghiani et al. (2010) argue that the itinerant fluctuations provide a link to mind wandering: "Itinerant fluctuations of this [ongoing brain] activity reflect the dynamic nature of the underlying internal model that does not remain locked in a stationary mode but remains malleable by continuously exploring hypotheses regarding future experience and action" (p. 10).

[219] This section is a modified and extended version of section 4.3 in Pliushch & Metzinger (2015).

attitudes 'pop-out' undefended, or how an epistemic agent model changes. In this section, I will first describe the properties of self-deceptive attitudes and then come back to experiments testing self-deception and suggest how they could be improved in virtue of the results of the second chapter.

I argue that the following properties of self-deceptive phenomenal representations can vary (see also Pliushch & Metzinger, 2015):

1. The degree to which they are *conscious* (e.g., Demos, 1960, Billon, 2011, von Hippel & Trivers, 2011a,b),
2. The phenomenology of *certainty*, i.e, the phenomenal experience of their subjective epistemic status (e.g., Lynch, 2012), and
3. Phenomenal *transparency* (realness and mind-independence).

As for the degree of consciousness, it cannot offer a satisfying explanation of self-deception, at least not on its own, because it cannot satisfactorily explain *all* the explananda. This is because a disunity by unity constraint explained by degrees of consciousness of different attitudes as determinants of behavior is only very weakly satisfied: there is no unity at all between conscious and unconscious processes. If behavior is determined by unconscious attitudes that would not even be justified, if questioned by others, then this would be a very weak case of self-deception – only one explanandum explained, namely inconsistency, which in this case is behavioral. I assume that there is no tension between conscious and unconscious attitudes, because for tension too, some unity would have to be violated and unity can be assumed only between attitudes on the personal level. So, even if one were to argue that something can be called 'self-deception' even if only some behavioral and phenomenological characteristics are fulfilled and not all, the more are fulfilled, the stronger the case for self-deception would become.

As for uncertainty of attitudes, it has been argued to explain behavioral inconsistency and tension of self-deceivers (section 1.2.2). I noted that there is a vicious circle of anxiety in that if some inconsistency triggers anxiety then anxiety itself should trigger corrective measures (section 1.1.2.1). This is the function of metacognitive feelings (see sections 2.1.3 and 2.2.2.2). I then asked which kind of uncertainty actually varies – psychological or epistemic and argued that in the self-deceptive literature so far a causal connection between epistemic and psychological certainty has been assumed and both would vary simultaneously (section 1.2.5). In short, the assumption would be that if one wanted to feel certain (psychological certainty) in one's epistemic certainty (namely that the evidence warrants a certain kind of attitude), then one would need to bias the evidence, since on the personal level we are constrained by the rules of logic and rationality. But what if our epistemic grounds for believing something became transparent? This has been argued to be the case for intuitions (Metzinger & Windt, 2014) and I have transferred this claim to self-deceptive attitudes: If psychological and epistemic certainty were to diverge and the phenomenal signature of knowledge were to become transparent, then self-deceivers would have the feeling that their self-deceptive attitudes are correct and justify them ad hoc, if needed, but this would be no actual arguments for us, observers, to accept their self-deception (section 2.1.1). I think that the following experiment demonstrates a case in which the phenomenal signature of knowing becomes transparent through repeated explanation giving. Khemlani & Johnson-Laird (2012) conducted experiments in which participants had to detect logical inconsistencies. Their results suggest that the additional fact that participants actively construct explanations of the inconsistencies makes it harder for participants to detect those afterwards. Khemlani & Johnson-Laird (2013) further constrained the interpretation of their previous study as far as inconsistencies were harder to detect only if participants had to explain both consistent and inconsistent sets of assertions.

The transparent phenomenal signature of knowing case stands in the middle between self-deceivers' attitudes being completely non-transparent and them becoming part of our world/self-model. The latter case is the one that I want to focus now. Recently, the term "transparency" has been used in explanations of self-deception (Galeotti, 2012, Bagnoli, 2012, Marraffa, 2012). Its meaning, however, is different from the notion of "phenomenal transparency" introduced by me in the previous section. These authors use term transparency either in the sense of cognitive lucidity[220] or in the sense of inaccessibility of whole processes or mental operations.[221] We understand transparency as *inaccessibility of earlier processing stages* that leads to a specific kind of phenomenology – the one of perceptual presence, realness, and mind-independence. Thus, the use of the transparent–opaque distinction is not a substitute for the conscious-unconscious distinction. On the representational level, the transparency constraint allows for the possibility of unnoticeable self-misrepresentations (Metzinger, 2003, pp. 431-432). If a certain conscious content is transparent, then it is experienced as real and the subject is certain of its existence.

So, how far does the notion of phenomenal transparency help clarify the phenomenon of self-deception? First, the difference between a) transparent mental representations and b) mental representations that have transparent mental representations as their content has to be considered. This distinction is important to differentiate the analysis of the formation and the functional role of self-deceptive phenomenal representations from the process of their justification:

a. Transparent representations, due to their being perceived as real, form a background model that serves as a basis for building simulation models (cf. Metzinger, 2003).Those transparent representations contain transparent parts of the phenomenal world- and self-model, the epistemic status of which we never question, due to their not being experienced *as* representations, but as the real world and the real me (e.g., we simply see the book in our hands and perceive both the book and our hands as real, cf. Metzinger, 2003, p. 165-166).
b. We can form thoughts *about* the content of such transparent representations, e.g. "*de re*-beliefs" about this book in our hands. Conscious thoughts, however, are (mostly) *opaque*, experienced as simulations, mental representations that might be true or false. One argument of the self-model theory is that this causally enables "modal competence", the ability to distinguish between what is real and what is only possible and to compare self-generated models of reality with a pre-existing default model, continuously marked out as real.

[220] Transparency as cognitive lucidity: "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.*" (Galeotti, 2012, p. 55; italics are our own emphasis)

[221] Transparency as inaccessibility of mental processes: "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. 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." (Bagnoli, 2012, pp. 100-101; italics are our own emphasis)

Typically, self-deceptive phenomenal representations are assumed to be beliefs that we take to be true. Thus, the most straightforward suggestion would be that they are opaque. Yet in Pliushch & Metzinger (2015) we argue that, like thoughts in mind wandering and dreaming episodes (Metzinger, 2003, p. 172), self-deceptive thoughts are often exceptions to the rule – they may be transparent, which is the case when they have been integrated into the *reality model.* The point is that there may actually be "mind-independent thoughts", thoughts that cannot be experienced *as* thoughts anymore, because self-deception often occurs exactly when cognitive contents have acquired the experiential status of "mind-independence." Phenomenal transparency leads to a different kind of subjective feeling of epistemic reliability, which is not to be equated with the epistemic status of beliefs, on the one hand, and the feeling of epistemic reliability accompanying opaque mental representations, on the other. Further, the transparency of self-deceptive representations is compatible with them having to be justified by the self-deceiver, in which case a self-deceiver forms an opaque thought *about* a now transparent self-deceptive representation.

At least four points speak in favor of phenomenal transparency as a frequent property of self-deceptive representations. First, it is an established fact that emotions, which may oscillate between transparency and opacity, play a role in self-deception (Mele, 2000, 2001; Damm, 2011). This directs attention to the necessity to carefully consider transitions between transparency and opacity of self-deceptive representations. Second, Gendler (2007) also points in the same direction, as far as she conceives of the self-deceiver as somebody who make-believes/imagines/fantasizes that something is the case. We argue that it is the transition from imagination to *being-the-case* that is relevant. Third, anosognosia is often mentioned in the literature as a case of self-deception (e.g. selected chapters of Bayne & Fernández, 2009, Ramachandran, 1996). However, the semantic overlap of our concepts of delusion and self-deception (their motivated nature and resistance against available evidence) does not imply an overlap of *phenomena*, or that all motivated delusions are instances of self-deception. Thus, remaining neutral to the question whether anosognosia is an instance of self-deception, we want to argue that another reason, apart from motivation and resistance against counter-evidence (which makes the ascription of self-deception to anosognostics so popular in the literature) is that in anosognosia, or denial of illness, the misrepresentation that one is healthy or that one's arm is not paralyzed has become a transparent part of their PSM. It is now a robust part of one's phenomenal body-model, experientially mind-independent, and thus makes it resistant to counter-evidence. Fourth, transparency provides a better explanation for real-world cases of self-deception, for example the air crash example described in Trivers (2010) in which a pilot and a co-pilot argue about the appropriateness of weather conditions for take-off. The co-pilot's behavior changes from first pointing out the danger of the weather conditions to eventually agreeing with the pilot, boarding the plane and taking off. How can the Korean co-pilot agree to board the plane, given his knowledge of ice on the wings, his warnings to the pilot and so on, if it means gambling with his life? The self-deceptive representation is now not part of the opaque, "doubtful" cognitive self-model anymore, because it has become a part of the transparent, external situation model, and, therefore, is experientially mind-independent. The speculative hypothesis voiced in Pliushch & Metzinger (2015) is that the transition from the opaque self-model into the transparent model of external reality sometimes takes a "detour" through the transparent layers of the somatic and emotional self-model, i.e., first establishing "mind-independence" and "realness", and then gaining the status of "externality" (see for example Thagard, 2006 for self-deception as a result of an emotional skewer). It would then be the *emotional* state of the self-deceived person that is responsible for the phenomenological projection of a cognitive misrepresentation into the outside world. Thus, as such it is not questioned (it *cannot* be questioned), and there are no simulations performed according to which it is false. Accordingly, the question is not one of entertaining a certain belief *as a belief* (being aware that it can be false), but rather one of being gradually drawn into a certain *model of reality*. Thus, the point is that after a belief is adopted and "owned" as part of the cognitive self-model, its truth-value is (usually) not questioned, because it has subsequently "escaped" to become part of the transparent PSM, or the external situation model, in this way guiding behavior *and* shaping our external phenomenal reality as well.

Moritz et al. (2014) have tested whether overconfidence in errors of schizophrenic patients can be attenuated. They asked participants to walk from an ego-perspective in virtual reality (VR) setting and pay attention to the environment (p. 268). Subsequently, there were asked to recall identity, location and affect (happy, neutral, angry face) of pedestrians in VR and got accuracy feedback for their responses. They were asked to fill out a Paranoia Checklist (frequency subscale) before and after the experiment which consistent of three such VR-walks, one for practice and two for the purpose of the experiment. The results indicate that the severity of paranoid symptom can be reduced by means of the given VR-experiment. Given that such cognitive biases as jumping to conclusions and overconfidence in errors might be seen as characteristic for self-deception, planting the "seeds of doubt" (Moritz et al., 2014, p. 267) to avoid jumping to conclusions and overconfidence in errors might be seen as the strategy to reduce self-deception too. Interestingly, in Moritz's et al (2014) study acceptance of contradictory information (to one's own conclusion) was achieved in a non-motivated manner.

What conclusions can be drawn from the virtual reality experiment? Our perception of reality can be manipulated. This is a change to a world-model and it changes our cognition. So, if certain attitudes were to become transparent such that we do not perceive them as attitudes, but as parts of our transparent phenomenal world-model, then this would change the epistemic agent models we possess. If such world-model changes were to occur in a motivated manner, they would explain the behavioral inconsistency and justificatory actions of self-deceivers. Absence of tension would be more probable for this type of self-deception to occur and would only make it more robust.

To sum up the results in chapter 2 so far, I have argued against intuitions as a basis for setting explananda of self-deception. Instead, I argued that behavioral and phenomenal criteria should be elaborated. In section 2.1.2 I named as such inconsistency and justification, as well as tension, insight and, in section 2.2.1, counterfactual goal-directed pull as explananda. I have further argued that subpersonal goal representations should be viewed as motivation for self-deception, that dolphin model of cognition best explains the phenomenology of self-deceptive process and that self-deceptive attitudes are characterized not only by degree of consciousness and certainty, but also transparency. In chapter 1 I set out four constraints on explanations of self-deception: disunity in unity, phenomenological congruency, parsimony and demarcation. Particularly the demarcation criterion is hard to answer, since self-deception seems to have become a cluster concept (section 1.1.3). As for the first two, in the absence of clear demarcation, the more behavioral and phenomenological constraints a theory of self-deception can explain, the better.

Chapter 3 will be again, like section 1.3, concerned with the empirical literature. Non-evolutionary and evolutionary theories will be introduced, that propose which kind of a function self-deception serves for. This will not only make an understanding of self-deception more profound, but new conducted experiments on self-deception will also be discussed. Baghramian & Nicholson (2013, p. 1025) argue that there is a disagreement in the literature on how to define self-deception, with which I agree and that empirical experiments are difficult to conduct. In section 1.3 I have discussed two paradigms for testing self-deception – voice recognition and pain endurance – and warned of two problems for these paradigms: first, the confusion between personal and subpersonal levels of description and, second, the content fixation problem. In the remaining part of the section I will contrast a philosophical example of self-deception with some empirical ones that are to be discussed in the following chapter. I will then propose ideas on how to improve the testing for behavioral and phenomenological explananda that I set out.

The kind of self-deception elaborated in the philosophical literature typically is one in which there is an inconsistency in the phenomenal self-model (PSM) that is globally accessible for attention, cognition and action. Take, for example, Amelie Rorty's (1988) hypothetically constructed example of self-deception: A cancer specialist denies that she has cancer, yet exhibits behavior consistent with the proposition that she has cancer, e.g. writing farewell letters to friends. In this case, the self-deceiver can plausibly be asked to explain the reasons for her behavior and the inconsistency can be made available for conscious cognition. Take another often constructed example of a wife who denies that her husband cheats on her, but avoids visiting locations that would confirm this being the case (see section 1.1.2.5). It seems that in this case, too, the aim was to construct a hypothetical example of self-deception in which the inconsistency can be made introspectively available for attention and cognition, which can only be the case if it can become part of the phenomenal self-model.

If one looks at the empirical psychological literature on self-deception, however, it will often suffice to assume that there is an inconsistency between the phenomenal and the unconscious self-model. To demonstrate this point, let me consider some of the examples interpreted as cases of self-deception by von Hippel & Trivers (2011b). As one example they describe a study by Ditto and colleagues which shows that "when people are led to believe that color change is a good thing, they wait more than 60% longer for the test strip to change color than when they believe color change is a bad thing" (von Hippel & Trivers, 2011b, p. 8). Another example they describe is a study by Epley & Whitchurch who found that, if asked to identify their own photo, "[p]articipants were more likely to choose their photo morphed 10% with the more attractive image than either their actual photo or their photo morphed with the unattractive image" (von Hippel & Trivers, 2011b, p. 5). A third example I want to mention is a study conducted by Snyder and colleagues. They found that if participants are given the choice of sitting either next to or away from a disabled person, and if the television in front of the two open seats shows the same program, they sit next to the person, presumably to demonstrate to self and other that they were not prejudiced against disabled people. However, if there are two different television programs, participants choose the seat further away (von Hippel & Trivers, 2011b, p. 10). What these studies clearly show is that not every self-related bit of information and not every self-related preference represented in the system becomes part of the PSM, and that overt action may often be guided by goal-representations that are not part of the phenomenal self-model. Yet, what such studies do *not* show is that there actually is some phenomenally represented inconsistency in the self-deceiver. To give one more example, Greve & Wentura (2010) describe self-immunization as "personal/consciously experienced stability versus subpersonal/cognitively adaptive dynamics" (where "stability" refers to trait *ascription* and the "dynamics" is related to trait *description*; p. 728) and note that self-deception is not resolvable purely on the personal level. This is not only the case for self-immunization, but for many other cases described in the psychological literature as well. Thus, while in philosophical thought experiments self-deceivers can often plausibly be asked to justify their inconsistent behavior, because the inconsistency is in principle globally available for attention and cognition, in empirically plausible psychological cases the inconsistency is often accessible to the third-person observer only, but not to the self-deceivers themselves, because it is not a part of the PSM.

Is every bias self-deceptive? Is every misrepresentation self-deceptive? Is every goal-directed misrepresentation self-deceptive, the one whose content can be manipulated by triggering different goals (compare the color change example mentioned above)? Does self-deception involve a certain phenomenology, e.g. an unspecific, vague feeling of uneasiness? Why does this feeling not lead to the revision of the misrepresentation[222]? Still, do self-deceivers *permanently* experience this feeling of uneasiness? The answer to the first three questions is: If one has set out fine-grained enough behavioral and phenomenological explananda whose presence could be tested, then yes, I would call it a case of self-deception. Concerning the vague feeling of uneasiness, I think that one could improve the questionnaires for asking participants to assess their phenomenology in order to try to disentangle feelings of uneasiness that are ascribed not only in case of self-deception, but e.g. cognitive dissonance (see section 3.1.1). One could, for example, add a question to test for counterfactual goal-directed pull. In the experiments described above that pull would be absent, I think. This, one potential phenomenal criterion would not be fulfilled. Regarding the last question, permanent feelings of uneasiness would be unhealthy though (Trivers, 2011). This is one of the reasons that I think that there has to be a mechanism by which tension-loaded self-deception becomes tension-free and to which tension itself could contribute.

Apart from a more fine-grained phenomenological questionnaire, I think that it would be interesting to experimentally test how the *type* of personal-level justification that self-deceivers have to construct actually changes their unconscious self-model in order to stabilize and strengthen their self-deception. In order to do that, the artificially induced experimental self-deception, as described above, might be too weak. An alternative is to make the experiment in pairs: a self-deceiver and her relative, friend or a person who knows her. This is because the close person could actually point to the kind of self-deception that she thinks the self-deceiver engages and then this kind of self-deception would not be artificial. Last, certain characteristics can be tested and explored in clinical groups that one assumes to be present also in self-deception. For example, anosognosics have been argued to be similar to self-deceivers in that they distort reality in a motivated way. But they do not experience tension. Why? An answer to this question might explain how self-deception becomes tension-free. Another interesting group is the one exhibiting pregnancy denial (section 1.1.3). Insight is a characteristic that one could study there. More virtual reality experiments for studying the impact of context change on self-deceivers would also be desirable.

[222] See Cooper (2007b) for a short description of Schachter & Singer's (1962) misattribution of arousal paradigm.

3 Functions of self-deception

> Without a prior account of the *function* of self-deception – the place of self-deception in human (and nonhuman) life – the efforts of Mele and the philosophers and cognitive psychologists he criticizes are analogous to attempts to understand how a chair is made without first understanding that chairs are made for sitting.
> (Rachlin & Frankel, 1997, p. 124)

This chapter will elaborate on the function of self-deception. Using distinctions from McKay & Dennett's (2009) paper on misbeliefs I will introduce the reader to the debate. In section 3.1.1 I will consider, first, non-evolutionary functions of self-deception and then, in 3.2, evolutionary functions of self-deception. Let me start with some distinctions that might become useful. McKay & Dennett (2009) differentiate between three kinds of processes by which misbeliefs (false beliefs) might be generated:

I. those that arise as a result of a breakdown of belief-forming processes and
II. those that arise as a result of the "normal course" of those processes where the normality is determined by the proper normal function in Millikan's teleological sense and may be statistically rare (p. 496)
III. the intermediate possibility of "doxastic shear pins" or system components "*designed to break* (in certain circumstances) so as to protect other, more expensive parts of the system" (p. 501). Defensive, protective function of delusions, pointing out their motivated nature, leaves the question open whether they are *psychologically* or *biologically* adaptive (p. 502).

McKay & Dennett (2009, p. 498) further differentiate two ways in which misbeliefs (and self-deception as one kind of such misbeliefs) may be adaptive:[223]

- by being a result of an adaptive *process*
- or by being adaptive *themselves*.

Here, they point out that one should be cautious about whether inferences drawn from behavior indicate a possession of a certain *belief* or merely that of an *action policy* and exemplify this using error-management theory's basic assumption that a system is prone to commit the less costly error, be it false positive or false negative (p. 500; compare this assumption to that of Mele's theory):

> The issue here is what doxastic inferences can be drawn from behaviour. After all, we always look before crossing a road, even where we are almost positive that there is no oncoming traffic. Our actions in such a case should not be read as reflecting a belief that there is an oncoming vehicle, but rather as reflecting a belief that there *might* be an oncoming vehicle […]. (McKay & Dennett, 2009, p. 501)

They further argue that "positive illusions" – "*unrealistically positive* self-appraisals and beliefs" (McKay & Dennett, 2009, p. 504; reference to Taylor 1989) are adaptive misbeliefs, if they are fitness-enhancing and not epiphenomena (p. 505). Positive illusions have been described as self-deceptive (see section 3.1.3.1). It will be useful to introduce the last distinction of McKay & Dennett (2009) that concerns the way in which positive illusion might be fitness-enhancing:

- *directly* by sustaining health: The idea is that *positive illusions* help avoid the long-term activation of stress-induced coping behavior which is detrimental (p. 506). Misbeliefs are

[223] McKay & Dennett's (2009) criteria for adaptiveness are: "(1) The object is a specific focus of deliberate design (not a mistake or a by-product); (2) the object appears, from a certain perspective, to be malfunctioning or limited insofar as it misrepresents information to the consumer of that information; and (3) such misrepresentation is actually beneficial to the consumer of that information" (p. 495).

adaptive not "by virtue of their falsity," but "because of the contingent facts about the world" (p. 508). Interestingly, the authors deny adaptivity to placebos though, singling them out from positive illusions. *Placebo misbelief*, is for them a case of a by-product of an adaptation, because they hold placebo to be a result of "economic resource management" (= Placebos activate immune ressources, because there is less need to be cautious with immune ressources in the presence of doctors, p. 507).

- *indirectly* by leading to adaptive actions: The authors deny that in this case misbeliefs are adaptive. The reason is that despite a "belief *system* geared toward forming illusional positive beliefs" (p. 506), the misbeliefs resulting from the working of such a system might not be necessarily adaptive themselves, or as adaptive as smoke detectors biased towards false alarms are: If a detector came out that would identify all alarms correctly without false alarms, it were to be preferred (p. 506). This is a critique directed at the explanations of adaptivity of misbeliefs in the framework of error-management theory.

To sum up, among the questions that need to be answered in a functional analysis of self-deception is (1) whether it is a breakdown of normal functioning, (2) whether its process or the resulting attitude is adaptive, (3) whether it is adaptive directly or indirectly and (4) what it is actually adaptive for, e.g. health or maybe even deceiving others (see section 3.2.2 for the latter proposal). Self-deception has been argued not to be a breakdown of *normal* belief forming processes (Bayne & Fernández, 2009), even if it may be a breakdown of *rationality* (see also section 1.1.2.3 for the introduction of different kinds of rationality). There are no distinctions between the beneficiality of the process and the resulting attitude that I am aware of. Directness of self-deceptive influence is also an issue. In section 3.1 I will discuss four non-evolutionary ways of solving the fourth question: (1) Self-deception serves to reduce cognitive dissonance (section 3.1.1); (2) it serves to preserve the stability of the self-concept (section 3.1.2); (3) it is the perseverance of self-esteem as a central feature of the self-concept that stands in the foreground (section 3.1.3); (4) defence of self-concept and self-esteem are proximal means of reducing the anxiety of death (section 3.1.4). The last solution – that of terror-management theory – builds a bridge to evolutionary solutions in virtue of possible evolutionary implications. In section 3.2 I will, then, explore three evolutionary functions offered for self-deception: that it serves the deception of other individuals (section 3.2.2), that it is a by-product of rational capacities of the human mind (section 3.2.3.1) and that is an exaptation – a by-product that acquired the function to uphold positive self-perception (section 3.2.3.2).

3.1 Non-evolutionary theories of self-deception

In this subchapter I analyze the explanations of self-deception in terms of cognitive dissonance (section 3.1.1), self-concept (section 3.1.2), self-esteem (section 3.1.3) and terror-management (section 3.1.4). Each kind of explanations narrows down the next in a certain respect: Cognitive dissonance explanation sees self-deception as a general inconsistency between representations, self-concept explanations focus on inconsistencies that threaten the self-concept. The explanations, that equate self-enhancement with (one kind of) self-deception, might be seen as those that ultimately concern self-esteem and terror-management theory ascribes to the defense of the self-esteem the function to relieve anxiety of death. Thus, one could argue that self-deception serves the function to relieve anxiety of death either directly or indirectly: indirectly via self-esteem and directly via the idea that Varki & Brower (2013) proposed, namely that self-deception evolved to relieve anxiety of death, because anxiety of death is the psychological evolutionary barrier against the development of full theory of mind (see section 3.1.4). On this latter idea the awareness of our death is an evolutionary hindrance – such individuals would be scared to fight for mates if aware of death – thus, such individuals could spread their genes only if they were somehow to self-deceive away the awareness of death.

Proulx et al. (2012) argue that at least cognitive dissonance and terror-management theory have a common schema (p. 287): *inconsistency* (= result of prediction error) triggers *aversive arousal* (usually described as uncertainty, anxiety) which leads to different *compensatory responses* (assimilation or reinterpretation of experience consistently with given priors, accommodation or revision of expected relationships, affirmation or commitment to alternative expected relationships etc.). More generally, "ego-defense" is argued to be an instance of an inconsistency compensation process (Proulx et al., 2012, p. 287). Self-concept and self-esteem defense is, then, also an instance of such inconsistency compensation process. Proulx's et al (2012) is similar to Baumeister & Newman's (1994) argument that, when the goal to reach a favorable conclusion is not achieved, negative affect leads to repeated, but biased personal level belief forming process that results in self-deception. The theories, that I will review, differ in the necessity of aversive arousal as a phenomenological component and in the proposed mechanisms of its reduction, but in virtue of its simplicity it might be helpful to keep this schema in mind.

3.1.1 Cognitive dissonance

In this section I will first make a short review of the cognitive dissonance theory and then show how it has been applied by Scott-Kakures (2009) to explain self-deception. Scott-Kakures argues that it is the drive to reduce inconsistency in general that can explain self-deception, while accounts presented in the next section (3.1.2) assume a more specific inconsistency – threat to self-concept – to motivated self-deception.

The structure of this section is as follows: I will introduce the general idea of cognitive dissonance, which centers on inconsistency and affective arousal and sketch the different triggering conditions for inconsistency proposed by successor theories of cognitive dissonance. One of such successor theories is *self-affirmation* whose main idea is that upon affirming their self-worth people are less defensive about accepting threatening information about their self-concept. It has a direct relation to recent proposals of how the presence of self-deception can be tested. While traditional philosophical theories of self-deception more often than not compare it with *wishful thinking* (see chapter 1), psychological theories focus more on the relation between self-deception and *self-enhancement*. Von Hippel & Trivers (2011b), for example, argue that self-enhancement is a kind of self-deception and the presence of self-deception can be tested by the self-affirmation paradigm (see section 3.2.2.3). Scott-Kakures (2009), on the other hand, attempts to explain self-deception as inconsistency reduction in general. The main question then is: how narrow/wide should the scope of self-deception be assumed? The trivial answer is that our folk-psychological ascriptions of self-deception are wider than those of self-enhancement, yet there is no unifying empirical testing paradigm of self-deception. This leads to the focus of empirical research on the paradigms that are available. The most fruitful way out of this dilemma is to develop a new testing paradigm for self-deception that builds on other assumptions than those presented in section 3.1. Certainly, this needs only to be done if presented theories are found unsatisfactory, so, let me continue with the review of cognitive dissonance.

Cooper (2007b) has recently written a book summarizing 50 years of research on cognitive dissonance. I will rely on that book, as well as on Harmon-Jones' (2012) recent encyclopedia contribution to the topic for a short introduction to cognitive dissonance, a theory that has been developed by Festinger (1957) and since then underwent several changes. The main idea is that people are "*driven* to resolve" inconsistencies (Cooper, 2007b, p. 3). The units, among which inconsistency arises, are cognitions which are "any 'piece of knowledge'" (p. 6), actions and attitudes included, because they all are argued to have a "psychological representation" (p. 6).[224] Cognitions are inconsistent, "if one cognition follows from the obverse (opposite) of the other," (p. 6) but only if those cognitions are *relevant* to each other (p. 6). Such an inconsistency creates *tension* which is drive-like and needs to be reduced (p. 7). Degree of inconsistency determines the magnitude of cognitive dissonance according to the following formula $DISSONANCE\ MAGNITUDE = \frac{SUM\ (all\ discrepant\ cognitions\ x\ importance)}{SUM\ (all\ consonant\ cognitions\ x\ importance)}$ (Cooper, 2007b, p. 9). There are different ways of reducing dissonance (changing discrepant, consonant cognitions and/or their importance), yet, there is one particular assumption about the changeability of different cognitions: attitudes are easier to change than behavior (Cooper, 2007b, p. 8).[225] Thus, according to Festinger's (1957) original account the notion of cognition is very wide to encompass action and attitude, there are little constraints on the conditions under which cognitions are inconsistent (relevancy) and the same idea as in self-deception is present that inconsistency leads to tension.

Self-consistency (Aronson)	**Self-affirmation (Steele)**	**New look (Cooper & Fazio)**
Discrepancy between behavior and one's view of oneself as *moral, rational and competent* Cooper 2007b, p. 94) • Inconsistency that involves some *expectations* regarding the self has the potential to evoke dissonance • Mechanism of dissonance reduction is *local*	Protection of the self-concept by means of preserving self-esteem (Cooper 2007b, p. 91) • Mechanism of dissonance reduction is *global*	- High *decision freedom*, coupled with - *commitment* to behavior - that leads to *aversive* consequences which are - *foreseeable* (Cooper, 2007b, p. 73)

Table 22. Cooper: cognitive dissonance → difference in triggering conditions
Distinctions from Cooper (2007b).

Local vs. global mechanism of dissonance reduction: "In the self-consistency view, the repair must be directed at the specific inconsistency that caused the problem. We can change our view of ourselves as moral and competent people or, more typically, we can change our attitudes or other relevant cognitions to make ourselves feel more worthy. For self-affirmation, the repair can be general. As I pointed out earlier, anything that reaffirms the integrity of the self-system will do." (Cooper, 2007b, p. 98)
Self-standards model of dissonance: Cooper (2007b) proposes a way to integrate the three approaches. The idea is that behavior triggers cognitive dissonance and that the accessibility of different standards (personal vs. normative) is the factor that decides how dissonance is reduced. If the behavior is inconsistent with *normative* standards, then self-esteem is not a moderator. If the behavior is inconsistent with *personal* standards, then self-

[224] Definition of cognition as psychological representation: "An action is different from an attitude which, in turn, is different from an observation of reality. However, each of these has a psychological representation – and that is what is meant by cognition" (Cooper, 2007b, p. 6).

[225] Among the paradigms of testing cognitive dissonance are (Cooper, 2007b):

- *Free choice paradigm*: Given choice among two similarly ranked (by the participant) items, the level of attraction decreases for not chosen item and increases for the chosen item and, as a consequence, the items spread apart in the degree that the participant likes them. The dissonance that such a decision between items create is the one between advantages and disadvantages of both items.
- *Induced-compliance paradigm*: If a participant is induced to act against his attitudes by the experimenter, the dissonance between the act and the attitude leads to attitude change.
- *Effort-justification paradigm*: If participants spent effort on some matter, they will change their attitude of the given matter: e.g., find a boring task more interesting.

esteem moderates the mechanisms of cognitive dissonance in the following fashion that is dependent on the accessibility of self-attributes. On the one hand, positive and *relevant* self-attributes increase dissonance as in self-consistency theory, because one's expectations of oneself are made salient (Cooper, 2007b, p. 110). Thus, high self-esteem should lead to more dissonance.[226] On the other hand, positive and *irrelevant* self-attributes decrease dissonance as in self-affirmation theory, because they protect self-esteem in a global way (p. 110). Thus, low self-esteem should lead to more dissonance, because people with low self-esteem do not have a buffer of positive qualities to self-affirm themselves.[227] Recently, Harmon-Jones (2012) has also proposed an *action-based account* of cognitive dissonance, according to which, 'cognitions' are viewed as action tendencies and dissonance arises if conflict-free action is impeded.

As an answer to Bem's self-perception theory, which states that self-perception happens by observing one's own behavior (Cooper, 2007b, pp. 35-41), there have been attempts to prove that there is actually an *internal tension or arousal* present in the dissonance reducer (Cooper, 2007b, pp. 42-61):

- *Indirect measures:* Taking the misattribution paradigm[228] as a starting point, Cooper et al. conducted studies to prove that dissonance reduction would not happen if arousal due to dissonance can be attributed to some external stimulus. *Negative valence* of arousal is assumed to trigger the reduction of cognitive dissonance (Cooper, 2007b, p. 52).
- *Physiological measures*: Dissonance leads to "heightened physiological arousal" measured by non-specific skin conductance response (SCR) which is "a measure of the body's increased production of skin moisture." (pp. 50-51) There is a certain similarity here to Gur & Sackeim's attempt to prove self-deception by arguing that skin-conductance response indicates the presence of self-deceptive beliefs (see section 1.3.2.1). One should be cautious, though, to make inferences from physiological to psychological states (see Drayson, 2014; section 1.3.4).
- *Questionnaire*: Participants were asked whether they were "*uneasy, uncomfortable,* and *bothered*", following the inducement of the state of cognitive dissonance. They gave affirmatory responses. Moreover, Cooper argues that it is possible that the *magnitude* of cognitive dissonance can be affected by artificially inducing arousal in participants via physiological substances or in other words: there is a correlation between level of arousal and level of subsequent change in cognitions in response to dissonance (Cooper, 2007b, p. 59).

Notice that the description of cognitive dissonance is very general – feeling of uneasiness, anxiety – and that the same is the case for self-deceptive tension. Similarity of the generality of the phenomenological description may have made cognitive dissonance a viable candidate for explanations of self-deception, yet, this is only the case if this similarity is not the result of insufficiently fine-grained testing procedures of phenomenology. If it is not and several processes cause the same kind of phenomenology, which, in virtue of such a fact, can be misinterpreted by the person to result from one or another process, one can, then, still ask whether the possibility of misinterpretation is the *enabling condition* for self-

[226] Harmon-Jones (2012) formulates this predictions as follows: "One of the primary predictions derived from this revision [self-consistency theory of cognitive dissonance] is that low and high self-esteem individuals should respond with less and more dissonance reduction (e.g., attitude change), respectively, because in dissonance experiments high self-esteem individuals are induced to act in ways that are more discrepant from their positive self-views. Experiments testing this prediction have produced mixed results" (p. 545).

[227] Interestignly, Cooper's interpretation of self-affirmation theory is such that people with high self-esteem need to less reduce dissonance than those with low self-esteem, because the former have a "reservoir of valued attributes to think about that can help them buttress their self-system" (p. 103).

[228] *Misattribution paradigm*: Schachter & Singer (1962) proposed a two-factor theory of emotion: emotion = arousal + attachment of a cognitive label. Thus, a study was conducted that showed that the injection of adrenalin can be interpreted as happiness or anger, depending on the situation – the action of a confederate (Cooper, 2007b, pp. 44-47).

deception or whether the other process with the same phenomenology (in this case – cognitive dissonance), really *is* self-deceptive.[229]

Another question is whether cognitive dissonance paradigm is *incomplete* as an explanation. Kunda (1992) argues that there are three reasons for why motivated cognition cannot be explained by cognitive dissonance: 1) the goals and means of triggering cognitive dissonance are too narrow (Kunda, 1992, p. 338); 2) if ever, cognitive dissonance paradigms explain the *conditions* under which dissonance arises, but not the *mechanisms* through which biased hypothesis testing takes place (Kunda, 1990, 1992); 3) although the directionally motivated phenomena (self-serving biases, failures of probability estimates and base rate estimates) may be explained by Festinger's original cognitive dissonance account (Kunda, 1990, p. 491) as *tension between inconsistent beliefs*, the presence of (physiological) *arousal* is an argument in favor of a noncognitive motivational account which for her the resolving of inconsistencies *per se* is not (p. 484). The first critique point follows from Kunda's assumption that inconsistencies *per se* do not lead to cognitive dissonance, but that it is e.g. the refuting of threats to the self that might arise through processing of information inconsistent with a given self-image that does lead to cognitive dissonance. The latter is Aronson's self-consistency account – a reformulation of Festinger's original cognitive dissonance account (Kunda, 1990, 1992; Scott-Kakures, 2009). Scott-Kakures (2009) argues against this point. The second criticism about mechanisms has come to be accepted in that Cooper (2007b) acknowledges the possibility of the mechanisms of motivated reasoning identified by Kunda to be the ones by which also cognitive dissonance reduction is carried out (pp. 85-86, 181). The last point about what can be a motivating force is an empirical question. Typically, goal representations are assumed to exert motivational forces onto different mental processes in psychological theories (see section 2.2.1).

Scott-Kakures (2009) takes up the challenge of showing that self-deception can be explained by Festinger's (1957) original cognitive dissonance paradigm (see figure 14), namely that inconsistent cognitions lead to an aversive psychological state that triggers mechanisms of cognitive dissonance reduction. Scott-Kakures (2009) adds that inconsistencies influence the dissonance effect, if they are *accessible* to the person (p. 82) and can be influenced by desires (p. 79) so that "the existence of dissonance will depend on the contents of the more global cognitive-motivational perspective or constitution of the subject" (Scott-Kakures, 2009, p. 79). Thus, Scott-Kakures (2009) explanation is a folk-psychological, personal level one, into which he incorporates Kruglanski & Webster's (1996) psychological concept of the need for "cognitive closure." Kruglanksi & Webster (1996) define the need for cognitive closure as the "individual's desire for a firm answer to a question and an aversion toward ambiguity" (p. 264) and argue that it affects the motivated (goal-directed) knowledge-construction process (p. 263). This kind of need is assumed to be "a "hypothetical construct" knowable only indirectly via its effects," but which is real, because "a scientific construct is real if its effects are real" (Kruglanski &

[229] In favor of the latter view studies, mentioned by Cooper (2007b), that give evidence in favor of the hypothesis that cognitive dissonance can be aroused by "observing *other people* act inconsistently with their own attitudes" (p. 117) if these other people are members of the in-group and the level of identification with the in-group is high. This speaks for cognitive dissonance, if Trivers' (2011) arguments are accepted in favor of viewing in-group/out-group biases as self-deceptive. Further, Cooper (2007b) assumes that cognitive dissonance is a phenomenon that is present across cultures where culture influences the standards that affect the mechanism of dissonance reduction. If self-enhancement is a kind of self-deception and it is found across culture, then the latter hypothesis would also speak in favor of the view that cognitive dissonance can indeed offer an explanation of self-deception.

Webster, 1996, p. 279). Applied to Scott-Kakures' (2009) explanation of self-deception, the hypothetic nature of the need for "cognitive closure" leads to the conclusion that, though his aim is probably to give an account of how self-deceivers' process of *explicit deliberation* or conscious reasoning occurs, in its stead he proposes a hypothetical folk-psychological description of a reasoning process in which cognitive closure receives the main explanatory weight as an element that has been tested to bias personal level hypothesis testing,[230] but whose operation is probably of a subpersonal level kind. In other words, Scott-Kakures' (2009) explanation consists in the description of how certain desires operate in the self-deceiver.

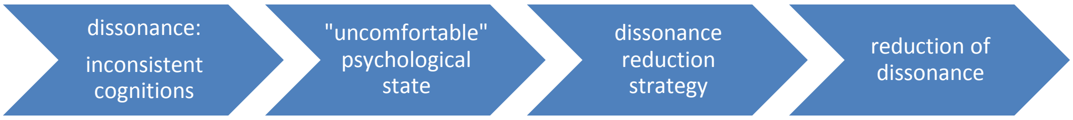

dissonance:
inconsistent cognitions
"uncomfortable" psychological state
dissonance reduction strategy
reduction of dissonance

Figure 14. Cognitive dissonance paradigm
Distinctions from Festinger (1957) and Scott-Kakures (2009, p. 77).

Scott-Kakures' (2009) general idea is that due to the *unspecific* nature of aversive arousal, that follows cognitive dissonance, one source of dissonance can be confounded with another. According to him, there are two sources of dissonance which are not distinctive of

[230] It is personal level motivation that Kruglanski & Webster (1996) have in mind, as it is visible from the following quotation: "Finally, recall that most effects of the situational demands were replicated by means of our individual-differences measure of the need for closure. Most of the items in that scale (26 of 42) have clear motivational flavor (e.g., terms such as 'I like,' 'enjoy,' 'hate,' 'dislike,' or 'prefer')" (p. 281).
The *conditions* for the need for cognitive closure to arise may be such that the task is uninteresting, there is noise interference or the individual is tired (p. 264). The *consequences* are the general tendencies of urgency ("inclination to 'seize' on closure quickly", cf. p. 265) and permanence ("(a) to preserve, or 'freeze' on, past knowledge and (b) to safeguard future knowledge", cf. p. 265). Depending on the time when this need arises in the *decision process* (prior or posterior to the decision), seizing occurs when decision has not yet been met (deliberation mind-set), while freezing occurs in the implementation mind-set after the decision (p. 266).The underlying assumption, thereby, is that "people under a heightened need for closure experience its absence as *aversive*. They may, therefore, wish to terminate this unpleasant state quickly (the urgency tendency) and keep it from recurring (the permanence tendency)." (p. 265; my emphasis)
Differentiating the need for cognitive closure from other concepts, the Kruglanski & Webster (1996) mention that the need for cognitive closure is not constrained to be a personality trait, but that "need for closure is determined by perceived benefits or costs of closed or open states as influenced by *situational*, cultural, or personality factors." (p. 267; my emphasis) Moreover, it is not constrained to particular belief contents and it biases (=influences the extent and the process of) the knowledge acquisition. The assumption that the need for cognitive closure guides the decision process implies that the difference between heuristic and systematic processing is one of degree and not of a kind (p. 268).
Among the biases, Kruglanski & Webster (1996) mention the amount of alternative hypotheses that are generated: Influenced by the need for cognitive closure, participants generate less alternative hypotheses, yet this leads them to have a higher confidence in their decision (p. 269). Moreover, the authors also mention experiments in which the need for cognitive closure is tested in *interpersonal decision making*, thus, in group decision making and emphasize that the given need "may affect how an individual thinks, feels, acts toward and speaks about socially significant others." (p. 266) In interpersonal interaction, the effect of the need for cognitive closure would be:

- Seizing condition: the participant is easy to persuade and looks for individuals that can persuade him
- Freezing condition: the participant is difficult to persuade (p. 272, 277).

self-deception and which can be confounded: the uncertainty whether *p* or not-*p* (1+2), as well as the awareness of this uncertainty that is conjoined with the desire to have an answer to the question whether *p* or not-*p* (3+4) (pp. 94-95). The latter one is the desire for "cognitive closure" which Scott-Kakures (2009) adopts into his explanation as an empirical example of how desires can influence the belief-forming processes:

> a high need for closure can affect the number of hypotheses generated, a subject's confidence in a conclusion, whether and how an inquirer seeks diagnostic or prototypical information, as well as making likely the use of cues discovered early in the course of her effort to settle a question. Motivation supplied by the need for closure, they suggest, can supply the motivational foundation of such apparently unrelated phenomena as impressional primacy effects, anchoring effects, and the correspondence bias. (Scott-Kakures, 2009, p. 92)

Thus, according to Scott-Kakures (2009) hypothesis-testing may lead to two sources of dissonance and reducing dissonance associated with 3+4 biases hypothesis-testing via mechanisms associated with "cognitive closure," but at the same time, the reduction of dissonance associated with 3+4 also reduces the dissonance associated with 1+2 (p. 95), because the way by which dissonance is reduced is coming to a conclusion regarding the matter in question (see figure 15).

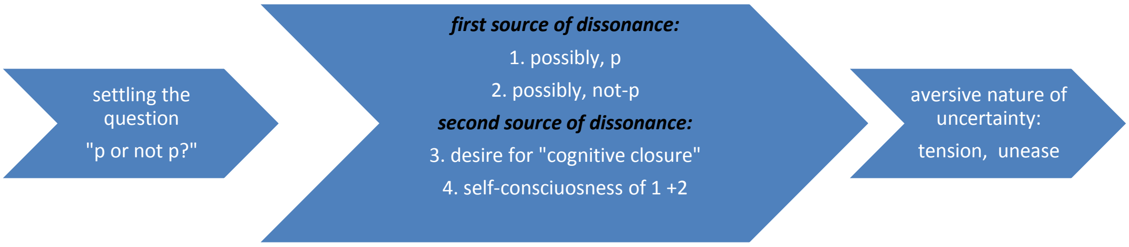

settling the question
"p or not p?"
first source of dissonance:
1. possibly, p
2. possibly, not-p
second source of dissonance:
3. desire for "cognitive closure"
4. self-consciuosness of 1 +2
aversive nature of uncertainty:
tension, unease

Figure 15. Scott-Kakures: sources of cognitive dissonance
Distinctions from Scott-Kakures (2009).

In self-deception, the desire that *p* triggers cognitive dissonance which subsequently biases hypothesis-testing (Scott-Kakures, 2009, p. 98) via the generation of an additional cognitive dissonance: the one between the openness of the question and the *expectations* that are generated by that desire (p. 101):

> If, by her current lights, the question "Do I or do I not have a typically lethal cancer?" is open, it can hardly be the case that "It is far more likely that I will send my daughter to college than that I will not send my daughter to college," and so on. (Scott-Kakures 2009, p. 100)

The merits that the authors assumes his account to have are, first, that it has a "clear psychological rationale," which is aversiveness, for the person's effort to settle the question to respond to cognitive dissonance, while such rationale is absent for Mele's error minimization[231] account (Scott-Kakures, 2009, p. 104) and, second, that it is parsimonious:

[231] Criticism of error minimization as without a psychological rationale: "Rather, it's problematic chiefly because, while we have a clear psychological rationale for why it is that attitude change, in the context of the effort to settle a question, is responsive to dissonance (it is aversive), we

if accuracy motivation were different from the motivation to reduce errors, then one would need to postulate two different kinds of motives that influence hypothesis-testing – avoid particular errors or seek the truth (p. 105). On Scott-Kakures' (2009) dissonance reduction account, however, there is only the epistemic motive to settle the question (p. 105).

Scott-Kakures' account has actually a lot in common with Mele's and other deflationary accounts that use certain psychologically tested constructs – a folk-psychological description of the self-deceiver is given so that the psychological construct (need for closure or error minimization) takes over the role of changing/biasing in whatever manner the personal level mental states of the self-deceiver that are then take part in the reasoning process of the self-deceiver. It is unclear how it should be tested which folk-psychological states precede the biasing and in which manner they interact, because they do not carry the main explanatory weight in virtue of not possessing the causal powers to change the reasoning process of the self-deceiver on their own, without the empirically tested psychological construct.

3.1.2 Stability of the self-concept

I have noted in the previous section that one of the ways to restrict the conditions under which cognitive dissonance has been aroused is to argue that it is the case when one's self-concept is in danger. Self-concept has been defined as a set of schemas defining oneself (Goleman, 1985, p. 96).The idea that schemas organize knowledge and experience is an old one (see e.g., Goleman 1985, pp. 74-77). Schemas are also said to guide attention (p. 79) and the possibility is allowed for *emotions* to trigger schemas (p. 81). Goleman (1985) has, in this vein, argued that the failure to update a schema is an instance of self-deception (p. 97), as well as a biased choice of a schema where this choice itself might be guided by a meta-schema (p. 107). The idea, on which I will focus on in this section, will be that self-deception preserves the stability of the self-concept via failure to incorporate inconsistent information (section 3.1.2.1) or via the redefinition of the self-concept (section 3.1.2.2).

3.1.2.1 Ramachandran's hemispheric specialization

According to Ramachandran (1996), self-deception, which he holds to be a kind of psychological defense (McKay et al., 2009, p. 176), takes place during the attempt "to create a coherent belief system in order to impose stability in one's behavior" (Ramachandran, 1996, p. 351). His theory is an alternative to Trivers' evolutionary account of self-deception (3.2). While Trivers advocates that self-deception evolved to facilitate deception of others (by fooling deception-detection mechanisms), Ramachandran holds the dissenting opinion that self-deception evolved to stabilize the self through the creation of a coherent belief system.[232] The idea that the self-concept needs to be stabilized, but through

have no such clear, psychologically compelling account of why an inquirer's effort to settle a question is responsive to PEDMIN concerns" (Scott-Kakures, 2009, p. 104).

[232] Ramachandran's (1996) critique of Trivers goes as follows: "When you lie to someone else, your purpose is to withhold information that you don't want the other person to know. For example, suppose a chimp (Chimp A) sees where a zoo keeper places a big bunch of bananas. Chimp A now points Chimp B in the wrong direction, so that he can have all the bananas to himself. Now, following Trivers argument, suppose Chimp A wanted to make sure that Chimp B doesn't detect the lie, he engages in self-deception, i.e. he really believes that the bananas are in the place that he points to. But if this were true, then Chimp A himself would also go look

other means, is also guiding Greve & Wentura's (2010) analysis that will be topic of the next section. Three elements can be distinguished in Ramachandran's theory: 1. the idea that self-deception serves the aim to stabilize the self-concept; 2. the function of maintaining stability and incorporating inconsistent information can be mapped onto the two hemispheres; 3. self-deception can be undone by caloric vestibular stimulation[233] (CVS). I will argue that it is the last element that is most interesting in the light of recent empirical evidence that points into the direction that anosognosia and optimism, which have both been argued to be kinds of self-deception (1.2.7), may share the same *subpersonal, interoceptive* mechanism.

Ramachandran (1996) developed his idea of hemispheric specialization (see table 23) on the basis of data obtained through experiments with anosognosic patients that have been paralyzed. He observed an asymmetric relation between the damaged hemisphere that led to paralysis and the occurrence of anosognosia: the latter occurred only when the right hemisphere was damaged (Ramachandran, 1996, p. 350).

Ramachandran's theory of hemispheric specialization	
Left hemisphere	Right hemisphere
Aim: impose consistency on behavior Strategy: "psychological defense mechanisms" (p. 350) Ex. denial, repression etc.	*Aim*: the "anomalous detector" generates a paradigm shift, i.e. incorporates inconsistent information (p. 351-351) Method: there should be some threshold to overcome (p. 352)
On-line processing	Dream sleep
If the left hemisphere is intact and the right hemisphere is damaged, this can result in **anosognosia** (p. 352)	If the right hemisphere is intact, but there is something wrong with the consistency mechanism, **depression** might follow (p. 360)

Table 23. Ramachandran: theory of hemispheric specialization
Distinctions from Ramachandran (1996).

From the asymmetry in the occurrence of anosognosia Ramachandran assumed an asymmetry in the functions of the hemispheres: the left hemisphere tries to impose consistency, while the right hemisphere tries to incorporate inconsistent information. The information that is not incorporated is argued to be rendered unconscious. Ramachandran supports his theory through the observation of the influence of cold-water nystagmus (caloric vestibular stimulation or CVS) on anosognosic patients. While under the effect of nystagmus, anosognosic patients temporary admitted their paralysis:

> It seems as though the *access* to these memories is ordinarily blocked, but the cold water removes the block. The memories then come to the surface and the patient 'confesses' her paralysis. And yet after the effect from the water wears off, the patient flatly denies her earlier admission of paralysis - as though she were completely rewriting her 'script'. (Ramachandran, 1996, p. 355)

There are alternative subpersonal explanations of anosognosia (Fotopoulou, 2013b; Mograbi & Morris, 2013). Mograbi & Morris (2013) argue that it is an open question

for the bananas in that false location. This would defeat the whole purpose of deception and be obviously maladaptive!" (p. 350-351)

[233] Eliciting a nystagmus – rapid eye movements - with the help of cold water is a procedure designed to test the vestibular apparatus. Cold water is applied to the ear canal of the patient. In the given case, it was the left ear canal, because influence on the damaged right hemisphere was desired (Ramachandran, 1996, p. 355).

whether "implicit awareness" in anosognosics (e.g., denying the paralysis, but adjusting the grasping process of object so that it is manageable with only one hand; denying the paralysis, but reacting emotionally to stimuli related to one's physical condition) is the result of a *motivated defensive psychological process* or just of a *cognitive unconscious process* (p. 187). They are open to the possibility that predictive coding might explain anosognosia and offers a *filter*-metaphor for its working: Higher-levels filter information that is forwarded by lower levels[234] (pp. 192-193; for more on the changes of description from agentive to subpersonal explanations see different selection mechanisms in section 2.2.1).

Though hemispheric specialization should not be viewed as a simple dichotomy (Serrien et al., 2006), the idea of hemispheric specialization still enjoys popularity. There have been cognitive, as well as emotional ascriptions of hemispheric specialization. Mihov et al. (2010) connect hemisphere specialization to *creative thinking* by conducting a meta-analysis that supports the idea that the right hemisphere dominates in the domain of creative thinking, as well as that "global thinking style, context-dependent thinking style and figural processes are significantly more characteristic to result in right hemisphere dominance than in left-hemisphere dominance" (p. 445). The right hemisphere is argued to be better at exploring new possibilities, while the left hemisphere – at applying what has been already learned (Mihov et al. 2010, p. 443). Goel et al. (2007) argue that there is hemispheric specialization in the prefrontal cortex (PFC) such that reasoning about *completely* specified information is determined by the left PFC and reasoning about *incompletely* specified (ambiguous) information is determined by the right PFC.[235] They hold this idea to be compatible with Gazzaniga's description of the left hemisphere as the "interpreter" (p. 2248), as well as with Johnson-Laird's mental model theory: In case of indeterminate information multiple mental models need to be combined into one model that preserves the uncertainty and ambiguity of representation (a role of right PFC), else "the interpreter" (left PFC) will impose a certain interpretation prematurely (p. 2249).

Vauclair & Donnot (2005), on the contrary, argue for a hemispheric specialization of *emotions*, namely that it is the right hemisphere which is more concerned with the perception of facial emotions. Mihov et al. (2010) also cite research in favor of the hypothesis that the activity in the right hemisphere is responsible for the processing of affective information (p. 445). Preuss et al. (2014) have tested that right ear CVS (left hemisphere processing) improved categorization of pictures independent of their emotional valence, leading them to conclude that "activating left hemispheric vestibular areas by means of right ear CVS may interact with the prefrontal emotional network and, therefore, improve affective control, specifically for emotionally positive stimuli" (p. 137). Preuss et al. (2014) have further tested the *valence specific hypothesis* (VSH) [236] that right hemisphere is associated with the processing of negatively and left hemisphere – positively

[234] Klein et al. (2013) argue that an impairment of error awareness contributes to anosognosia. Klein et al. (2013) hold that their account is compatible with predictive coding, because "[w]hether error awareness itself is a product of another higher-level predictive-coding mechanism that, for example, compares the predicted task performance with the accumulating prediction error evidence remains to be investigated" (p. 2).

[235] Goel et al. (2007) tested this idea by using a transitive inference task in which participants had to determine the logical validity of transitive statemets, e.g. [A>B, B>C, thus A>C] (p. 2247). Valid arguments are argued to be always determinate, while invalid can be determinate (false, like A<C in the example above) or indeterminate, e.g. [A>B, A>C, B>C] (p. 2247).

[236] Preuss et al. (2014) mention with respect to VSH the 'sticky switch' model according to which depression (right hemisphere activation) and mania (left hemisphere activation) are two opposed hemispheric activation endpoints (p. 134).

valenced stimuli (p. 134). This would even fit to Ramachandran's hemispheric specialization hypothesis, if it is assumed that inconsistent information is negatively valenced and consistent information – positively valenced. Preuss' et al (2014) results suggest that for emotionally positive stimuli affective control increased during right ear CVS and decreased during left ear CVS (p. 137), but there was no corresponding effect for emotionally negative stimuli, e.g. left ear CVS increasing performance (p. 138).

Thus, a host of both cognitive and emotional hemispheric specialization hypotheses have been proposed. Particularly interesting for self-deception, I think, is Preuss' et al (2014) study that links CVS to the processing of affective information. This is so not only because anosognosia is susceptible to left ear CVS, but also because optimism is susceptible to left ear CVS in the same way. Given that anosognosia and unrealistic optimism are conceptually similar (impaired belief updating in the face of negative disconfirming evidence) and involve the same brain region (deficient coding of negative information or lesion in right inferior frontal gyrus - IFG), McKay et al. (2013) have tested whether optimism is susceptible to CVS, as asonognosia is. Participants estimated the risk of contracting illnesses after CVS with the result that risk estimation was significantly higher after left-ear CVS, suggesting that *left CVS reduces unrealistic optimism*. McKay's et al (2013) conclusion is that a unitary mechanism underlies anosognosia and optimism. Given the connections of CVS to affective processing, I think that the affective basis of this mechanisms needs to be explored.

There are further connections between CVS and experiments conducted to test self-deception. Ferrè et al. (2013) have found that left cold CVS decreases the tactile threshold (= the threshold for perceiving touch), but increases the pain threshold. They argue that left cold CVS influences these thresholds independently of each other and further that there is a direct vestibular-somatosensory interaction, but not an interaction mediated by attention. Increased pain threshold has been found by participants in Quattrone & Tversky's pain endurance study: upon being motivated, participants held their hand longer in the cold water. Lopez et al. (2012), citing beneficial effect of CSV on anosognosia (p. 1831), have tested CSV to influence the body schema (tactile stimuli perception was elongated and the perception of the width and length of the left hand increased). Rubber hand illusion, as CVS, also has an analgesic effect (Hegedüs et al., 2014). Hegedüs et al. (2014) have tested rubber hand illusion to increase the heat pain threshold. While this is evidence of how the rubber hand illusion via changes to bodily awareness influences affect, the connection between bodily awareness, affect and cognition is the one of interest.

Interim conclusion: Vestibular sense provides an additional kind of information to the proprioceptive, interoceptive and exteroceptive ones. It has been even called the "sixth sense" (Lenggenhager & Lopez, 2015). Vestibular sense has been argued to influence affective processes (Mast et al., 2014). *Anosognosia* (Turnbull et al., 2014) and *unrealistic optimism* (McKay et al., 2013) are (temporary) attenuated by caloric vestibular stimulation (CVS). CVS is also argued to "modulate the alternation rate in binocular rivalry" (Mast et al., 2014, p. 8; for the connection between binocular rivalry and self-deception see 4.2.2). Turnbull et al. (2014), on the premise that vestibular sense influences both affective and spatial processing of information, view anosognosia as a "dynamic, emotional by-product of a cognitive deficit" (p. 24) in "veridical spatial cognition" (p. 21). They argue that anosognosics perceive the world in an egocentric fashion (how one wants the world to be). This opens up the discussion about the influence of vestibular sense on higher-order motivated cognition in general and self-deception in particular, not least in virtue of conceptual similarities between anosognosia, unrealistic optimism and self-deception (motivated nature and biased processing of information; see section 1.2.7). First question to ask is by which means vestibular sense might influence self-deception and a possible answer is that it happens via the influence of interoception. Lenggenhager and Lopez (2015) voice a hypothesis that vestibular stimulation might influence interoception[237] (p. 14). Not only affective, but spatial influence might be at work as well, similarly to Turnbull's et al (2014) anosognosia explanation. For this speaks the fact that emotional processing is tied to the spatial aspect also independently of vestibular sense: Emotional attentional blink, which is the phenomenon that affective target precludes detection of a subsequent target for a short period of time, needed to disengage from it (Schwabe et al., 2011, McHugo et al., 2013), is tied to a *spatially specific location* (McHugo et al., 2013). Spatial specificity can now, as Turnbull et al. proposed, be linked to attentional selectivity, in order to conclude that emotional attentional blinks allows for selection of information that might aid in deceiving oneself. Emotions have been so far argued to play different roles in self-deception, e.g. cause, function, content, mechanism (1.2.5). I think that in case of vestibular affective regulation we are talking about a kind of a mechanism that might *influence* or *substitute* the self-deceptive process that occurs in the absence of vestibular influence. If this is the case, then this kind of influence is more similar to that of *cognitive penetration*, e.g. a certain affective process influencing a certain other, maybe belief-forming one, or affective information might be more tightly embedded into the self-deceptive process, e.g. if one accepts Jackendoff's (2012) claim that every thought possesses an affective component. Summing up, vestibular sense might provide an affective kind of regulation on higher-order motivated cognition that influences potential affective consequences of belief-like states. In virtue of enabling such regulation, vestibular sense provides an effortless form of (momentary) *control* without a sense of agency. Wu (2011) argues that, at least for bodily actions, "[a]gentive control is typically a dynamic process essentially characterized by perceptual selection and compensatory movement in a way that must conform to the agent's current goals" (p. 66).

[237] It should be noted that, as vestibular sense seems to provide a regulatory tool for interoception, interoception itself provides another kind of a regulatory tool. Barttfeld et al. (2013) describe interoception as attention to internal states and metacognition - as attention to internal thoughts and feelings and they have tested the brain circuitry in both to overlap. Herwig et al. (2010) argue that already an awareness of being in an emotional state can attenuate arousal. Füstös et al. (2013) have tested the relationship between interoception, measured via heartbeat perception, and emotion regulation, as self-reported amount of reappraisal of emotionally loaded photos. The results are similar to Herwig's et al. in that interoceptive awareness aided in down-regulating affect-related arousal. Füstös et al. (2013) follow that interoceptive awareness might be a protective factor for the psychological well-being. To mention is also that self-deception has been also argued to constitute a protective factor for the psychological well-being as well (e.g., Taylor, 1989).

There is, on the contrary, empirical evidence *against* a successful regulatory quality of interoceptive awareness. Domschke et al. (2010) have reviewed literature on the relationship between interoceptive sensitivity (hypervigilance of somatic sensations, e.g. heartbeat) and anxiety and have come to the conclusion that the evidence favors an "increased cardiac interoceptive sensitivity in anxiety-related phenotypes" (p. 7). Dunn et al. (2010) have tested the link between interoception (heartbeat perception) on the one hand and intuitive (understood as automatic, emotional judgment) decision making on the other. The task was akin to the Iowa gambling task: choose profitable from nonprofitable card decks over a certain amount of trials. Their results indicate that "superior interoception actually hindered successful intuitive learning when somatic markers favored maladaptive choices" (Dunn et al., 2010, p. 1842).

In the context of the discussion above, it is to be taken into account that the distinction between the following notions (see Garfinkel et al., 2015) is not always clearcut

- interoceptive accuracy (performance on objective tests, e.g. heartbeat detection)
- interoceptive sensibility (self-reported assessment of interoceptive ability)
- interoceptive awareness (correspondence between confidence and discriminative ability).

If all those interoceptive effects of CVS are the result of the same mechanism responsible for cognitive effects seen in anosognosia and optimism, then this mechanism is of a *subpersonal, interoceptive* kind. Further, if in virtue of conceptual similarities with anosognosia and optimism (1.2.7) self-deception is also argued to underlie the workings of this mechanism, then this offers additional support for the claim that the analysis of self-deception should be given in subpersonal terms, as well as that affective information may play a more important explanatory role than it has been assumed so far.[238]

I see two ways in which the given regulatory function of vestibular sense might be incorporated into predictive coding:

- By working with *counterfactual richness* on which the *sense of presence* is argued to depend (Seth, 2015b). Counterfactual richness is the amount of relations between potential actions and their expected sensory consequences. If affective consequences are also part of sensorimotor contingencies, attenuation of unrealistic optimism by vestibular stimulation, thus, might be explained as a case in which the participants' sense of presence in the current environment is attenuated (by means of attenuating the affective connections to it) so that another cognitive perspective on different circumstances is made possible.
- By working with *beliefs about action* (control) that possess content (expectations) and uncertainty (precision) and are argued to constitute the *sense of agency* in virtue of the claim that the latter arises out of prior beliefs that agents will minimize the divergence between controlled and desired outcomes (Friston et al., 2013, p. 6). If vestibular stimulation provides a subpersonal form of control without the sense of agency and may substitute forms of control with a sense of agency, then the question arises of how one can distinguish control with and without sense of agency in predictive coding.

All three of these aspect might be related: feelings might arise if transition probabilities between control states are violated and feelings might depend on the presence of certain counterfactual states (see section 2.1.3). I will say more about this in chapter 4.

3.1.2.2 Redefinition of the self-concept

> Whatever one's self actually "is" – a structure or a process –
> it is a compromise between stability and change.
> (Greve & Wentura, 2010, p. 721)

In the previous section I introduced the idea that self-deception serves to defend the self-concept by means of selective updating of incoming information. Greve & Wentura (2003, 2010) offer an account of self-immunization that similarly involves the issue of self-concept change. Changing the *content* of the self-concept is though only one possibility in which self-concept can be influenced. Slotter et al. (2010), for example, argue on the example of romantic break-up for three kinds of changes to the self-concept:

- *Self-concept content change* – a reconstruction of the content of the self-concept.
- *Diminuition in self-concept clarity* – degree to which certain self-aspects are held with certainty and conviction; Reduced concept-clarity further uniquely contributed to emotional distress.
- *Self-concept constriction* – reduction of self-concept size.

[238] I do not want to draw any similarities between hypnosis and self-deception, but it is interesting that results of Bryant & Kourch's (2001) study indicate that hypnosis can inhibit emotional responses (feelings). Another curious piece of data is that self-enhancement seems to influence cardiovascular function (a low-level physiological reactionWhy and Huang (2011). Affective information influencing higher-level cognition, is an idea orthogonal to the one that self-enhancement (as a high-level cognitive phenomenon) influences low-level physiological reactions.

Greve & Wentura (2010) concentrate on self-concept change, yet self-concept clarity may play a more important role in self-deception. Particularly interesting is the fact that in paranoia, which like self-deception has been argued to have a defensive function of self-concept protection, instability of the self-concept might play a more important role than its content, especially in the light of the fact that the hypothesis that paranoia would be associated with positive conscious self-concepts has not been supported (Tiernan et al., 2014, pp. 310–311). Greve & Wentura (2010), on the basis of the idea that for creation of a coherent set of beliefs about oneself two different needs compete among each other – need for stability and need for change, argue that paradox-free self-deception[239] is to be explained as *self-immunization*. The idea that "the self-concept probably needs not be as accurate as possible, but certainly as realistic as necessary" (Greve & Wentura, 2010, p. 721) can be found in Taylor (1989).[240] Self-relevant information is, according to the authors, susceptible to the process of self-immunization, which in itself is a self-defense mechanism (Greve & Wentura, 2010). This self-defense mechanism has, according to them, an aim of reconciling the reality with the pleasure principle. It protects self-esteem (p. 725) without the price of disregarding reality, the latter being the case in denial.[241] One of the characteristics of being a successful defense mechanism is, according to Greve & Wentura (2010), that self-deceivers cannot be asked about them, having biased their self-concept directly (p. 723). *Self-immunization* is the phenomenon, which consists in participants changing the description of a certain psychological trait that is important to them, in order to be able to further ascribe this desirable trait to themselves in the cases in which their prior description of the trait does not fit anymore. Adherence to reality is secured by modifying the diagnosticity of a particular skill for a certain domain (Greve & Wentura, 2010, p. 723) in the face of one's subjective competence in performing that skill (p. 724). Though Greve & Wentura do not define diagnosticity, their use of the term suggests that they mean by it the subjective rating of importance/centrality of a trait for a certain domain:[242]

[239] Greve & Wentura (2010) argue that self-deception is a sample case of a general problem of the *pleasure/reality dilemma of self-defense* which can be solved by self-immunization, instead of self-deception: "Thus, self-stabilization can be attained without ignoring realities (with respect to concrete skills or experiences) and without running into any "paradoxes" of self-deception" (p. 726).

[240] This idea is similar to Taylor's (1989) idea that self-deception as self-enhancement is possible as long as it allows the incorporation of important negative information (for more information see the subchapter below about Taylor's theory of self-deception). Negative information is important, if it is diagnostic (enduring to some life domain) or pervasive (relevant to different life domains) according to Taylor. For further elaboration see next section.

[241] As Greve & Wentura (2010) differentiate three lines of defense: rejecting (denial), neutralizing (techniques of reality negotiation: reappraisal, rationalization etc.) and self-immunization and argue that the first two are prone to deception (p. 722). They do not define self-deception, so it remains open (but highly likable) that they think self-deception to be (at least in part) denial. Concerning the neutralizing line of defense, I cannot see the crucial difference between it and self-immunization, because I hold that both neutralizing and self-immunization are techniques for the interpretation of reality. The only possibility to distinguish these two could be that rejection and neutralization concern denial or reinterpretation of *evidence*, while self-immunization concerns accepting the evidence and modifying the *concept* that is supported by this evidence.

[242] Central traits of the self-concept tend to be kept despite possible contradictory evidence: "If a trait was central for the self-concept, a significant priming effect was found for those skills that participants believed themselves to be good at, but not for those skills that participants believed themselves to be poor at" (Greve & Wentura, 2010, p. 725). This idea resembles what Sanford has written about self-deception two decades ago: "It is more difficult to accommodate my

> If a decreased level of performance in a certain skill […] cannot be denied, but there is a need to maintain the self-concept with respect to a related general domain or trait […], reducing the *diagnosticity* of that particular skill for the general domain solves the problem. (Greve & Wentura, 2010, p. 723; my emphasis)

Michel & Newen (2010) see self-immunization fit into a dual rationality framework: "In *dual rationality*, subjects will keep their rational standards in general and, at the same time, introduce a new standard of evaluation with respect to particular matters of subjective importance." (p. 741; my emphasis) Michel & Newen (2010) argue that a satisfactory account of self-deception has to be psychologically viable, explain the frequent occurrence of self-deception[243] and the persistence despite contradictory criteria evidence.[244] They reject the dual-belief requirement, because having a suspicion is enough (p. 734) and argue that motivational bias is not enough to explain self-deception, because biases involve subintentional manipulation of data, yet, it is the manipulation of the evidential quality of data that is decisive (p. 738). Given these constrains, Michel & Newen (2010) argue that self-deception is a special kind of motivational dominance[245] and that "loose, ordinary folk-talk" definition of self-deception is underdetermined, because "it does not specify how the motivational impact on acceptance is accomplished" (p. 732). The solution to the under-determinacy is, according to them, the specification of the notion of *criterial evidence*:

> Counter-evidence is sufficiently strong if, under a neutral evaluation, the not-*p*-data are clear enough to destroy *p*-confidence. The kind of evidence we have in mind is best characterized as "*criteria evidence*" against *p* […] If **S** evaluates information as criteria evidence against *p*, **S** entertains a background-belief that this type of information makes the truth of *p* strongly unlikely. (Michel & Newen, 2010, p. 734)

Self-deceivers believe, despite criterial evidence, in the contrary and through this, violate rationality standards. Michel & Newen (2010) emphasize rationality in humans. Thus, for explaining self-deception they need some explanation of how it is possible to maintain rationality and to engage in self-deception. They do not postulate a division within the self, but a *division within the standards of evaluation of self-relevant information*. Acknowledgement of the ego-centric belief-system (p. 743) allows Michel & Newen

actual behavior to my self-image than to regard myself as temporarily blind to my real chances" (Sanford, 1988, p. 168).

[243] Self-deception as a familiar, but not paradigmatic belief condition: "Self-deception occurs frequently enough to be familiar, but it still seems the exception rather than a regular condition of belief" (Michel & Newen, 2010, p. 733).

[244] Definition of self-deception as auto-manipulation in subjects who possess the property of being evidence-sensitive: "From what has been said thus far, a core challenge for a theory of self-deception has become clear. Being a commonplace phenomenon, self-deception is a form of auto-manipulation that has to work in rational and evidence-sensitive subjects. At the same time, self-deception, as a matter of fact, requires that evidence be strongly asymmetrical." (Michel & Newen, 2010, p. 734)

[245] "Pattern of Motivational Dominance (PMD)

(1) **S** is of normal intelligence and has normal rational capacities.
(2) **S** is strongly motivated to evaluate *p* as true.
(3) **S** has access to a sufficient amount of relevant information about the subject matter SM and this information is supportive of not-*p*.
(4) **S** honestly accepts the proposition p.
(5) If **S** had been neutral towards SM, the same evidence would have led **S** to accept not-*p* instead of *p*." (Michel & Newen, 2010, p. 732)

Michel & Newen (2010, p. 732) define acceptance as a liberal notion of belief.

(2010) to differentiate between two standards of data evaluation – rational (intersubjective) and pseudo-rational (ego-centric). A rational evaluation differs from a pseudo-rational evaluation, insofar as the first is made on the basis of truth, whereas the second – on the basis of subjective importance of some matter (p. 741). Not every subjectively important matter is susceptible to the pseudo-rational evaluation, but only the matter that provides space for a reinterpretation, thus, is under-determined (p. 740). Irrationality and self-deception of such a judgment consists in the fact that two inconsistent standards of evaluation are held at the same time (p. 741).

I think that Michel & Newen's (2010) a strong rationalist of Greve & Wentura's (2010) concept of self-imunization can actually be substituted by subpersonal selection mechanisms (2.2.1). Knowledge selection is the favorable alternative in the psychological literature. Ghosh & Gilboa (2014) argue that schemata (and the psychological definition of the self-concept is a set of schemata) have a network structure, are formed on the basis of multiple experiential episodes, lack detail to allow a certain degree of generality and are adaptable to accommodate new episodes of experience. They may have a chronological or hierarchical organization, have subschemata that they share with other schemata (cross-connectivity) and induce a certain behavior in subjects (see table 24). vmPFC is the region involved in their activation, which is also supported by the fact that damage to vmPFC causes *confabulation*.

Necessary features	Additional features
Associative network structure	Chronological relationships
Formed from multiple episodes	Hierarchical organization
Lack of detail	Cross-connectivity
Adaptability	Link to contextually appropriate behavior

Table 24. Ghosh & Gilboa: schema-features
Distinctions from Ghosh & Gilboa (2014).

To summarize, in this section I introduced the idea that self-deception might serve to protect the stability of the self-concept by *redefining* it and warned against the danger of overrationalizing the concept of self-immunization. This is one constraint for any convincing theory of SD that it must avoid overrationalization. Section 1.3.4 was devoted to it and, as a result, I argued in the second chapter (2) that self-deceivers are motivated by *subpersonal* goals and that the epistemic agent models, that they construct, should not be equated with the third-person description that observers impose on self-deceivers' epistemic agent models, e.g. that some sort of *reasoning* goes on.

3.1.3 Self-esteem – a central feature of the self-concept

> Research on self-enhancement has followed three distinct waves. The first wave emphasized the relevance of self-belief accuracy in adaptive functioning (Jahoda, 1958). The second wave of research highlighted overly positive self-evaluations as a marker of adaptive functioning (Taylor & Brown, 1988). In contrast, the third wave of research has been based on the premise that both realistic and inflated self-beliefs can qualify as correlates of adaptive functioning (Aspinwall & Taylor, 1992; Robins & Beer, 2001; Sedikides & Rudich, 2002).
> (Sedikides et al., 2002, p. 602)

The second and the third waves will be in the focus of this section. I will first introduce the view that "positive illusions" (self-enhancement, optimism and illusion of control) are self-deceptive and have an offensive function of leading to happiness (section 3.1.3.1). Then I will review the self-enhancement literature (self-enhancement has also been argued to be a kind of self-deception, see section 3.2.2) in which protection of self-concept has been argued to occur via preservation of self-esteem[246] as its central feature (section 3.1.3.2). It might be noticed that there is tension between the offensive function to bring happiness and the defensive function to protect self-esteem. In the next section I will, then, review the idea that defense of self-esteem may actually serve another function - to avoid the anxiety of death (section 3.1.4).

3.1.3.1 Positive illusions lead to happiness

Shelley Taylor was probably one of the first to combine the self-inflation bias and self-deception,[247] but her work is still cited in the literature and is widely referenced.[248] Greve & Wentura's (2010) formulation that "the self-concept probably needs not be as accurate as possible, but certainly as realistic as necessary" (p. 721) best describes Taylor's understanding of self-deception as interpretation of threatening events, coupled with a recognition of the threat.[249] It is salient that her definition tries to capture *tension* present in self-deception and that she accepts that at some level the self-deceived person will possess contradictory information. Taylor does not clearly differentiate between positive *illusion* and *self-deception* (Van Leeuwen, 2009, p. 107, footnote 1) and often uses the terms interchangeably.

According to Taylor (1989), it is important that illusion/self-deception is not possible in respect to any subject, but only in respect to one's *social attributes*, because no objective referents are available for them (p. 242). Illusions further create a *self-fulfilling prophecy* due to "creating the world that we already believe exists" (p. 244). She distinguishes three categories of *illusions*: self-enhancement, exaggerated belief in one's personal control[250] and unrealistic optimism (Taylor, 1989, p. 6; for the distinction between biases, illusions and denial see table 25).

[246] Yamada et al. (2013) suggest though that there is not correlation between self-esteem and "superiority illusion" or "above-average effect" and that the extent of the superiority illusion is determined by resting state functional connectivity: "To further characterize the study subjects, we also measured trait anxiety and self-esteem using the State-Trait Anxiety Inventory […] and Rosenberg Self-Esteem Scale […], respectively. Neither of these measures was correlated with the superiority illusion […]. These findings suggest that the *superiority illusion* measures beliefs, *independent of anxiety or self-esteem*" (Yamada et al., 2013, p. 2; my emphasis).

[247] I write "combine", because it seems not clear if Taylor identifies them.

[248] For example von Hippel & Trivers (2011a, p. 55) cite her in their current article on self-deception; Van Leeuwen (2009) has recently written a critical article about it.

[249] In her words: "In many ways, the healthy mind is a self-deceptive one, as I will attempt to show. At one level, it constructs beneficient interpretations of threatening events that raise self-esteem and promote motivation; yet at another level, it recognizes the threat or challenge that is posed by these events" (Taylor, 1989, p. xi).

[250] The exaggerated illusion in one's control happens only under specific circumstances: in the *implementational* mindset, as contrary to the *deliberative* mindset (Gendler, 2007, pp. 245-247).

Error and bias	Illusions	Denial
"The *terms error and bias*, which one might employ instead, suggest short-term accidental mistakes and distortions, respectively, that might be caused by some careless oversight or other temporary negligence. The *term illusion* in contrast, reflects a broader and more enduring pattern of beliefs." (my emphasis; Taylor 1989, p. 44)	Illusions 1. concern the *self*, 2. refer to the way how the beneficial outcomes are achieved (*illusion of control),* 3. promote mental health due to their *specific content* (beliefs about the self, one's mastery and the future) and not due to the optimism that may be present in them. (Taylor 1989, pp. 43-44)	Denial 1. invokes threat or anxiety, 2. has mixed (positive and negative) consequences, 3. makes no use of the utility of negative information and 4. does not associate with mental health. (Taylor 1989, pp. 131-132)

Table 25. Taylor: distinctions between illusions, bias and denial
Distinctions from Taylor (1989).

Taylor's aim with respect to self-deception/illusion is to show that it is not only *normal*[251] but also *adaptive*, though her definition of adaptive ("promoting rather than undermining mental health", Taylor, 1989, p. 47)[252] is different from the one used in evolutionary psychology.[253] She holds that to prove the adaptivity of illusions, one has to prove that

1. they are positive in their consequences and that
2. they allow making appropriate use of negative information (Taylor, 1989, p. 123).

The conclusion to which Taylor seems to come regarding the first point is that self-deception, as opposed to repression and denial, is not a defensive, but an *offensive mechanism* with the aim for humans to express themselves and court pleasure:[254]

> Thus, with regard to the mental health criteria of happiness and positive self-worth, those with positive illusions seem genuinely to be happy and to think well of themselves. In contrast, those who cope with life via repression report an absence of negative experience, but no corresponding feelings of happiness and self-worth; moreover; maintaining the absence of negative emotion seems to require active management. (Taylor, 1989, p. 129)

With respect to the second point Taylor holds that illusion/self-deception is responsive to the negative information (Taylor, 1989, p. 132), however not to every kind of negative information, only to specific kinds of it,[255] namely

- *diagnostic information*: the one that concerns an enduring quality of the self;
- *pervasive information*: to different domains of life (Taylor, 1989, p. 144).

[251] Her argument for *normality* of self-deception: "The widespread existence of these biases and the ease with which they can be documented suggests that they are normal" (Taylor, 1989, p. 46).

[252] Taylor indentifies several *criteria for mental health* in her book (referring to Jahoda), such as ability to be happy, to hold positive attitudes toward oneself, capacity to develop an autonomous self-regard, mastery in work and social relationships and integration of the forces of personality (Taylor, 1989, pp. 47-48).

[253] At the end of her book she brings up the question of *evolutionary adaptivity* of self-deception, but does not bring an elaborated argument. Taylor's (1989) considerations are as follows: "The fact that positive illusions are typically so much stronger in children than in adults argues against the idea that they are learned adaptations to life" (p. 245).

[254] She is referring to Harold Sackeim on this matter (Taylor, 1989, p. 127).

[255] Self-deception distortes information only as little as possible, but as much as necessary: "*negative information* that is *diagnostic* or *pervasive* in its implications is represented faithfully in consciousness, but as benignly as possible" (Taylor, 1989, pp. 144-145; my emphasis).

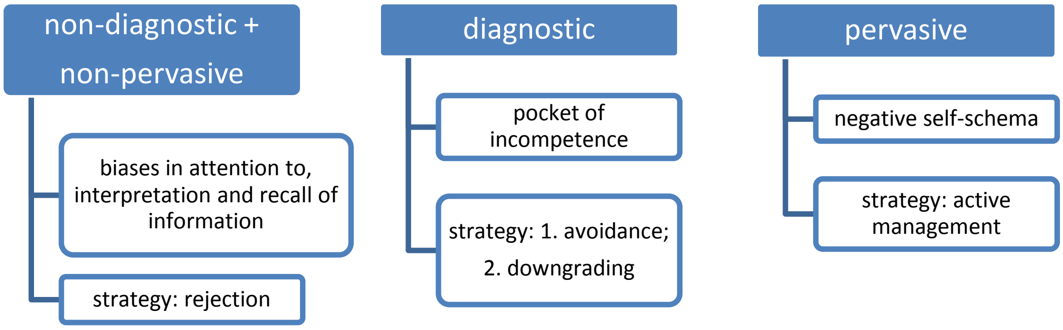

non-diagnostic + non-pervasive
biases in attention to, interpretation and recall of information
strategy: rejection
diagnostic
pocket of incompetence
strategy: 1. avoidance; 2. downgrading
pervasive
negative self-schema
strategy: active management

Figure 16. Taylor: information division
Distinctions from Taylor (1989).

According to Taylor (1989), at least three types of information should be differentiated: information that is both non-diagnostic and non-pervasive, diagnostic information (pervasive or non-pervasive) and pervasive information (diagnostic or non-diagnostic; see figure 16). Only negative non-diagnostic and non-pervasive information is argued to have the possibility of being rejected by biases of attention, interpretation and recall of information.[256] Negative diagnostic and/or pervasive information calls for some sort of acknowledgment of the threat: a creation of a pocket of incompetence (narrowing the negative information on a certain *life domain*[257]) or of a negative self-schema (organizing beliefs about the self on a certain *topic* which can penetrate different life domains in one set[258]). This life domain is mostly not a central one (Taylor, 1989, p. 156). Both a pocket of incompetence and a negative self-schema have in common one feature that they help a person to cut weaknesses from the positive part of one's self-concept (Taylor, 1989, p. 154). They differ in a way of the management. While negative pervasive information necessitates an emergence of the negative self-schema and requires active management[259] (Taylor, 1989, p. 154-155), negative diagnostic information due to its restriction to a certain domain and thus creation of a pocket of incompetence allows

1. an avoidance strategy and
2. the strategy of downgrading the importance/significance of the domain (Taylor, 1989, p. 153).

[256] The workings of biases is dependent on the content that is to be biased – important or not: "Biases in attention to information, the interpretation of information, and the recall of information are successful only if the negative or challenging information itself is not ultimately useful or meaningful" (Taylor, 1989, p. 152). It is to be emphasized that this view is different from the one that self-deception would succeed on any, even life threatening content, e.g. remember the example of a self-deceiving cancer specialist (see section 2.1.2).

[257] Definition in her words: "A *pocket of incompetence* is a domain of life in which a person readily acknowledges a lack of talent and consequently avoids the domain altogether" (Taylor 1989, p. 152; my emphasis).

[258] Definition in her words: "the concept of *self-schema* was described as an organized set of beliefs about the self concerning some trait (such as being witty or kind) or life domain (such as being a musician or historian)" (Taylor 1989, p. 154; my emphasis).

[259] Usefulness of a schema is in the ability to act quickly: "Having a schema for a negative attribute enables people to identify schema-relevant situations quickly and then to prepare for them" (Taylor, 1989, p. 155).

Referring to Greenwald, Taylor (1989) writes that *selective attention and selective memory*[260] are mechanisms through which benign self-deception is possible[261] during which important negative information is able to be recognized (p. 157). To prove the adaptivity of self-deception, Taylor uses the result of Sackeim's and his colleagues' study that points to a negative association between depression and self-deception, the latter being measured using Sackeim & Gur's self-deception questionnaire (Taylor, 1989, p. 159). In a recent article Sheridan et al. (2015) have also tested whether self-deception (measured using self-denial scale of BIDR) influences covitality, which is a construct that encompasses six psychology variables: optimism, hope, self-efficacy, grit (perseverance for long-term goals), gratitude, subjective life-satisfaction. Their results suggest that "[s]elf-deception could be considered a healthy psychological defense mechanism that enhances the effectiveness of these covitality psychological traits by allowing one to maintain a sense of well-being when confronted by the uncertainties of the future" (p. 13). Van Leeuwen criticizes Taylor's idea that self-deception can contribute to happiness. Taylor advocates for the adaptivity of illusions by showing that they promote mental health which has happiness as one of its components. With regard to the relationship of happiness and illusion Taylor (1989) writes:

> Happy people have higher opinions of themselves; they are more likely to make self-serving causal attributions; they are more likely to demonstrate an illusion of control; and they show more unrealistic optimism. Perhaps the point is not even worth discussion. (Taylor, 1989, p. 49)

For Van Leeuwen, the point is worth discussion and even intuitively plausible that the contrary holds (yet Van Leeuwen (2009) does seem to agree that *self-enhancement* is a kind of self-deception, p. 122, footnote 20). Van Leeuwen argues that the only type of happiness that can serve as a candidate for being achieved by self-deception is *matrix happiness* (only positive sentiments without worthwhile external goods).[262] What he argues for is an even more restricted view that a self-deception is not conducive to happiness at all:

> Having true beliefs enables people to accomplish things. I need to believe what my options are before I can choose among them. Having beliefs about what the outcomes of chosen options will be guides choice. If any of the beliefs mentioned in this general schema is false – either about the options or outcomes – then I'm likely to end up dissatisfied. So false belief tends towards dissatisfaction. But self-deception leads to false belief. So self-deception leads toward dissatisfaction. Dissatisfaction tends away from happiness (both Matrix and choiceworthy). So self-deception tends away from happiness. (Van Leeuwen, 2009, p. 120)

[260] Selective attention is preconscious and selective memory is conscious according to Taylor (1989, p. 158). For Taylor, selective memory is possible due to its *organization by association:* "With positive information stored through multiple pathways and associations for remembering and negative information stored with perhaps a few out-of-the-way associations, it is easy to see how memory actively fosters the self-aggrandizing self" (Taylor, 1989, p. 158).

[261] Selective attention and memory as mechanisms for self-deception: "Through the twin mechanisms of selective attention and selective memory, it is possible to self-deceive not only successfully but adaptively" (Taylor, 1989, p. 157).

[262] Van Leeuwen categorizes types of happiness using two variables – positive sentiments and the lack of worthwhile external goods. Consequently, four types of happiness arise with *choiceworthy-happiness* (positive sentiments plus worthwhile external goods) on the top as the most desirable (Van Leeuwen, 2009, pp. 108-111).

According to Van Leeuwen, both if situation is controllable and if it's not, honesty is better than self-deception, because only knowing ones limitation can help to compensate for them. Furthermore, self-deception hinders the recognition of the type of the situation – controllable or not – due to the neglect of evidence (Van Leeuwen, 2009, p. 124). If one also takes into account that his concept of self-deception involves epistemic tension, then self-deception understood as illusion will, probably, never be an option to reach happiness, independently from the situation (controllable or not). Van Leeuwen's (2009) argumentations depends heavily on the conceptual distinctions that he makes (see table 26). For him, even if a representation deviates from reality, as long as it is helpful in practical actions, it is not an illusion, for *imaginings* help to plan actions successfully:

> By lumping under the single label of 'positive illusions' both self-inflated positive beliefs and imagined future outcomes that are useful for planning, Taylor has done a disservice. Imagining can be valuable, even in cases in which the contents of one's imagining depart widely from reality. (Van Leeuwen, 2009, p. 117)

Another debatable conceptual distinction is that Van Leeuwen further holds that self-deception cannot become a self-fulfilling prophecy. Young (2013), for example, argues that empirical evidence that positive illusions lead to psychological well-being (e.g., happiness), physical well-being (e.g., terminally-ill living longer) and ability to be adaptive (e.g., facilitation of achieving goals in the implementational mindset) is at best inconclusive. Among conceptual criticism Young (2013) mentions is that given that optimistic beliefs are directed at the future, it is difficult to determine whether they are true or false, exactly due to the possibility of a self-fulfilling prophecy (p. 550).

Self-deception vs. self-inflation bias	**Self-inflation bias vs. sense of self-worth**	**Illusions vs. self-fulfilling confidence**	**Illusions vs. future-oriented imaginings**
Distinction criteria: 1. *Specificity* of SD; 2. *Epistemic tension* present in SD.[263] • *Self-inflation bias*: an overly positive phenomenon; • *Self-deception*: requires epistemic tension which logically	• All humans have a sense of self-worth, independently of their faults, and are justified in believing that to be the case (p. 119); • Only the sense of self-worth is conducive to happiness,	• *Falsity*: Even if one reckons that illusions cause positive consequences (and not merely correlate with them[264]), they cannot be self-fulfilling, because an illusion is something that does not turn out true (p. 115);	• *False* beliefs: Even if one envisions unrealistic outcomes that are fulfilled only to some degree, that state cannot be called an illusion, because it does not need to contain any false beliefs, but only future-oriented imagining (p. 115-116); • *Honestly attending to evidence* is the crucial difference between imagining and illusions.

[263] Here the difference between self-inflation bias and self-deception in Van Leeuwen's words: "This [self-inflation bias] is simply a *general* tendency that many or most people have to see themselves in a comparatively more positive light than evidence would justify. […] The self-inflation bias seems to me to involve a **fixed *frame* for interpreting incoming information** in a self-flattering way. Self deception, on the contrary, involves motivated selection of different frames. Furthermore, self-deception and self-inflation bias differ in that the former phenomenon, but not the latter, involves an **epistemic tension** in the mind of the agent between the evidence cognized and the product of self-deception; in the self-inflation bias, on the contrary, the agent is rather wholehearted in holding the flattering picture of herself" (Van Leeuwen, 2007a, p. 425, footnote 15; my bold emphasis).

[264] In her book Taylor actually holds that causality between positive illusions and happiness is bidirectional (Taylor, 1989, p. 52).

has to come from acknowledging somehow the negative information.	while at the same time not being illusive.	• *Practical setting*: If something is used for planning of an action, it can be non-conforming to reality, but still not an illusion either[265] (p. 116).	Imagining involves attending to all aspects of the evidence and judging them by usual epistemic standards, then building a happy end state and a route to this state via imagining.

Table 26. Van Leeuwen: conceptual distinctions between phenomena related to SD. Distinctions from Van Leeuwen (2009).

It is worth mentioning how much influence terminological distinctions have on the supposed function of self-deception. Metcalfe (1998), for example, upon accepting Fingarette's intentional theory of self-deception (1.1.1.3), argues that self-deception has an ego-enhancing or stress-reducing function. Instead of self-deception, Metcalfe (1998) explains phenomena of overconfidence by a memory-based processing heuristic (pp. 105-107), exemplified by the fact that repetition increases the judged validity of sentences (Metcalfe, 1998, p. 107):

> By this view, to the extent that people do not have perfect memory, and the imperfections include both gaps and mistaken information, or even information that is close to correct but not exactly correct, their metamemory judgements are bound to be both inaccurate and systematically biased. (Metcalfe, 1998, p. 106)

Cummins & Nistico (2002), on the contrary, argue that positive illusions (termed by them 'positive cognitive biases') are responsible for generally high level of life satisfaction due to their non-specific (not concerning specific skills) and empirically unfalsifiable nature, e.g. consider judging (according to which objective criteria?) the case of being more friendly or interesting than others (p. 51). The authors further argue that positive illusions are created by cognitive schemata that serves as a basis for accommodating new information (p. 60).

All in all, the claim that self-deception is conducive to happiness depends on the degree to which one allows negative information to be acknowledged in self-deception: if not at all, then self-deception might be argued to lead to unhappiness (Triandis, 2009, p. 18). If self-deception is defined as involving tension and contradictory knowledge, then per definition it would be effort consuming, because effort would be needed to deal with the contradiction and inconsistency. It is much easier, on the contrary, to envision a tensionless and, thus, blissful self-deceiver that achieves happiness in virtue of her self-deception. Trivers (2011), whose evolutionary theory of self-deception is to be discussed in section 3.2, pins down as the general rule: "the earlier during information processing that self-deception occurs, the less its negative down-stream immunological effects". So whether self-deception sustains health and happiness is a definitional issue. Consider the claims that repression decreases immune function (Trivers, 2011, p. 123), denial of homosexuality is HIV-progression-

[265] Reservoirs of representations as setting-dependent storages of knowledge and beliefs: "individuals operate with more than one reservoir of representations, and they switch, depending upon what kind of practical setting they're in. But the tendency to label as 'illusion' every reservoir that's not entirely reality-tracking is a mistake" (Van Leeuwen, 2009, p. 116). It is to be emphasized that the assumption that reservoirs of representations are independent from each other is a tool that can be used to analyze inconsistent behavior: there is not only *one* set of representations governed by rationality constraints, but several sets that guide action depending on context. This does not solve the memory problem though – subjects would remember inconsistent behavior and would need to explain in nevertheless.

costly, writing about trauma improves immune function, but also that positive affect and music are immunologically beneficial (Trivers, 2011, chapter 6). Whether self-deception is healthy or not depends on our assumption about how information is argued to be processes: Sustained inconsistency not only leads to tension, but also it is something that is immunologically costly.

3.1.3.2 Self-enhancement and self-protection

It is not only Taylor (1989) who holds that self-enhancement is a kind of self-deception. Von Hippel & Trivers (2011a,b) have also recently argued this to be the case, as well as self-enhancement to be susceptible to the self-affirmation paradigm (see section 3.2.2.3). Self-enhancement is an instance of self-evaluation motivation where the latter includes processing and evaluation of self-relevant information (Sedikides & Alicke, 2012, p. 303). In this section I want to briefly put self-enhancement into context by introducing other kinds of self-evaluation motivations. After this I will mention the biases that self-enhancement has been argued to evoke and the intra- and interpersonal restraints that it is susceptible to. Finally, I will point out the evidence that self-enhancement may serve to protect self-esteem.

Five self-evaluation motives have been distinguished: self-enhancement, self-protection, self-assessment, self-verification and self-improvement (see figure 17). Self-protection and self-enhancement are actually two sides of the same coin:

- *self-protection*: “avoiding, minimizing, misinterpreting, or discarding information that has unfavorable implications for the self”;
- *self-enhancement*: “pursuing, magnifying, overinterpreting, or fully endorsing information that has favorable implications for the self.” (Sedikides & Skowronski, 2012, p. 239)

Self-enhancement/self-protection strivings are susceptible to self-threat and self-affirmation paradigm, which means that under self-threat one tends to enhance one’s qualities and under self-affirmation biased processing of information is attenuated (Sedikides & Alicke 2012). This is where self-enhancement gets its defensive connotation from. The other kinds of self-evaluations are argued to be less important (Sedikides, 2007; Sedikides et al., 2004), because e.g. people use highly diagnostic questions to determine whether they possess positive, but not negative attributes; they show poorer recall of negative and important self-related attributes and people with a negative self-concept also show superior recall of positive self-related attributes, in comparison to those with a positive self-concept.

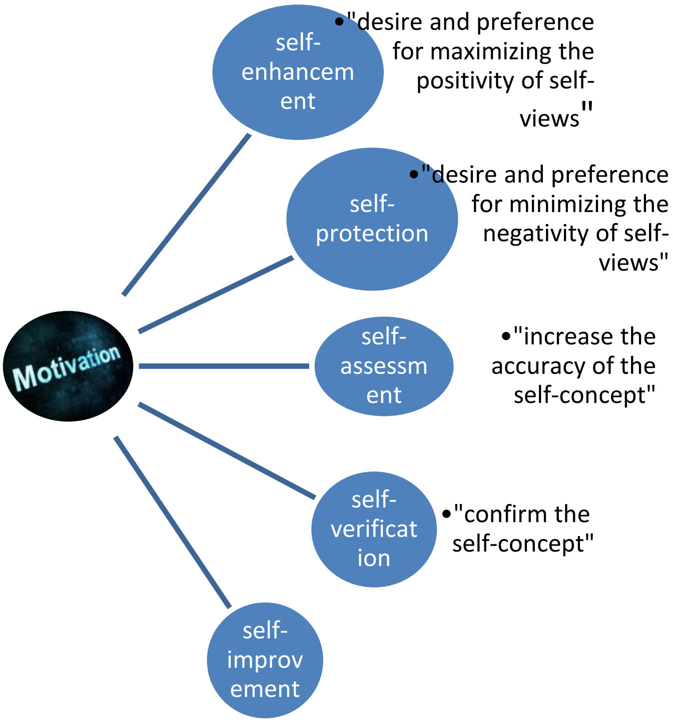

Motivation
self-enhancement
•"desire and preference for maximizing the positivity of self-views"
self-protection
•"desire and preference for minimizing the negativity of self-views"
self-assessment
•"increase the accuracy of the self-concept"
self-verification
•"confirm the self-concept"
self-improvement

Figure 17. Sedikides: kinds of self-evaluation motivation
Distinctions from Sedikides & Alicke (2012, p. 316), Sedikides et al. (2004).

Similarly to Baumeister & Newman (1994) who in the context of self-deception argued that the unfulfilled goal to reach a favorable conclusion evokes negative affect, Sedikides (2007) also holds that self-enhancement/self-protection motivation has *affective consequences* – positive or negative, dependent on whether one's goal has been achieved (pp. 2-3). This is probably a general assumption about goals in general: if they are unfulfilled, they trigger negative affect. Similarly to Taylor (1989) who argued that positive illusions lead to happiness (see last section), Sedikides & Alicke (2012) argue that self-enhancement and self-protection motivation are beneficial for the psychological health, e.g. promote problem solving, optimism (a kind of "psychological resource," Sedikides, 2007, p. 9), positive mood, absence of depression and distress (pp. 314-316).

Further, those authors who accept that self-enhancement is a kind of self-deception, also accept biases that self-enhancement motivation can bring about (see table 27) as self-deceptive. Trivers (2011), for example, mentions self-handicapping, out-group derogation, moral hypocrisy also as self-deceptive phenomena. Van Leeuwen (2013b) has criticized Trivers for this broad application of the notion "self-deception," yet, the fact that the abundance of biases can be brought about by self-enhancement/self-protection strivings shows that the root of the problem, emphasized by Van Leeuwen, lies in the categorization of self-enhancement as a kind of self-deception. If *every* kind of self-enhancement is an instance of self-deception, then all the biases that can be brought about by self-enhancement are also self-deceptive. What kind of restriction could be usefully made here, without it being artificial? As I argued in chapter 2, one would need to refine the behavioral and phenomenological profile of self-deceivers and test whether these biases satisfy the given profile.

i. *positivity embracement:* self-enhancement related to evaluation of feedback and success expectations, e.g.
    - *self-serving attributions for success*
    - *engagement in self-promoting social interactions*
    - *enhanced remembering of favorable feedback in comparison to unfavorable one*

ii. *favorable construals*: emphasis on flattering self-characteristics in social situations, e.g.
    - *positive illusions*
    - *better-than-average effect* (tendency to "regard the self as superior to others in many domains of functioning", Sedikides & Alicke 2012, p. 305); factors that constrain the BTAE:
        - *attribute valuence and controllability:* whether an attribute is positive and valuable
        - *attribute importance*: interpersonal or personal importance of an attribute
        - *attribute verifiability/ambiguity*: whether there are criteria to objectively verify the ascription of an attribute
    - *favorable interpretations of ambiguous/negative feedback*

iii. *defensiveness:* self-protection from threat, e.g.
    - *self-handicapping*
    - *defensive pessimism*
    - *out-group derogation*
    - *moral hypocrisy*
    - *self-serving attributions for failure*

iv. *self-affirming reflections:* in the face of (potential) negative outcomes positive self-views are still held, e.g.
    - *downward counterfactual thinking*
    - *temporal comparison*
    - *focusing on strengths, values, or relationships*
    - *selective self-memory* ("disadvantageous recall for negative as opposed to positive feedback," Sedikides & Alicke 2012, p. 305)

Table 27. Sedikides: categorization of self-enhancement/self-protection strivings
Distinctions from Sedikides & Skowronski (2012, p. 240), Sedikides & Alicke (2012, p. 305), Sedikides (2007)

Sedikides & Alicke (2012), as well as Sedikides & Skowronski (2012) adopt results of Hepper et al.'s factor analyses that categorize self-enhancement/self-protection strivings (or a "repertoire cognitions, emotions and behavior" characteristic for self-enhancement/self-protection) into four "families" or categories

The limitations that self-protection/self-enhancement strivings are susceptible to also relate in a straightforward way to those discussed in the self-deception literature. Self-enhancement argued to be restrained by relational and social context, as well as by self-focus, introspection, flexibility in the construction of the self-concept and variation in behavior-based ascription about the self-concept (see table 28). To draw the parallel to self-deception, Sloman et al. (2010) argued that self-deception requires vagueness (see section 1.3.2.3), while self-enhancement is also argued to be restricted by social context such that only attributes difficult to verify may be subject to it in order to avoid accountability and identifiability. Sedikides et al. (2002) cite empirical reference that self-enhancement is controllable (p. 593) and argue that *accountability* reduces relative self-enhancement (difference between self-evaluations of accountable and unaccountable participants, p. 594) where accountability is defined as "participants' expectation that they will be called on to explain, justify, and defend their self-evaluations to one or more others (termed 'audience')" (p. 593). In this particular experiment self-enhancement was measured as the difference between the grades that accountable/unaccountable participants gave to their own previously written essays. A further experiment proposed the subjective perception of *identifiability* (p. 597) as a mechanism that was responsible for the effect and that was defined as "whether participants believe that their self-evaluations can be linked to them personally" (p. 596). The difference between accountable/identifiable and accountable/unidentifiable conditions is that in the latter condition a reviewer of the essay would not know the name of the participant who has written it. A follow-up experiment clarified that *evaluation expectancy* defined as expectancy "to be *evaluated* by a (high-status) audience" (Sedikides et al., 2002, p. 598) influences the amount of self-enhancement and leads to the focus on one's weaknesses in the self-evaluation. Which conclusion should one draw from these findings with respect to self-deception and self-enhancement? Is ambiguity of interpretation a similar characteristic of self-deception and self-enhancement or is it that, because of the psychological understanding of self-deception as self-enhancement, all attributes associated with self-enhancement are now transferred to self-deception? To remind you, in Sloman's et al (2010) study participants' motivation was to find out that they are intelligent. It may even be the case that both are true. Justification of behavior is a conscious, controlled activity such that all kinds of information *accessible* to the agent will be taken into account. Independently of what phenomenon one wants to test – self-deception or self-enhancement – in case of the need to justify themselves participants will behave in a similar manner. If this is true though, then this is additional evidence that agentive self-deception is hardly possible, since self-deceivers would need to bypass conscious deliberation as agents in a controlled manner. The connectionist alternative favored in the psychological literature is knowledge selection (see section 2.2.1):

> For example, parallel distributed processing conceptions of self-representations suggest that self-concepts are recomputed each time a self-judgment is required, and such recomputations are influences by those portions of the self-evidentiary base made *accessible* by situationally triggered *constructs* (Van Overwalle & Labiouse, 2004). (Sedikides & Skowronski, 2012, pp. 243-244; my emphasis).

The idea that attitudes in general and beliefs in particular are constructed on the spot from the available context has also been recently brought up as an alternative to dispositional accounts by Michel (2014; section 1.2.3). Sedikides & Skowronski (2012) widen the scope of the constructionist idea to self-concepts. What the constructionist idea does not explain, though, is the appearance of *controllability* in self-deception – that one does not self-deceive about *every* kind of content, but only *particular* kinds of content (see selectivity problem in section 1.1.3). Priming, for example, seems to make every kind of content accessible. Experimental results suggest that in case of *subliminal priming* with traits "helpful" or "dishonest," those participants that had to invent a story and that subsequently rated the degree to which they possessed the primed traits, judged themselves to be more "helpful" and less "dishonest" than did observers who viewed videos of other participants' telling invented stories (Sedikides & Skowronski, 2012). At the same time priming also led to the effect that "[p]articipants who were primed with the construct *dishonest* (vs. helpful or neutral) rated themselves higher on dishonesty traits than on either helpful or neutral traits, and so did observers" (Sedikides & Skowronski, 2012, p. 250). Following Mele (2001), one might say that not priming per se, but *motivated* priming leads to self-deception (or in this case self-enhancement). In every case, though, I do not think that unsuitability of priming per se as a subpersonal means to evoke self-deception should be taken as an arguments against the suitability of *every* kind of subpersonal mechanism to cause self-deception.

- *Interpersonal limits:*
  - *Relational context:* Self-enhancement may be reduced with respect to close others.
  - *Social context:* Attributes to self-enhance must be difficult to verify because of possible accountability and identifiability.
- *Intrapersonal limits:*
  - *Self-focus*, Internal focus diminishes self-enhancement, presumably because of highlighting one's inner standards and discrepancies.
  - *Introspection* as a special case of *self-focus:* Introspection reduces self-enhancement, presumably because it reduces self-certainty and heightens accessibility of negative traits.
  - *Flexibility in self-thought/construct accessibility:* Self-concept is constructed each time self-evaluation is being made on the basis of accessible knowledge.
  - *Variation in behavior-based inferences about the self:* Self-enhancement/self-protection strivings change self-related inferences about traits and dispositions from behavior, if they can "bypass *deliberative* processing of self-behaviors," e.g. via priming (p. 251; my emphasis).

Table 28. Sedikides & Skowronski: restrains on self-enhancement/self-protection. Distinctions from Sedikides & Skowronski (2012).

In the first part of this section I was concerned with the relation between self-enhancement and self-deception. I have shown that there are similarities between the two, though it is unclear whether those similarities are similarities between *phenomena*, or whether they point to the fact that our *understanding* of those phenomena is being assimilated. In the second part of this section I want to explore the claim that self-enhancement is about the protection of *self-esteem*. Sedikides (2007) argues that valuation motives (self-enhancement and self-protection) are *pancultural*, if a distinction between individualistic and collectivistic traits is taken into account and that results of administering measures of global self-esteem to participants – using Rosenberg Self-Esteem Scale - could serve as a proof for that claim.[266] Beer et al. (2013) have recently argued that "self-enhancement is the most parsimonious explanation of social comparisons (with an average peer) in the face of *self-esteem* threat as well as conditions of limited cognitive ressources" (my emphasis; p. 710). Rosenberg et al. (1995) also claim that "[m]aintenance of self-esteem leads to *self-protective motives*, *self-enhancement processes*, and a variety of coping processes" (p. 145; my emphasis). Crocker & Park (2004) state that self-protective or self-enhancing strategies, susceptibility to self-affirmation manipulations are indirect ways to measure the presence of self-esteem goals in the face of absence of direct measures (p. 394). Tesser et al. (1996) also assume that *self-defense or self-protective* mechanisms are those that concern *self-esteem/self-evaluation*[267] (p. 50). Their aim is to show that these defensive mechanisms have the same goal of maintaining self-esteem (question about "one versus many self-esteems", cf. 56). If it is the same self-esteem that these processes maintain, then they should be interchangeable. This is what results of the experiments, they mention, point to, specifically that dissonance processes can be substituted not only by self-affirmation, but

[266] A common way to measure self-esteem is using Rosenberg Self-Esteem Scale that has been analyzed to contain "two components – self-confidence and self-deprecation" (Rosenberg et al., 1995, p. 142). Concerning panculturality of self-enhancement, the results indicate that "[c]onsistently with a universal perspective on the valuation motives, (a) all nations scored above the midpoint of the scale, manifesting positive self-evaluations, (b) in all nations self-esteem was correlated with the same variables (i.e., extraversion, neuroticism, romantic attachment styles), and (c) the factor structure of the scale was virtually identical across nations" (Sedikides, 2007, p. 9).

[267] They define self-esteem as "chronic individual differences in feelings about the self" and self-evaluation as "acute, situationally induced differences in such feelings" (Tesser et al., 1996, p. 64).

also by SEM[268] processes (Tesser et al., 1996, p. 59). Moreover, they hypothesize, based on the results of experiments that used the misattribution paradigm, that *affect*, whose origin is unknown, mediates between these defensive processes (p. 63). Schachter & Singer's (1962) misattribution paradigm was developed to show that participants attribute emotions to themselves on the basis of cognitive interpretation of arousal (3.1.1). I already noted that self-deception and cognitive dissonance are argued to possess similarly vague phenomenology – that of uneasiness and anxiety (3.1.1). Similar phenomenology (negative affect) is also ascribed in cases when goal acquisition fails (see section 2.1.3 and 2.2.1). If affect mediates between different defensive phenomena and self-deception is one of them, then this should make us cautious of labelling different mechanisms as 'self-deceptive' only in virtue of them being triggered by negative affect. Tesser et al. (1996) root their assumption about the substitutability of self-esteem processes in research on goal-tension and goal completion (pp. 51-55). They mention Lewin's hypothesis that "goal sets up a tension state, which remains until the goal is satisfied" (p. 51) and studies of Lewin's student Bluma Zeigarnik that attempt to prove this hypothesis:

> [I]f (as Lewin suggested) the adoption of a goal sets up a tension state that remains when the task is interrupted, this tension should have kept the *accessibility* of the incompleted tasks high, and subjects should have had a better memory for the incompleted tasks than the completed tasks. This was exactly what Zeigarnik found. Indeed, a better memory for incompleted tasks has come to be known as the "Zeigarnik effect." (Tesser et al., 1996, p. 51; my emphasis)

Tesser & Cornell (1991) further hypothesize that self-related processes may build a hierarchy, given their results "that self affirmation affected SEM and that SEM affected dissonance reduction" (p. 520). Not every kind of self-esteem (see figure 18) may be related to these defensive mechanisms. Maintenance of high *implicit* self-esteem has been argued to be one candidate[269] (Zeigler-Hill, 2006). Along these lines, Zeigler-Hill (2006) applies hierarchical multiple regression to participants' scores on measures of *explicit* (Rosenberg Self-Esteem Scale), *implicit self-esteem* and *self-esteem instability*. Self-esteem instability is defined as the deviation across time of explicit self-esteem scores (p. 129) and is argued to reflect the difference between state and trait properties of self-esteem (p. 137). Zeigler-Hill (2006) wanted to verify the association between discrepant self-esteem (high explicit and low implicit self-esteem) on the one hand and *narcissism*[270] and self-esteem instability on the other. The results point into the direction of the association between discrepant high self-esteem, narcissism and self-esteem instability. The conclusions that the author draws from these results is that discrepant high self-esteem may be coupled to the possession of *self-doubt* and *insecurities* by the individual and further that the results complement the

[268] The *self-evaluation maintenance (SEM) model* developed by Tesser is described as follows: "When a performance dimension is very important (i.e., relevant to one's self-definition), the self is threatened by comparison to a psychologically close other who is performing well; when a performance dimension is not particularly self-relevant, the self is augmented by basking in the reflected glory of a psychologically close other who is performing well" (Tesser et al., 1996, p. 49).

[269] High implicit self-esteem has been argued to make self-deception obsolete: it protects one's self-concept when the latter is threatened (Zeigler-Hill, 2006, p. 121).

[270] Trzesniewski et al. (2008) also hold self-enhancement, defined as "the tendency to hold unrealistically positive beliefs about the self," to be one manifestation of *narcissism* and operationalize self-enhancement as "the discrepancy between participants' ratings of their intelligence and more objective indicators of their intellectual ability (SAT scores and grade point average, or GPA)" (Trzesniewski et al., 2008, p. 182).

findings of individuals with high discrepant self-esteem exhibiting self-enhancing and self-protecting tendencies (p. 136). Self-doubts and insecurities also belong to the standard conceptual toolkit of self-deception and have resulted in the tension requirement (see chapter 1). In case of self-deception I argued that tension is an *affective*, but not a *cognitive* requirement (2.1.3). This conclusion may also be valid in case of individuals with discrepant self-esteem. More importantly, there may be at least two kinds of relation between self-esteem and self-deception: certain kinds of self-esteem may *cause* self-deception or it may itself be the content of self-deception. For example, high explicit, but low implicit self-esteem, or fragile self-esteem may be argued to be the reason why self-deception occurs in the first place. Self-esteem may also be argued to be the content of self-deception, if what the self-deceiver errs about is how good she is at certain tasks. I think that self-esteem instability may be a candidate for the latter case, not least because of the instability.

- *global self-esteem*: "the individual's positive or negative attitude toward the self as a totality" (Rosenberg et al,. 1995, p. 141)
- *specific self-esteem:* attitude towards certain aspects of the self.
- *secure self-esteem:* "positive attitudes toward the self that are realistic, well-anchored, and resistant to threat" (Zeigler-Hill, 2006, p. 120)
- *fragile self-esteem:* "feelings of self-worth that are vulnerable to challenge, need constant validation, and frequently require some degree of self-deception" (Zeigler-Hill, 2006, p. 120)
- *implicit self-esteem*: "nonconscious, automatic, and overlearned self-evaluations" (Zeigler-Hill, 2006, p. 120)
- *explicit self-esteem*: "result of conscious interpretations or, in many cases, reinterpretations" of one's self-evaluations (Zeigler-Hill, 2006, p. 123)

Figure 18. Kinds of self-esteem
Distinctions from Rosenberg et al. (1995), Zeigler-Hill (2006).

In this section I first argued that self-enhancement and self-deception share similar features, but that the reason for this may also be rooted in the approximation of the concept of self-deception to that of self-enhancement in the psychological literature. Thereafter, I have argued that self-esteem may be the feature that self-enhancement serves to preserve, if self-deception were to have a defensive function at all. In the next section I will explore the possibility that self-deception serves as a buffer against anxiety of death. This hypothesis can be supported either directly by terror-management theory, or by Varki & Brower's evolutionary hypothesis about the role of self-deception from the development of full theory of mind.

3.1.4 Terror-management theory

> Greenberg, Pyszczynski, and Solomon (1986) suggest that
> the need for self-esteem may originate in the terror associated with death.
> (Tesser et al., 1996, p. 49)

In section 3.1.3.2 I was concerned with the relation between self-enhancement, self-deception and self-esteem. Von Hippel & Trivers (2011b) state that in the case that self-deception serves the goal of self-enhancement, susceptibility to self-affirmation manipulation indicates the presence of self-deception (p. 7). Unwarranted degree of self-esteem is one kind of self-enhancement often mentioned in the literature and sometimes these notions are even used interchangeably:

> As noted earlier, people who are high in explicit but low in implicit *self-esteem* are the most defensive and narcissistic (Jordan et al. 2003), and thus it seems likely that *self-enhancement* only brings benefits to the degree that it is believed both consciously and unconsciously. (von Hippel & Trivers, 2011b, p. 15; my emphasis)

Pysczcynski et al. (1999) argue that Steele's self-affirmation theory, e.g. defended in Sherman & Cohen (2002), describes "midrange defensive strategies," because the kinds of defenses tested by it are "more proximal than the symbolic terror management defenses in which cultural beliefs and values are used to protect one from concern about death, which have no semantic or logical ties to the problem of death" (p. 843). As I promised at the beginning of section 3.1, in this section I will, first, discuss the possibility that defense of self-esteem serves the function to relieve anxiety of death and, afterwards, I will introduce Varki & Brower's (2013) idea that anxiety of death is the psychological evolutionary barrier against the development of full ToM and that self-deception developed as a means to overcome this barrier.

Terror-management theory, developed by Pysczcynski et al. (1999), deals with the way in which "concerns about human mortality" (p. 835) affect human behavior. The two basic hypotheses of TMT are that anxiety, associated with thoughts about mortality, motivates people to 1. maintain positive self-images and 2. defend their cultural worldview. They define cultural world-view and self-esteem as follows:

> Along with the evolutionary emergence of cognitive abilities that enabled members of our species to comprehend our own mortality, our ancestors developed a solution to the problem of death in the form of a dual-component cultural anxiety buffer consisting of (a) a cultural worldview – a humanly constructed symbolic conception of reality that imbues life with order, permanence, and stability; a set of standards through which individuals can attain a sense of personal value; and some hope of either literally or symbolically transcending death for those who live up to these standards of value; and (b) self-esteem, which is acquired by believing that one is living up to the standards of value inherent in one's cultural worldview. (Pysczcynski et al., 1999, pp. 835-836)

Self-esteem is, thus, seen as a "protective shield" against death-terror and that it is argued to possess anxiety-buffering properties (Pyszczynski et al., 2004a,b). The authors quote studies that confirm the prediction that "[t]o the extent that self-esteem provides protection against anxiety, then increasing self-esteem should make one less prone to anxiety when later exposed to threatening material" (p. 438). Thus, TMT predicts that high self-esteem individuals are less defensive, although some studies also mentioned by them point towards the conclusion that the opposite is true (p. 439). TMT predictions are tested by means of *mortality salience (MS) hypothesis*: Every structure that helps relieve anxiety induced by reminder of death will be protected.[271] Pyszczynski et al. (2004) argue that "MS leads to cognitive self-esteem bolstering in the form of *self-serving biases*" (p. 448; my bold emphasis). As one of the examples they quote a study by Dechesne, Janssen, and van Knippenberg (2000) which shows that after MS induction positive, but not neutral feedback was seen as more valid (Pyszczynski et al., 2004, p. 448). Results of Burke's et al. (2010) meta-analysis on the effects of MS induction suggest that death is a unique threat in that

[271] Definition of the mortality salience hypothesis: "The *mortality salience (MS) hypothesis* states that to the extent that a psychological structure provides protection against anxiety, then reminding people of the source of their anxiety should lead to an increased need for that structure and thus more positive reactions to things that support it and more negative reactions to things that threaten it" (Pysczcynski et al., 1999, p. 836; my emphasis).

MS effect does not depend on whether a neutral or another not death-related threatening condition has been taken as control condition (p. 186).

In the following I will discuss the question about the importance of self-esteem, point to some conceptual unclarities in terror-management theory and then discuss Varki & Brower's (2013) hypothesis. Is self-esteem a goal that should be pursued? Crocker & Park (2004) deny this. Crocker & Park (2004) argue that increases in *state self-esteem* are pursued because of the goal of validating *self-worth* which is a higher-order goal in the goal – hierarchy (pp. 393 - 394). Self-worth is based according to these authors not on domains we believe we are good in, but those "in which, if they [we] could succeed, they [we] would feel safe and protected from the dangers they perceived in childhood" (attachment style dependency, p. 394). Crocker & Park (2004) emphasize that the short-lived benefits of self-esteem (see table 29) follow only the successful pursuit of the goal, the unsuccessful pursuit, on the contrary, brings the reverse – "intensely negative emotions, increased anxiety, feeling of being at risk of social rejection" (p. 396). The solution to the costs that Crocker & Park (2004) propose is the shift from self-esteem centered goals to those that "are larger than the self or are good for others and the self" (p. 406):

> Self-esteem depends on *perceived* success or failure in those domains on which self-worth is contingent; success and failure in these domains generalize to the worth and value of the whole person […]. (Crocker & Park, 2004, p. 393; my emphasis)

DuBois & Flay (2004) criticize this solution by citing studies pointing that high self-esteem is adaptive, given that self-esteem is pursued "in ways that represent a good fit with the individual's surrounding environment." (p. 417) They conclude that apart from self-esteem per se, the *processes by which it is formed and maintained* (the importance of which Crocker & Park have emphasized), as well as *context* has to be taken into account when considering the consequences of self-esteem for health and well-being (DuBois & Flay, 2004, pp. 418-419).

High self-esteem benefits	Costs of unstable self-esteem
• emotional benefits (positive affect, e.g. pride), • anxiety reduction as terror management theory predicts, • feeling of social inclusion as sociometer theory predicts, • perceived competence and optimism (here the Crocker & Park cite Taylor & Brown), and may also be pursued because of • assumed "other benefits, such as professional or financial success, love or fame" (p. 396)	• *costs to autonomy* – less autonomy and self-determination, where autonomy is the "sense of being the causal origin of one's behavior" or being "the source of motivation," (p. 399) • *costs to learning and competence*: discounting mistakes precludes learning on them, • *costs to relationships*: pursuers of high self-esteem are (perceived) as less supportive and caring by others, • *costs to self-regulation* which involves "restraining impulses to engage in behaviors that have known costs to the self […], as well as the ability to pursue goals that have future benefits,"(p. 402) • *costs to physical health*: anxiety, stress and unhealthy coping behavior • *costs to mental health*: risk factor for depression

Table 29. Crocker & Park: self-esteem - benefits and costs.
Distinctions from Crocker & Park (2004).

Sheldon (2004) also argues against Crocker & Park by taking a self-determination theory (SDT) perspective with respect to the goal of pursing self-esteem. Sheldon (2004) accepts the following conceptual tools from SDT concerning goal evaluation:

- goal's content: intrinsic vs. extrinsic (e.g. money, fame)
- goal's motivation: autonomous (e.g., interest, conviction) vs. controlled (e.g. pressure, guilt).

Given that autonomous motivation and intrinsic content are valuable, Sheldon (2004) argues for self-esteem being beneficial in cases when the goal of achieving self-esteem is not "focal in awareness," he calls it the "sidelong" approach to self-esteem (p. 423). Sheldon (2004) also mentions the point mentioned in the previous section that different kinds of self-esteem may have different value:

> Crocker and Park (2004) seemed to vacillate on this important question, in some places suggesting that only some kinds of self-esteem striving are problematic (i.e., involving contingent, unstable, or ego-involved self-esteem), but in other places suggesting that all self-esteem strivings are problematic. (Sheldon, 2004, p. 422)

Pyszczynski & Cox (2004) defend the importance of self-esteem by pointing to the importance of pursuing self-esteem in the TMT as "penultimate goal in people's self-regulatory hierarchies, subordinate only to the goal of managing fear" (p. 427) and propose pursuing "a sense of self-worth based on a core sense of one's inner identity" as a solution to the costs of self-esteem (p. 428). They also emphasize the centrality of self-esteem for the self-concept by equating the pursuit of self-esteem with *maintaining a positive self-concept* (Pyszczynski & Cox, 2004, p. 426). Crocker & Park (2004) respond to the three commentaries that given their definition of the "*pursuit of self-esteem* as the intention to validate self-worth by proving or demonstrating the qualities that the self does and does not have" (pp. 430-431) it is unlikely that the costs of pursing self-esteem will be eliminated by the proposed alternatives[272] (p. 432). Further, several researchers disagree that the (only) function of self-esteem is defense against death awareness (Crocker & Nuer, 2004; Leary, 2004; Ryan & Deci, 2004) by pointing out that

- death may be a "source of energy, resolve, and enthusiasm" (Crocker & Nuer 2004, p. 471);
- the distinction between contingent and true self-esteem (Ryan & Deci, 2004) where contingent self-esteem is the one which is influenced by anxiety of death and true self-esteem "is based in ongoing satisfaction of needs for competence, autonomy, and relatedness" (p. 473);
- findings that speak against the claims of TMT, as for example that "people with low self-esteem, who, according to TMT, ought to be most afraid of death, are more likely to commit suicide than people with high self-esteem" (Leary, 2004, p. 480).

Pyszczynski et al. (2004a) answer that goal of self-esteem is a superordinate one which means that other instrumental goals can still be pursued, given that there is no death awareness:

> In this sense, one could say that TMT posits a set of interrelated functions for the self-esteem motive and that facilitating the pursuit of particular life goals – or the engagement with life – is one of them. TMT provides an explanation for how those functions fit together: People pursue particular life goals to maintain self-esteem; self-esteem provides a buffer against existential anxiety; when anxiety is released because of drops in self-esteem, this signals the need for behavior to bring oneself back in line in pursuit of one's goals or to disengage from those goals and pursue other avenues for self-esteem; this increased effort or disengagement from specific goals occurs because one needs the protection from existential anxiety self-esteem provides. (Pyszczynski et al., 2004a, pp. 485-486)

[272] Interestingly, Crocker & Park (2004) equate self-enhancement with self-deception by referring to "self-protective or self-enhancing strategies" mentioned by DuBois & Flay (2004, p. 416) as "self-protective or self-deceptive strategies" (Crocker & Park, 2004, p. 432).

Concluding the discussion on the value of self-esteem, I want to point out that the two positions – whether the pursuit of self-esteem is beneficial or not – ground on different premises about defense in general: is it something that can be successfully accomplished and lead to positive affect or is it something that leads to mental and physical costs in virtue of its incomplete nature such that the negative information endangers the system from within and leads to instability (see table 29, section 1.3 and 3.1.3.1)? Trivers (2011), for example, argues that insofar as truth suppressing is an instance of self-deception, it is costly to the immune function (p. 117), and thus, that the general rule is "the earlier during information processing that self-deception occurs, the less its negative downstream immunological effects" (p. 134). Thus, he is against the view that there is a psychological immune system that includes defensive mechanisms along Freudian psychodynamic theory that protects individual's happiness (p. 68-69). Trivers' elaborations suggest that the self-deceptive defense is never fully successful, but not every kind of self-enhancement may be seen as implying such kind of defense, e.g. compare this to the account of positive illusions (3.1.3.1). As such it is a conceptual issue about the nature of defense. I think that both kinds of defense are possible and, thus, in such debates one has to pay particular care to not attack a "straw man." I will shortly demonstrate the difference between these two kinds of defense on the example of Surbey's (2004) account (see table 30).

What all types of SD have in common: They conceal information from consciousness and result in distortions of reality.			
Kind of definition	Definition	Instances	Measurement
narrow definition of SD	two contradictory beliefs whose level of consciousness is motivated (Gur & Sackeim)	repression, denial	GSR-activity
broad (within social cognition)	motivated avoidance of threatening information	positive illusions or self-enhancing biases	paper-and-pencil, fMRI

Table 30. Surbey: types of self-deception.
Distinctions from Surbey (2004, p. 119-121).

Surbey differentiates between two definitions of self-deception – broad and narrow. They correspond to the intentionalist (agentive defense) and deflationary (subpersonal selection) understanding of self-deception. Surbey states that concealment of information from consciousness is a common characteristic of all types of self-deception and holds repression and denial, as well as positive illusions and self-enhancement to be instances of self-deception. In section 1.3 I have already mentioned that Paulhus et al. (1998) argue that self-enhancement and denial are the two factors (or types) of self-deception (p. 1036). Notably, Surbey holds that different forms of self-deception could possess *different functions* (Surbey, 2004, p. 121) such that certain kinds of self-enhancement may be conducive to mental health (Taylor et al., 2003):

> The present investigation uses multiple assessments of self-enhancement and multiple indicators of mental health to examine whether self-enhancement is positively related to mental health, as the positive illusions position maintains (Taylor & Brown, 1988), negatively related to mental health, as the defensive neuroticism position maintains (e.g., Colvin et al., 1995; Paulhus, 1988; Shedler et al., 1993), or related in curvilinear fashion to mental health, as the optimal margin of illusion position maintains (Baumeister, 1989). Little support was found for either the defensive neuroticism or the optimal margin of illusion position. Instead, across multiple measures and indicators, the relation of self-enhancement to mental health was largely linear and positive, as the positive illusion position predicts. These results cannot be accounted for by the responses of a few highly

> anxious, depressed, or otherwise distressed participants, because screening procedures precluded such individuals from participating in the study. (Taylor et al., 2003, p. 173)

An argument against this claim could be, for example, that self-enhancement "may involve invalid and undesirable distortions of the self-concept" (Baumeister et al., 2003, p. 5). One could also argue that self-enhancement, like positive illusions, promotes perseverance while acquiring a certain goal (Krebs & Denton, 1997, pp. 29–33). Another unclear issue in terror-management theory is the nature of accessibility of death-related thoughts. Pysczcynski et al. (1999) argue for the "accessibility of death-related thoughts *rather than* the direct emotional experience of fear, anxiety" (p. 836; my emphasis). Thus, conscious *negative affect* is not necessarily associated with MS induction. But it is not the *conscious* accessibility of death-related thoughts either (p. 838). Pysczcynski et al. (1999) relate this kind of accessibility to Wegner & Smart's concept of "deep activation, when death-related thought is highly accessible but not in current consciousness, focal attention, or working memory" (p. 839) and by stating that such thoughts are "on the fringe of consciousness" or "highly accessible but not in current focal attention" (p. 840). An example of accessibility induction in one of the experiments might be useful for clarification purposes: "Subliminal presentation of the world *death* led to significantly higher levels of death-thought accessibility than subliminal presentation of the neutral world" (Pysczcynski et al., 1999, p. 841).

Terror-management theory proponents further argue for a *dual system theory* explanation (2.2.2.2) such that there are "two distinct information-processing or memory systems, one for dealing with recall and rational manipulation of declarative information about the world and the other dealing with the acquisition of skills, procedures, and behavior patterns" (p. 837). Pysczcynski et al. (1999) hypothesize that defense against mortality can occur either at the same (lower) level by some denial-evoking techniques guided by "rational processing" and the "rules of logic" or by pursing (higher) level goals of maintaining self-esteem and cultural worldview (pp. 837-838). The first kind of defense is called by them the *proximal defense*. Some of the examples they give for proximal defense are also mentioned in the self-deception literature: Quattrone & Tversky's (1984) cold-water experiment, from which they draw the conclusion that "people are willing to undergo higher levels of pain if they are led to believe that high pain tolerance is *associated with a long life*" (p. 839; my emphasis), Kunda's (1987) experiment that coffee drinkers deny bogus evidence about coffee drinkers being more susceptible to certain diseases (p. 839). They speak about proximal defenses as "rational in the sense that they entail logical, albeit biased analyses of information" and count Kunda's (1990) motivated cognition model which states that goals to arrive at particular conclusions bias information-processing to be a proximal kind of defense (p. 839). *Distal defenses* (those of pursing the two abstract goals) needn't be conscious (Pysczcynski et al., 1999, p. 838). The appeal to consciousness and susceptibility of distal defense to cognitive load conditions (p. 842) in differentiating between proximal and distal defense has led the authors to test a further hypothesized distinction between the two, namely that distal defenses are executed after a *delay and distraction*, while proximal ones are employed right after the threat (p. 840). Thus, proponents of the terror-management theory overrationalize the biasing literature (see section 1.3). As a consequence, the mechanisms, by which terror-management (or at least their descriptions) is achieved, is a merit a further analysis, but the general idea is still enticing.

This may have been part of the reason why Varki & Brower (2013) have combined Trivers' idea that self-deception evolved in the service of deceit and terror-management theory into a hypothesis about evolutionary conditions of the development of full theory of mind (ToM). The authors' main idea is that anxiety of death is the psychological evolutionary barrier against the development of full ToM (acknowledging that others have beliefs or second-order intentionality, p. 86), so that only those that have the mechanism for coping with such anxiety could develop uniquely human abilities associated with full ToM. This mechanism is reality denial (p. 116) or "[a]n unconscious defense mechanism used to reduce anxiety by denying thoughts, feelings, or facts that are consciously intolerable" (p. 17).

One can imagine the following evolutionary scenario: our ancestors without full ToM did not differentiate between self/other and had no anxiety about their own death. When full ToM develops, the distinction between self and other will lead to the awareness that the self might/will die. This will render full ToM-capable individuals incapable of fighting for mates and incapable of transferring their genes. Thus, full ToM capability will be a liability, unless one has a capacity to deceive oneself in order to be able to risk one's life in order to spread one's genes, which would lead full ToM to spread in the population.

Anosognosia (p. 122) and unrealistic optimism (p. 123), as well as overconfidence (p. 160) are seen as results of the workings of such a reality denial mechanism by the authors, but also religion, or at least some sense of spirituality in non-religious humans (p. 144ff), climate change (p. 221ff) and, perhaps unexpectedly, meditation[273] ("the common goal of most meditation methods is to deny reality by deliberately eliminating extraneous thoughts and focusing on only a few," p. 253). Thus, though the authors label the phenomenon 'denial,' their description shows that it encompasses phenomena labelled 'self-deception.' The benefit of self-deceiving might have been, according to Varki & Brower (2013), the sexual selection or increased mating possibilities – an idea borrowed from Trivers (p. 166ff). The consequence of such a theory about denial is that it is seen as an "essential skill for us to function normally in the world" (p. 218), that there might be an innate capacity for it (p. 217) and that we should not try to fully escape it (p. 243, 249). The metaphor that the authors use for the description of this psychological barrier is borrowed from computer science:

> In other words, the cognitive evolution of warm-blooded birds and mammals may have proceeded through several peaks and valleys until it reached the optimal peak of self-awareness with rudimentary ToM. But getting across to the next optimal peak (full ToM) may have required going through a very deep and dark valley: the full recognition of reality and mortality. (Varki & Brown, 2013, p. 133)

Varki & Brower (2013) hypothesize that culture and learning might substitute the second phase of the Baldwinian mechanism of explaining evolution of complex behavior where the two phases are 1. learning something new from other co-specifics and 2. genetic "fixation" of the property (p. 207). This would mean that there need not be a genetic fixation of certain properties. The same might be the case for certain kinds of self-deception. Varki & Brower (2013) idea is namely that a certain ability for self-deception had to evolve to be able to cope with death-anxiety, but that on top of this basic ability other more specific self-deceptive abilities have followed. Whether *these* are genetically

[273] Meditation and mindfulness usually are said to extinguish self-deception. Levesque & Brown (2007), for example, argue in the context of dual theory that mindfulness – paying attention to one's thoughts and feelings – is a moderator between implicit motivation and day-to-day behavior.

fixed is in question. This is the bridge to the next topic – evolutionary explanations of self-deception.

Before that a summary of the results of section 3.1 is to be presented. In this section I have presented theories by which self-deception has been explained in such an order that each subsequent theory precisized the function that self-deception might serve: reduce dissonance as in cognitive dissonance, secure the stability of the self-concept, of which self-esteem is a central feature. Self-esteem, then, might serve the ultimate aim of reducing the anxiety of death. Independently, Varki & Brower suppose that the claim that self-deception developed as a means of reducing death-anxiety can be followed directly from self-deception's courage-promoting effects on our behavior.

What is my take on this development? First, if self-deception is a cluster concept (1.1.3), then there need not be only *one* function that self-deception serves. Depending on the behavioral and phenomenological profile, e.g. presence of tension and counterfactual goal-directed pull, the way self-deceivers justify their self-deception, one would suppose that different kinds of information processing underlie it and with it – different implications for hypothesized function and health-aspect would follow. The general idea that self-deception serves to reduce death anxiety is appealing, because it is neutral on the information processing aspect and with it – on whether self-deception is immunologically healthy or not. The specification of how, e.g. self-esteem defense, would then lead to hypothesized costs of self-deception, e.g. permanently suppressing some information – permanent stress – being unhealthy, because is exhausts system's resources for dealing with invaders. In this case, inconsistency in the system would be an invader endangering the system from inside. Inconsistency will also stand in the focus of predictive coding, according to which prediction errors between expectation and outcome are those that the systems attempts to reduce, either by changing the hypothesis about the causes that generated those errors, which would mean changing the world/self-model, or by action that would, then, change the evidence available to the systems and through it also the prediction errors. How could immunological costs be modeled with predictive coding? When would inconsistency lead to such costs and when not? This is for future research to determine.

3.2 Evolutionary theories of self-deception

In this chapter, I will review evolutionary explanations of self-deception. First, I will consider the constraints on the validity of an evolutionary explanation independently of its application domain. Then, I will present three different possibilities for an evolutionary theory of self-deception: Trivers' adaptationist view which will take the largest part, van Leeuwen's spandrel view and Lopez & Fuxjager's exaptation view. These are theories that represent the three major kinds of evolutionary explanations, namely that a trait is:

- ➢ an adaptation,
- ➢ a spandrel which is a by-product of an adaptation,
- ➢ an exaptation which is an adaptation that has changed its function.

The first kind of explanations is by far the more popular, one argument being for example the degree of functional complexity of the organization of the mind (Carruthers, 2006, p. 9). After considering these three possibilities and in the light of the results in the previous section I will argue that self-deception may be an evolutionary adaptation to bypass anxiety of death that acquired another (additional) function of deceiving others.

3.2.1 Constraints on the validity of an evolutionary psychological explanation

The aim of this section is to present the possible approaches of giving an evolutionary explanation of biological phenomena and to choose a suitable one for self-deception. The checklist introduced in this subchapter will, then, be applied to subsequent evolutionary explanations of self-deception that will be presented. To achieve the given aim, I will review the work of Richardson (2007) who evaluated the applicability of three approaches from evolutionary biology to evolutionary psychology: reverse engineering, forward engineering and the comparative approach. Richardson (2007) focused his attention on Cosmides and Tooby's theory that social exchange is essential to human reasoning (Richardson, 2007, pp. 91-92), yet the intended scope of his argumentation embraces all kinds of evolutionary psychological theories. It also embraces a possible evolutionary theory of such a psychological phenomenon as self-deception. I will come to the conclusion that a viable evolutionary theory of self-deception has to answer the criticism of under-determination and that one has to be cautious about the scope of its application (evolutionary or current environment).

First, let me talk about the explanandum of an evolutionary theory of self-deception, then about its scope by discussing the environment (external context) and, last, come to different kinds of evolutionary explanations. What does it mean to talk about a psychological mechanism as an evolutionary adaptation? First of all, an evolutionary psychological explanation should be grounded in evolutionary biology[274] (Richardson, 2007). For self-deception to be an evolutionary psychological process one has to prove that it is a biological adaptation to Pleistocene conditions maintained by natural selection.[275] As Richardson focuses on Tooby & Cosmides' work, I will briefly present their phenotypic view of evolution[276] that Crawford summarizes it as follows:

> Technically speaking, when talking about the formation and functioning of psychological mechanisms in the context of evolution by natural selection, the discussion is about psychological adaptations. They are innate specialized information-processing mechanisms realized in the neural hardware of the nervous system. When activated by appropriate problem content from the external or internal environments, they focus attention, organize perception and memory, and call up specialized procedural knowledge that leads to domain-appropriate inferences, judgments, and choices (Cosmides & Tooby, 1989). (Crawford, 2004, p. 9).

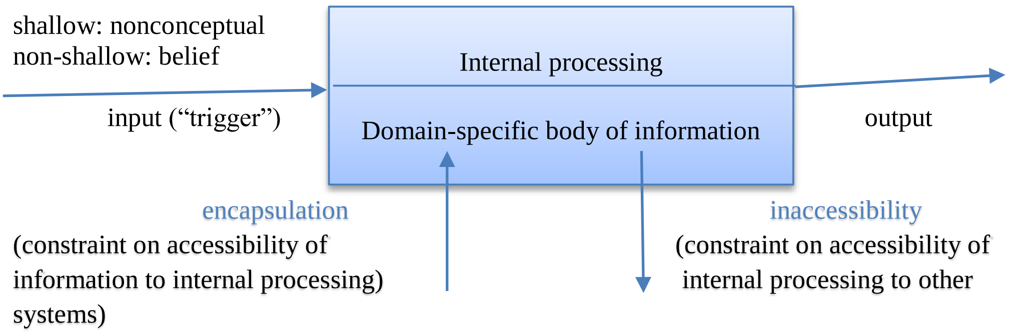
Richardson (2007) states that while more evolutionary psychologists prefer the view that these information-processing mechanisms are specialized (Tooby & Cosmides are among these), there are also those who view the mechanisms as general (Richardson, 2007, p. 91, 141). Thus, clarifying what content triggers the information-processing mechanisms, what kind of mechanisms these are, how they influence information-processing and subsequent

[274] Evolutionary biology as basis for evolutionary psychology: "It is evolutionary biology that defines the context in which the adaptive claims of evolutionary psychology should be assessed. The standards we should use are evolutionary standards" (Richardson, 2007, p. 37).

[275] Richardson (2007) states that given Cosmides & Tooby's evolutionary function of human reasoning to facilitate social relations, the initial goal of evolutionary psychology "thus, is explaining psychological processes as biological adaptations to Pleistocene conditions, adaptations that have been shaped and maintained by natural selection" (p. 20).

[276] Definition of a phenotypic view of evolution: "This view might be called the *phenotypic view of evolution*, because the focus is on the anatomical structures, physiological processes, or the behavior patterns that helped solve the problems faced by the organism. Although at least some of the phenotypic variation must be heritable for natural selection to occur, genetic concepts are not referred to explicitly" (Crawford & Salmon, 2004, p. 25; my emphasis).

actions is an essential part of the postulation of an evolutionary hypothesis of a psychological adaptation, at least if this understanding of a psychological mechanism is accepted. Unfortunately, the debate about the criteria along which a certain process might be individuated (see figure 19) has been inconclusive so far. Due to the generality of the matter at hand, namely that self-deception is a kind of phenomenon that could potentially influence *all* kinds of information available to the organism, it is difficult to impose encapsulation constraints and that integration of processing streams (cognitive binding) might be a more interesting question to tackle (see section 2.2.2.3).


shallow: nonconceptual
non-shallow: belief
input ("trigger")
Internal processing
Domain-specific body of information
output
encapsulation
(constraint on accessibility of information to internal processing) systems)
inaccessibility
(constraint on accessibility of internal processing to other

Meanings of domain-specific:
a) content-specific, e.g. only nonconceptual input;
b) function-specific: distinguished by tasks they perform, e.g. dissociation between tasks implies dissociation between modules

Criterion for the distinction between modules:

a) specificity of (innate) information: explicitly represented information
b) specificity of the kind of processing/mechanism: information might be implicitly represented in the algorithm (Carruthers 2006)

Computational tractability requires either

a) encapsulation:
   a. narrow-scope: inaccessibility of *determinate* information to operations of the system
   b. wide-scope: *bottleneck* on amount of information *at a time*, refusal of exhaustive search
b) frugal (information-constrained) search heuristics

Independent evolvability is a criterion distinguishing between mosaic (independently evolvable) and connected traits.

Figure 19. The notion of a module.
Distinctions from Carruthers, (2004, 2006, 2007), Prinz (2006), Samuels (2006), Woodward and Cowie (2004).

Localization is different from **domain specificity**: If mental functions are realized by *overlapping networks*, then locations of modules cannot be equated with anatomical regions (Prinz, 2006).
There is doubt that so-called central processing, e.g. reasoning, is modular:
a) Criteria for distinguishing among cognitive mechanisms on the basis of *domain specificity* are not clear (Samuels, 2006). Another view would be that distinct *functions* require distinct *mechanisms* (Carruthers, 2004).
b) Challenge of explaining *conceptual integration* and relationships between mechanisms (Samuels, 2006).
c) An alternative to specialized *mechanisms* is that of specialized *bodies of knowledge* (Samuels, 2006). A critique of the latter possibility, though, has been argued to consist in the necessity of explicit representations in a case where domain-general mechanisms process information based on different sets of knowledge (Carruthers, 2006).

Both the ultimate (why) and the proximal (how) level of explanation plays an important role in evolutionary psychological theories[277] (Crawford & Salmon, 2004, p. 43). For both the distinction whether self-deception contributed to ancestral inclusive fitness or is contributing to the current inclusive fitness is important[278] (Crawford & Salmon, 2004, p. 38). Two independent arguments, for each kind of environment, are in order to prove that self-deception has been beneficial in the past and stays beneficial in the current environment.[279] Crawford, for example, holds, that self-deception both contributed to ancestral well-being and continues to contribute to the current well-being (see table 31). Further, proving that self-deception enhances reproductive fitness in the current environment only is not enough to prove that it *has* an evolutionary function (Crawford & Salmon, 2004, p. 39). It may be a starting point for an argument that self-deception *will acquire* an evolutionary function in the future, given that evolutionary explanations assume that a certain amount of time is needed to pass for a (potentially) evolutionary mechanism to influence inclusive fitness.

Another question with respect to the environment in which a certain function operates is to which *degree* it is genetically predetermined and to which degree it is influences by the environment. Woodward & Cowie (2004) hypothesize that there might not be a folk-psychology module, but dispositions to *create* a certain kind of environment in which beliefs and desires are the units of communication, understanding and knowledge acquisition (p. 318). A similar idea is advocated by Barrett (2007) that modules are *phenotypic structures* that develop differently depending on the kind of information they get. As Barrett (2007) argues, evolution produces *types* which are underspecified, but not *tokens.*[280] Universality on this picture is not an argument for innateness, given that universal environment would lead to the development of the same token (p. 214). Sterelny (2007), also along the similar lines as Woodward & Cowie and Barrett, argues for an *informational niche construction* (engineering of environment by humans). Most interestingly for the subsequent discussion of Trivers' theory of self-deception, he argues that to uncover deception *sensitivity to a wide range of information* is necessary (p. 220). The consequence of this is for Sterelny (2007) that at least some heuristics, that humans often employ, have to be *informationally demanding*.

[277] The distinction between proximate and ultimate expalanations: "*Proximate explanations* refer to the immediate factors, such as internal physiology and environmental stimuli that produce a particular response. *Ultimate explanations* refer to the conditions of the biological, social, and physiological environment that, on an evolutionary time scale, render certain traits adaptive and others nonadaptive (Mayr, 1961)" (Crawford & Salmon, 2004, p. 42; my emphasis). Thus, proximate function is a mean for the ultimate one: "Its [morality] *proximate functions* involve biological, psychological, and sociocultural variables. Its *ultimate function* is evolutionary: to pass on genes by means of strategies that are executed using these proximate variables" (Holcomb, 2004, p. 90; my emphasis).

[278] Assumption about evolved psychological means as determining current behavior: "Evolutionary psychology is based on the assumption that psychological mechanisms that evolved to solve problems encountered by human ancestors in their long-gone environment are involved in producing current behaviors and institutions" (Crawford, 2004, p. 13).

[279] Ancestral and current environment may set different constraints: "Because ancestral and current environments may differ, adaptations or behaviors that contributed to ancestral fitness may no longer contribute to current fitness. Similarly, behaviors that may contribute to fitness now may not have contributed to it in the past" (Crawford & Salmon, 2004, p. 38).

[280] Types are underspecified in that a tendency to act in a certain way in a set of situations may be evolutionary selected, but the specific action token is dependent not only on the (pheno-) type of the agent, but also on the input from the current environment: "As noted, proper domains and outcomes are defined by history of selection. Actual domains and outcomes, on the other hand, are defined in terms of an interaction between the input criteria and operations of the system, and the state of the current environment" (Barrett, 2007, p. 210).

Contribution To Ancestral Fitness	Contribution To Current Well-being	
	No	**Yes**
No	True pathologies: e.g., schizophrenia	Quasi-normal behaviors
Yes	Pseudopathologies	Adaptive-culturally variable behaviors: e.g., **self-deception**

Table 31. Crawford: distinction between well-beings
Modified from Crawford (2004, p. 14).

Crawford (2004) defines the concepts of true pathologies, quasi-normal behavior, pseudopathologies and adaptive-culturally variable behavior with respect to their contribution to ancestral and current well-being. Thus, true pathologies are those kinds of behavior that contribute neither to ancestral, not to current well-being, for other three concepts he gives analogue definitions (pp. 13 - 18). In the case of adaptive-culturally variable behavior Crawford (2004) justifies the concept name by holding that this kind of behavior is adaptive, because it still serves the function it had in the ancestral environment and it is variable, because it manages to fulfill the function under the changed circumstances of the current environment (p. 17).

The argumentation in favor of the assertion that self-deception either contributed to ancestral inclusive fitness, current inclusive fitness or both has to be conducted with respect to the appropriate kinds of *environmental factors* – those that can affect fitness differentially. Richardson (2007) discriminates three kinds of environment – external, ecological and selective – of which only the latter can be used in the explanations concerning inclusive fitness (see figure 20). External environment is partitioned on the basis of physical or biotic factors, ecological environment - on the basis of differences in the performance of organisms and selective environment - on the basis of *relative* differences in performance of organisms (Richardson, 2007, p. 69). In other words, external environment can be characterized in terms of such factors like temperature or amount of predators (p. 69). If the inclusive fitness of the same type differs across space and time, then different kinds of ecological environment are at work (p. 69). If among the same type the fitness of particular individuals differs, then it serves as a criterion to differentiate the selective environment (p. 69).

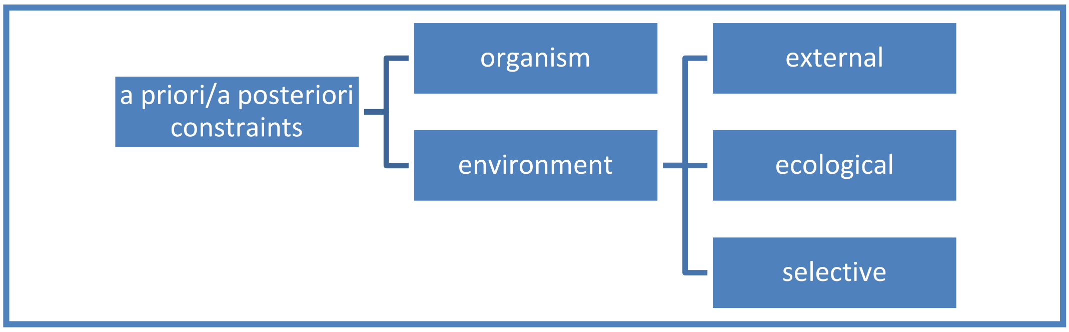


Figure 20. Kinds of environmental constraints
Distinctions from Richardson (2007, pp. 65-75).

Richardson (2007) argues that design constraints – those on the applicability of a certain evolutionary explanation to a certain structure/mechanism – can be formulated either a priori or a posteriori concerning either the organism, or the environment. His use of the a priori/a posteriori distinction is different from the philosophical one. The philosophical distinction concerns the fact whether something can be known independent of experience, while for Richardson (2007) it concerns the fact whether something can be known from the causes, or has to be inferred from the effects:

"The key issue is whether the constraints defining what counts as optimal form or behavior are *prior* to the assessment of fit between form and function or whether the constraints are inferred from, and *posterior* to, the assessment of fit between form and function. That is, we can begin with the constraints in place, or infer the character of the constraints from other factors." (p. 66)

So far, I have pointed out the need to consider ultimate and proximate kinds of explanations[281] regarding the ancient and current selective environment in order to give an evolutionary explanation of a psychological phenomenon, in our case - self-deception. Different kinds of strategies can be applied to achieve this. Richardson (2007) distinguishes three kinds of such strategies: reverse engineering, forward engineering and comparative method (see table 32). The first two differ in the direction of inferences that are drawn – whether they are drawn from effects to causes in the case of reverse engineering or from causes to effects in the case of forward engineering. With respect to self-deception, it would mean to either take the phenomenon self-deception as a starting point and trying to establish its evolutionary causes in the case of reverse engineering or to start with the analysis of Pleistoscene conditions and constructing a historical explanation of self-deception in the case of forward engineering (Richardson, 2007, p. 96). The choice, whether to apply the first one or the second, should be made according to the answer to the question, whether more is known about the phenomenon under discussion, or about the ancestral environment it is supposed to have developed from, because the starting point of the argument has to be known best. Self-deception is a controversial phenomenon. Thus, instead of taking as a starting point for an evolutionary explanation the definition of self-deception which is still unclear and debated, favoring the second type of approach and proceeding from evolutionary causes to the explanation of the phenomenon in question would make the given evolutionary approach applicable to different kinds of definitions. I think that this is also the approach undertaken by Varki & Brower (2013) – they identified the evolutionary problem (anxiety of death) such that the functional role for the solution of the problem fits the description of self-deception.

Richardson (2007) exemplifies how the second type of approach (forward engineering) could be accomplished by citing Brandon’s five conditions for an ideal adaptation explanation. Fulfilling Brandon’s conditions with respect to self-deception means showing that self-deception has been selected for (1) due to certain ecological factors (2). It can be selected for only if it is heritable (3). The information about the population structure (4) and the genetics of the trait (5) completes the list. An argument against the claim that an evolutionary explanation of self-deception has to fulfill those criteria could be that, as Cooper (2007a) argues, it is rarely the case that all of Brandon’s conditions are fulfilled (p. 189). Yet, as Reuter (1999) points out, the point of an ideal model is to use it as an evaluation standard for proposed explanations (p. 9).

The third type of approach takes phylogenetic information into account for *comparison* purposes (p. 154). It means that related evolutionary lineages are compared considering the question whether the trait under discussion contributed to inclusive fitness in that lineage (Richardson, 2007, p. 148). If this question is answered affirmatively, this means both that the trait contributed to inclusive fitness in related lineages and that the function it had was relevant for the lineage (pp. 148-149). Applying this kind of approach to self-deception would mean comparing self-deception in humans with that in apes – both belong to the class hominids (pp. 158-159). This presupposes that self-deception in animals is possible. If it is not, then nothing can be gained from the application of such a phylogenetic approach, because no meaningful comparison between lineages will be possible, if humans are the only ones who can self-deceive. Summing up, forward-engineering, reverse-engineering and lineages comparison are the three possible ways to construct an evolutionary explanation in the case of self-deception. The decision which approach to favor depends

[281] Lockard (1997) calls them proximal (sensory, concerning memory, cognitive) and distal (evolutionary) (pp. 120-121).

on what one takes to be the most reliable, fine-grained and precise information on self-deception – its definition or the ancestral environment under which it developed. Whether there is animal self-deception for lineages comparison depends on how one defines the phenomenon.

Type of approach toward empirical evaluation of evolutionary explanations	**Kind of inference**	**Criticism/problem**
“reverse engineering”	Inference from effect to cause: effect → cause	- Underdetermination/ambiguity of structure (R 2007, p. 52) - One-to-one assumption: one function – single mechanism (W & C 2004, p. 316)
Adaptationism	Inference of effect from the relevant causes: cause → effect Application of Brandon’s 5 conditions of being an adaptation explanation (R 2007, pp. 99 - 105) 1. selection (character and extent of variation) 2. ecological factors (explanation of selection) 3. heritability 4. population structure (**homo-** vs. heterogeneous) 5. trait polarity (primitive vs. **derived**)	Absence of critical kind of information concerning the historical contingencies (R 2007, p. 139)
comparative method (focus on descent)	Compare a trait to phylogenetically related ancestors and conditions	Uncertainty of ecological information (R 2007, pp. 157-158, 166)

Table 32. Evidential strategies in favor of an evolutionary explanation. Distinctions from Richardson (2007), Woodward & Cowie (2004).

Note that Richardson (2007) argues that “reverse engineering” infers function from structure (p. 44), while Woodward & Cowie (2004) – structure from function (p. 314). *Under-specification* might lead to an *ontological filling-in* so that open parameters are filled explicitly or implicitly (Barrett 2007).

Summarizing, applying these three kinds of approaches to self-deception means

- reverse engineering approach:
    - starting with the definition and properties of self-deception and inferring which factors could have led to it acquiring a given evolutionary function;
    - special care should in this case be given to the fact that there could be multiple options for possible causes
    - and, thus, argument for the dismissal of rivalry explanations are necessary;
- forward engineering:
    - starting with the ancestral environment and inferring that *selection* for self-deception has taken place which implies showing that the trait of self-deception was *heritable* and possessed a certain amount of variance among individuals and that certain *ecological factors* explain the selection; furthermore the *homogeneity* of the population would be conducive to such an evolutionary explanation to prevent the effects of chance, as well as self-deception being a *derived trait* – novel among the descendants;
- comparison approach:
    - comparing the advantage of the presence/absence of the trait of self-deception in humans and in phylogenetically related ancestors (apes), which presupposes an independent argument for the possibility of self-deception in animals.

Consequently, examining the history of the self-deceptive trait plays an important role especially for the second kind of approach, because according to this kind of approach the reconstruction of this history (based on the contemporary situation and limited information about the past) determines the explanation for the trait (Richardson, 2007, p. 98). The amount of information for taking such a kind of approach, though, according to Richardson (2007), is unsatisfactory:

> Lacking knowledge of the forms of communication present in our ancestors, lacking knowledge of the sorts of social organization present, lacking knowledge of the sort of ecological problems that these ancestors confronted, lacking even knowledge of which ancestors are the proper focus of investigation, we lack the ingredients for seriously explaining our psychological capacities in evolutionary terms. We are left either with empty generalities or unconstrained speculation. (Richardson, 2007, p. 171)

Interim conclusion: An evolutionary theory of self-deception has (at least) to provide an argument in favor of a hypothesis that it is a heritable trait that has been selected for due to certain ecological factors and consider the critique of the absence of satisfactory information for its proof. In doing this is has to avoid two traps of searching for an evolutionary explanation: one regarding its *functionality* that consists in the fact that the function of the trait can change from the ancestor to the current environment and another regarding its status or that a trait can persist because of it being a *byproduct* of another adaptive trait (Sedikides et al., 2004, p. 70). Trivers' theory denies that the function of self-deception changes and that it is a byproduct. Van Leeuwen's theory is silent on the first question and argues for self-deception being a byproduct. I will argue that self-deceptive function may have been widened from defense against anxiety of death to other deception in section 3.3.

3.2.2 Trivers' evolutionary theory of self-deception

I will analyze Trivers' argumentation in favor of self-deception having an evolutionary function to facilitate other deception as follows: First, I will consider the distal (why self-deception evolved? – to deceive others) and proximal (how was it possible? – genetic and psychological mechanisms) levels of explanation, as well as the definition and types of self-deception according to Trivers. Last, I will go into details about the evidence in favor of the given theory. I will argue that according to Trivers self-deception is a widely encompassing phenomenon that possesses different proximal mechanisms and that while the existing evidence does not contradict Trivers' hypothesis, it is not sufficient. Thus, overall, Trivers' theory would benefit from further clarification of the target phenomenon in order for the evidence to gain more explanatory weight.

3.2.2.1 Definition and types of self-deception

In this section I will first review Trivers' definition and categorization of self-deception and then discuss the question of the success-aptness and adaptivity of misrepresentations. Trivers definition of self-deception is "active misrepresentation of reality to the conscious mind" (Lynch & Trivers, 2012, p. 491; Trivers, 2000, p. 114).

Self-deception is an active misrepresentation to the conscious mind.		
Active	**Misrepresentation**	**to the conscious mind**
(1) Self-deception is a social, personal and subpersonal level phenomenon.	(2) Self-deception is a kind of successful misrepresentation where successful is understood as enhancing inclusive fitness.	(3) The subject who is self-deceived is not conscious of his/her self-deception.

Table 33. Trivers: definition of self-deception.
Distinctions from Trivers (2011).

Being active is the property of agents. In his popular book about self-deception, Trivers states that the conscious mind is "*devoted* (in part) to constructing a false image" (Trivers 2011, p. 27; my emphasis). Von Hippel & Trivers (2011) also accept the existence of what they have named the classic form of self-deception, namely "*convincing* the self that the lie is true" (p. 10; my emphasis). According to Mele (2001), an agency view on self-deception presupposes that the self-deceiver *intentionally* biases her reasoning process. Trivers also states in his popular book on self-deception that one of its hallmarks is the "*unconscious* running of selfish and deceitful ploys" (Trivers 2011, p. 27; my emphasis), but that truthful information has to be processed and stored somewhere:

> Of course it must be advantageous for the truth to be registered somewhere, so that mechanisms of self-deception are expected to reside side-by-side with mechanisms for the correct apprehension of reality. The mind must be structured in a very complex fashion, repeatedly split into public and private portions, with complicated interactions between the subsections. (Trivers, 1985, p. 461)

In the review of the literature on self-deception I have already emphasized the difficulties of the agentive view on self-deception, in general, and divisionist strategies that suppose agentive subpersonal units, in particular (1.1.3). I have further argued that different kinds of selection can be of an agentive and a non-agentive type, depending on the description one prefers (2.2.1). I favor non-agentive selection and that the intuition, that self-deception is agentive, can be explained otherwise than postulating agentive kinds of selection. The point, namely, is that, apart from tension (which can also be explained in a non-agentive way, see section 2.1.3), it is the appearance of *control* in self-deception that has led the authors to arguments that self-deception is agentive, conscious or personal. Phenomenal Self-Model (PSM) or a mental model of the system itself is globally available for cognition, attention and *control* of behavior (Metzinger, 2003, pp. 210-211). Its content is the content of the conscious self (p. 299). Agency presupposes that "a process of *selecting cognitive contents for further processing* is represented and integrated into the PSM" (Metzinger, 2003, p. 406). That an epistemic agent model of self-deception is constructed or involves feeling of agency directed at cognitive activities related to self-deception has been questioned in section 2.2.1 and 2.2.2. The sensitivity of self-deceptive mechanisms to only certain kinds of information, conducive to the achievement of a specific goal, leads the observer to suppose agency exactly *because* of another intuition that self-deception can happen about *every* kind of content. Either one has, then, to postulate a host of subpersonal mechanisms each of which deals with a specific kind of information (see modularity in section 3.2.1), or one postulates one mechanism that is, in either case of self-deception, specific to different kinds of information, but selectivity in the latter case is the easiest to explain by appeal to agency. An evolutionary explanation is not tied to the assumption that every function has to be fulfilled by exactly one mechanism (this would be a modularity assumption). The assumed *function* is, then, the unifying element in the explanation, not the kind of *mechanism*(s) by which it has been brought about. The unification in terms of mechanisms may be the result of the assumption that “the world is presumed to have this kind of multilevel structure, of mechanisms within mechanisms” (Craver, 2015, p. 8). One unexplored explanatory perspective on self-deception is that it is an emergent property – a property that cannot be explained in terms of parts of the mechanism (for the latter definition see Craver, 2015, p. 21).

The phenomena, assumed by Trivers to fulfill the given function of deceiving others, have been argued to be too broad by Van Leeuwen (2013b). I will first present Trivers’ categorization and then discuss Van Leeuwen’s criticism. According to Trivers, there are two overarching types of self-deception: self-driven and imposed (see figure 21).[282] Self-driven self-deception is not here restricted to self-induced self-deception in the sense of intentionally causing an unintentional deception (Scott-Kakures, 2012, p. 23), for example writing something misleading in the diary in hope of forgetting that this was false later. While self-induced self-deception has been argued to be brought about by the individual himself, imposed self-deception is being inflicted on the individual in question by another individual, often a relative (Trivers, 2000, p. 122). Trivers (2000) further supposes that cases of imposed self-deception arise when the interests of the individual are in conflict with those of the one who inflicts the self-deception (pp. 122-123). He brings an example of a parent, imposing his wishes upon his child to the extent of abusing them (p. 123). I think that it is reasonable to suppose that induced self-deception does not enhance one’s inclusive fitness. If this is the case, then an argument should be given for the claim that the percentage of cases of induced self-deception is small with respect to those cases that enhance fitness, else self-deception would not possess an adaptive value in the present environment (an analogue argument should be provided for the ancestral environment as well). If this were the case, it would affect Trivers’ general argument that self-deception does possess an evolutionary function of deceiving others.

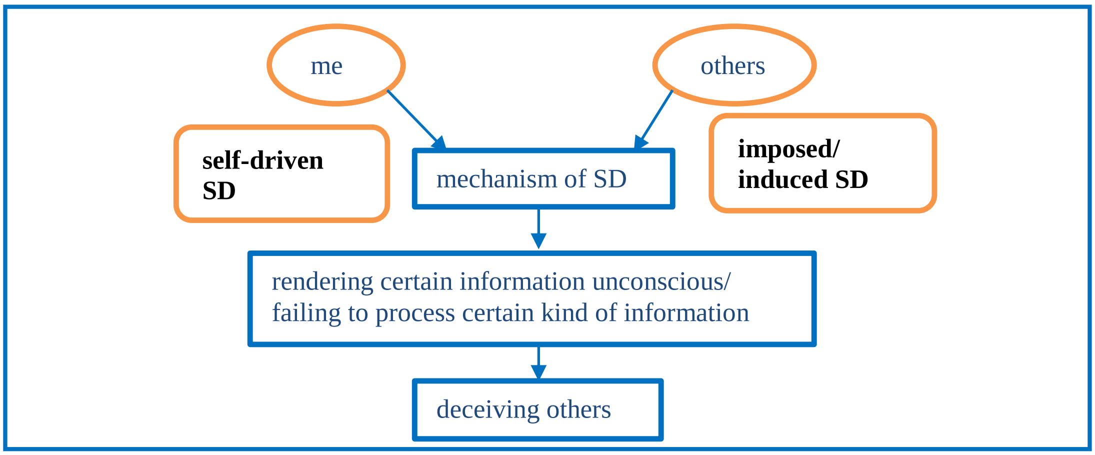

me
others
self-driven SD
mechanism of SD
imposed/ induced SD
rendering certain information unconscious/ failing to process certain kind of information
deceiving others

Figure 21. Trivers: self-driven and imposed self-deception
Distinctions from Trivers (2011).
Arrows are meant to point causal connections.

As mechanisms by which self-deception is accomplished Trivers (2011) proposes different kinds of biases: self-inflation, derogation of others, in-group/out-group associations, biases of power, moral superiority, illusion of control, construction of biased social theory, creation of false personal narratives (Trivers, 2011, pp. 15-27), as well as thought

[282] Trivers does not use the notion “self-driven self-deception”, but to differentiate it from the imposed kind, I gave it this name in order to be able to better discriminate the two.

suppression (p. 56), temporary fantasy[283] (p. 104) and the placebo effect (p. 70). It is consistent with Trivers' claim that biased processing of information is the hallmark of self-deception:

> In summary, the hallmark of self-deception in the service of deceit is the denial of deception, the unconscious running of selfish and deceitful ploys, the creation of a public persona as an altruist and a person "beneffective" in the lives of others, the creation of self-serving social theories and biased internal narratives of ongoing behavior, as well as false historical narratives of past behavior that hide true intention and causality. The symptom is a *biased system of information flow, with the conscious mind devoted (in part) to constructing a false image and at the same time unaware of contravening behavior and evidence*.[284] (Trivers 2011, p. 27; my emphasis)
>
> A hallmark of self-deception is *bias*. Mere computational error is not enough. Such error is often randomly distributed around the truth and shows no particular pattern. Self-deception produces biases, patterns where the data point in one direction - usually that of self-enhancement or self-justification. (Trivers 2011, pp. 147-148; my emphasis)

This categorization encompasses a very broad range of self-centering biases.[285] I think that most of them have a common feature – they are focusing on distinguishing self and in-group and attributing more positive qualities to them than to others. I will call this feature *self-centering* on the personal and *in-group centering* on the social level (see figure 22). My argumentation for this unifying feature is as follows: Self-enhancement and derogation of others serve the same aim – to portray oneself in a better light than others. Possible products of self-enhancement and other-derogation are false personal narratives (Trivers, 2011, p. 25). Thus, self-enhancement/derogation of others might lead to false personal narratives. An analogue connection could be given between two social categories of self-deception that Trivers names: in-group/out-group biases and biased social theory. While I understand self-enhancement/derogation of others and false personal narratives as exhibiting personal level influences, I think that in-group/out-group biases and biased social theories are social phenomena:

- personal: self-enhancement/derogation of others → false personal narratives;
- social: in-group/out-group bias → biased social history.

Moreover, I hold that it is plausible to assume that in-group/out-group biases play a causal role in constructing biased social theories, at least given Trivers understanding of the latter biasing as constructing "a consistent, self-serving body of social theory" (Trivers, 2011, p. 24). In-group/out-group biases are self-serving in the sense of portraying the in-group in a more favorable light than the out-group.

[283] Trivers (2011) introduced temporary fantasy in his popular book as a kind of self-deception by briefly citing an anecdotal episode of women fantasizing during the time of ovulation (p. 104). Note that counting temporary fantasies as kinds of self-deception poses the question about the connection between self-deception and mind wandering where the latter is understood as "engaging in cognitions unrelated to the current demands of the external environment" (Schooler et al., 2011, p. 319).

[284] The same characteristics of self-deception can be found in Trivers (2000, p. 118).

[285] As the notion of self-serving bias is used in psychology for a certain kind of attribution style – external in the case of a failure and internal in the case of success – I used the term self-centering bias to encompass all biases that depict the self – or the in-group in the case of in-group-centering – in a favorable light.

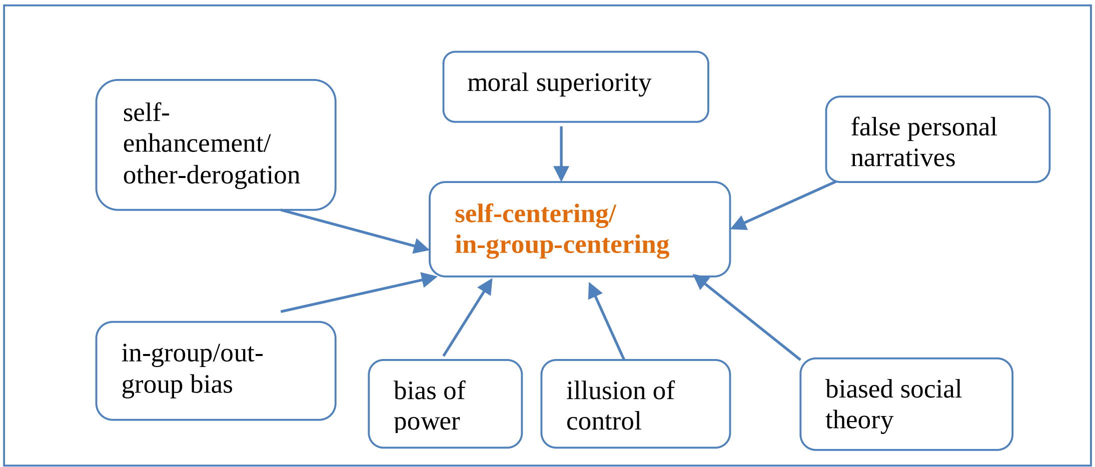

moral superiority
self-enhancement/ other-derogation
false personal narratives
self-centering/ in-group-centering
in-group/out-group bias
bias of power
illusion of control
biased social theory

Figure 22. Trivers: categories of self-deception. Distinctions from Trivers (2011).

In black are the categories. In orange is the property that I think all the categories share. The arrows mean "have a property of being."

I think that other three categories that are left (biases of power, moral superiority and illusion of control) are also self-centered. Biases of power induce, according to Trivers (2011) "blindness toward others" (p. 21). Moral superiority implies seeing oneself as more fairly than others (p. 22). These two categories subsequently fit the description of being self-centering. Illusion of control induces the belief of possessing more control of the outcome than one actually has (p. 23). My argument is reinforced by Trivers' assertion about the product of self-deception, namely that it is a certain belief or even a belief system that contains self-serving biases (Trivers, 1985, p. 416). Though Trivers' categorization points into the direction of self-deception being a result of biases that present the self and the in-group in favorable light, he mentions that in certain cases self-diminution can also benefit the individual (Trivers, 2011, p. 167). Yet, Trivers connects cases of self-diminution to induced self-deception,[286] for example an imposed self-deception of a victim in the case of domestic violence in which the victim takes the view of the abuser (Trivers, 2011, p. 63). Thus, the value of self-diminuition for self-deception is questionable. All in all, the question whether self-diminuition is also a kind of self-deception is unclear and depends on whether one accepts imposed self-deception as one kind of self-deception.

I think that the main merit of Trivers' categorization is that it emphasizes not only the personal, but also the *social* level. Yet, self-deception is supposed to be a *robust* phenomenon: as long as the motivation does not change, one is motivated to remain self-deceived. Yet, some in-group biases are *fragile*. The own-race biases, for example, is attenuated after positive emotion induction (Johnson & Fredrickson, 2005). Own-race bias can be seen as a kind of in-group bias such that recognition of faces of the in-group is better than that of the out-group, e.g. black vs. white faces.

Van Leeuwen (2013b) criticizes Trivers' recent popular book about self-deception (2011) and argues that it is flawed in two respects: Trivers 1. uses the term self-deception in an

[286] For example with respect to an in-group induced self-deception Trivers (2011) writes: "The strong suggestion, then, is that it is possible for a historically degraded and/or despised minority group, now socially subordinate, to have an implicit self-image that is negative, to prefer other to self – indeed, oppressor to self – and to underperform as soon as they are made conscious of the subordinate identity. This suggests the power of imposed or induced self-deception – some or, indeed, many subordinate individuals adopting the dominant stereotype regarding themselves" (p. 65).

imprecise manner and 2. does not offer a proper argument for his evolutionary account of self-deception. The first criticism will be topic of discussion in this section. Van Leeuwen (2013b) holds that implicit in-group favoritism, stereotype threat (as an instance of imposed self-deception), placebo effects, false emotions, temporary fantasies, male overconfidence, false historical narratives are argued by Trivers to be instances of self-deception despite their differential nature. Therefore, the opponent of the latter concludes that the absence of common neural or psychological mechanisms precludes their treatment as a unified class, namely that of self-deception. Van Leeuwen's criticism raises the question about the subsumption of different phenomena under the concept of self-deception. In the literature on self-deception, the differentiation between our ascription practices of self-deception and the phenomenon of self-deception is often not met. This fact may have its explanation in the impossibility to analyze self-deception from the first-point of view *during* its occurrence, because per definition the self-deceiver does not know at the time of self-deception that he is self-deceiving himself. The tacit assumption is that by considering the ascription practices of self-deception we will learn about the phenomenon of self-deception (see also Bortolotti, 2010 for the defense of an analogue claim in the case of delusions and section 1.2.6). Van Leeuwen (2013b) mentions that the category of self-deception should be useful for scientific investigation. If Trivers subsumes different phenomena that, for van Leeuwen at least, have nothing in common, then it makes sense to ask whether Trivers could be treating self-deception as a cluster concept. For Wittgenstein, there are no necessary and sufficient criteria for the ascription of a certain cluster concept (1.1.3). Van Leeuwen's (2013b) worry is that if the category of self-deception encompasses a too large range of phenomena, then the concept of self-deception would lose its scientific value. This criticism concerns first and foremost Trivers' claim that bias is the hallmark of self-deception. The scope of the term "bias" is very wide: distortion of an information processing according to a certain criterion, often acquisition of truth. It is important to notice that, first, the flaws just mentioned are the ones that not only Trivers' account of self-deception suffers. Then, I will look more closely into Trivers again to search for factors that would narrow down the ascription of self-deception, because I agree that making the concept of self-deception too wide would be a weakness for a theory of self-deception. Mele's account of self-deception, like that of Trivers, also focuses on biases, namely motivated biases or those that "may be triggered and sustained by desires in the production of *motivationally* biased beliefs" (Mele, 2012, p. 7). Helzer & Dunning (2012) hold that "motivated reasoning emerges as a paradigmatic case of self-deception", if the contradictory belief requirement is abandoned (p. 390). Thus, Van Leeuwen's objection could be posed also with respect to Mele's account and Helzer & Dunning's (2012) claim: if every cold bias could be triggered by motivation, then every bias is – potentially – self-deceptive. Furthermore, Crawford & Salmon speak of self-deception as a process that renders information unconscious (Crawford & Salmon, 2004, introduction to part II of the collected volume, p. 97). Self-deception is said to tap processes by which information can become concealed from consciousness on different levels (Krebs, Ward, & Racine, 1997; Surbey, 2004, p. 121). The question now is how actually Van Leeuwen's criticism of the ubiquity of biases can be overcome by an appropriate restriction:

> *Van Leeuwen's definition criticism - usefulness*: It is questionable whether the concept of self-deception is a useful one if every bias is to be subsumed under it.
>
> *Restriction of self-deceptive biases - conditions*: If concealment of information from consciousness is the only criterion for the ascription of self-deception then every (subpersonal) bias would fulfill it.

Von Hippel & Trivers (2011b) hold that those information processing biases are self-deceptive that “favor welcome over unwelcome information in a manner that reflects the individual’s goals” (p. 2). Trivers (2011) mentions in his popular book that in order to discern self-deceptive biases from not-self-deceptive ones one has to “scrutinize our biases to see which ones serve the usual goal of self-enhancement or, in some other fashion, deception of others, and which ones subserve the function of rational calculation in our direct self-interest” (p. 148). As I have already noted in section 3.1.3.2, even if self-deception is restricted to self-enhancement, still, a huge variety of biases (among them a lot of those that Trivers names here) are susceptible to self-affirmation/self-protection strivings by which self-enhancement is tested. It is, then, often an issue of a definition whether one accepts some phenomenon as self-deceptive or not. Let me consider, for example, Jopling’s (2013) arguments against placebo being a case of self-deception. Trivers (2011) supposes that placebo is a case of *rationalization* (p. 72). Jopling (2013), on the contrary, denies the self-deceptive quality in placebos for the following reasons.

1. The fact that provided definitions of placebo “are too vague, too wide, too narrow, or guilty of smuggling in theories of causation and mechanism, thereby, conflating explanation with definition” (Jopling, 2013, p. 1207);
2. The fact that placebo and “narrative truth” or “the internal coherence of the interpretation” can be considered as alternative explanations, so that narratives in placebo are truth-tracking, while in self-deception they are not, but are instead “[f]reed from the constraints of objective historical and psychological fact” (Jopling, 2013, p. 1211);
3. *Open-label placebos* (those known to be placebos by patients) are not cases of self-deception, because patients had reason to believe in their effect, because “they were told, crucially, that placebo pills ‘have been shown in rigorous clinical testing to produce significant mind-body self-healing processes’” (Jopling, 2013, p. 1219).

It is worth mentioning how at least the second point – whether self-deception is in a certain respect reality-constrained or not – has been subject to debate. Let me consider the framing effect. Is it a case of self-deception, whether one thinks that the glass is half-empty or half-full? Tversky & Kahneman (1981) have discovered that phrasing a choice in terms of wins or losses changes the decision and the risk aversion in an intuitive manner: If a choice is framed in terms of wins, then the subjects are more prone to undergo risks. What if I would present a story about an individual whose risk aversiveness has been self-deceptively manipulated, so that this individual engaged in risky behavior, and would argue that Tversky & Kahneman’s heuristic is a subpersonal mechanism that has made such a case of self-deception possible?

In the remainder of the section I want to explore the second point of Trivers’ definition of self-deception – that it is a misrepresentation that has success conditions.[287] Smith (2014) has recently presented a teleofunctional non-intentional account of self-deception. He argues that self-deception is success-apt, or, in other words, that one could succeed or fail to self-deceive and proposes to explain success-aptness in self-deception as a case in which self-deceiver possesses a mechanism that causes the self-deceiver to form misrepresentations:

> O is self-deceived iff O possesses character C with purpose F of correctly representing some feature of its world, and character C* with purpose F* of causing C to misrepresent

[287] The following quotation highlights the agential component and that of information processing: “We *hide reality* from our conscious mind the better to hide it from onlookers. We may or may not store a copy of that *information* in self, but we certainly act to exclude it from others” (Trivers, 2011, p. 9; my emphasis).

> that feature, and it is in virtue of performing F* that C* causes C to misrepresent. (Smith, 2014, p. 191)

The proper function of this mechanism is said to be the one of inducing a failure of the representational apparatus (p. 192), but Smith (2014) sees his view also compatible with both Trivers' and Van Leeuwen's theories regarding the function of self-deception (evolutionary function vs. by-product; see section 3.2), insofar as self-deception might "acquire its purposeful character in consequence of being reproduced cognitively and behaviorally, rather than genetically" (p. 197). Smith (2014) further assumes that the proper function that Mele ascribes to self-deception is the one of minimizing costly errors (p. 194). I think that two kinds of success conditions could be distinguished in self-deception (in analogy to the proximal and distal levels to be elaborated below): 1. that one succeeds to cause one's representational apparatus to fail, to borrow Smith's terminology; 2. that one succeeds to fulfill the function of self-deception by means of 1, e.g. that one succeeds in deceiving others by succeeding to deceive oneself. I think that the conditions for one are easier to formulate.

In the case when one accepts the description of self-deception as a self-enhancing misrepresentation acquired on the basis of ambiguous information, Zehetleitner and Schönbrodt's (2014) conditions for a successful misrepresentation[288] might be used: presence of noise, asymmetry in the cost of errors and presence of indicator representations. They argue that misrepresentations are successful if there is a certain amount of information loss and either a non-uniform a priori frequency of events or an asymmetric utility matrix. To borrow the example from the article under consideration, if you have to decide whether to eat a blue or a green snake, knowing that the ones with the blue color are poisonous while the green ones are not, then, if you cannot always correctly represent the color of the snake, misrepresenting more greenish ones as blue would have adaptive benefit, because the utility matrix is non-uniform – eating a snake has less value than being poisoned. What such a definition of a successful misrepresentation is presupposing is a tight relationship between hits and false alarms: False alarms are better than misses, so a false alarm is almost as good as a hit. Interestingly. Hesselmann et al. (2010) argue against an evidence accumulation model (bias towards true hits or false alarms) for a predictive coding model (bias only towards true hits or correct rejections) on the basis of fMRI pre-stimulus activity levels in the case of perceptual (visual and auditory) decisions. In the motion experiment subjects had to recognize coherent motion and in the auditory experiment – target sound despite noise. The assumed pre-stimulus activity levels for the evidence accumulation model (hits + false alarms > misses + correct rejections) were not observed, while those for predictive coding (hits + correct rejections > misses and false alarms) were confirmed.

In this section I presented Trivers' definition and categorization of self-deception and the desiderata for an account of self-deception that would follow from it. I, furthermore, embedded Trivers' account into the discussion context and argued that not only Trivers' account, but also Mele's and Surbey's account suffer from the modified van Leeuwen's criticism that enlarging the concept of self-deception to encompass every bias is not scientifically useful. I further argued that in the case of evolutionary account it is the supposed function that provides the unit of unification and determines the usefulness and not the supposed proximal subpersonal mechanisms that serve to fulfill the given function.

[288] Zehetleitner & Schönbrodt (2014) posit that to have a representation is to be in a state that has a property to have content, where having content is a mapping from representations to external states.

3.2.2.2 Distal level of evolutionary explanation

In the previous sections I presented Trivers' definition and categorization of self-deception. The current section will build on those and display his argument that self-deception has the function to deceive others. I will argue that Trivers' hypothesis that self-deception evolved to deceive others is plausible, but emphasize that all contemporary psychological means by which it is currently tested are about its function in the *present*, but not in the *ancestral* environment. Last, I will present two models – that of Byrne & Kurland (2001) and Byrne & Kurland (2001) and Johnson & Fowler (2011) – that have tested Trivers' hypothesis by means of evolutionary game theory.

The rationale of Trivers' evolutionary argument is as follows: Animal communication involves a frequency-dependent co-evolutionary struggle between deception and its detection which has led to a new kind of deception – self-deception (Trivers 1985, p. 395) whose function is to deceive others in the absence of cues normally associated with deception (von Hippel & Trivers 2011b, pp. 2-4). These cues are nervousness, suppression of physical indicators of nervousness (e.g. high pitch), cognitive load[289] and idiosyncratic signs. Thus, self-deception is argued by von Hippel & Trivers to be favored by natural selection and to enhance the inclusive fitness of the self-deceiver.[290]

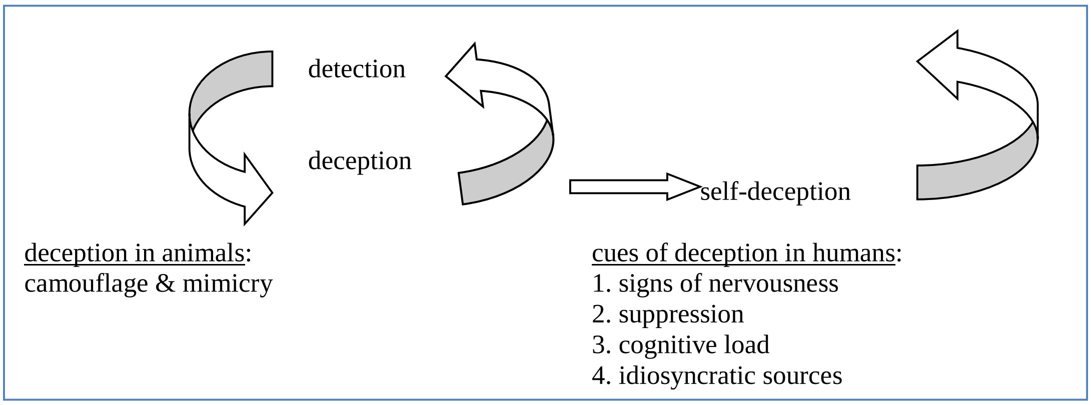


Figure 23. Trivers: detection-deception circle
Distinctions from Trivers (1985, p. 395) and von Hippel & Trivers (2011b, pp. 2-3).

The thick arrows demonstrate the dependency of detection and deception: the better the deception, the more pressure there is to develop countermeasures. The same relationship is between self-deception and its detection. The thin arrows indicates the evolutionary process that led to the development of self-deception.

The act of deception in question encompasses not only gross cheating which is a failure to reciprocate, but also subtle cheating which is reciprocating to a lesser degree than the altruist has reciprocated. The distinction between gross and subtle cheating is given in Trivers (1971, p. 46) and its connection to self-serving biases in general and self-deception in particular is being further elaborated in the recent literature:

> It was expected that this 'subtle-cheating bias' would have been adaptive because it would motivate individuals to repay less than what was actually required. I proposed that this bias could involve *self-deception*, as people might not realize they are subtly cheating

[289] Cues as pausing or simple sentence structure give away the fact of two types of content being processed at once. This processing leads to cognitive load (von Hippel & Trivers, 2011b, p. 3).

[290] Inclusive fitness should not be confused with reproductive success in this context: "Natural selection favors not the individual that maximizes her reproductive success but the individual that maximizes her inclusive fitness, where this includes personal reproductive success plus effects on relatives, devalued by the appropriate degrees of relatedness." (Trivers, 1985, p. 65)

> others, and vehemently deny it if they were caught. Wright (1994) mentioned some similar ideas in this popular book about evolutionary psychology and morality. Wright suggested that self-serving biases in social accounting are ‘a corollary to the theory of reciprocal altruism’ (p. 277). He further remarked that ‘humans seem unconsciously compelled to give a bit less than they get’ (p. 277). (Janicki, 2004, p. 63; my emphasis)

I recapitulate von Hippel and Trivers’ argumentation in favor of the selectivity of self-deception as follows: Selection pressure on deceptive abilities is high because of the ability of others to detect deception which leads to a high selective pressure on the evolution of self-deception (von Hippel & Trivers, 2011b, pp. 3-4). Thus, Trivers’ argument in favor of the selectivity of self-deception is the usefulness of deception, on the one hand, and availability of means to detect deception, on the other hand. In order to substantiate this claim, von Hippel & Trivers (2011b) enumerate weaknesses of studies that suggest humans detect lies poorly (p. 3). Following von Hippel & Trivers and on the premises of dual process theory, ten Brinke et al. (2014) have tested whether implicit (using IAT) or unconscious (using sematic-classification task) lie detection is more accurate than conscious one. The rationale was that “corrective” deliberate processes distort implicit lie detection (ten Brinke et al., 2014, p. 3). The aim of those was measuring the association of supraliminally/subliminally presented images of people who either did or did not steal 100$ with categories “truth” or “lie”. The result of ten Brinke et al’s (2014) supported the claim that unconscious lie detection is more accurate. Remember that proving that self-deception had an evolutionary function in ancestral environment is not the same as proving that it possesses it now. Thus, independently of whether humans are good or bad lie detectors in the *present*, the question is open whether it was the case in the *ancestral environment*. In this respect, proving the selectivity condition depends on the elaboration of the environmental conditions under which self-deception is supposed to have evolved.

If self-deception evolved to deceive others this mean that one needs to show that self-deception promoted inclusive fitness under Pleistocene conditions and may promote it in the current environment. In other words, Trivers has to answer the objection mentioned in section 3.2.1 about the absence of critical information about the Pleistocene conditions in order to be able to substantiate the part of his claim about the ancestral fitness. Bandura (2011) formulated the question in his commentary to the recent article of von Hippel & Trivers (2011b) on the topic as follows: “Given the evolutionary billing of the article under discussion, what were the ancestral selection pressures that favored self-deception?” (Bandura, 2011, p. 16) This question can be taken as a criticism or as an appeal to the researchers to find the answer in the future (Gangestad, 2011, p. 24). Dunning (2011) argues against Trivers that in small groups self-deception does not work. Dunning’s thesis presupposes that ancestral groups were small. He supports his claim by citing *contemporary* psychological studies that prove that boasting individuals are not liked by others (p. 19). Von Hippel & Trivers (2011a) reply that at this point Dunning conflates the terms boasting and self-enhancement and that self-deception happens only to a plausible extent (p. 43).

Apart from the question about the adaptivity of self-deception in the ancestral environment, if the claim is argued for that it is adaptive in the present, then another kind of concern has to be answered: If self-deception’s functional role is to serve as a substitute of deception, then it should not be useful in cases where others recognize it. In this case, either one constrains the definition of self-deception to cases where it is not recognized, or one has to present the argument in favor of the position that, in at least the majority of cases, self-deception is not recognized, which allows it to exert its function (Van Leeuwen, 2007b, p. 335). Von Hippel & Trivers seem to favor the second way to resolve the given problem with the plausibility constraint, implying that the evidence presented in favor of the beneficial role of self-enhancement in their article accounts for the majority, yet more evidence would be needed to substantiate this claim.

Let me now discuss two implications of Trivers' distal function of self-deception. One implication thereof is that self-deception is a *social* phenomenon (see also previous section). Trivers argues that if happiness, optimism and confidence can count as self-deceptive phenomena, they are also of interpersonal nature.[291] In his earlier work on self-deception Trivers did not assign as much weight to the social dimension of self-deception, as he does now and argued that positive illusions can benefit the individual independent of their influence on others (Trivers, 2000, p. 125). This emphasis on the social dimension stands in contrast to other (defensive) theories of self-deception (von Hippel & Trivers, 2011b, p. 14) and I agree with Trivers that the social dimension of self-deception merits further analysis. Trivers holds that insofar as self-deception allows reality-shaping in an interpersonal context, it is an *offensive* strategy. It is questionable though, whether this kind of strategy needs to be intentional, as von Hippel & Trivers (2011a) advocate for, since it might be argued that intentions are folk-psychological concepts that should not be postulated at the unconscious level for reasons of parsimony:

> If someone else managed to miss the point, let us state it clearly – we certainly believe that self-deception is intentional, in the sense that the organism itself intends to produce the bias, although the *intention could be entirely unconscious.* That is, humans have been favored by natural selection to self-deceive for its facilitative effect on the deception of others. (von Hippel & Trivers, 2011a, p. 42; my emphasis)

I argued in section 2.2.1 that self-deception is goal-directed and that those goal representations might be subpersonal. Context-sensitivity is not necessarily a property of persons as a whole. The fact that others from a third-person perspective can plausibly ascribe a certain kind of intention to the self-deceiver and the fact that the self-deceiver himself can rationalize her behavior to a certain degree do not provide insight on the way the information is processed and subpersonal goal representations are bound into the epistemic agent model (see section 2.2.2.3).

The second implication of Trivers' theory of self-deception, that I want to mention, is that animals could be self-deceived.[292] Trivers (2010) argues that there is non-verbal, as well as verbal self-deception (p. 382). The first one occurs, at least, in the context of aggressive conflict and courtship, while the second one – in the context where language allows communicating false information: "Once you have language, you have an explicit theory of self, social relationships and the world, ready to communicate to others" (p. 382). Scott-Kakures (2002) argues against the claim that animals could be self-deceived by holding that reflective reasoning is an essential element of self-deception and, thus, the latter is a distinctively human capacity (p. 585). According to him, the self-deceiver must actively participate in his self-deception (p. 591), as he is a hypothesis-tester that reflectively reasons for a certain conclusion (p. 586). Moreover, the reflective reasoning condition allows the differentiation between self-deception and wishful thinking (p. 591, p. 600). It

[291] Self-deception as interpersonal: "We disagree with his characterization of happiness, optimism, and confidence as strictly internal. Instead, we regard all three of these states as being of great *inter*personal importance, given that they signal positive qualities about the individual to others" (von Hippel & Trivers, 2011a, p. 42).

[292] Trivers (1985, p. 413) hypothesizes that an adult male chimpanzee could be self-deceived about his object of interest – fingernails – when being chased away from a female in estrus by another adult male.

further satisfies the intuition that self-deceivers embrace a certain form of irrationality – violate their own standards of evaluation (p. 591) - and accounts for cases in which reflective reasoning itself is the biasing mechanism that shapes the attitudes of the self-deceiver (p. 592). Mele (2012) accepts the proposition of Scott-Kakures to add reflective reasoning to his set of jointly sufficient conditions for entering self-deception that were discussed earlier (see table 14). In chapter 2 I argued against such an overrationalized description of the self-deceiver.

Till now, I have reviewed arguments in favor of the hypothesis that self-deception has the function to deceive others. The unelaborated question is whether this is the *only* evolutionary function that self-deception possesses. Surbey (2004) holds that facilitation of deception is only one of the following other potential *evolutionary functions*:

- reduction of cognitive load;
- facilitation of deception;
- promotion of family bonds;
- facilitation of reciprocal altruism;
- maintenance of mental and physical health;
- sequestering of threatening thoughts and memories.

According to Surbey (2004), self-deception serves not only the function of the facilitation of deception, but also the function of cognitive load reduction, because e.g. *unconscious* processing accounts for the information that is distracting and interfering with daily functioning (p. 123). Self-deception is further argued to strengthen family bonds and promote cooperation among kin and non-kin (reciprocal altruism), because e.g. overly positive partner images strengthen the relationship and cooperation is positively tied to scores on the SDQ if tested using hypothetical scenarios based on the iterated Prisoner's Dilemma game[293] (Surbey, 2004, pp. 126–127; Surbey & McNally, 1997). Surbey also holds that self-deception promotes health maintenance, because e.g. it leads to happiness and the absence of depression (pp. 129-131) and that it serves the defensive function of threat suppression, as e.g. it negatively relates to anxiety and negative emotions in general (pp. 131-133). She names several *selective pressures* that led self-deception to develop given functions are the following: cost of consciousness, detection of deception, establishing harmonious family relationships, mutual cooperation with non-kin and fitness-enhancing goals (p. 122). Trivers' argument contra Surbey would probably be that all the other functions of self-deception mentioned by Surbey can be reduced to the facilitation of other-deception. I think so because in response to Pinker's (2011) objection that promotion of optimism and health is a strictly internal function (p. 36), von Hippel & Trivers (2011a) disagree, emphasizing their interpersonal importance, "given that they signal positive qualities about the individual to others" (p. 42).

The comparison between Trivers' and Surbey's argumentation highlights the underdetermination/ambiguity of the function of mental traits as a result of the absence of sufficient historical information that makes the clarification and substantiation of an evolutionary account of self-deception difficult and in need of an additional empirical proof, granted the assumption that given proof can be acquired. For self-deception having a distal evolutionary function of facilitating deception of others, it has to be a heritable, preferably derived trait that has been selected for given certain ecological factors in an environment with a preferably homogeneous population.

[293] Iterated Prisoner's Dilemma is a game paradigm with a certain payoff matrix that determines the highest reward for the case that one defects while the partner cooperates and the lowest for the case that one cooperates and the partner defects (Surbey, 2004, p. 127). Cooperation on the first move in necessary for a Tit for Tat strategy in the Prisoner's Dilemma (p. 127-128). The Tit for Tat strategy is retaliatory insofar as defections are punished (p. 128).

Given the criticism that the absence of relevant knowledge about ancestral environment precludes any defending of an evolutionary hypothesis about mental traits, modeling of self-deception in an evolutionary game can be seen as an answer to the given worry. Though this modeling does not replace the absence of evidence about ancestral environment, still, such successful modeling can be counted as *hypothetical* evidence in favor of a given evolutionary hypothesis. It is hypothetical in the sense that certain assumptions about the ancestral environment are met, then modeling takes place and inferences are drawn about whether a certain trait *could have evolved* or not.

The evolutionary game best suited to model self-deception is, according to Trivers (2010), the *iterated Prisoner's Dilemma* in which each of two players in a series of encounters has to decide whether to cooperate or defect given a certain payoff-matrix or the *Ultimatum Game* in which a certain split of money is proposed by one player to the other and can be either rejected or accepted (p. 389-390). The reason for the choice of these two evolutionary games is that trust is here a key variable and deception is possible (p. 390). Responding to Trivers' appeal to flesh out an evolutionary model of self-deception (Trivers, 2010, p. 391), some researchers have taken up the challenge and modeled the evolutionary development of self-deception. I will consider the modeling attempts of Byrne & Kurland (2001) and Johnson & Fowler (2011).

Byrne & Kurland (2001) test Trivers' evolutionary hypothesis that self-deception evolved the better to deceive others in an *asymmetric hawk-dove game* (p. 461). I will, first, describe the procedure and, then, answer the question of how Trivers' hypothesis is accounted for by this kind of modeling. The game procedure is the following: For several generations, a member of a population is matched in a two-player hawk-dove game against another member. The results of these encounters determine the composition of the population. This composition is compared across generations (p. 461). The hawk-dove encounter is described by them as follows:

> Two people simultaneously discover some food and each tries to intimidate the other and take the food. The hawk strategy (*H*) is to fight if the other player does not back down. The dove strategy (*D*) is to display fighting prowess but always retreat rather than engage in an actual fight. (Byrne & Kurland, 2001, pp. 461-462)

The payoff matrix for encounters between two subjects depends on the fighting cost and the value of the resource (for the payoff matrix see Byrne & Kurland, 2001, pp. 462). Byrne & Kurland use the modularity hypothesis as a basis for their model. They define the modularity assumption as follows:

> In cognitive neuroscience, the term *'modular mind"* refers to a collection of semi-autonomous decision-making and computational devices, each with its own algorithmic agenda […] (bold emphasis added; Byrne & Kurland 2001, p. 458). Thus, the mind is conceived as a collection of task-specific mental organs, that are designed by evolution to solve particular cognitive problems like language acquisition, color perception, social cheating and inferential reasoning […]. (Byrne & Kurland 2001, p. 459; my emphasis)

These modules compete with each other and are responsive both to internal psychological mechanisms and external stimuli while trying to regain control over the agent (Byrne & Kurland, 2001, p. 459). If one of the modules is stronger in some particular situation, no inter-module conflict will arise. If two or more modules are nearly equally strong (with respect to a threshold ε, below which ambivalence arises), then some strategy should be used to resolve the conflict. Byrne & Kurland (2011) take as a starting point the definition of self-deception as asserting “a belief *B* despite evidence that *not B*” (p. 458) and hold self-deception to be an inter-module conflict resolution mechanism[294] in the sense described above (p. 474). The authors have selected the simplest non-trivial modular mind to test whether players possessing the self-deceptive type would outcompete the ones possessing the non-self-deceptive type.

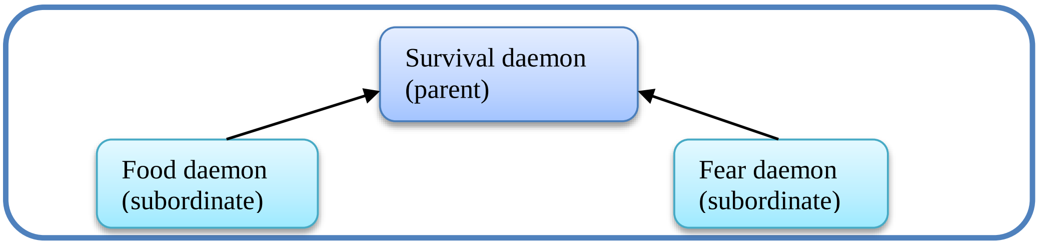


Figure 24. Byrne & Kurland: Simplest Non-Trivial Modular Mind. Distinctions from Byrne & Kurland (2001).

The simplest non-trivial modular mind consists, according to Byrne & Kurland (2001), of one parent daemon with two subordinate daemons. Byrne & Kurland define daemons as follows: “we and Pinker use *daemon* to denote infinitely stupid devices, which are mechanistically poised to detect highly restricted stimuli and then generate highly defined responses.” (Byrne & Kurland, 2001, p. 459) The food daemon assesses the value of the food, the fear daemon assesses the cost of a fight (p. 462).

Self-deception is modeled in a situation that concerns survival: try to get the food or avoid being injured in a fight - on the premise that only one module can keep control of the player at a time: if the survival daemon is ambivalent about the decision mentioned above, then the survival daemon (a module) cannot keep control of the player. Self-deception is, according to Byrne & Kurland (2001) a means of resolving such ambivalence by devaluing the value of the one of the variables (to zero), food or fear. It happens by making updates to the devalued variable – temporary, for the time of an encounter – impossible (p. 465). The assessment phase of an encounter of the two players consists of the initial assessment of the value and cost, then the display of this assessment to the opponent, the update of one’s own assessment (p. 464). Self-deception can occur either before or after the update (p. 463). Byrne & Kurland (2001) differentiate between two types of self-deception: SDF (fear is suppressed) and SDH (hunger is suppressed). The suppressed information is kept unconscious for the time of the encounter with the agent who is to be fooled (p. 466). The authors note that the kind of suppression noted above is supposed to capture Gur & Sackeim’s criteria for the ascription of self-deception that have already been discussed in the previous chapters (Byrne & Kurland, 2001, p. 461).

Given this framework, how is Trivers’ hypothesis reflected in it? First, Byrne & Kurland (2001) equate “player signals with player beliefs” (p. 465). The equation of signals with beliefs assures, according to them, that “any deception other than that resulting from self-deception is ineffective” (p. 465), because it will be detected. Second, they introduce the variable λ which measures the susceptibility of the opponent where the latter is the extent to which the opponent believes the sender’s signal (p. 465). If the opponent is highly susceptible, he will not test the self-deceptive signal for reliability:

[294] It is important for ambivalent cases involving survival daemons. In the Minsky’s model such a conflict is normally resolved through losing control by the parent daemon, but when survival is at stake, the parent daemon should keep control. For this it uses self-deception (Byrne & Kurland, 2001, p. 463).

> Self-deception is not a strategy that is chosen by a player but rather it is a phenomenon occurring automatically within a player and independent of the *susceptibility* of the opponent. Therefore, the net benefit of self-deception must be measured over a spectrum of encounters across which λ varies. In this way, one can discuss an expected cost, if the signal is tested with some probability. [295] (Byrne & Kurland, 2001, p. 467)

The modeling results are that the non-self-deceptive type is eliminated from the population, while the mix of SDF and SDH remains for all values of ε and λ, except for the extreme ones (p. 470). At the extremes, for high ratio λ/ε the resulting population consists of pure-SDF, for low ratio it consists of pure-SDH. The result that the authors draw is that self-deception would have evolved according to this model.

The potential problem with the way that Byrne & Kurland defend Trivers' evolutionary theory of self-deception is the modularity assumptions. The daemons that the mental architecture of Byrne & Kurland consisted of are modules. Von Hippel & Trivers (2011a,b) argue against the modularity view and in favor of a distinction between different levels of consciousness (see the following section). Can the given modeling attempt be counted as evidence for Trivers' hypothesis despite this fact? Von Hippel & Trivers (2011a) reject the modularity view by arguing that it does not provide an easy solution, because it does not allow omitting the concept of the self or the distinction between different levels of consciousness (p. 45). Thus, at first glance this is not the case that their theory is incompatible with the modularity assumption. Let me consider certain features of the modularity view with the question in mind whether these features make von Hippel & Trivers' and Byrne & Kurland's accounts incompatible. The modularity view by itself does not answer the following questions:

1) whether the modules have a neuronal correlate in the brain or are just virtual features (Byrne & Kurland, 2001, p. 459);
2) whether there is some central executive (Byrne & Kurland, 2001, p. 463): Though Byrne & Kurland hold that in the radical version of mental modularity there is no central executive of the self (p. 458), holding the modularity view does not equate to holding a radical version of the modularity view;
3) whether modules use rational decision-making strategies: In Byrne & Kurland's model, the players do not update their beliefs in the Bayesian fashion (pp. 473-474)

Neither Byrne & Kurland, nor von Hippel & Trivers, adopt a position on the first issue. Considering the second issue, von Hippel & Trivers (2011a) argue that the absence of a central self disrupts sustained goal pursuit, because the modules would be each time activated by different external events (p. 45). Byrne & Kurland differentiate between the parent and subordinate kind of daemons (modules), but do not argue explicitly for architecture with or without a "central self". It is an empirical question whether, if an architecture modelling a central self is chosen, the results would also speak for an evolution of self-deception, granted that there is agreement on how to model the concept of a central self in the first place. The third question concerns how the information is processed in the mind. Byrne & Kurland's (2001) players perform belief updates in non-Bayesian fashion (pp. 473-474). Von Hippel & Trivers do not commit oneself to a certain position regarding this question. Yet, if one wanted to flesh out their account using, for example, the predictive coding approach, which views the mind as a hierarchical Bayesian prediction error minimizer (Clark, 2013b; Hohwy, 2013b), then the given modeling results might not be the appropriate kind of evidence, because they are too simplistic. But they are a good starting point. I want to conclude this discussion on the modularity assumption in Byrne &

[295] *susceptibility (λ)* = the degree to which a player's signal affects the beliefs of the opponent (Byrne & Kurland, 2001, p. 467)

Kurland's modeling attempt by quoting Dennett on the utility of Minsky's agents (homunculi or demons) that served as the basis for the given modeling attempt:

> Homunculi – demons, agents – are the coin of the realm in Artificial Intelligence, and computer science more generally. Anyone whose skeptical back is arched at the first mention of homunculi simply doesn't understand how neutral the concept can be, and how widely applicable. Positing a gang of homunculi would indeed be just as empty a gesture as the skeptic imagines, if it were not for the fact that in homunculus theories, the serious content is in the claims about how the posited homunculi interact, develop, form coalitions or hierarchies, and so forth. (Dennett, 1991, p. 261)

Johnson & Fowler (2011) analyze overconfidence – that is, according to them, an instance of self-deception, because no bluffing was involved (p. 319). Their model is similar to Byrnes & Kurland's, but is a more general model than the Hawk-Dove game insofar as it allows *intermediate values* on the continuum confident – not confident.[296] The encounter procedure between hawk-players and dove-players is similar: resources are claimed, a fight determines the winner who receives a benefit and a loser who suffers the cost. The decision whether to fight is met on the basis of the assessment of the capability of the opponent. Whereas in the model of Byrne & Kurland there was the susceptibility term that determined the degree to which one believes the opponent's signal, Johnson & Fowler instead introduce the error term that measures the uncertainty about the opponent's signal. Johnson & Fowler come to the conclusion on the basis of their modeling results that given a certain benefit/cost-ratio (sufficiently large benefits compared to the cost) and a certain amount of uncertainty about the competitors' capability holding incorrect beliefs about one's own capability can be rewarding (p. 317).

Are the presented models appropriately complex to be the models of self-deception? Both models view players as non-Bayesian, quite simple reasoners. The mental architecture is underspecified. The situation is constrained to incorporate only one goal – to claim the resource suffering minimal cost. Can such kind of modeling be the evidence for the function of self-deception? In the case of Byrne & Kurland' model, self-deception is modeled by simply devaluing the food or fight variable and allowing the opponent to detect such devaluation depending on how big the susceptibility variable is. In the case of Johnson & Fowler's model, self-deception is modeled by introducing an error term about the capability of the opponent.

All in all, it might have been better if these evolutionary games had equipped their players with Bayesian reasoning capabilities (4.3). Further, given that our knowledge about the ancestral environment is limited, we cannot constrain the environment in evolutionary games accordingly to the ancestral environment. The underdetermination will remain. What the results of the two modelling attempts provide are possible formal and ecological validity *constrains* under which self-deception might develop. Whether these constraints were really part of the ancestral environment, is another question. Moreover, the conceptual connection between impaired updating of information as a model of the process of self-deception in evolutionary games, as well as impaired updating in predictive coding as an explanation of delusions (1.2.7) is to be pointed out. To remind the reader, prediction errors are updated according to Bayesian rules and too much or too little weight on either the

[296] Hawk-Dove game as a special case (with discrete values) of the overconfidence model: "We find that the Hawk-Dove game is a special case of our model, in which the only possible strategies are to be infinitely overconfident ([…], Hawk) and therefore always claim the resource, or infinitely underconfident ([…], Dove)" (Johnson & Fowler, 2011, p. 319).

prediction error or priors expectations would lead to either over-updating or under-updating (Hohwy, 2013; Frith & Friston, 2013).

3.2.2.3 Proximal level of evolutionary explanation

In the previous section I have presented Robert Trivers' distal explanation for why self-deception evolved, namely to enable better deception of others. In this chapter I will be concerned with the hypothesized proximal mechanisms by which means it is made possible for self-deception to fulfill its evolutionary function.

The implicit assumption in the hypothesized function of self-deception to deceive others is that there is a conflict of interests. According to Trivers, the entities, between which conflict arises, can not only be individuals, but also certain their *parts*. I think that his view can be seen as a further elaboration of Davidson's (1.1.1.1) and Pears' (1.1.1.2) views. They explained self-deception by making divisions into subsystems for information storage (Davidson) or even agency (Pears). Trivers, on the contrary, offers some mechanisms as to how such a division can take place. There are two kinds of proximal mechanisms according to him that may create conflicting behavior: genetic and psychological. Genes that are differentially related to each other may, in virtue of different interests, lead to conflicting personal level behavior. Psychological dissociations (e.g., conscious-unconscious, automatic-controlled) may also lead to contradictory behavior. How exactly the dissociations on the genetic level lead to dissociations on the psychological level, remains an open question. Thus, Trivers is mostly concerned with the first of the two behavioral features of self-deception, namely processing of inconsistent information that can differentially influence behavior (the other would be justification, see section 2.1.2) and leaves open the question about the phenomenology of self-deception.

The general premise of interactions on the genetic level is that the genotype influences the phenotype[297] via the following chain of transformations: DNA → RNA → protein.[298] Genetic relatedness is the key for understanding the dissociations on the genetic level (Trivers, 2011, p. 77). How can degrees of relatedness be calculated?

> The general rule for calculating a degree of relatedness is obvious. As we trace a genealogical connection from one individual to another, we *multiply* the r's that connect the various individuals. If two individuals are related through more than one genealogical connection, we *add* the values together that are obtained for each genealogical connection. Note that it is not necessarily true that one's individual's degree of relatedness to another is the same as the other way around. […] Thus, we want to be careful always to specify the *direction* of relatedness (A's relatedness to B) and, in calculating this r, to begin with A and end with B. (Trivers, 1985, p. 46)

Trivers (1985) argues that evolution favored the discrimination of degrees of relatedness in individuals (p. 46). This happens by (subpersonal) learning some standards of comparison

[297] Definition of genotype and phenotype: "We can list all of an individual's genes; we call such a list the individual's *genotype*. The resulting structure, physiology, and behavior we call the individual's *phenotype*" (Trivers, 1985, p. 89).

[298] The way in which genes determine traits: "Since DNA → RNA → protein, each gene may be thought of as a portion of DNA that creates a single protein. In this sense a single gene determines a single trait (protein). But since a protein can have many different effects on other parts of the body, each gene typically affects many traits. Likewise, each trait typically requires the action of many different genes" (Trivers, 1985, p. 93).

by the individual.[299] What is the relationship between degrees of relatedness and self-deception? Trivers' general argumentative line is the following: Different degrees of relatedness between individuals lead to conflict, because natural "selection favors altruism whenever Br > C, where B is benefit and C is cost" (Trivers, 1985, p. 65). Yet r, which is the degree of relatedness, may be different for every participant of the social interaction. This kind of conflict is reflected in the analogical kind of conflict between parts of the individual's genome according to the inheritance rules for these parts (Trivers, 2000, p. 123), for example mitochondrial DNA (mDNA) is passed only from mother to offspring (Trivers, 2000, p. 123). Self-deception is, then, the means to resolve the conflict.

The different degrees of relatedness that Trivers (2000) focuses on are those between the individual and its parents, but I presume that the argumentation is extendable to all other kinds of relatives as well. The parental conflict arises because parents are related to their children only by a half and to other relatives in the different degree than children themselves.[300] Trivers speaks of the set of the parental genes as the parental self, as well as of the set of the maternal genes as a maternal self. The parental conflict expresses itself as a conflict within the individual between paternal and maternal genes (the analogical conflict is present in behavior). To clarify the impetus to self-deceive in parental conflict, I will exemplify the case of parental domination. In the case of parental domination the "conflict may most easily be handled by rendering negative feelings toward the parent unconscious. Thus, we also expect *self-deception* in these interactions, rendering some facts unconscious on both sides the better to deceive others" (Trivers, 1985, p. 165). Self-deception on the genomic level is achieved via

- Genomic imprinting;
- Sex antagonistic genes.

Genomic imprinting or parent-specific-gene-expression refers to the difference in the degree of maternal/paternal gene expression that depends on the sex from which these genes were inherited[301] (Trivers, 2000, p. 123). *Sex antagonistic genes* are those genes that affect reproductive fitness in opposite ways depending on the sex that they are found in (Trivers, 2000, p. 125). Thus, in the case of genomic imprinting the expression of a certain gene changes depending on the sex of the parent *from which they were inherited* and in the case of sex antagonistic genes the difference in the effect of the gene on the organism is dependent on the *sex that they are found in*. This difference in the expression of paternal vs. maternal genes can also be found on the level of tissue: some tissue allowing maternal genes to express themselves more and vice versa (Trivers, 2000, p. 124). According to Trivers, genetic imprinting and sex antagonistic genes can evoke self-deception. Here are

[299] Discrimination of degrees of relatedness: "We can conclude that genes do not recognize themselves directly in other creatures, bur rather, in many animal species, individuals are able to measure relatedness by learning some standard of comparison, such as the self, by which others are then discriminated" (Trivers, 1985, p. 134).

[300] Conflict between parents and offsprings in the case when the cost of an altruistic act is bigger than the benefit, but smaller than twice the benefit:
"Consider parents with two offspring. Imagine that the first offspring is considering an altruistic act toward the second. It is only selected to act altruistically whenever B > 2C. The parents are equally related to the two offspring and would therefore enjoy a gain in reproductive success whenever one offspring acted altruistically toward the other at B > C. […] Since there must exist situations in which B < C < 2B, we expect to see conflict between parent and offspring over the selfish tendencies of the offspring." (Trivers, 1985, p. 162)

[301] Conflict at tissue-level: "Thus, it is possible that there are conflicts at the level of tissues in which one can also imagine *selves*-deception, that is, deceitful signals sent out from one tissue, overemphasizing one parent's interests whose signals are devalued by another tissue, overemphasizing the opposite sexed parent's interests." (Trivers, 2000, p. 123)

two examples. In the first genetic imprinting influences individual's stance toward inbreeding and in the second sex antagonistic genes suppress negative traits. This suppression allows, according to Trivers, the genes to remain in the genotype, because their cost has been eliminated due to it (p. 125).

> You are related on the maternal side and will thus enjoy an increase in relatedness to any resulting offspring by inbreeding on the maternal side, but the paternal genes will enjoy no increase in relatedness though they will suffer any inbreeding depression associated with the inbreeding. We can imagine your maternally active genes urging you to consider the inbreeding while your paternally active genes might take a moralistic posture and emphasize the biological defects thereby generated. (Trivers, 2000, p. 124)
>
> It is easy to imagine an interaction between this mechanism and parent-offspring conflict, since parents may help you locate—and encourage you to suppress—such negative traits, but due to imperfect overlap in self interest, they may encourage you to think a trait negative to yourself when it is in reality only negative to themselves. Similarly, it is conceivable that paternally active genes (for example) may attempt to suppress maternally active ones (or vice versa) by pretending that it is an organism-wide negative phenotypic trait that needs to be suppressed. (Trivers, 2000, p. 125)

Surbey (2004) interprets Trivers as suggesting that intragenomic conflict leads information to be processed "at different levels of consciousness" (p. 134). The same may be argued about the dissociations postulated by von Hippel & Trivers (2011a,b). Von Hippel & Trivers (2011b) hold that there exist different levels of consciousness (p. 13) and a variety of dissociations between mental processes[302] (p. 6). The function of the overlapping kinds of dissociations is namely argued to consist in limiting the "conscious access to the contents of their own mind and to the motives that drive their behavior" (von Hippel & Trivers, 2011b, p. 6). Thus, ensuring the processing of information at different levels of consciousness is seen as essential (and evolutionary older) to self-deception (Trivers, 1991, p. 179). The three kinds of distinguished dissociations are implicit/explicit memory, implicit/explicit attitudes, automatic/controlled processes (von Hippel & Trivers, 2011b). Von Hippel & Trivers (2011b) define explicit memory as the one which includes information that can be *consciously* recollected, while implicit memory is the one which includes information that cannot be consciously recollected (p. 6). The definition of explicit vs. implicit attitudes is analogue. Controlled processes are those that "involve *conscious* effort, awareness, and intention and can be stopped at will" (my emphasis, p. 7), while automatic processes are those for which this is not the case (p. 7). The authors name vivid imagining (as part of the process of self-inducing false memories), prejudice and engagement in socially desirable behavior (while at the same time holding socially undesirable goals) as examples for the three kinds of dissociations respectively (pp. 6-7). They demonstrate the workings of these dissociations in evoking self-deception, in von Hippel & Trivers (2011), mainly on cases of self-enhancement. In the discussion of the levels at which the subjects can be self-deceived, they differentiate three kinds of self-deception: the one where the "truth" is prevented from being encoded, self-enhancement and classic self-deception (see table 34). Here though, only the conscious-unconscious level is considered (p. 13).

[302] Trivers (2011) states that "the key to defining self-deception is that true information is preferentially excluded from consciousness and, if held at all, is held in varying degrees of unconsciousness" (p. 9).

	prevention from truth-encoding	self-enhancement	classic self-deception (specific deception)
conscious	yes	yes	Yes
unconscious	yes	yes (if no, individuals act defensive and narcissistic)	No

Table 34. Von Hippel & Trivers: Levels of consciousness to be deceived at. Distinctions from von Hippel & Trivers (2011b, p. 13).

Self-enhancement is considered an instance of self-deception and it is assumed that cases in which truthful information is prevented from being encoded can also be cases of self-deception. The difference between the *self-enhancement* and *classic self-deception* is explained as the difference between self-deception with a general goal (self-enhancement) and with a specific goal (convincing somebody of a specific fiction) (von Hippel & Trivers, 2011b, p. 8).

To sum up, though in von Hippel & Trivers (2011) the authors concede an important role to dissociations in memory, attitudes and processes, they

a. name only two levels, namely the conscious and unconscious ones, at which subjects can be self-deceived in the discussion of levels at which the subject can be self-deceived and
b. hold that only in what they call classic self-deception the truthful information is stored at the unconscious level.

Further, though Trivers (1991) writes that “the hallmark of self-deception is a biased system of information transfer from conscious to unconscious and back” (p. 180), dissociations between the information available at the conscious and unconscious levels are argued to have a role only in the case of classic self-deception and, to a limited amount, in self-enhancement. If the unwanted information has never been attended to, then there is no need for a dissociation. This brings us to another aspect – the notion of *potential awareness* which is brought about to explain personal level inconsistency in self-deception:

> The flexible nature of this information-gathering bias also reveals that people have some awareness that upcoming information may be inconsistent with what they have already discovered. Thus, biases of this sort are consistent with classic definitions of self-deception that emphasize simultaneous knowing and not-knowing, in the sense that the individual consciously knows the welcome information that has been gathered but also has some awareness that unwelcome information could be around the next corner (we refer to this as *potential awareness*). In this case, however, true *knowing* of unwelcome information is precluded because the individual ends the information search before anything unwelcome is ever encountered. (von Hippel & Trivers, 2011b, p. 2)

Von Hippel & Trivers (2011) concede the possibility of tension in self-deception and seem to adopt Mele’s strategy of solving the tension-requirement which is understood as a cognitive inconsistency and not the feeling of uneasiness. I think that what Mele called an *uncomfortable suspicion* about the unwelcome information could be compared with von Hippel & Trivers (2011) notion of *potential awareness*, but that both are susceptible to criticism that I am about to present. Mele (2001) holds that satisfying his jointly sufficient conditions for entering self-deception, “may *often* involve considerable psychic tension,” yet that this tension is not conceptually necessary (p. 52). Against the view that only believing *p* and believing not-p can explain the tension, Mele holds that believing *p* and believing that there is a significant chance that not-*p* is enough (2001, pp. 67-73, 2010, p. 749). By arguing that conscious belief into the significant chance is enough, instead of conscious and unconscious beliefs (Mele, 2010, p. 749), one can avoid that the conscious-unconscious distinction bears the most explanatory weight in the analysis of self-deception. Von Hippel & Trivers (2011b) define potential awareness as “awareness that potential information could be around the corner” (p. 2). It is important for explanations of potential awareness to avoid the circularity that for somebody to be potentially aware of something he has to encode the given information in some way. Do von Hippel & Trivers want to say that the individual *has* the awareness that contrary evidence might be around the corner or that the individual *could have (had)* the awareness that contrary evidence might be around the corner? The second account would be a dispositional kind of account. Absence of representation is not to be equated with the representation of absence (Brentano, 1971[1911]; Dennett, 1991, p. 359).

Let me look in more detail into the possibilities of having different kinds of awareness with respect to contradictory evidence. The problem with the intentionalist accounts was that they seemed to postulate the awareness of contradictory evidence or the awareness that the evidence *is* around the corner. Mele weakens this requirement to the awareness that contradictory evidence *might be* around the corner (this is my interpretation of suspicion). In combination, we have four possible types of awareness (see table 35).

	is$_{\textbf{evidence}}$	**might be$_{\textbf{evidence}}$**
is$_{\textbf{awareness}}$	There **is** awareness that evidence **is** around the corner.	There **is** awareness that evidence **might be** around the corner.
might be$_{\textbf{awareness}}$	There **might be** awareness that evidence **is** around the corner.	There **might be** awareness that evidence **might be** around the corner.

Table 35. Possible types of awareness of contradictory evidence in SD.

Mele's awareness is of is$_{\text{awareness}}$ - might be$_{\text{evidence}}$ type. Greenwald (1988) also seems to favor this kind of awareness, as he explains the example of a doctor not recognizing a terminal illness by arguing that the self-deceiver "indeed did have knowledge of possibility, which he avoided converting into certainty" (p. 122). What is von Hippel & Trivers' type of awareness? I argue that if it is of might be$_{\text{awareness}}$ type, then it cannot solve the tension requirement, because it means that for the self-deceiver in the question there is no awareness, but from the omniscient perspective of the observer the self-deceiver is in the situation in which it is likely that he will obtain awareness in the near future. Is$_{\text{awareness}}$ is susceptible to another kind of criticism: Why does the self-deceiver end the hypothesis-testing despite having the awareness that the evidence is or might be around the corner? It is to be noted that tension understood as the phenomenological requirement does not necessary have this problem due to the different interpretations and misinterpretations that are possible of affective content.

So far, I have elaborated Trivers' general argument that the function of self-deception is to deceive others, as well as the dissociations on the genetic and psychological level that enable self-deception. I argued that both concern the unconscious-conscious distinction and that the notion of potential awareness presented by von Hippel & Trivers cannot explain tension, neither phenomenological nor cognitive. In the following I will evaluate the two paradigms by which self-deception is argued by the aforementioned authors to be tested – self-affirmation and cognitive load.

According to von Hippel & Trivers (2011a,b), self-deception can occur on different stages of information processing: at the stage of information search, at the stage of semantic interpretation (differentiation between attitude-consistent and attitude-inconsistent information; von Hippel & Trivers, 2011b, p. 9), remembering of information (misremembering of attitude-inconsistent information; von Hippel & Trivers, 2011b, p. 9), rationalization (reconstructing a more socially acceptable motive behind a certain behavior;

von Hippel & Trivers, 2011b, p. 10) or convincing oneself of a certain lie. This segmentation of the belief-forming process is similar to that of Balcetis (2008; see section 2.2.2.1).

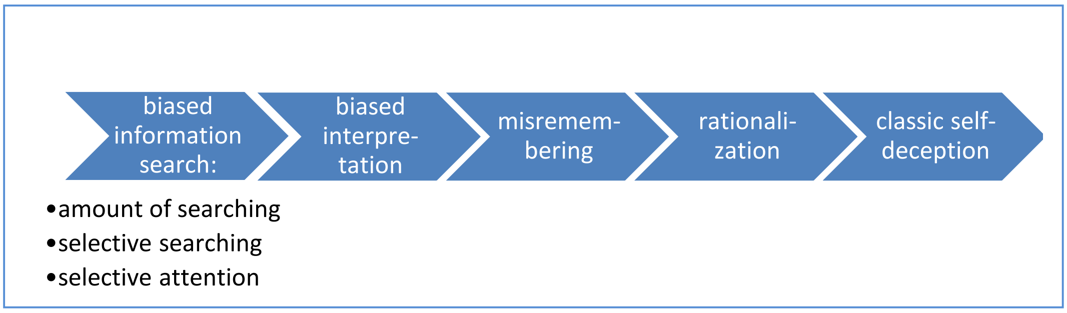


Figure 25. Von Hippel & Trivers: Information processing biases.
Distinctions from von Hippel & Trivers (2011b).

Von Hippel & Trivers (2011a, b) hold that the self-deceptive nature of biases can be uncovered either by means of self-affirmation manipulations, or by means of cognitive load manipulations (von Hippel & Trivers, 2011b, pp. 7-8). These are, according to the authors, the two paradigms for testing self-deception in virtue of their possibility to test potential awareness (see table 36). I have already criticized the notion of 'potential awareness' above for being no personal level awareness at all. I think that what the paradigms show is just their (motivated) influence on the behavior and avowals of participants.

Self-affirmation	Cognitive load
If one is being responsive to *self-affirmation manipulations*, then the given bias is motivated and could have been avoided (the requirement of potential awareness is fulfilled)[303] (von Hippel & Trivers, 2011b, pp. 7-8).	If one is being responsive to *cognitive load manipulations* – attenuated under the condition of cognitive load -, then the bias indicates that potential awareness was given, because processing under cognitive load is more effortful (von Hippel & Trivers, 2011b, p. 8).

Table 36. Von Hippel & Trivers: paradigms of testing self-deception
Distinctions from von Hippel & Trivers (2011b).

I want to mention two potential points of conceptual confusion about the two paradigms: the first concerning the effortfulness of bias that is presupposed in the cognitive load paradigm and the second concerning the ambiguity of self-enhancing attributes in the self-affirmation paradigm. The first is that notion of 'bias' can be understood in, at least, a two-fold manner – is biased processing more effortful or more automatic? In the first case it would mean that individuals self-deceive when not under cognitive load, in the second – that they do so under cognitive load. The cognitive load procedure implies that the first case is correct, or, in other words, that self-deceptive information processing is more effortful.[304] This question is important insofar as it would mean a possible restriction of the

[303] The motivated nature of a bias is determined by the possibility to eliminate or attenuate it in the absence of such motivation (or its compensation): "If the bias represents a self-deceptive process that favors welcome over unwelcome information, then this bias should be eliminated or attenuated by self-affirmation […] Such an effect for self-affirmation not only provides evidence for the role of motivation in the information-processing bias but also indicates that the person had the potential to process the information in a less biased or unbiased fashion." (von Hippel & Trivers, 2011b, pp. 7-8)

[304] Cognitive load as indicator for the effortfulness of a certain process: "Although manipulations of cognitive load do not address the *motivational issues* that underlie self-deception, they do address the issues of *cost and potential awareness*. If cognitive load leads to the elimination or attenuation of a particular bias, then the evidence suggests that the biased processing was

studies that can be considered as evidence in favor of the given evolutionary theory of self-deception.

I will now consider the notion of bias in two studies quoted by von Hippel & Trivers in order to show that these studies deliver contradictory results with respect to the question whether it is the bias which is eliminated under cognitive load or whether cognitive load leads to biased processing. One study cited by von Hippel & Trivers (2011b, p. 9) as evidence confirming the fact that selective skepticism is weakened by cognitive load states in the abstract preceding the study:

> Study 1 found that inferences drawn from favorable interpersonal feedback revealed a correspondence bias, whereas inferences drawn from unfavorable feedback were sensitive to situational constraint. Study 2 showed this *sensitivity to the quality of unfavorable feedback* to disappear under cognitive load. (Ditto et al., 1998, p. 53; my emphasis)

As one can see, the referenced study holds that positive information processing invokes a correspondence bias, whereas negative information processing is sensitive to the source of the feedback and, thus, more reality oriented. This sensitivity is according to Ditto et al. a *semi-automatic* process (1998, p. 61) and vanishes under cognitive load. Thus, in the given case cognitive load does not *eliminate* the bias, rather it *leads* to it, namely to the unconditional acceptance of information independent of its quality. Moreover, Ditto et al. (1998) argue that self-deception is not present in the majority of cases of motivated reasoning if Quantity of Processing View (QOP) is accepted (see section 1.3.3 for the comparison between the quantity and the quality view on motivated processing). QOP implies that preference-consistent information leads to *less effortful* cognitive processing than preference-inconsistent information (p. 54). Thus, according to QOP, it is the processing *extent* that changes and not the processing goal, which in the case of motivated processing is often assumed to be "intentional distortion of reality" (p. 53):

> Second, positing *processing extent* rather than *processing goal* as the key difference underlying treatment of preference-consistent and preference-inconsistent information removes much of the self-deceptiveness from motivated reasoning (Ditto & Lopez, 1992). Because previous explanations of motivated bias view individuals as intentionally pursuing the goal of reaching a desired conclusion, some level of self-deception is necessary in that the illicit nature of the goal driving the process must go unrecognized by the individual (Erdelyi, 1974; Kruglanski, 1996). The QOP view, on the other hand, sees motivated bias as arising from a more passive, less intentional process in which preference-inconsistent information provokes a critical analysis of its validity, whereas the validity of preference-consistent information is accepted, unthinkingly, at face value. *Because intention is removed from the process, no self-deception is implied*. (Ditto, 1998, p. 65; my emphasis).

Thus, while von Hippel & Trivers interpret this study as evidence for self-deception, the authors of the study do not share this opinion. Moreover, it is questionable whether positive information processing or negative information processing which involves the sensitivity to the quality of negative feedback is the biased one.

Von Hippel & Trivers (2011b) refer to another study (p. 10) as evidence that cognitive load leads to the disappearance of self-deception (in which case self-deception is understood as rationalization). Von Hippel & Trivers state that hypocrisy is a bias which stems from

actually more effortful than unbiased processing." (von Hippel & Trivers, 2011b, p. 8; my emphasis) rationalization and is eliminated by cognitive load. The study conducted by Valdesolo & DeSteno (2008) aims to answer the question whether or not hypocrisy derives from automatic or volitional biases. Valdesolo & DeSteno conclude that hypocrisy is attenuated by cognitive load and, thus, is an instance of effortful processing:

> Hypocrisy readily emerged under normal processing conditions, but disappeared under conditions of cognitive constraint. Inhibiting control prevented a tamping down or override of the intuitive aversive response to the transgression. Of import, these findings rule out the possibility that hypocrisy derives from differences in automatic affective reactions towards one's own and others' transgressions. Rather, when contemplating one's own transgression, motives of *rationalization and justification* temper the initial negative response and lead to more lenient judgments (Valdesolo & DeSteno, 2008, p. 1337; my emphasis).

Comparing the use of the cognitive load in these two studies, I think that the rationale of the cognitive load paradigm is that it eliminates processing that is more effortful than its counterpart with which it is compared. But does bias require more or less effortful cognitive processing? If bias is understood as a deviation from standards of rationality, then in the first study cognitive load has led to a confirmation bias (sensitivity to the quality of feedback has been eliminated), while in the second study rationalization bias has been eliminated under cognitive load. While the first study suggests that biased processing is not that much effortful, the second one suggests that it is.

The difficulty for von Hippel & Trivers' account is not only the equivocal use of the term 'bias.' But it is also the fact, that, if self-deceptive bias is attenuated by cognitive load, then truthful information processing happens first, which would mean that storage of motivation inconsistent information and the role of dissociations should more strongly come into the explanation which, as mentioned previously, is not the case, as von Hippel & Trivers hold that in the case of self-enhancement there is no dissociation. Consequently, applicability of the cognitive load paradigm to self-deception needs further clarification.

Moreover, apart from biases that are uncovered by self-affirmation and cognitive load paradigms, there are other proximal mechanisms that are mentioned in Trivers' book: denial, projection and cognitive dissonance (2011, pp. 147-156). To my understanding, Trivers indentifies the placebo effect as an instance of self-deception exactly because it is consistent with the cognitive dissonance theory.[305]

Let me come to the second point concerning self-fulfilling prophecies. If self-enhancement is a kind of self-deception and self-deception serves the function to deceive others, then self-enhancement has to be *unrecognized* (van Leeuwen 2007, p. 335). This argument can be extended if an additional premise is accepted that self-enhancement occurs in cases where there is *ambiguity* in attribute ascription. A dilemma of self-enhancement for its service in the deceit of others follows:

1. If self-deception concerns a certain domain that allows vagueness in attribute ascription (see section 1.2.7, 1.3 and 3.1.3), then instances of a self-fulfilling prophecy have to be differentiated from instances of self-deception or an argument should be given why a self-fulfilling prophecy can be self-deceptive. It further makes a difference if the circumstances are vague from the perspective of the self-deceiver or the observer. Deciding a priori how to distinction self-deception and self-fulfilling prophecy does not seem possible. Generally, this means that the characterization 'self-deception' can be changed in retrospect to 'self-

[305] Placebo as rationalization: "The general rules of placebo effect are consistent with the cognitive dissonance theory […] – the more a person commits to a position, the more he or she needs to rationalize the commitment, and greater rationalization apparently produces greater positive effects" (Trivers, 2011, p. 72).

fulfilling prophecy.' As such, different temporal windows need to be considered for different kinds of self-deception in order to decide whether it really is self-deception.

2. If an instance of self-enhancement is not a self-fulfilling prophecy, then, the fact that the magnitude of self-enhancement is often measured by comparing the evaluation of the individuals with those of knowledgeable observers,[306] leads to the conclusion that all instances of self-enhancement, operationalized in this way, are recognized by others and, thus, are not conducive to the deception of others.

Taylor (1989), whom von Hippel & Trivers (2011) cite, states that there are three kinds of positive illusions – self-enhancement, unrealistic optimism and exaggerated belief in one's personal control (Taylor, 1989, p. 6) and that they are self-fulfilling (p. 244). If von Hippel & Trivers share this premise, a question would arise whether it is self-deception that this paradigm proves or a self-fulfilling prophecy. Can self-deception be a self-fulfilling prophecy? In each case, whether it is and is not a self-fulfilling prophecy, it should per definition go undiscovered:

> But self-deception on certain topics tends not to be discovered. For example, social propriety rules out extensive negative feedback from others concerning one's character traits and idiosyncrasies of personal behavior, so that one tends not to discover one's own self-deception on these matters, though of course *one discovers that of others*. (Schmitt, 1988, p. 199; my emphasis)

Last but not least, I want to mention two studies that point to *social weaknesses* that self-deceivers might possess. The idea that self-deceivers accomplish to deceive others implies that they are superior in certain kinds of social interaction, yet results of Lynch & Trivers' (2012) study show that self-deceivers have difficulty appreciating humor and Lamba & Nityananda (2014) argue that self-deceivers are poor at detecting deception of others. Recently, Lynch & Trivers (2012) published an article that develops an argumentative line with respect to the conclusion that self-deception inhibits laughter. The reasoning is as follows: humor appreciation either involves the identification of anomaly or the ability to process a potential threat as harmless (p. 492). Self-deception hinders the processing of anomaly (p. 491), as well as ability to recall threatening words (p. 492). The instrument that has been used in the studies above has been the denial scale of the BIDR. Thus, Lynch & Trivers used BIDR's SDQ to measure the correlation between the comedy appreciation and the results on the given paper-and-pencil test. The results indicated that self-deception hinders the appreciation of humor. See section 1.3.1 for the discussion of the BIDR questionnaire. Apart from Lynch & Triver's (2012) study there is a study by Lamba & Nityananda (2014) who also argue that self-deceivers (defined as overconfident or underconfident individuals) are overrated/underrated by observers. The setup of this study was that students in small groups weekly interacted with each other and had to subsequently evaluate each other. Interestingly, the authors argue not only that self-deceivers can deceive others, but also that they themselves are *poor at detecting when others deceive them*: overconfident individuals overestimated this quality in others and vice versa for underconfident individuals. This is the case when participants had to grade others, but not rank them (Lamba & Nityananda, 2014, p. 3). An open question is how particularly the last fact is reconcilable with the benefits of the function of self-deception to deceive others. If self-deception makes self-deceivers vulnerable to the deception of others, then its benefit in the arms race where everybody tries to deceive everybody is limited.

[306] Kinds of definitions of unrealistic positive evaluations (Paulhus, 1998, p. 1197):
- *self-aggrandizement*: degree to which an agent rates herself to positively differ from others;
- *self-enhancement*: degree to which one's evaluation is different from a "credible criterion."

In this section, I argued that dissociations presented by von Hippel & Trivers (2011a,b) concern the unconscious-conscious distinction and that the notion of potential awareness cannot explain tension, neither phenomenological nor cognitive. I have further drawn the reader's attention to two conceptual unclarities: that of effortfulness of bias in the cognitive load paradigm and that of the self-fulfilling prophecy in the self-affirmation paradigm.

3.2.3 Criticism of Trivers' evolutionary theory of self-deception

Trivers' theory has been subject to diverse criticism. In the given subchapter I will first present van Leeuwen's (2007b) criticism of Trivers' theory and then sketch van Leeuwen's spandrel hypothesis, as well as Lopez & Fuxjager's (2012) exaptation hypothesis as alternative evolutionary theories of self-deception.

Van Leeuwen criticizes Trivers in two articles (2007b, 2013b). In his recent review of Trivers popular book, Van Leeuwen (2013b) argues that Trivers does not present an argument for his theory, because presenting an argument would amount to developing specific hypotheses and empirically testing them. I do not see in how far the development of an argument is to be equated with developing specific empirical hypotheses that follow *from* an argument. It is true, though, that Trivers' theory would merit from more empirical data. In another article, Van Leeuwen (2007b) holds that Trivers advocates for two independent evolutionary adaptive functions of self-deception – deception of others and positive orientation towards future (p. 334) which in the light of von Hippel & Trivers (2011a,b) article on self-deception and Trivers' (2011) popular book on self-deception is clearly not the case, as Trivers argues that there is only one evolutionary function of self-deception – facilitation of the deception of others and other possible functions (possibly intrapersonal) are only intermediary and serve the given interpersonal goal. Thus, I will present van Leeuwen's criticism on self-deception having the function of deceiving others.

First, van Leeuwen (2007b) argues on the assumption that reliable information about the environment is evolutionary beneficial (p. 334) that it is doubtful that self-deception would give a greater benefit than truthful information encoding or lying (p. 337), if individuals with these different phenotypes were to compete. As we have seen in the section about the distal level of evolutionary explanation, there have been successful attempts to model self-deception using the hawk-dove evolutionary game that point to another conclusion. Thus, van Leeuwen's first argument is in fact an argument that can be tested with empirical means, but not with conceptual ones.

Second, he argues that cases in which humans engage in self-deception do not map on cases in which they deceive each other (p. 335), for example if a quarterback deceives himself into thinking that his coach will not be angry if he plays poorly, then he does not try to deceive his coach, but only himself (p. 335). The second instance of criticism is related to the third, namely that self-deceivers are often the only ones who believe the content of their self-deception (p. 335). These two arguments have 1. to differentiate between ancestral and current environment because Van Leeuwen's counter-examples are about the current environment, but the ancestral one is more important when generally speaking about the evolution of self-deception (see section 3.2.1) and 2. to seek empirical support that in the majority of the cases self-deception and other-deception are unrelated and that in the majority of the cases self-deception is recognized by the observers. Merely pointing out that there are cases in which this is not the case is not enough to undermine an evolutionary argument.

Thus, the conclusion, that can be drawn, is that the evidence to support Trivers' theory is inconclusive. Yet, the conclusion that it cannot be the case, that Trivers' theory applies to the target phenomenon, does not stand.

3.2.3.1 Van Leeuwen's spandrel hypothesis

Self-deception comes from rationality in the context of a finite desiring mind.
(Van Leeuwen, 2008, p. 207)

To repeat the results so far, in section 3.2 I was concerned with the presentation of an evolutionary theory about the development of self-deception, namely Trivers' hypothesis that self-deception evolved to deceive others. I have presented his definition of self-deception, the distal (evolutionary games as support for the thesis) and proximal (genetic and psychological means of self-deception) explanatory levels. I came to the conclusion that at the given point in time the hypothesis cannot be either refuted nor unequivocally accepted, but needs further elaboration. In this and the following section I will present alternative evolutionary perspectives on self-deception, namely that it is a byproduct (of another phenomenon possessing an evolutionary function) and that it is an exaptation (that it *acquired* an evolutionary function that it did not evolve for). Trivers' evolutionary theory implies that self-deception *has* an evolutionary function and cannot be a spandrel of rational capacities as Van Leeuwen (2007b) argues. Trivers points against Van Leeuwen the widespread nature of self-deception (personal communication).[307]

So far the rationale has been the following: if self-deception is normal and adaptive, there has to be a reason for this and an evolutionary account would offer such a reason. Neil Van Leeuwen raises this question to a new level - whether self-deception is a byproduct of mental traits.[308] In this way he also omits the rationality-problem:[309] mental traits, that self-deception is a byproduct of, are themselves rational. So, one can keep the view of a human being as rational,[310] but is also able to integrate self-deception into this view (Van Leeuwen, 2008). Consequently, Van Leeuwen's theses are that self-deception is a product of abilities that enable rationality "given finite minds" (Van Leeuwen, 2008, p. 192) and that self-deception is not a product of natural selection, but a spandrel[311] (p. 192). He gives the following definition to self-deception:

[307] Compare also Trivers' argumentation about the genetic components of reciprocal altruism: "There is no direct evidence regarding the degree of reciprocal altruism practiced during human evolution nor its genetic basis today, but given the *universal and nearly daily practice* of reciprocal altruism among humans today, it is reasonable to assume that it has been an important factor in recent human evolution and that the underlying emotional dispositions affecting altruistic behavior have important genetic components" (Trivers 1971, p. 48; my emphasis).

[308] To my best knowledge the idea that self-deception is a by-product of some rational capacities of mind is not new in the self-deception literature. As A. Rorty (1986, p. 119) puts it: "Self-deception and *akrasia* are by-products of psychological processes that make ordinary rational action possible."

[309] Recapitulating, the rationality-problem is the difficulty to reconcile the assumption that individuals are rational to a certain degree with the assumption that they self-deceive often.

[310] Van Leeuwen (2008) offers a working definition of rationality as conducive to truth and coherence in a belief set and in practical planning (p. 193).

[311] Van Leeuwen (2007b) acknowledges that his spandrel view is compatible with what he calls weak adaptationism: "I should note before closing that the spandrel view of self-deception is not logically inconsistent with a *weaker* form of adaptationalism than the modularist account that Trivers seems to hold. This form of adaptationism about self-deception would hold that

> *An agent is in a state of self-deception if and only if*
> *(i) she holds a belief,*
> *(ii) that belief is contrary to what her epistemic norms in conjunction with what evidence she has would usually dictate, and*
> *(iii) a desire, with content appropriately related to the belief formed, causally makes the difference to what belief is held in an epistemically illegitimate fashion.* (Van Leeuwen, 2008, 195)

The closest to Van Leeuwen's spandrel view is Mele's deflationary view of self-deception (Van Leeuwen, 2008, p. 195, footnote 9). Mele (2012) argued that the acquisition of a certain belief and, thus, the occurrence of self-deception, depends on the confidence threshold. The confidence threshold for belief acquisition depends on the aversion to specific costly errors and the information cost where the information cost is the resource and effort required to acquire the information (p. 8). This is Mele's conceptualization of what he calls the Friedrich-Trope-Liberman (FTL) model which combines two models of hypothesis testing – that of Friedrich and that of Trope & Liberman. Van Leeuwen cites the byproduct/spandrel thesis and the absence of the FTL model as the most important differences between him and Mele. Van Leeuwen rejects the FTL model because, according to him, it can't explain self-deceptive beliefs that have a high subjective cost (Van Leeuwen, 2008, p. 198, footnote 17). Accepting the FTL model would mean, according to him, that cases in which one acquires a belief that possesses a high subjective cost are precluded. What is implicit in this remark of Van Leeuwen against Mele is that according to the former, error minimization is not doing justice to the *freedom* of the self-deceiver to self-deceive about *every* kind of content. The solution that Van Leeuwen proposes is a personal level account of how practically rational human beings might become self-deceived. I have already argued against personal level explanations of self-deception (see chapter 2), which will not be able to distinguish it from other irrational phenomena (see section 1.2.6). In the following I will nevertheless go into the details of his account as an alternative to Trivers' account for completeness reasons, as well as for the central role that he ascribes to goal representations, selectivity in the process of self-deception, as well as to its phenomenology that I also hold to be important (see section 2.2). His kind of description is a personal level one and one possible inference from it might be that the conscious belief-forming process of the self-deceiver really occurs in this manner. I disagree in this respect (2.2.2.3) and argue for the application of goal representations and selectivity on the subpersonal level such that the phenomenology of the self-deceiver urges him to *search* for the explanation of his uneasiness and anxiety on the personal level and to given inconsistent justifications for his behavior.

Van Leeuwen (2008) offers the following types of self-deception:

- willful self-deception
- wishful self-deception
- dreadful self-deception

What is common to all types of self-deception is the presence of motivation (deceptive element) and the product which is the belief that *p* in the face of contradictory evidence that points to not-*p*. The difference between the types of self-deception is the one of motivation. Wishful self-deception is motivated by the desire that *p* is the case, willful – by the desire to believe that *p* and dreadful – by the desire that not-*p* is the case. According to Van Leeuwen, intentionalist and deflationary views can be combined: wishful self-

the capacity for self-deception is indeed a structural byproduct, but one whose positive fitness value prevented it from being selected against (and hence extinguished)." (p. 342) deception and dreadful self-deception can be compared to Mele's straight and twisted self-deception (Van Leeuwen 2008, p. 196, footnote 13), while willful self-deception is intentional.[312]

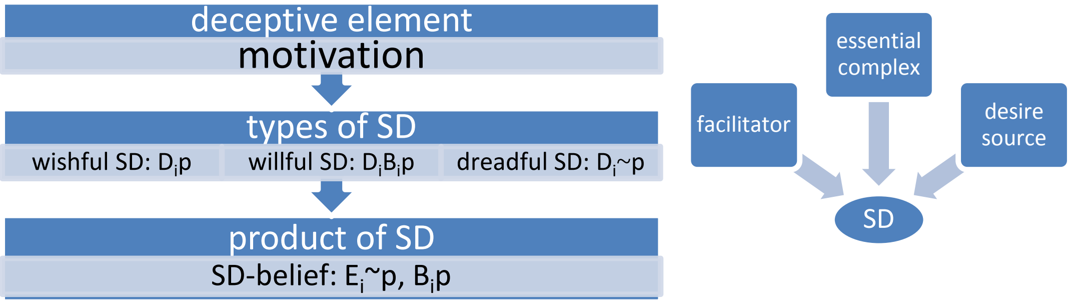


Figure 26. Van Leeuwen: categorization of self-deception. Distinctions from Van Leeuwen (2008).

i = the agent, b = belief, D = desires, E = evidence in favor of a proposition (Van Leeuwen, 2008, p. 196).

Deceptive element is the state or attitude that prevents normal belief formation process, whereas the **product of self-deception** is the mental state that self-deception has brought about (Van Leeuwen 2008, p. 194).

The **figure on the left side** is to be interpreted as follows: The content of the deceptive element (motivation) can be either the desire that p is the case, the desire to believe p or the desire that not-p is the case. The product of self-deception is the belief that p in the face of contradictory evidence that not p.
The **figure on the right side** is to be interpreted as follows: The combination of features of the mind that belong to the essential complex, facilitating factors and certain desire sources causes self-deception.

As has been stated above, van Leeuwen argues that self-deception arises as a byproduct of rational functions of certain features of the human mind. He describes these features as essential to emphasize their importance to rationality where rational comportment is understood as the one "conducive to goal attainment given finite limits" (van Leeuwen 2008, p. 198). These essential features in the presence of facilitators and desire sources explain the occurrence of self-deception. Facilitators and desire sources are defined by van Leeuwen as follows:

> A *facilitator* is a mental feature that makes easier the failure of epistemic norms and evidence in self-deception. A *desire source* gives rise to desires that are apt to cause self-deception. (Van Leeuwen, 2008, p. 200)

He does not make clear which features he thinks are necessary for self-deception and which are not.[313] Those essential features seem to be jointly sufficient to cause self-deception, but only under certain circumstances. Otherwise, facilitators (which are themselves by no means sufficient to self-deception) are needed as one can see from the following quotation regarding the inertia-facilitator:

> The essential complex may not always be sufficient by itself to cause a cuckold to hang on to the belief that his wife is faithful against the weight of the evidence. But given inertia

[312] In his other article Van Leeuwen (2007a, p. 424) differentiates also goal-driven self-deception as a fourth type. In the newer article (Van Leeuwen, 2008), he incorporates the idea underlying the distinction in his first desire source (see below).

[313] Mele (2001), whose account is close to that of van Leeuwen, also does not speak about necessary conditions of self-deception, only about jointly sufficient conditions.

> of his already existing web of beliefs, self-deception becomes possible. (Van Leeuwen, 2008, p. 201)

In the following I present the essential features, facilitators and desire sources, through which self-deception arises. Emphasis will be first, on how these factors lead to self-deception and, second, what role they play in normal, practically rational reasoning according to Van Leeuwen (2008). He names seven essential features that play a role in self-deception (pp. 198-200):

1. the phenomenological component of desires (sting) that occurs in cases of anticipation of their satisfaction or evidence of non-satisfaction
2. selective attendance to evidence
3. inclination to avoid discomfort
4. structural organization of evidence in the mind
5. consideration of evidence and conformity to epistemic norms in belief formation
6. pleasure that is gained from desire satisfaction
7. search of pleasure

These features are essential to practical rationality, because they lead to goal attainment (1,3,6,7) and compliance to one's own epistemic norms (5) given that the human mind is restricted with respect to the amount of information that it can process and store (2, 4). Van Leeuwen (2008) calls the latter restriction "finite limits" (p. 198) and, thus, I will call the last function which is to respect the given finite limits, finitude management.

I understand it as that Van Leeuwen (2008) suggests the following connection between *essential features*: in order to *attain goals* humans experience pleasure at evidence that their desires have been satisfied (EC6) and/or negative "characteristic sting" at evidence that they have been not satisfied (EC1). These phenomenological features in conjunction with the inclination to seek pleasure (EC7) and/or avoid discomfort[314] (EC3) lead to selective attendance of evidence (EC2). Attending to evidence is itself necessary because, as *finite creatures*, humans have to form beliefs on the basis of certain pieces of evidence and in accordance with certain epistemic norms (EC5) and these acquired beliefs subsequently have to be structurally organized in order to facilitate evidence-retrieval (EC4). The two key points are finitude of humans and goal-attainment which together lead to *selective evidence* that can become self-deceptive in certain circumstances and under a certain motivation that is a desire other than the desire for truth (Van Leeuwen, 2008, p. 202).

[314] Avoiding discomfort has the function of avoiding harm (Van Leeuwen, 2008, p. 199). Though Van Leeuwen does not explicitly state why avoidance of harm is conducive to goal-attainment I think his implicit assumption is that it is human vulnerability that makes harm-avoidance necessary.

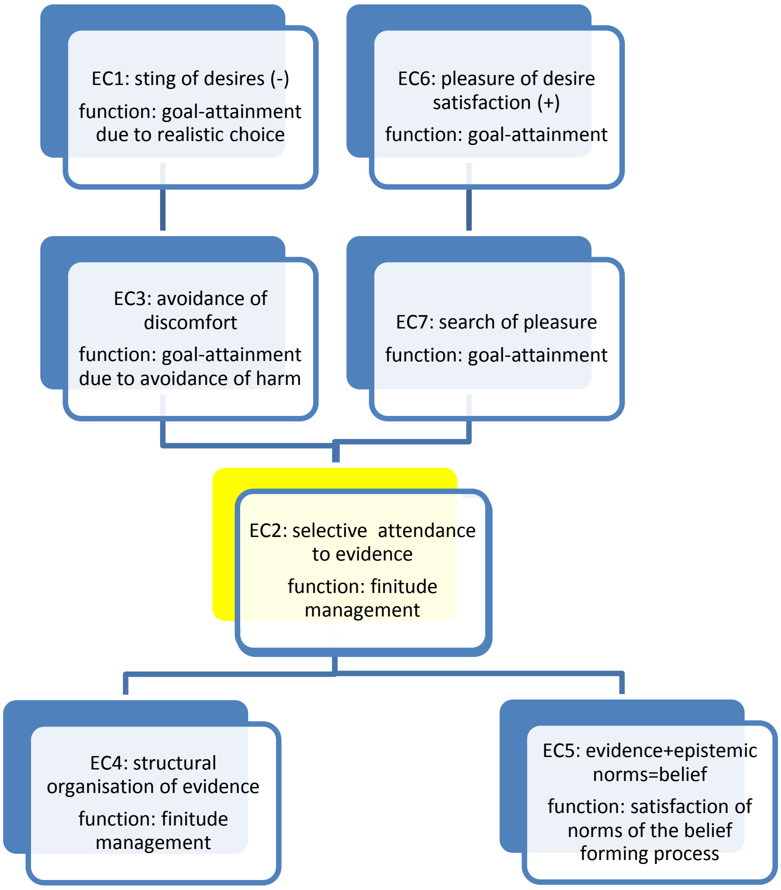

EC1: sting of desires (-)
function: goal-attainment due to realistic choice
EC6: pleasure of desire satisfaction (+)
function: goal-attainment
EC3: avoidance of discomfort
function: goal-attainment due to avoidance of harm
EC7: search of pleasure
function: goal-attainment
EC2: selective attendance to evidence
function: finitude management
EC4: structural organisation of evidence
function: finitude management
EC5: evidence+epistemic norms=belief
function: satisfaction of norms of the belief forming process

Figure 27. Van Leeuwen: Essential features and they function in practical rationality Distinctions from Van Leeuwen (2008).

EC – essential complex, a complex that consists of essential features (7 features) that aid in the occurrence of self-deception. Selective attendance to evidence is, according to me, the key features and, thus, has been marked yellow.

Thus, on the one hand, the acquisition of beliefs is based on evidence, happens in accordance with one's own epistemic norms (EC5) and these beliefs are subsequently structurally organized (EC4). These essential features of human reasoning, by themselves, are in accordance with the norms of theoretical reasoning that one could formulate. On the other hand, other phenomenological features lead the reasoning process in a certain direction, the direction of goal attainment. These are seeking pleasure (EC7), avoiding discomfort (EC3), as well as the subsequent experience of pleasure when goals have been attained (EC6) or discomfort[315] if this is not the case (EC1). Given that humans are limited in the extent of their information processing, selective attendance to evidence seems necessary. Consequently, I hold that the key in Van Leeuwen's elaboration of the way essential features enable self-deception is selective attendance to evidence.

Essential features of human reasoning sometimes are not enough to trigger self-deception according to Van Leeuwen. Facilitators and the presence of certain desire sources help accomplish this task. Van Leeuwen (2008) mentions the following *facilitators*[316] (the list

[315] At this point I interpret the "characteristic sting" that accompanies the failure to satisfy one's desire as phenomenological discomfort.

[316] Van Leeuwen's facilitators have some similarities with Mele's input-control strategies insofar as most of them have an input-control function.

of which according to him is not complete, p. 200): inertia of the web of beliefs[317] (F1), generation of thoughts by desires (F2), differences in degree of skepticism (F3) and suppression of memories (F4), which all themselves have rational functions. These rational functions consist, accordingly, of achieving coherence in a belief set, causing the appearance of certain representations in thought,[318] screening out falsehoods and minimizing distraction. Facilitators, on the contrary, contribute to self-deception due to the fact that they can evoke confirmation bias, lower the threshold for accepting a belief to a degree that is unjustified, reclassify the evidence in a certain manner that is unjustified[319] and willfully suppress[320] memory (Van Leeuwen 2008). Though the first two factors (confirmation bias and adjustment of threshold levels), that can lead to self-deception, resemble Mele's ones, the last two definitely do not fit into a deflationary account.

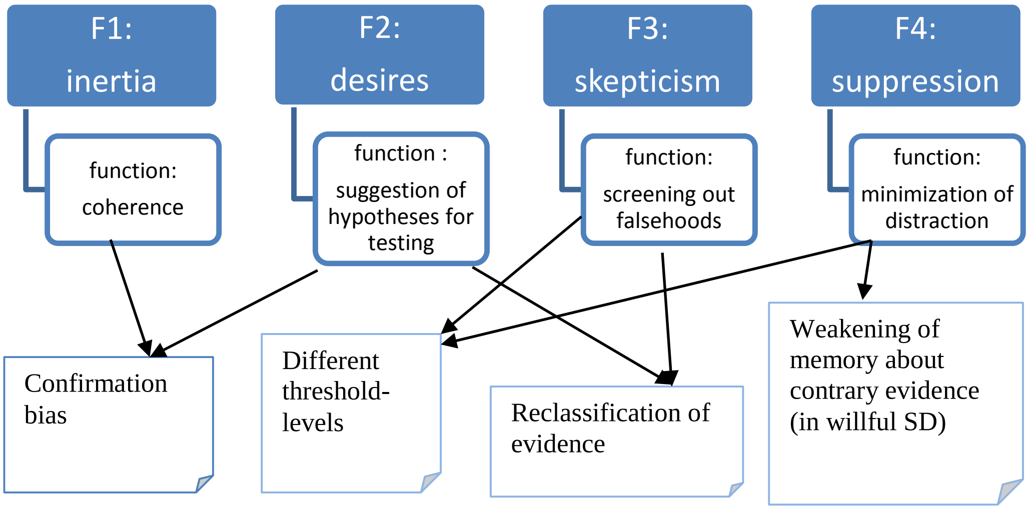

F1: inertia
function: coherence
F2: desires
function : suggestion of hypotheses for testing
F3: skepticism
function: screening out falsehoods
F4: suppression
function: minimization of distraction
Confirmation bias
Different threshold-levels
Reclassification of evidence
Weakening of memory about contrary evidence (in willful SD)

Figure 28. Van Leeuwen: Facilitators of self-deception.
Distinctions from Van Leeuwen (2008).

F – facilitator. The given four facilitators play, on the one hand, the listed function in practical reasoning, but they skew, on the other hand, the normal belief forming process due to the elicitation of the factors, at which the arrows point.

According to Van Leeuwen (2008), not only essential features and facilitators have a rational function, but also *desire sources* that contribute to self-deception, on the one hand, and practical reasoning, on the other: intentions to achieve a certain goal (DS1) and cognitive dissonance (DS2). Both desire sources have a rational function: aid to accomplish projects. Both can generate self-deception: long term intentions can generate self-deception that the goal is obtainable, whereas cognitive dissonance can generate the self-deception about the extent of one's positive qualities (see table 37).

[317] Van Leeuwen takes the account of inertia from Quine and Ramachandran (Van Leeuwen, 2008, p. 200, footnote 25)

[318] Appearance of representations in thought is beneficial to practical reasoning insofar as "we can´t form beliefs about contents that haven't occurred to us" (Van Leeuwen, 2008, p. 201).

[319] Van Leeuwen mentions (Van Leeuwen 2008, p. 201) that reclassification of evidence can be conscious. In another part of the article (Van Leeuwen 2008, p. 202), he does not specify if he has conscious reclassification of evidence in mind. That´s why I will suppose that for him every reclassification of evidence is conscious.

[320] Van Leeuwen (2008) speaks at this point about suppression and not repression (p. 202), but he does not discuss whether there is any difference between the two.

In order to attain goals, an *intention* – the first desire source - should include the following components: means-end reasoning, screening of options and tracking the goal, every one of which has its own subsidiary desires. The rationale is that to realize the given intention options have to be considered, means have to be established that help achieve the goal and the goal itself has to be kept in mind. This is accomplished due to the desire to get rid of obstacles, the desire to realize means to an end and the desire for information. The desire to realize means to an end is the one that, in certain cases, skews reasoning and contributes to the occurrence of self-deception (p. 203). When intentions are long-term (so that the result is not directly measurable) and the evidence is mixed, self-deception about the possibility to achieve means to an end arises:[321]

> But often, especially in the case of *long-term intentions*, subsidiary desires of intentions will be for things not immediately tangible but still felt to be important. If a parent intends to send a child to a good college, this intention will be accompanied by a subsidiary desire for the child to do well when he gets to high school – this is a *means to an end*. But suppose this subsidiary desire arises when the child is badly underperforming in eighth grade. In this sort of situation, the subsidiary desire concerning the child´s intelligence may be the sort that gives rise to self-deception, since *evidence* about a person's intelligence is often *mixed enough*, even if it's heavily weighted to one side, to enable the kind of *selective attending* involved in the essential complex. (Van Leeuwen, 2008, p. 203; my emphasis)

As the second desire source Van Leeuwen names *cognitive dissonance*. It is a phenomenon of cognitive psychology that has found its elaboration in Festinger's cognitive dissonance theory (3.1.1). The idea is that, in the cases where there is cognitive inconsistency, phenomenological feeling of discomfort or cognitive dissonance is aroused and it is a driving force that enforces consistency.[322] Van Leeuwen (2008) describes the phenomenon as follows: "Cognitive dissonance is the discomfort people feel when their behaviors don't conform to their conceptions of themselves as moral, competent, and consistent." (p. 203) According to him, cognitive dissonance has as a desire source one rational and one self-deceptive function: namely, due to discomfort it can motivate a self-change or a self-deception. I think that Van Leeuwen's options that cognitive dissonance offers bear a certain similarity to the two options in the evolutionary hawk-dove game – fighting or running (3.2.2.2). The interpretation of cognitive dissonance that Van Leeuwen gives is similar to how Scott-Kakures describes Aronson's interpretation of cognitive dissonance – namely as arising out of the pursuit of three goals: consistent, competent and morally good self (Scott-Kakures, 2009, p. 83). According to Scott-Kakures, who cites Kunda in this respect, this interpretation of the cognitive dissonance theory is too narrow to explain self-deception (Scott-Kakures, 2009, p. 83). The idea was that self-deception can be explained by cognitive dissonance in general, instead of explaining it by self-consistency as one of the further developments of cognitive dissonance theory. Self-affirmation and new look are the two other prominent further developments of cognitive dissonance (3.1.1).

[321] For a more detailed view on constraints of self-deception see also Balcetis (2008).

[322] The striving towards cognitive consistency was emphasized in social psychology from the 1940s to the 1970s and cognitive dissonance theory was not the only one that was based on it (Helzer & Dunning, 2012, p. 380).

DS1: INTENTION (Van Leeuwen, 2008, p. 202-203)		
Means-end reasoning	**Screening of options**	**Tracking the goal**
Desire for realization of means to an end	Desire to be rid of obstacles	Desire for information
➢ **SD by long-term intentions**		
DS2: COGNITIVE DISSONANCE (Van Leeuwen, 2008, p. 203-204)		
Rational function: accomplishment of practical projects due to motivation to change incompetent, inconsistent or immoral behavior		
➢ **SD by desire to get rid of cognitive dissonance**		

Table 37. Van Leeuwen: Desire sources.
Distinctions from Van Leeuwen (2008).

Van Leeuwen differentiates between three components of intentions: means to end reasoning, screening of options and tracking the goal each of which has a subsequent desire that is generated by it. Only the first desire can aid in bringing about self-deception.

Comparing Van Leeuwen's theory of self-deception to other accounts, his view is not only similar to Mele's deflationary theory, the fact he himself acknowledges, but also to Rorty's intentionalist divisionist approach. Rorty (1988) points out that self-deception is the by-product of functional structures and strategies that attempt to integrate different subsystems of the self (p. 21). Notice also the similarity between Van Leeuwen's theory and Andersen's explication of control theories description of Craver & Scheier:

> The process [control taking] involves (1) activation of relevant end states; (2) activation of beliefs about appropriate opportunities and means to successfully pursue these end states; (3) actions taken to remove whatever (negative) discrepancy exists between one's current standing and the desired state; and (4) monitoring of how close one is to attaining the end state (Craver & Scheier, 1999). (Andersen et al., 2007, p. 144)

Moreover, Van Leeuwen holds his account to encompass inflationary, as well as deflationary type of self-deception, because he differentiates between willful (intentional), wishful and dreadful self-deception. The latter two kinds are the kinds that Mele distinguishes, although he names them straight and twisted self-deception. Van Leeuwen points out that his focus in Van Leeuwen (2008) has been on wishful self-deception, but that it can be extended to the other two types:

> Extending this model to cases of willful and dreadful self-deception will involve examining how the kinds of desire constituitively involved in *those* types can trigger the sorts of selective attention and other processes here identified. (Van Leeuwen, 2008, p. 204, footnote 28)

Van Leeuwen's explanation of the mechanism of self-deception is, taken more generally, that certain kinds of desires and emotions (pleasure, dissatisfaction) evoke selective processing of the evidence. He puts weight on elaborating what kinds of desires these are and why they are conducive to practical rationality, but explaining precisely the connection between these desires and the sorts of selective processing mechanisms he mentions that should have been the focus of his analysis if he wanted to explain how self-deception is accomplished. At this point, the selectivity problem mentioned by Bermúdez stays unexplained, namely why motivation not in every case triggers self-deception, but only in certain cases.

Two more points concerning Van Leeuwen's theory are worth mentioning, the first one on his emphasis of goal directedness and practical rationality of humans, the second on his emphasis of the finitude of human beings. Van Leeuwen focuses on the importance of *personal* goals. How big of an influence *subpersonal* goals are is yet still subject of debates (2.2.1). Finitude of human beings is being employed by the author to explain the trade-off between the information needed by the reasoning process to produce accurate results and the limited information that the process actually gets in order to finish in an appropriate amount of time. Thus, according to such a rationale, if infinite processing power were given, humans would reason differently, but instead reason in a suboptimal manner. The termination point of searching for new information dependent on the goals of the individual. In postulating a goal/desire – dependent termination point for gathering new information during a belief-forming process Van Leeuwen's and Mele's view are similar. Recent computational accounts – predictive coding – argue that human information processing is Bayes-*optimal* though (see section 4.3).

Summarizing, van Leeuwen offers an alternative for Trivers' evolutionary theory of self-deception which consists in viewing self-deception as a spandrel of (practical) rational capacities of the human mind. The emphasis is on the finitude of the human mind that figures as a premise in the development of certain mental features that, on the one hand, have rational functions and, on the other hand, can lead to self-deception.

The criticism that I see Van Leeuwen's spandrel theory of self-deception to be susceptible to is exactly the one that Van Leeuwen points out with respect to Trivers' theory: a spandrel view, like the adaptationist one, has to offer a proper evolutionary argument in favor of the hypothesis that all those features of which the phenomenon is a spandrel have developed as a result of evolutionary processes. In the case of Van Leeuwen's theory this would mean giving an argument of why and how the mental features that he considers conducive to rationality have developed and enhanced inclusive fitness.

3.2.3.2 Lopez & Fuxjager's exaptation view

Lopez & Fuxjager (2011), on the one hand, defend Trivers' argument from Van Leeuwen's criticism and, on the other hand, propose that even if not evolutionary adaptive, self-deception could be an exaptation. They give a spandrel-adaptation argument for this:

> In this way, it might appear that self-deception is similar to an *exaptation*: a trait that evolved for one function that gets co-opted for a new one (Farmer, 1997). Here, we are considering the possibility that self-deception was originally a spandrel that later got co-opted to help individuals come to the adaptively beneficial positive self-perception. If this is right, self-deception is not an exaptation *per se*, but something fairly similar. (Lopez & Fuxjager, 2011, p. 322)

The authors come to this conclusion by the following reasoning: answering van Leeuwens' critique of von Hippel & Trivers, the authors state that Van Leeuwen's definition of self-deception is more general than that of von Hippel & Trivers and, thus, should be favored (Lopez & Fuxjager 2011, p. 316). They take Van Leeuwen's (2007b) conditions that have to be fulfilled for the subject to be self-deceived to be: 1. a belief against one's own epistemic norms in conjunction with the available evidence and 2. a desire for a certain state of affairs to occur that can play a causal role in skewing the belief forming process (p. 317). Lopez & Fuxjager (2011) argue that a two-step argument (see table 38), that is given in Trivers' (2000) work, can prove that self-deception has the ability to increase the individual's reproductive success (p. 316).

The two-step argument is as follows: 1. self-deception leads to *positive self-perception*; 2. positive self-perception leads to adaptive benefits. They argue in favor of the first claim that self-deception leads to positive self-perception on the basis of the fact that the majority of cases presented in the self-deception literature, or, in other words, the standard cases, involve, as a result, positive self-perception. This, as well as the results of Quattrone & Tversky's cold water experiment (1.3.2.2), is taken as empirical evidence for the claim that self-deception leads to positive self-perception. To remind you, Quattrone and Tversky's cold water experiment shows that depending on the desirability of the outcome, subjects hold their hand in cold water longer if they think that it will be diagnostic of good health, rather than bad health. Holton's (2001) argument that self-deception is necessarily about oneself is accepted by the authors without further consideration (Lopez & Fuxjager, 2011, p. 318-319).

The authors argue that positive self-perception leads to adaptive benefits on the basis of the winner effect which they define as "an increased ability to win fights and social conflicts following prior winner effect" (Lopez & Fuxjager, 2011, p. 319), because the winner effect helps acquire a higher position in the social hierarchy. The winner effect has been demonstrated in humans in a study by Yee et al. (2009) where in a virtual environment subjects were given either short or tall avatars. The height of the avatars modulated the winning experience in encounters between subjects, so that even in subsequent encounter outside of the virtual environment individuals, that previously have been given taller avatars, behaved more aggressively (p. 321).

Self-Deception: a two-step argument	
1. SD > **positive self-perception**	If SD **more often than not** results in self-enhancement, then it should lead to positive self-perception (p. 318). The number of standard examples in the SD-literature, as well as Quattrone and Tversky's cold water - experiment can be seen as supporting evidence for the given claim.
2. positive self-perception **> adaptive benefit**	An argument is needed that not only is self-deception beneficial for humans in current social environment, but also that it was beneficial in the evolutionary past. The **winner effect** provides this evidence, given that it is present in other species, as well as is able to evolve (as a mechanism to facilitate smooth hierarchy development, see the avatar study by Yee et al. (2009)).

Table 38. Lopez & Fuxjager: two-step argument for exaptation
Distinctions from Lopez & Fuxjager (2011).
> stands for "leads to"

On this basis, Lopez & Fuxjager conclude that even if self-deception's evolutionary origin was being a spandrel, it has adaptive value because of the winner effect to which it leads (p. 322). A possible point of criticism is that Lopez & Fuxjager do no differentiate between positive self-perceptions and self-enhancement, whereas von Hippel & Trivers (2011) explicitly focus on self-enhancement as a type of self-deception, not on positive self-perceptions. Lopez & Fuxjager argue that the winner effect leads to a shift in the perceived abilities and the assessment of cost and benefit considering future encounters, thus, leading to a positive psychological effect of winning (p. 321). On the one hand, this does not clarify how positive self-perceptions and self-enhancement relate to each other. On the other hand, this does not exclude the possibility of the winner effect causing a self-fulfilling prophecy. As noted above, whether a self-fulfilling prophecy could be an instance of self-enhancement, is a separate matter to consider (3.2.2.3).

3.3 Comparison of the adaptation, byproduct and exaptation explanations

Let me compare the three theories elaborated in the previous sections – whether self-deception is an adaptation, byproduct or exaptation – against each other. Trivers argues that the ultimate cause of self-deception is to deceive others and the proximal means are the cognitive biases that constrain information-processing. Van Leeuwen holds self-deception to be a byproduct of practical rational capacities of the human mind that undermines knowledge (Van Leeuwen, 2013a) such as inertia of beliefs (leads to confirmation bias but has a function of imposing coherence in a belief-system) or memory suppression (leads to adjustment of thresholds of accepting beliefs but has a function of minimizing distraction). Lopez & Fuxjager agree with Van Leeuwen that self-deception is a byproduct, but argue that it is one that has gained a function of its own, namely via positive self-perceptions to lead to the positive psychological winner effect which consists in the shift of one's perceived abilities in dependence of previous situations in which one has won in a fight or a social conflict. Such an assessment is beneficial if it leads to the subsequent winning (a self-fulfilling prophecy) and may also be congruent with Trivers' function of self-deception to deceive others. In the following I will apply the checklist of critical points brought by Powell & Clarke (2012) against both adaptationist and byproduct explanations in the discussion of the evolutionary origins of religion[323] to the discussion on the evolutionary origins of self-deception (see table 39).

Adaptationist explanations	Byproduct explanations
1. Need to distinguish current trait-utility from its original utility.	1. Need to avoid "just so stories" in explaining the origins of the adaptations that brought about the given byproducts: If those adaptations arise out of *modular* structures, one needs to justify the modularity claim too.
2. Reliance on "just so stories" or plausibility alone. Worst case: disguising a "just so story" as an inference to the best explanation. Possible solution: identify proximal mechanisms. Danger: another "just so story" on the level of mechanisms.	2. Complexity of the byproduct explanation: a) Demonstrate *causal* link between adaptation and byproduct. b) Explain why the byproduct *persists* despite the absence of function, e.g. that the byproduct is connected to the adaptation in a certain way and/or it could not be eliminated by mutation.
3. Failure to consider other factors other than adaptation.	

Table 39. Powell & Clarke: methodological concerns - adaptation vs. byproduct Distinctions from Powell & Clarke (2012).

Adaptationist explanations 1. need to be careful about the scope of their claim (ancestor or current environment), 2. need to avoid telling a "just so story" and 3. need to take into consideration other possible factors apart from adaptation, e.g. learning through culture. To 1: Trivers' hypothesis that self-deception has an evolutionary function of facilitation of deception of others implies that it enhances the inclusive fitness in the ancestral and current environment. It is impossible to *test* evolutionary predictions in the ancestral environment (3.2.1), which should make one cautious about the scope of the evolutionary prediction. Another open question concerns the time that natural selection needs to operate (Powell & Clarke, 2012, p. 16), because the answer to this question constrains the possibilities of the development of self-deception: is it the usefulness in small hunting-gathering groups that needs to be shown or in bigger, more anonymous societies? Of importance is that it is actually not only the ancestral environment and the current environment that have to be

[323] Powell & Clarke (2012) argue that religious byproducts, if ever, are products of functional cognitive *module-like structures* (p. 12).

considered, but also *all the environments in between.* Because, if there were in-between environments in which self-deception was not beneficial, then one has to give an argument why it did not cease to exist or even why it developed anew (!). Given the impossibility to acquire evidence about the development of self-deception in the ancestral environment and the possibility to test it only in the current environment, what should we hold about Trivers' hypothesis that self-deception evolved to deceive others? It is important to realize that the usefulness of these two claims (self-deception being beneficial in ancestral and current environment) can be discussed separately, as also the usefulness of the answers to the ultimate (self-deception evolved to deceive others) and proximate (self-deception operates via self-enhancing biases that can be explained by self-affirmation theory) causes of self-deception. Thus, a more cautious claim would be that the environment, one needs to focus one's attention on, is of the current kind. If via experimental testing of proximal mechanisms the beneficiality of the ultimate cause of self-deception (which is other-deception) can be proven for the current environment, it would already show the usefulness of the evolutionary explanation. Even if an evolutionary hypothesis has to be narrowed down to be adaptive in the current environment, it still can be useful in providing ideas for new experiments such that constraints can be set on the conditions in which the phenomenon in question arises. Evolutionary theory here serves as a conceptual framework for and constraints on interpreting different phenomena (McGuire et al., 1992). Offering proximal mechanisms has been argued to be a way to avoid a "just so story" (see table 39). What does this mean to accuse en evolutionary explanation to be a "just so story"? It means that given our insufficient knowledge about ancestral environment (see section 3.2.1 for liabilities of evolutionary explanations), a story about how a certain feature or phenomenon evolved might be invented that is not testable. This is why I in section 3.2.2.2 presented evolutionary games that try to prove Trivers' theory. Even if the exact conditions of the ancestral environment are unknown, modelling can be made on the basis of certain premises about that environment. In this way, if the modelling results are positive, one might be one step further away from a "just so story," because the premises under which the phenomenon would evolve have been clarified and an evolutionary theory is then to be regarded as dependent on those premises.

To 2: Trivers considers biases as the main proximal mechanism of self-deception. The critique that Van Leeuwen offered against Trivers' theory was mostly critique about the proximate mechanisms via which self-deception operates (imprecision of the definition of self-deception, ubiquity of biases etc.). But are biases the *mechanisms* by which self-deception is executed or are those only the *constraints* that limit the execution of self-deception? Powell & Clarke (2012) cite the latter as one of the possibilities for the case of religion, namely that religion is constrained by "pervasive psychological biases" and the latter might serve as "attractors in the cultural evolutionary landscape" (p. 5). Whether these biases have to be *modular* is also in question. If the structures, that fulfill different functions, are overlapping, they cannot be modular and the notion of modularity per se is also unclear (3.2.1).

To 3: Another important factor, apart from genetics, is culture. An alternative idea to that of genetic fixation of traits is that the cultural environment enforces their development in individuals via learning. Varki & Brown (2013) have argued that this might be the case in self-deception (3.1.4).

What about van Leeuwen's byproduct idea? Van Leeuwen has elaborated on the reasons why practical rational capacities can be beneficial both in acquiring truth-conducive and biased judgments. Thus, he has made the connection between rational capacities as adaptations and self-deception as its byproduct. Yet, given that a byproduct explanation is more complex, more evidence would be needed to justify the given claim, starting with the evidence that the rational capacities he names evolved in the first place. It is also the case that the more complex the phenomenon itself is, the more explanatory work needs to be done to explain how it could *accidentally* arise as a byproduct. Self-deception is a complex phenomenon, since it has the potential to influence perception, judgment and behavior in a certain way. If the modularity thesis for practical rational capacities is rejected, then alternative evolutionary theory about their development has to be offered. The problem with modularity and evolutionary *psychology* in general is that a computational (information-processing) perspective is not enough, but it has to be backed up by neuroscientific data (Panksepp & Panksepp, 2000). Panksepp & Panksepp (2000) voice against a massive modularity thesis the concern that "most of the higher aspects of the human brain/mind arise largely from the interaction between general-purpose neural systems of the multimodal cortical association areas and the very basic life experiences encoded by more ancestral emotional/mind systems" (p. 112). A reason should also be given why self-deception *preserved* being a byproduct over time. For explaining the latter, Lopez & Fuxjager' exaptation explanation (self-deception acquiring the function to promote winners) might be accepted. Yet, it is susceptible to Van Leeuwen's imprecision criticism against Trivers' theory (3.2.2.1): is self-deception only about one's self-perception and/or about one's self-perception at all? Is the winning effect the only evolutionary benefit it offers? As a result, the byproduct/exaptation hypotheses are, by no means, superior to Trivers' evolutionary adaptation view. It should be noted, though, that the hypothesis that self-deception evolved to deceive others allows different (temporal) kinds of relationships between self-deception and other-deception (see table 40): self-deception may precede other-deception, or other-deception may be followed by self-deception justifying it, as well as that self-deception may be directed at past or future circumstances.

Other-deception (OD) as *by-product* of SD		**SD as *means* of OD**	**SD as *justification* for intentional OD**
About future plans: Implementational mindset	*About past mistakes*: Cognitive dissonance after fiascoes	Motivation: *persuading others* in the truth of something	Motivation: self-serving justification of public dishonesty, e.g. as strategic lie
The Bay of Pigs fiasco in April 1961: CIA agent's self-deception that invading Cuba to aid anticommunists would be successful. One of the other-deceived was Kennedy who has supposedly told after the fiasco: "How could I have been so stupid?" (p. 10)	*Intervention in Iraq by the Bush administration in 2003:* Instead of the danger of weapons of mass destruction, democratization has been taken as the reason for the intervention. ⇨ Redescription *ex post* to protect psychological well-being	*Tonkin Resolution in 1964:* The supposed presence of North Vietnamese torpedoes was used as a pretext to accept little evidence in favor of starting the Vietnam War.	*Cuban Missile Crises in October 1962:* R. Kennedy and Soviet Ambassador A. Dobrynin trade Cuban missiles of the Soviet Union for the American missiles in the Turkey. The secrecy of the deal might have served the self-deceptive belief that toughness in times of the crisis is the best strategy.

Table 40. Galeotti: typology of self-deception
Distinctions and examples from Galeotti (2014).

So, I present the results of the discussion on the evolutionary theories of self-deception. In the section on non-evolutionary theories of self-deception I argued that each kind of explanation presented was connected to the next. Cognitive dissonance explains self-deception as a *general* inconsistency between representations, but it could also be argued that only inconsistencies that threaten the self-concept necessitate such a solution. Self-esteem concerns may be argued to be most threatening to the self-concept and terror-management theory ascribes to the defense of the self-esteem the (possibly evolutionary) function to relieve anxiety of death. This indirect route to ascribing self-deception the function to relieve anxiety of death falls together with the direct route that Varki & Brower (2013) proposed, namely that self-deception evolved to relieve anxiety of death, because anxiety of death is the psychological evolutionary barrier against the development of full theory of mind. Combining evolutionary and non-evolutionary explanations of self-deception and in the light of the fact that self-deception might have multiple functions (Surbey 2004), I want to propose a hypothesis that self-deception may have evolved to relieve the anxiety of death, but in virtue of the changes of social relationships in the current environment may have acquired an additional function of other deception. Let me call it the *combined function hypothesis*. This is a hypothesis that requires empirical testing and, best, if BIDR questionnaire were not the only one criterion for distinguishing self-deceivers, but also my suggestions at the end of section 2.2.3 were considered. This is because in virtue of its weakness – intentional object of measurements not being clear (1.3.1) – how is one supposed to evaluate a study in which self-deception was established by BIDR measurements and it was argued that self-deceivers are less confident and credible liars (Wright et al., 2015)? The answer is: if BIDR does not measure self-deception, or only a certain type of self-deception, then, it can either be the case that the results are invalid for inferences about self-deception in general, or for inferences about this type of self-deception in particular. The latter possibility leaves it open whether only a certain type of self-deception, namely tension-free self-deception, enables deceiving others. Last point to notice about the discussion of the function of self-deception is as follows: instead of internal states as demarcation criteria for self-deception (chapter 1) or behavioral and phenomenological profile (chapter 2), one could also set self-deceptive *function* as a demarcation criterion such that everything that fulfills this function, e.g. deceiving others, is a kind of self-deception. I think that, whatever function a phenomenon possesses, it can certainly be considered in *addition* to the phenomenological and behavioral profile, but that a function alone can be fulfilled by different, even if related, phenomena.

4 Modelling self-deception

In the previous chapters I have reviewed literature on self-deception, offered my own definition of the phenomenon, set constraints and discussed different proposals regarding its function. I set four constraints on the satisfactory explanation of self-deception in the first chapter. My take on them was as follows:

- *Parsimony*: No internal states or kinds of attitudes should be postulated in the explanation of self-deception just in virtue of explaining self-deception (sections 1.2.5, 2.1.1 and 2.1.2);
- *Demarcation*: Self-deception should be distinguished from other phenomena by setting more precise behavioral and phenomenological constraints (section 2.1.2);
- *Disunity in unity*: Behavioral profile of a self-deceiver (inconsistency that is being justified upon demand) requires that there is a certain kind of disunity, which again requires that there would be a sort of a unity that non-self-deceivers possess and self-deceivers lack. Usually, the personal level serves as the unifying element, while there is a disagreement on how to pinpoint the kind of disunity that self-deceivers exhibit. *Personal* level *synchronic* disunity leads to paradoxes, whereas disunity between the personal and subpersonal level, e.g. stereotypical behavior, would fulfill neither the justification criterion, nor the phenomenological profile. What I focused instead was the explanation of *diachronic* inconsistency of behavior on the assumption that self-deceiver's attitude oscillates. Two kinds of selection have been hypothesized to underlie self-deception: selection of the world/self-model and of the epistemic agent model (section 2.2.1). The second kind of selection, in virtue of cognitive binding filling the argumentative gaps (section 2.2.2.3), would lead to attitudes whose signature of knowledge has become transparent (section 2.1.1). Further, the constructed epistemic agent model might be characterized by counterfactual goal-directed pull (section 2.2.1). The first kind of selection, on the contrary, would lead to attitudes that have become transparent themselves and, thus, are experienced as real (section 2.2.3).
- *Phenomenological congruency*: Self-deceivers experience tension during self-deception and insight thereafter. Tension might play at least four roles in self-deception (section 1.2.5 and 2.1.3). The first pair of roles is: causing and resulting from self-deception due to the presence of inconsistent attitudes. This is paradoxical (section 1.1.2.1), unless one assumes that there are tension-cycles in self-deception: when it becomes tension-free and vice versa. This assumption per se is not new, since tension is argued to appear each time an inconsistency is detected on the personal level, e.g. an agent is confronted with contradictory evidence. What I proposed is that tension elicits new hypothesis testing cycles that may let it disappear, e.g. in virtue of changing counterfactuals underlying it (see second pair below). The second pair is: arising out of a self-deceptive process and changing that mentioned process. The first role of the second pair led me to the argument that self-deceptive tension is a kind of a metacognitive feeling with an indicator function (section 2.1.3). As for the second role of the second pair, here the phenomenal aspect might be an epiphenomenon such that the sheer fact that tension is being represented changes the subpersonal dynamics of the information flow and transitions to different kinds of hypothesis spaces. Tension may, namely, depend on the presence of certain counterfactuals (Dokic's idea; section 2.1.3).

Construction of a computational model of self-deception that considers all the constraints is the next step, if self-deception is to be tested empirically. Computational models (information processing models) have been argued to be useful in inferring mechanisms from *behavior* and *brain activity* (Stephan & Mathys, 2014). Importantly, a construction of a computational model requires a translation of psychological constructs into another, mathematical, form.[324] One has to be cautious about the use of terms "network" and "system" though, since they are used in a different manner by different disciplines[325] (McIntosh et al., 2001, p. 1235).

Predictive coding is the implementation of empirical Bayes where the latter is a procedure for representational learning and inference that is assumed to occur in the human brain (Friston, 2003, 2005). *Bayesian models* have at their core the premise that representational learning and inference abides by the Bayes theorem which is that a posteriori probability of a representation is equal to its prior probability times the likelihood. *Predictive coding* (PC) states that representational learning and inference in the brain occurs as a result of the propagation of *prediction error* along the hierarchy of levels at each of which it is compared with predictions of different degrees of granularity, so that the latter can be updated according to the Bayes' rule.

Let me briefly remind the reader about the open question that I promised to answer in this chapter. First, self-deception is a kind of *hypothesis* testing (see introduction to chapter 2). Second, self-deception is a *repeated* kind of hypothesis testing such that each hypothesis testing cycle differs from the next one in certain parameters, among them the control states (beliefs about future actions that substitute in the predictive coding framework goal representations) and counterfactuals (sections 2.1.2 and 2.1.3). *Transitions* between control states and hidden states (those modelling states of the world) or, more generally, the temporal dynamics between cognitive and affective states will need to be layed out for a model of self-deception (section 2.1.3 and 2.2.2.3).

In this chapter I will first review connectionist modelling solutions (section 4.1), then *personal* level error minimization accounts (section 4.2) that underlie Mele's view (1.1.2.3). Thereafter, I will differentiate four possible types of Bayesian accounts regarding the personal level implications that those might possess (section 4.3). This is because predictive coding is a *Bayesian* error minimization account, so, how error minimization, Bayes' rule and the phenomenal level interact is an important question to settle. I will then introduce the explanatory tools that predictive coding possesses and that will be helpful in explaining self-deception (section 4.4). Lastly, I will suggest what a computational model of self-deception might look like (section 4.5).

4.1 Connectionist models

In this section I will consider Thagard's constraint satisfaction account that argues that self-deception arises out of the simultaneous satisfaction of two sets of constraints – cognitive and emotional. Thagard's constraint satisfaction account is in accordance with the connectionist perspective on goal-directed knowledge selection (2.2.1) and its merit is the emphasis on the incorporation of affective information into the explanation of self-deception, yet may have to be replaced by a more neurologically plausible account – predictive coding that will be considered next (see section 4.2).

[324] McIntosh et al. (2001) argue that a mapping of a psychological construct to the brain "is contained in the dynamic operations of large-scale cortical networks" (p. 1233).

[325] Different definitions of a network: "In neuroanatomy, a network typically corresponds to a small collection of neural elements, usually neurons or neuronal ensembles, whereas a system is defined as a collection of anatomically related structures that serve a distinct function (e.g., visual system, vestibular system). In neural modeling, a network is defined by sets of equations that specify influences between modeled elements. In cognitive psychology, a network may be used in reference to a computational model or a conceptual model that relates cognitive processes or subprocesses." (McIntosh et al., 2001, p. 1235)

In connectionist models computations, which are algorithmic descriptions (Sun, 2008), are carried out in a network that consists of units that can influence each other by excitatory and inhibitory connections (Thomas & McClelland, 2008, p. 26). Connectionist models are *simplifications* of the computations of neural systems (p. 29) and may approximate some features of symbolic systems (p. 30). Knowledge can be *stored* in connection weights of connections between units and it may be *activated* at different points in time (p. 46). In case of self-deception one would take psychological constraints worked out in the previous chapters to craft a specific kind of network that satisfies these constraints (Sun 2008, p. 10). Greenwald (1997) holds that self-deception, e.g. nonverbal skin conductance response and the deviating verbal response (see section 1.3.2.1 on Gur & Sackeim's skin conductance experiment) can be explained by supposing *independent paths* through such a network (p. 64-65). His explanation of conscious cognition by means of a connectionist network is too simplistic though:

> The network model […] incorporates representations of two forms of conscious cognition. […] One of these – conscious cognition as network operation that boosts activation to resonantly stable high levels in subnetworks – corresponds to an interpretation of *conscious cognition as a focus of attention* on some thought or percept. The network's second representation of conscious cognition is its possibility of having verbal outputs that, by virtue of their connections to inner nodes ("hidden units") of the network, are able to report (in some sense) on internal network status. These verbal outputs correspond to an interpretation of *conscious cognition as a capacity for introspective report* (or "self-consciousness"). (Greenwald, 1997, p. 64)

Sahdra & Thagard (2003) propose another connectionist model of self-deception – a parallel constraint satisfaction account. Reasoning is assumed to be coherence-based and holistic (Thagard, 2006, p. 29). Coherence is achieved through (multiple) constraint satisfaction.[326] There are at least two sets of constraints – cognitive (activations) and emotional (valences), the difference between which might be demonstrated on the following example:

> But the coherence computation determines not only what elements will be accepted and rejected, but also an emotional reaction to the element. It is not just "go to Paris - yes" or "go to Paris – no," but "go to Paris – yeah!" or "go to Paris – yuck!" (Thagard, 2006, p. 21)

This distinction is taken by Thagard to reflect the one between "hot" and "cold" biases (Sahdra & Thagard, 2003, pp. 213-214). As such, it might also be compared with McKay's et al (2007, p. 938) distinction between constraints of *verisimilitude* and *motivation*.[327] Units in Thagard's (2000) connectionist network are argued to stand for cognitive representations, emotions, bodily states, coherence, incoherence, as well as special *metacoherence* units of happiness and anxiety; constraint satisfaction between these units allows one to assess which inferences a subject would build on a certain matter given the

[326] Kinds of coherence: "*Coherence* is determined by the constraints between representations, with different kinds of constraint and different kinds of representation for six kinds of coherence: explanatory, conceptual, perceptual, deductive, analogical, and deliberative. For example, in *explanatory coherence*, the representations are propositions and the constraints include positive ones that arise between propositions when one explains another and negative ones that arise between propositions that are contradictory or in explanatory competition" (Thagard, 2006, p. 161; my emphasis).

[327] McKay et al. (2007) note about Sahdra & Thagard's model: "The resulting system successfully self-deceived in that it yielded acceptance of false propositions, consistent with implemented preferences" (p. 938).

evidence, emotions etc. she possesses (p. 195). For the prediction of beliefs of the self-deceiver the second, emotional kind of constraints is particularly important. Distribution of valences (by means of emotional constraints) across the network results in an *emotional gestalt* or “gut reaction” (Thagard, 2006, p. 30):

> The result [of an emotional coherence that assesses activations and valences] can be an *emotional gestalt* consisting of a cognitively and emotionally coherent group of representations with degrees of acceptance and valences that maximally satisfy the constraints that connect them. (Thagard & Nerb, 2002, p. 276; my emphasis)

“[C]hanges in representations and their valences” make possible e*motional gestalt shifts* (Thagard & Nerb, 2002, p. 276). The emotional gestalt is specified by metacoherence units which are bound to give “an overall assessment of how much coherence is being achieved” (Thagard, 2000, p. 193). The rationale for the existence of metacoherence units is the following: given that it is impossible to know directly whether *explanatory coherence* has been achieved, *metacoherence* units provide a *conscious* assessment of the extent of this coherence (Thagard 2006, pp. 255-256) that finds its expression in certain kinds of feelings:

- *Explanatory coherence* leads to the feeling of happiness, through which the person *recognizes* that coherence has been achieved (Thagard, 2006, p. 255)
- *Incoherence* leads to a general feeling of anxiety that may trigger a search for new hypotheses (Thagard, 2006, p. 256)

The gut feeling may be valid, if it indicates “coherence with evidence,” or invalid, if it indicates “a different kind of coherence based on wishful thinking or motivated inference rather than fit with the evidence” (Thagard, 2006, p. 256). In the latter case an *emotional skewer* may be at work which is “a factor that is so vivid and affectively intense that it can produce an emotional gestalt that does not *reflect* the maximal satisfaction of the *relevant and appropriate* constraints. Such factors skew the coherence calculation by placing too much weight on *epistemically irrelevant factors* such as self-promotion” (Thagard, 2006, p. 256; my emphasis). I interpret this quotation as saying that evidence (forwarded in the network through activations) and emotions (forwarded through valences) might lead to a difference in assessment of coherence – a difference reflected in the degree of activation of coherence/incoherence units in comparison to happiness/anxiety units. Such an emotional skewer is argued to lead to self-deception as an unjustified acceptance of a belief.

Sahdra & Thagard (2003) tested their idea by constructing a network in which the units are propositions (of two types: evidence + hypotheses[328]), the connections among which possess activations (excitatory+inhibitory) and valences (positive+negative). The material for propositions and hypothesized connections was provided by Hawthorne’s *The Scarlet Letter* for building a HOTCO2 model.[329] After the network is specified, whether a specific proposition or a set of propositions (hypotheses) will be accepted or rejected might be tested by letting the activations and valences spread through the network until it settles and there is no more change (Sahdra & Thagard, 2003, p. 229). A proposition is accepted, if its activation is positive (p. 222). In case of self-deception it may be an inconsistent set, e.g it

[328] The division of propositions into hypotheses and evidence determines the sets of propositions that are free of contradictions. This is the case because, according to Sahdra & Thagard’s (2003) model, when several hypotheses explain the same piece of evidence, then they are coherent, but they can also compete, if they are not explanatorily connected (p. 220).

[329] In HOTCO activations influence valences, but not vice versa, while HOTCO2 the influence between activations and influences in bidirectional (for equations see Thagard & Nerb 2002). Thus, in HOTCO, “the activation of the *coherence nodes* depends on the units’ calculation of their degree of *constraint satisfaction*, [...] the activation of the *happiness and sadness nodes* depends on the units’ *activations and valences*” (Thagard, 2000, p. 197; my emphasis).

may contain propositions like "I am a good clergyman" and "I am hypocritical but selfless" (p. 222). The heavier the weight of the input valence, the more inconsistent the set can become,[330] leading the authors to conclude that "self-deception occurs via emotional biasing of people's beliefs, while people attempt to avoid or approach their subjective goals" (Sahdra & Thagard, 2003, p. 222). Self-deception is argued to be a result of *unconscious* emotional coherence (p. 226), as a result of which "anxiety or internal conflict typically, but not necessarily, accompanies self-deception" (p. 226).

Not only self-deception as *motivated inference* is argued to be explained by *gut overreaction*, but also *fear-driven inference* where Mele's "twisted self-deception" is taken to be a kind of fear-driven inference (Thagard & Nussbaum, 2014). Gut overreaction is the case in which one's emotional reaction *misinforms* one about the evidential value of a hypothesis, because the subjects *misinterpret* arousal. If arousal is misinterpreted to support a certain hypothesis, then this hypothesis may become a self-supporting one, e.g. "Othello feels bad" and holds that this supports the hypothesis that "Desdemona is cheating" which leads him to feel worse and to believe in the truth of the hypothesis even more, therefore building a *feedback loop* (Thagard & Nussbaum, 2014). Such feedback loops between evidence and emotion, where the latter encompasses integration of appraisal and bodily reactions, are argued to lead to amplifying reactions and acceptance of false propositions.

From the four characteristics of self-deception (inconsistency, justification, tension, insight) it is tension that the given connectionist model seems to focus on. I have offered to view tension as a kind of metacognitive feeling (2.1.3). In virtue of Thagard & Nussbaum's (2014) assumption that feelings might be misinterpreted and misdescribed it merits to mention that Proust (2015a,b) ascribes the property of *transparency* to feelings, which means that they can be redescribed, but their nature cannot be thereby changed, e.g. arousal cannot be felt as happiness or anger depending on the circumstances (see Schachter & Singer's (1962) dual factor theory of feelings for the defense of the latter assumption). On the other hand, in the discussion of self-enhancement as one possible kind of defense (see section 3.1.3.2) I mentioned that one unifying feature of defensive mechanisms has been argued to be *affect* of unknown origin that mediates between different kinds of defensive processes (Tesser et al., 1996, p. 63). It is further a feature of self-deceptive phenomenology that it cannot be distinctive, because else the self-deceiver would recognize self-deception every time a self-deceptive feeling arises (2.1.3). I think that that one can uphold Proust's transparency thesis and the possibility of misinterpretation of feelings if one allows for the existence of *indistinctive* feelings such as those present in self-deception. Thus, one could say that it is not *arousal* in feelings that is misinterpreted, but the *origin* of feelings.

Sahdra & Thagard's model goes against another, more central feature of Proust's theory of feelings that I applied to self-deception. According to their model, feelings arise as a result of constraints among units, but not as a result of a *comparison* of the validity of the belief-forming process with one's chosen criteria. Their idea is more in unison with Huebner & Rupert's assumption (2014) that every representation possesses a motivational force and that feelings indicate incongruency with goals (Baumeister & Newman, 1994) than with a *comparator* model that Proust has argued for (2.1.3). Baumeister & Newman's (1994) idea was that self-deceptive belief-forming process is repeated until it is goal-appropriate and, thus, self-deception can actually succeed to extinguish negative affect. I argued, on the

[330] Sahdra & Thagard's (2003) formulation for the trend: "The general trend was that the greater the valence input, that is, the more emotional the system, the easier (that is, faster) it was for the system to model self-deception" (p. 227).

other hand, that self-deceptive feelings indicate inappropriateness of the subpersonal belief-forming process of the self-deceiver, but their *intensity* depends on the degree to which the results of the process are incongruent with goal representations that the self-deceiver possesses, because in virtue of being inappropriate the self-deceptive belief-forming process does not succeed. Sahdra & Thagard's model suggests that feelings are not extinguished, but misinterpreted. The correct kind of description is an empirical question. Maybe several are true in different kinds of situations. I will come back to this question in the next section during the discussion of Anil Seth's interoceptive predictive processing model.

Sahdra & Thagard (2003) reject Bayesian approaches to modelling human thinking (and along it Talbott's (1995) Bayesian model of self-deception) in favor of *coherence*-based approaches by accepting the evidence that human reasoning often goes *against* the laws of probability theory[331] (p. 228). Thagard & Nussbaum (2014) also criticize conventional probability and utility theories of reasoning for not being able to explain the *interdependence* between probability and utility. These and other concerns have been argued to be overcome by a new subpersonal Bayesian model of perception, cognition and action – predictive coding, which I discuss in the next section. Summarizing, the function of this section was to introduce a connectionist model of self-deception that I think is on the right track in highlighting the importance of the affective dimension to self-deception.

Interim conclusion: In this section I introduced one connectionist model of self-deception that emphasizes the importance of affective dimension (every representation varying not only on the propositional, but also on the affective dimension). In terms of the four tension functions I would categorize Sahdra & Thagard's solution as implementing the 4th function, namely tension as part of the (subpersonal) self-deceptive process itself. Affective dimension does not serve an indicator function and the phenomenology of affect does not play a role (may be an epiphenomenon). What is crucial for this model is the influence of the affective dimension on changing the *weights* in the connectionist network.

Since I chose predictive coding over connectionist models, I will do a quick comparison of both in the remaining part of the section. First, let me consider the explanatory benefits of predictive coding (which is an implementation of empirical Bayes, see section 4.3) independently of connectionist models. Tenenbaum et al. (2006) argue for a computational-level Bayesian framework of "inductive learning and reasoning as statistical inferences over structured knowledge representations" (p. 309). A *computational-level* framework is for them a functional one (not the level of implementation, see section 4.3 for different types of Bayesian explanations). *Bayesian* means that inference in such a framework abides by the Bayes' rule (posterior probability of a representation equals prior probability times likelihood). Bayesian inference is *statistical* because it takes into account prior experience with the data. Lastly, *structured* knowledge representations is a characterization of how the probabilistic model that generates expectations (that are compared with the data) looks like, namely that it possesses a certain structure that may depend on different principles (Tenenbaum et al. 2006, p. 310). That such a model can be hierarchical allows for different levels of abstraction. That "Bayesian inference can account for knowledge acquisition at any level of abstraction" (p. 314) is a significant benefit. So, probabilistic models seem to provide answers to a variety of questions.

Second, the relation between probabilistic and connectionist models can be described as follows: with respect to the amount of structure in the model and the strength of inductive

[331] On the other hand, Thagard (2000) argues that ECHO can – a model only with activations, but with no valences – can be interpreted as a special case of a probabilistic network (p. 248).

biases connectionist models are more restrictive (they assume that both are weak), while probabilistic models allow for the variation on both scales, which means that the space of possibilities is larger (Griffiths et al., 2010, p. 358). As a result thereof, according to Griffiths et al. (2010), the approach to building a model differs for probabilistic and connectionist models so that probabilistic models are of a top-down ('function-first') type (they start with abstract principles), while connectionist ones are of a bottom-up ('mechanism-first') type (they start with the lower level, namely the neural one). Griffiths et al. (2010) use the more flexible nature of probabilistic models to support their claim that top-down probabilistic models are superior to bottom-up connectionist models, insofar as they allow for variation of the nature of representation and inductive biases. They emphasize the explanatory value of explicit representations (p. 359): "all models – including connectionist models – build in hypothesis spaces; probabilistic models simply make the space explicit" (p. 362).

Marcus & Davis (2013), on the other hand, claim that the evidence in favor of probabilistic models of higher-order cognition is inconclusive. They identify two problems:

- *Task selection*: a small number of tasks is compatible with probabilistic models in every domain (p. 2354)
- *Model selection*: the fit of the model depends on making experimental data-dependent assumptions (p. 2356), in the extreme the fit of the model is forced by interpreting how subjects understand the questions experimenters pose: "[a] response that can be rationalized is not the same as a response that is rational" (p. 2359).

Given these liabilities, Marcus & Davis (2013) conclude that probabilistic models of higher-order cognition "have not converged on a uniform architecture that is applied across tasks" (p. 2359). On the premise that inferences about cognitive processes on the basis of behavior are underdetermined (p. 313), McKay & Efferson (2010) also argue that there is a possibility for the implementation of belief-forming mechanisms being non-Bayesian due to e.g., Bayesian implementations being "expensive neurologically" (p. 315). They claim that empirical evidence is consistent with both the assumption about Bayesian and non-Bayesian implementation (p. 317), but point out that it is difficult to establish experimentally the Bayesian nature of cognition, because one does not know how much time is needed to change the posteriors of two participants with different backgrounds and it is natural to assume that participants in experiments *will* have a different background.[332] These pieces of criticism are primarily directed against personal level Bayesian accounts of human reasoning. Predictive coding offers not a *personal* level Bayesian explanation of cognition, but a *subpersonal* one. Still, for predictive coding too it has to be justified that it is not another "just-so" story such that given the data there can always be constructed a predictive coding model to fit it (see Clark, 2013b with comments and response for the discussion of the topic). New articles are currently published on predictive coding and the explanatory possibilities are still being expanded, so I do not see it wise to prematurely discard the theory.

[332] Gender-specific difference in socialization as influencing experimental results: "Men and women, however, are socialized in systematically different ways. As a result, they will have different evidence available to them as they proceed through life, and they may enter an experimental scenario with different prior beliefs about female sexual interest—beliefs that may be justifiable in each case, given the evidence they have, respectively, encountered." (McKay & Efferson, 2010, p. 318)

4.2 FTL model:[333] Error minimization account

That self-deception is explainable as a kind of error minimization is a view defended by Mele (2001; see section 1.1.2.3). Error minimization comes in at least two flavors – that of James Friedrich (1993) and that of Yaacov Trope & Akiva Liberman. Both are *personal level* accounts of error minimization. Predictive coding is a *subpersonal* level error minimization account, so discussing of the FTL model will serve for comparison purposes. The main claim of Friedrich's model of hypothesis testing is that it serves the goal of error minimization (Friedrich, 1993). A criterion for a good model of hypothesis testing is to him the focus on the minimization of mistakes and the production of a small number of errors (p. 301). [334] Errors are here inconsistencies with a chosen *model* and *mistakes* are "inaccuracies in judgments of real-world stimuli" (Friedrich, 1993, p. 301).

There are two types of errors according to him:

- False positives: coming to the conclusion that something is the case when it is not;
- False negative: coming to the conclusion that something is not the case when it is.

Friedrich (1993) uses the personal level descriptions and examples for false positive and false negative errors, as in the following example:

> If I want to purchase a car, for example, I am generally more concerned with the possibility of choosing a car that turns out to be a lemon (a false-positive error) than I am with overlooking a car that might have served me well. (Friedrich, 1993, p. 299)

Friedrich (1993) further denies that the distinction between accuracy motives (to establish the truth) and directional motives (to reach a certain conclusion) (for this he is criticized by Mele (2001), see section 1.1.2.3). In other words, according to Friedrich, there is only one kind of motivation – error reduction – that guides the hypothesis testing process. It follows that when people behave as if they were guided by accuracy motives or by directional motives, it is a mere *appearance* (see figure 29):

> Also, the more normatively appropriate behavior attributed to the operation of accuracy motives may simply reflect circumstances in which lay hypothesis testers are testing for more than one type of error (both false positives and false negatives). This would yield hypothesis testing that takes on a normative appearance without having been guided by concerns for accuracy in a formal, logical sense.
>
> One critical difference between a PEDMIN analysis and other motivational accounts is the former's prediction that changes in hypothesis-testing strategies should accompany changes in the focal or primary error, even when hypothesis testers lack the kind of personal involvement in the issues that would make them "want" to reach particular conclusions. (Friedrich, 1993, p. 300)

[333] The term 'FTL model' is used by Mele (2001) for his version of an error minimization account based on the two accounts to be presented here.

[334] Evolutionary theories of reasoning fulfill the given criterion. Friedrich cites Cosmides and Tooby and agrees with the following: "From an inclusive fitness perspective, strategies that minimized costly errors and mistakes in biologically significant contexts (e.g., social exchange, mate selection, and predator avoidance) should be selected for. Content-free, domain-general reasoning ability would be favored only to the extent that its application in these contexts resulted in fewer reproductively costly mistakes than did alternative strategies" (Friedrich, 1993, p. 301).

This quotation highlights that reduction of errors, though it is described in personal level terms, is not to be understood as a mental action of the subject. Errors are minimized by two cognitive testing strategies (Friedrich, 1993):

- *Plausibility testing* in cases when neither error is important: merely establishing the plausibility of a hypothesis is enough, use of heuristics and shortcuts (pp. 304-305). The desire to find out the truth, but be unbiased can only arise if one cares about the reduction of *both* errors, but not if one does not care about the reduction of *either* of them.
- *Sufficiency testing* in cases of saliency of specific errors: checking that a *selected* hypothesis produces a positive outcome, even if it is not accurate (pp. 305-308). Although sufficiency testing leads one to believe falsehood (because importance of the task disturbs the balance of gathering data) and to be overconfident in the truth of certain hypotheses, it is argued to be adaptive in avoiding costly errors (p. 305-306).

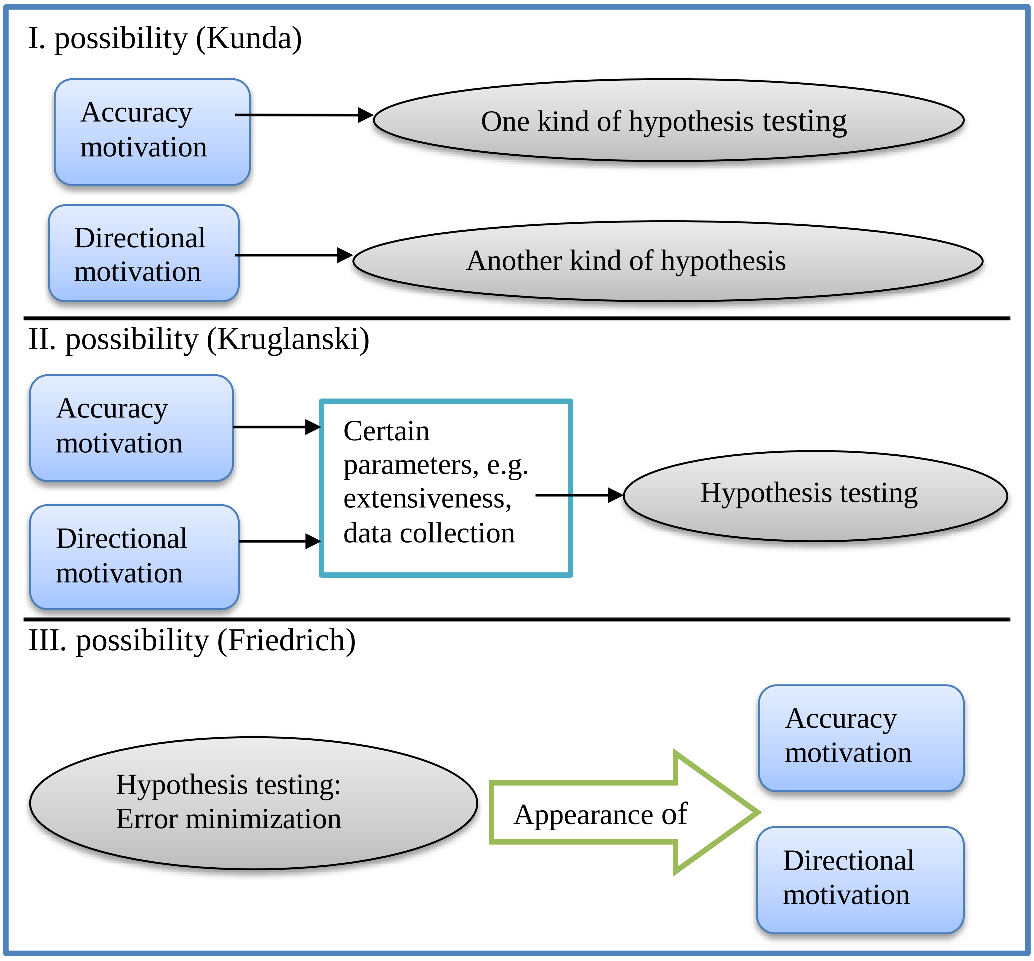

I. possibility (Kunda)
Accuracy motivation
One kind of hypothesis testing
Directional motivation
Another kind of hypothesis
II. possibility (Kruglanski)
Accuracy motivation
Directional motivation
Certain parameters, e.g. extensiveness, data collection
Hypothesis testing
III. possibility (Friedrich)
Hypothesis testing: Error minimization
Appearance of
Accuracy motivation
Directional motivation

Figure 29. Friedrich: How motivation could influence hypothesis testing. Distinctions from Friedrich (1993, pp. 299-300).

Which error is considered as important and the one to be minimized, is “a function of situational and dispositional variables” (p. 310). In other words, it is the context that determines the choice of the cognitive testing strategy. Self-serving motives, like enhancing one’s self-esteem, becomes a primary error because of the cost of psychological discomfort (Friedrich, 1993, p. 314). Cost-benefit ratio also determines the occurrence of self-deception:

> 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. When mistakes associated with self-deception are costly and salient, however, self-servingness should decrease markedly as test strategies shift to more effectively limit these alternative errors. (Friedrich, 1993, p. 314)

Perceived control over outcomes, one of Taylor's (1989) 'positive illusions' claimed to be self-deceptive (see section 3.1.3.1), determines according to Friedrich (1993) the primary errors in the sense that there is a tendency for primary errors to be controllable (p. 315). Self-deception, on the other hand, may concern errors which are *not* controllable, e.g. for low scorers on the IQ test (which is uncontrollable) the primary error would be "lowering self-esteem unnecessarily" (Friedrich 1993, p. 316) and possesses a defensive function of pain alleviation (p. 316). Friedrich (1993) points out that his hypothesis testing account – error minimization - is to be favored for parsimony reasons: it explains the variety of strategies that hypothesis-testers apply, as well as their striving for error minimization instead of truth-discovery (p. 317). Bayesian accounts may be seen either as rivals to this error minimization account in explaining mental processes (p. 298), or as completing each other. Predictive coding possesses the parsimony strengths that Friedrich mentions, but incorporates in addition the Bayes' rule as an explanatory element.

Trope & Liberman (1996) argue that the hypothesis testing process consists of two steps: the generation of a hypothesis and its evaluation (see figure 30). The generation of a hypothesis is goal-dependent (p. 249). The specificity[335] of the hypothesis, its probability of being true and its desirability affect its subsequent evaluation. The evaluation of a hypothesis consists of four steps:

- the derivation of conditionals about the hypothesis ("*if* some possibility is true, *then* some evidence is likely to exist", p. 241);
- seeking the evidence according to the conditionals;
- interpreting the evidence (deciding whether it is consistent with the hypothesis, whether it supports it);
- assessing the likelihood of the hypothesis being true (on the basis of the available evidence) (p. 241).

[335] Hypothesis as more specific than its alternatives: "Any given hypothesis is usually more specific than its alternatives. A hypothesis often refers to a single possibility (e.g., the target is a lawyer), whereas the alternatives may include a large number of possibilities (e.g., the target has some other occupation)" (Trope & Liberman, 1996, p. 247).

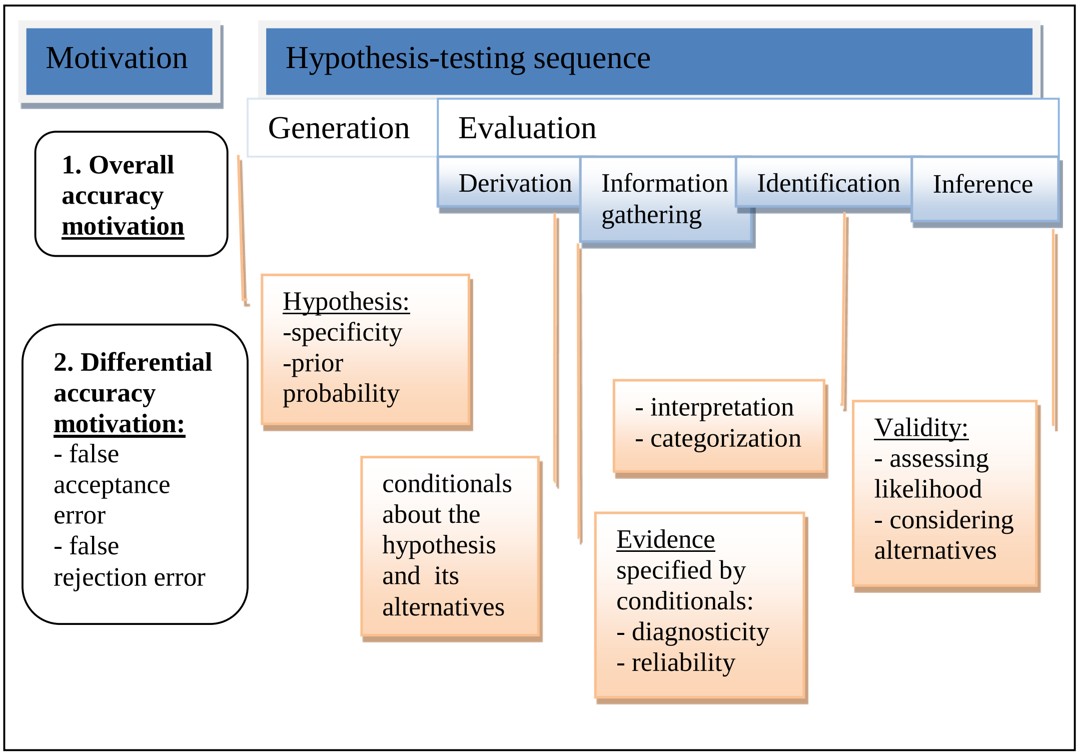


Figure 30. Trope & Liberman: hypothesis testing model
Distinctions from Trope & Liberman (1996).

Trope & Liberman (1996) differentiate between two hypothesis-testing methods each of which can influence any of the steps previously identified (p. 243):

- *diagnostic testing*: "takes into account the diagnosticity[336] and reliability of evidence, and the base rate probability of the hypothesis" (p. 241);
- *pseudodiagnostic testing*: If either motivation or cognitive resources or both are low, then people may rely on heuristics or sufficiency testing[337] (p. 242).

These two methods are similar to Friedrich's (1993) plausibility and sufficiency testing: plausibility to pseudodiagnostic and sufficiency to diagnostic. Trope & Liberman (1996) do add thought that diagnostic kind of inference may *reduce* the confidence in the truth of a hypothesis under consideration, if evidence is in accordance with alternative hypotheses discovered during diagnostic testing (243).

The hypothesis-testing sequence is initiated by the *motivation* for hypothesis-testing[338] which is equal, according to authors, to the decision criteria (= confidence thresholds, pp. 253-254):

- acceptance threshold (cost of false acceptance relative to the cost of information)

[336] Evidence is diagnostic, if it discriminates between the to-be-tested hypothesis and alternative hypotheses (Trope & Liberman, 1996, p. 241).

[337] Difference between diagnostic and pseudodiagnostic testing: "Whereas diagnostic testing considers both the sufficiency of a hypothesis (i.e., the probability of the evidence if the hypothesis is true) and its necessity (i.e., the probability of the evidence if the hypothesis is false), pseudodiagnostic testing assesses only the sufficiency of a hypothesis" (Trope & Liberman, 1996, p. 242).

[338] The authors restrict the scope of their theory insofar as it is not the process of how motivation arises that is to be explained, but the link between already existing motivations to hypothesis-testing: "Our model has little to say about the underlying motivations that lead people to test their hypotheses in the first place. What it provides are mechanisms through which any underlying motivation, biased or unbiased, can be translated into *common* epistemic terms of error minimization and guide the hypothesis testing process" (Trope & Liberman, 1996, p. 254).

- rejection threshold (cost of false rejection relative to the cost of information)

While asymmetric costs lead to biased motivation, symmetric costs do not secure that the motivation is unbiased or objective, according to the authors (p. 254). They bring as an example a scientific investigation where acceptance thresholds should be stricter for the investigation to be considered unbiased (p. 254).

Antecedents of hypothesis-testing motivation	Motivation	Consequences of motivation
- prior confidence - cost of information seeking - cost of the errors of false acceptance and false rejection	• overall motivation (to avoid errors in general) • asymmetric motivation (differential decision thresholds)	Hypothesis-testing sequence

Table 41. Trope & Liberman: link between motivation and hypothesis-testing. Distinctions from Trope & Liberman (1996, p. 240, 254).

Trope & Liberman (1996) differentiate between overall motivation to avoid errors (p. 240) and differential motivation that depends on the importance of false acceptance and false rejection errors (p. 240; see table 41). Apart from error costs, low prior confidence which authors equate with *uncertainty*, drives the hypothesis-testing process (p. 256). Thus, hypothesis-testing is partly motivated by the desire to reduce uncertainty (p. 254). Trope & Liberman (1996) further hold that if asymmetric motivation is present, then there are two ways by which this directional motivation can affect the hypothesis testing:

1. "by *cognitive mechanisms* that bias the hypothesis-testing sequence" (p. 258; my emphasis);
2. by setting the *decision criteria* at a certain level which lead to asymmetric information search and earlier termination of the process (pp. 258-260).

The usefulness of this distinction is questionable. It is namely unclear why Trope & Liberman (1996) bother to define two ways by which hypothesis testing might be skewed (setting of thresholds and biasing the hypothesis testing sequence), if Trope et al. (1997) state that evidence can always be interpreted either way:

> It appears, then, that the same evidence may be construed and reconstructed in different and even opposite ways, depending on the perceiver's hypothesis. This makes hypothesis testing a subjective process, whose outcome depends on the mental models one brings to bear on the incoming evidence. (Trope et al., 1997, pp. 115-116)

Interim conclusion: Trope & Liberman's (1996) approach is more elaborate than that of Friedrich with respect to which stages hypothesis testing involves and which factors, apart from error costs, influence it, e.g. priors, which links their approach to Bayesian ones (see the end of the previous section for the explanation of the Bayes' rule). Still, both are personal level approaches of error management which also Mele (2001) sees as such: "They [hypothesis testers] test hypotheses in ways that are shaped by error costs, which, in turn, are shaped by their desires" (p. 42). Desires (personal level term) do not influence subpersonal errors. That Trope & Liberman's approach is of a personal level kind is also supported by the fact that they provide a link to *mental models*. Trope et al. (1997) argue that "[a] hypothesis provides a mental model in terms of which evidence may be construed and retrieved" (p. 115) and that "there is no normative model stipulating 'appropriate' error costs and the thresholds associated with them" (p. 125). A personal level mental model is one that can be an epistemic agent model, e.g. "I, as a reasoner, am now directing my attention at certain pieces of (structured) evidence."

In virtue of the fact that the presented personal level error minimization account is not an account of the way humans usually reason (the epistemic agent models that they construct), this account can be seen as a hypothetical construct useful for the explanation of behavior. Later developments in the area of cognitive modelling indicate that predictive coding – a subpersonal error minimization account - might offer a more fruitful alternative explanation (see section 4.4).

4.3 Types of Bayesian explanations and their implications

Self-deception results from two types of selection processes: world/self-model selection and epistemic agent model selection (section 2.2.1). I argued in the previous section that the FTL model does not account for the epistemic agent model. In this section I will elaborate what different explanatory scopes Bayesian explanations might have. The clarification of this matter is helpful for explanations of self-deception in order to avoid the danger of overrationalization, namely, assuming the presence of certain epistemic agent models where there are none (see section 3.1.2.2). The classification, I am to present, applies not only to Bayesian explanation, but also to explanations in terms of *prediction errors* and *inferences*. As I have noted at the end of section 2.2.2.3, 'inference' can be used to denote personal (= epistemic agent model construction), as well as subpersonal processes. In the previous section I argued that FTL model describes a *hypothetical* kind of an epistemic agent model. Here, I will make a more nuanced distinction, for the reader not to confuse the explanatory scope of the terms 'Bayesian updating', 'inference,' 'prediction error' in later sections, as I describe predictive coding.

Apart from overrationalization, the second aim of this section is to start building up a model of self-deception by answering a meta-question of what is actually *suboptimal* about self-deception. In previous sections I introduced several different models of self-deception. Let us assume that model of self-deception contains representations (structures that store certain kinds of information; see Pitt, 2012), over which certain operations (processes) are performed such that a new representation, that influences behavior, arises, called 'self-deception' about a certain topic. Here the process of influencing one's own states is also called 'self-deception.' The representations can be of several types – beliefs, goal representations, hopes, wishes, fears. Each of the models, starting from a model of how non-biased (neutral) representations are acquired, introduces a certain way in which the aforementioned model can be *skewed* - biased in a certain way to acquire a desirable result, which is the self-deceptive representation (Lockie, 2003, p. 136). The mentioned non-biased acquisition is illustrated by the scheme when the agent gathers evidence and evaluates it in a certain manner that is devoid of influence by affective states (goal representations, hopes, wishes, fears etc.). I want to argue that one can analyze self-deception by using the predictive coding model which is a model of how Bayesian inference is performed. In the following section I will introduce several tools that this model offers for analyzing self-deception, but in this section I will already give a rough idea of what Bayesian inference as predictive coding minimization is. For this, I will first introduce each of the terms: Bayesian, prediction error, inference – and then ask: if that kind of inference were to produce optimal results (representations), what can go wrong to produce self-deception? It should be noted that the assumption that there *is* something that needs to go wrong, or in other words, that self-deception needs to involve suboptimality somewhere in the representation acquisition and/or maintenance process, is also in need of an argument. To recapitulate, there are four central themes in this section: first, to introduce Bayesian inference as predictive coding minimization; second, to point out several notions used in this model that can have a personal and a subpersonal usage; third, to discuss the optimality question (I posed it in section 1.1.2.3 when discussing Mele's account); fourth, to review how several ways in which suboptimality has been introduced in the previous models of self-deception can fit into the predictive coding picture.

Let me start with explaining what Bayesian inference as predictive coding minimization means. *Bayesian* updating is updating according to the Bayes' rule (posterior = prior*likelihood). For example, if you complete a diagnostic test about whether you have or do not have a certain disease, your doctor would (hopefully) compute the results according to the Bayes' rule: take into account the prior probability of how widespread the disease is in the population, as well as you results on the test such that if the disease is very rare, then, even if the probability is high according to the test, the posterior probability is still not as high. Prior probability is often neglected though, which is called the base rate fallacy (Tversky & Kahneman, 1993). *Inference* in the context of predictive coding is the optimization of probabilistic representations on the causes of sensory input (Friston, 2010), e.g. inferring that there is a red apple on the table that caused the respective sensory (in this case visual) input that one perceived. Bayesian inference is, thus, optimization of representations that takes into account the prior probability and the likelihood when considering the question on whether there is a red apple that one has seen. (What is the probability of seeing a red apple in a flat of a person that favors healthy food, contrary to a person preferring junk food? What appears like an apple might be a caramelized kind of sweets. Given the sensory input, how likely is it to see the red apple? One could touch and see whether the apple is sticky or not.) It is important to note that in this paragraph I have described Bayesian inference as a higher-order cognitive process that an agent might become aware of, but this is not a standard case: it is not the case that it is an *agent* that performs the inference, but the brain, and it is often not the case that those inferences[339] are, or even can become, part of the phenomenal model of the agent's cognitive processes – an agent just sees an apple, a thought just appears out of nowhere.

The kind of Bayesian inference described above can be implemented as *prediction error* minimization, that is, as recurrent message-passing between predictions about the causes of sensations and prediction errors (Friston, 2009). On the lowest level, predictions are compared to sensations and the (unexplained) error (difference) between those is then

[339] Which entities/units inferences are to be ascribed, has to be clarified. Davies & Egan (2013) speak about personal and subpersonal Bayesian *inferences* in their consideration of explanations for the occurrence of delusions. Moreover, the authors speak of "Bayesian inference in a perceptual module," or with other words that it is the module that assigns probabilities, which implies that modules are entities to which drawing of inferences can be ascribed. According to Friston (2005) though, inferences are the estimates of the parameters for a generative model, so every time the model is updated, an inference is drawn. A generative model describes the causal structure of a series of inputs and is fine-grained to a certain degree (Clark, 2013b). To capture the complexity of the causal structure, the generative models is usually assumed to consist of several hierarchical *levels* of abstraction such that each level tries to predict the activity at the level below (Clark, 2013b). The criterion according to which hierarchical levels in the model are distinguished (degree of abstraction) can be interpreted in slightly different ways. For example, Kilner et al. (2009) make a *content*-related distinction of levels in their PC account of the mirror neuron system: kimenatics, goals, intentions and context. This causal background for this distinction is that our movements depent on short-term goals, which depend on long-term intentions, which depend on the current experiential context. Their model of a PC account of higher-order cognition is criticized by Blokpoel, Kwisthout, van Rooij (2012) for introducing complex causal structures on sufficiently high levels that cannot be estimated in Bayesian fashion anymore (see 2.2.2.3 for more). Another way to characterize the kind of abstraction that hierarchical levels provide is in terms of *modalities*: lower levels are modality-specific, while higher levels – multimodal, as well as amodal and as such also include interoceptive predictions (Friston (2013a).

compared to predictions on the next level and so forth. To return to the apple example, there might be two competing hypotheses – apple or sweets. Our senses let us perceive the apple in a certain way, e.g. that the object has such and such form and boundaries (that we distinguish the apple – figure – from the background) may be an inference drawn at some intermediate level between perceiving red dots on the background and recognizing an apple as an apple. To draw the analogy to the Bayes' rule above, predictions correspond to prior probabilities and the prediction error from the lower level can be compared to the likelihood term of the Bayes' rule (Friston & Stephan, 2007, p. 442). I would like to stress that also here, prediction error is not the personal kind of error that we as agents may recognize and act to minimize, but our perception, thoughts and actions are the result of such (hierarchical) prediction error minimization, effortlessly performed by the brain.

At this point, I am going to present an example of the personal usage of the term 'prediction error' that is useful not only for discrimination purposes, but also because this example concerns metacognition, which is relevant for self-deception. To remind the reader, I argued that tension (feeling of uneasiness in self-deception) is a kind of metacognitive feeling (see section 2.1.3). The example will concern *comparator* accounts of metacognition. Proust (2013), for example, on whose theory I have based my argumentation regarding tension, assumed that feelings arise as a result of the working of comparators (2.1.3). This is a feature she shares with Shea's et al. (2014) account of metacognition. Shea (2013) argues with respect to the domain of decision-making that even if the *process* is unconscious, given that the *decision* itself is conscious, taking it and the neural evidence into account may help to distinguish between different subpersonal models of decision making. Shea et al. (2014) propose that *metacognition*, and among it feelings of uncertainty and doubt, can be captured by a dual system cognitive control theory. Metacognition is according to Shea et al. (2014) a form of cognitive control (p. 187) which can be intra-personal (by system 1 not requiring working memory) and supra-personal (by system 2 that requires working memory; p. 186). Supra-personal cognitive control requires system 2, because automatic learning of metacognitive information is deemed impossible in case of interactions between multiple agents (p. 188), so that a "space of shared metacognitive information" (p. 189) needs to be constructed. This account of metacognition, as that of Proust (2015), is based on dual processing theory and one of the assumptions of dual processing theory is that one kind of processes *controls* the other, if this other process signals, e.g. via certain kinds of feelings, that something has gone wrong (see section 2.2.2.2, also for the distinction between dual system and dual processing theories that I do not discuss here). In congruence, Shea et al. (2014) hold that metacognitive representations incorporate prediction errors generated by system 1; if they are large, they produce certain metacognitive feelings, whose significance system 2 then learns (p. 190).

Such comparator models are on a higher level of abstraction than predictive coding and furthermore they are restrictingly dichotomous (supposing two kind of levels). I agree with Hohwy (2014) that predictive coding is to be guided "away from simple dichotomies" (p. 8). Apart from that, it is questionable whether the description of the kind of control applies to self-deception. Self-deception might involve the *feeling* of control, but not personal control per se, or just subpersonal control without the feeling of agency, given that it stems from subpersonal selection processes. Thus, how one might have the feeling of metacognitive control without the actual control itself has to be accounted for too (see next section for how predictive coding accounts answer this question).

After introducing Bayesian inference as prediction error minimization, I want to emphasize that there are 4 different ways in which one could explain generation of certain epistemic agent models (phenomenal models of our thought processes) by means of Bayes or predictive coding (PC) and that it is important to distinguish them in order to avoid conceptual confusion about the explanatory scope of the given explanation:

1. Conscious reasoning process (= certain kind of an epistemic agent model) occurs in Bayesian way. *Scope*: phenomenology
2. Conscious reasoning process can be explained in terms of a Bayesian model in an "as if" sense, in distinction from the "'genuine approximation' to a true Bayesian scheme" (Clark, 2013b, p. 201). *Scope*: explanation of data (e.g. behavior and autophenomenological reports of subjects) by the model
3. The subpersonal processes that give rise to the conscious reasoning process are implemented by a certain Bayesian model. *Scope*: explanation of subpersonal processes in terms of *any* implementation of a Bayesian model
4. The subpersonal processes that give rise to the conscious reasoning process are implemented by PC. *Scope:* explanation of subpersonal processes in terms of PC.[340]

These distinctions are important in order to answer the question about the possibilities of incorporating inoptimality into predictive coding. As I did for prediction error, let me illucidate to the reader what a personal level use of 'Bayesian reasoning' would be. A personal and subpersonal Bayesian explanation of human reasoning is not to be confused (see previous section for examples of the first kind). There have been attempts to explain human reasoning in Bayesian terms (e.g., Talbott, 1995; Park & Sloman, 2014), as well as evidence of failure in probabilistic reasoning (e.g., Tversky & Kahneman, 1993; Bortolotti & Broome, 2008), yet care has not always been taken to distinguish among the four possibilities above. Failure in probabilistic reasoning (e.g., the base rate fallacy I talked earlier) can serve as a reason to dismiss a *personal* level Bayesian account, but not a subpersonal one (that the brain computes the information in a certain manner). As a contrast, consider Knill & Pouget (2004) who argue that "[t]he real test of the Bayesian coding hypothesis is in whether the *neural computations* that result in perceptual judgments or motor behavior take into account the uncertainty in the information available at each stage of processing" (p. 713; my emphasis). Also Yoshida et al. (2010) who have attempted to "establish the neuronal correlates of our [their] Bayes optimal model of cooperative play" (p. 10746), note that their model excludes the phenomenology and hold that their model is "purely mechanistic in describing how brain states can encode theoretical quantities necessary to optimize behavior" (p. 10750). The predictive coding approach is primarily a subpersonal level application of the Bayesian theorem, where the latter is employed at different levels of the hierarchy[341] (Frith & Friston, 2013)**.**

If one were to analyze self-deception by means of predictive coding (which I will do in the last section), then one has to answer the question how a predictive coding model can incorporate model breaking (on the assumption that self-deception involves such kind of breaking), or how a predictive coding model can explain phenomena that, at least, *seem* not optimal from the observer point of view (on the assumption that self-deception does not involve any kind of breaking).

Given the distinction between representations and operations over representations as parts of the model, agents act on the basis of the model of the environment that they possess.

[340] For more on the relationship between empirical Bayes, of which hierarchical predictive coding is an implementation, see Friston (2003, 2005).

[341] Predictive coding as a prima facie subpersonal theory: "All this happens at the sub-personal level. That is to say, we are not consciously aware of prior expectations, prediction errors, or updating except, perhaps, when the prediction error is large. In terms of neuronal representations, precision can be thought of as amplifying prediction errors associated with a high degree of certainty or reliability" (Frith & Friston, 2013, p. 5).

Break the inference (an operation over representations central to predictive coding) would mean that there would be exceptions to the rules according to which inference is usually computed, e.g. exceptions to the Bayes' rule). In the 3rd chapter (3) I cited different psychological views on self-deception each of which supposed different kinds of *biases* underlying it. Those would be also a kind of breaking from the perspective of unbiased information evaluation. These two cases – breaking the Bayes' rule and breaking unbiased information evaluation differ in an important respect: the function that those operations compute. In the first case it is optimal calculation of probabilities in case of restricted or uncertain information,[342] in the second some other normative function, e.g. that of rationality, may be assumed. In analogy, regarding representations, breaking could be defined as the possession of representations so that, when operations are applied to them, the resulting representation, e.g. a self-deceptive attitude, does not conform to the function that the operation is thought to compute. In section 1.1.2.3, I reviewed two distinctions: first, whether optimality concerns the *model* or *behavior* (Bowers & Davis, 2012, p. 398) and, second, that *inference* might be optimal, while the *model* is not, because the model might include suboptimal model parameters, e.g. aberrant subpersonal beliefs (Schwartenbeck et al., 2015, p. 110). Schwartenbeck et al. (2015) do not incorporate the operation – inference – into being a part of the model, but, apart from that, in this paragraph I just reinstated the distinctions made in the literature regarding the objects of optimality: models (representations and their connections), operations (Bayesian inference), behavior (determined by the resulting model after operation computation). Given the personal/subpersonal distinction though, an additional aspect is that subpersonal computations might be optimal, while personal ones are not (and vice versa). While on the personal level we as observers apply rationality in evaluating the behavior and beliefs of others, on the subpersonal level those behavior and beliefs might either result from a Bayes optimal computation or not. Though it is plausible to assume that a Bayes-optimal subpersonal computation would lead to a belief that we would categorize as 'rational,' it need not be the case. As when we look at a broken clock at the "right" time for it to show us the correct time, accidentally a suboptimal computation may also result in a rationally justifiable belief. This is to say that the personal/subpersonal distinction adds an additional degree of distinction to operations as objects of optimality constraints.

For self-deception specifically, in previous sections I reviewed accounts that stated that while self-deception is not optimal with respect to one function (theoretical rationality), it is optimal with respect to another (practical rationality). Talbott, for example (see section 1.1.1.6), argued that self-deceivers are Bayesian utility maximizers. They are Bayesian, because they compute probability in an optimal way, but the computations of probability have been changed so that representations possess not only *probabilities*, but also *desirabilities*. Also Van Leeuwen (section 3.2.3.1) argues that self-deceivers are *practically* rational. So, whether self-deceptive model/inference/behavior is optimal depends on whether one has *rational* optimality or *utility* optimality in mind. At the same time, if one assumes a divergence between theoretical and practical rationality in case of self-deception (1); argues that self-deception is practically, but not theoretically rational (2), as well as if one equates theoretic rationality with Bayes-optimality (3), then it would follow that self-deception (as a process) is not Bayes-optimal. For example, McKay & Efferson (2010) argue that *interestingly cognitively biased* individuals do not "have beliefs that are *theoretically* optimal given the available information" (p. 310; my emphasis) and that, thus,

[342] Bayesian beliefs are taken to be *grounded*, meaning that they are "appropriately founded on evidence and existing beliefs" (McKay & Dennett, 2009, p. 494).

they depart from Bayes-optimality. McKay & Efferson (2010) differentiate *utility theory* from *error-management theory* insofar as the former states that selection will favor *behavioral biases* that reduce error costs,[343] while the latter includes another claim that those behavioral biases are a result of an interestingly cognitive bias (p. 310). This is to say that for a cognitively interesting bias it is not sufficient to stem from error reduction that leads to utility maximization. The process by which that bias has been acquired has to be suboptimal.

Given the personal/subpersonal distinction though, the two could deviate – there could be personal errors (misrepresentations) and subpersonal non-errors, which one might argue is exactly the case in self-deception. One may assume that in self-deception there is a combination of personal level errors and subpersonal non-errors because personal level errors explain why observers do not agree with self-deceivers.[344] Further, in this thesis I will stay within the bounds of predictive coding so that I will not explore how one could *break* the inference, but only how one could change the parameters of the model (which are certain kinds of representations present in the model) so that self-deception could result. I set this constraint because I think that since predictive coding is a fairly new approach, one should first explore all the possibilities it offers, before going beyond.

In the following, I will review how representations and operations, used to describe self-deception on the personal level, have been explained on the subpersonal level by predictive coding: priors and goal-selectivity, as well as speed-accuracy and accuracy-complexity trade-offs.

1. *Priors*: Priors or, respectively, prediction representations, that I mentioned when I introduced Bayesian inference as prediction error minimization, determine the influence of the information that an agent gathered prior to the task at hand. McKay & Dennett (2009) have described Bayesian updating as a trade-off between accommodating new evidence and valuing the one previously acquired.[345] This is important because self-deception can also be understood as a trade-off between accepting new evidence on the one hand, e.g. the one presented by observers in arguments, and nurturing one's acquired self-deceptive attitudes on the other (see section 3.2.2.2). To remind the reader, also schemata and templates have been argued to take part in explaining self-deception (see chapters 1, 2 and 3). For example, one possible explanation for self-deceiver's conservatism in accepting new evidence that one could give is the presence of a schemata. Priors have been equated with templates, e.g memory templates (Clos et al., 2014) and as such provide a predictive coding explanation for the working of schemata.[346] All in all, it is to remember that the role ascribed to priors in personal and subpersonal processes is identical – incorporating previous information into the

[343] McKay & Efferson (2010) argue that there are three beliefs that influence the presence of behavioral biases in such a way that at least some of the beliefs are not-Bayes optimal, but the behavior is consistent with the utility theory (p. 313, 321):

- The belief that the current decision involves asymmetric costs;
- The belief about the specific magnitude of the asymmetric costs;
- The belief that the current state is the one in which a more costly error is possible.

[344] Why self-deceivers have the feeling to have known the truth all along might be seen as another reasons for the hypothesis that self-deceivers err on a personal level, but retrospective acknowledgement of error need not mean that the error was really there.

[345] "Bayes' therem is in a sense a prescription for navigating a course between excessive tendencies toward 'observational adequacy' (whereby new data is over-accomodated) and 'doxastic conservatism' (whereby existing beliefs are over-weighted)" (McKay & Dennett, 2009, p. 497). Anosognosia is at the doxastic conservatism end.

[346] There have been also attempts to relate *priming* (another biasing subpersonal psychological mechanism) and predictive coding. Predictive coding has been namely argued to be conceptually related to priming of any kinds Clos et al. (2014, p. 62), given that the Bayes' rule includes priors.

evaluation of the new one, but the way in which it is done on the subpersonal level in predictive coding is more complicated (see next section for more details).

2. *Goal representation congruency*: Self-deception is motivated by goal representations (see section 2.2.1) and optimism bias is explained in predictive coding via the influence of goal-representations. An additional benefit of this paragraph is that, since self-deception has been argued to share a lot of conceptual similarities with delusions and optimism bias (see section 1.2.7), it will not only give a foretaste on how goal representations could be argued to influence perception (and cognition) on the subpersonal level in predictive coding (for more detail see following section) in general, but also how goal representations influence perception in a specific case of optimism bias. Furthermore, in the light of the optimality discussion above, it is useful to note that optimism bias is incorporated into predictive coding as a bias toward the realization of one's goal that does *not* violate its principles (Friston et al., 2013). Optimism is argued to arise because of the mutual dependency between *transition* probabilities of control states (= beliefs about future action that stand for goal representations) and hidden states (= representations that model the assumed states of the world). I will talk more about transition probabilities, control states and hidden states in the next section. Here it suffices to say that a model of the causes of our sensations includes different kinds of states, among them hidden states and, sometimes, control states. There might be complex goal representations that need a sequence of other goal representations to be fulfilled first. Those simple goal representations are connected by transition probabilities to reach from one desired state to another and, importantly, there is a circular dependency so that not only what is perceived determines the goal representations, but also the other way around. What this means is that goal representations determine also those of our cognitive processes that we would judge as unbiased on the personal level. This general notice does not solve the prominent selectivity problem for self-deception though: why do goal representations other than those directed at finding out the truth exhibit such a strong influence in case of self-deception? So, the incorporation of goal representations into predictive coding is only one piece of a puzzle in analyzing self-deception.
3. *Speed-accuracy trade-off in case of uncertainty*: Priors and goal representations are not the only kinds of representations that can determine the decisions that agents make. Each of the parameters of Bayesian updating possesses *precision*, which is the estimation of the uncertainty so that prior*evidence = posterior is precision-weighted (Hohwy, 2013a). The precision of a prior can be compared to the learning rate for that prior, which is the *extent* to which past history influences the current case (Daunizeau et al., 2010b). As for the loss function, it may be needed in case that there is some sort of a conflict, for example when one has to make a quick decision given uncertain information. It is the case when one has to make a quick perceptual decision about whether what one sees is a picture of a face or a house and there is an additional signal (sound presented before the picture) that, with varying probability, indicates what it is to be seen (Daunizeau et al., 2010b). If one knew the extent to which the auditory signal is reliable in indicating the perceptual category, then one would knew to which extent one should incorporate the prior information. The time factor is the additional constraint. One can argue that decisions under *uncertainty* are Bayes-optimal (Daunizeau et al., 2010a,b). As a consequence, forgetting effects (how quick prior information is discounted and, hence, forgotten) are also Bayes-optimal, because they depend on the learning rate (Daunizeau et al., 2010b, p. 18). Daunizeau et al. (2010b) further hold that given the choice of the loss-function as reflecting a speed-accuracy conflict, "*categorization errors are optimal decisions* if the risk of committing an error quickly is smaller than responding correctly after a longer period" (p. 6; my emphasis). As for the role of the speed-accuracy trade-off for self-deception, it is to mention that a similar asymmetry in the cost of errors has been argued to be a condition for a successful misrepresentation by Zehetleitner & Schönbrodt (2014) (see section 3.2.2.1 in Trivers' evolutionary theory) and that self-deception has been argued to occur in cases when the evidence is ambiguous (pain endurance paradigm, see section 1.3.2.2). One option for the definition of conditions under which a misrepresentation is successful is the asymmetry of error costs in cases of ambiguity.

Ambiguity has been independently claimed to be a requirement for self-deception in one of the experiments using the pain endurance paradigm (see section 1.3.2). This is relevant for the question whether self-deception is Bayes-optimal that I discussed above, because if the speed-accuracy trade-off plays a role in self-deception and this trade-off is Bayes-optimal, then to this degree self-deception is also Bayes-optimal. Certainly, the fact that predictive coding can accommodate speed-accuracy trade-offs and learning rates, is not an argument for why it should be a preferred analysis framework in case of self-deception, but still, if several tools for analysis can be incorporated into predictive coding, then it offers a richer general basis for analysis.

4. *Accuracy-complexity trade-off over time*: The link between the speed-accuracy trade-off and self-deception depends mainly on one's analysis of the possibility for misrepresentations, although it is not clear why self-deceivers are supposed to always be in a time pressure when making decisions, which is to say that the preference function in their case has to be more complex. The link between self-deception and the accuracy-complexity trade-off may be more subtle. I will construct it to show how one more trade-off, that can be incorporated into predictive coding, can at the same time be applied to analyze self-deception. Optimism bias, understood here as the diminished incorporation of new information in elderly people, has been claimed to be a case of self-deception by (von Hippel & Trivers, 2011b). Moran et al. (2014) argue that with age the *complexity* of the Bayesian model (of the causes of sensory input) decreases over time. Here, the idea is the following: to infer the causes of our sensations, a simpler or more complex model (with a higher number of parameters) can be constructed. Complexity can be roughly understood as the number of parameters needed to model those causes. The accuracy of the model (how good the model predicts the data) is not the only quantity that is minimized during prediction error minimization, but it is actually accuracy minus complexity[347] (for the exact definitions of accuracy and complexity see Friston, 2010). This is to say that complexity is a penalty term. It is a penalty, because a model should be generalizable to different pieces of data, which are not necessarily consistent with each other. Imagine that you have a pool of different data instances and a model can predict every that instance perfectly. When it encounters a new pieces of data and cannot predict it, the necessity of a perfect prediction would enforce the change of the model so new parameters are introduced and a model is made more complex. Thus, according to Moran et al. (2014), Occam's razor, which occurs here in the need to avoid overfitting, leads to attenuated learning with age. This means that complexity reduction leads to the reduction of *short-term* Bayesian updating (e.g., every time encountering an instance a model cannot predict, change the model) and a shift to enhanced top-down processing (Moran et al., 2014, p. 6). The authors voice an optimistic conclusion that "as we age, we converge on an accurate and parsimonious model of our particular world […] whose constancy we actively strive to maintain" (Moran et al., 2014, p. 1). After this introduction into the terminology, let me consider the application of the trade-off to self-deception. Von Hippel & Trivers' (2011b) claim that optimism bias *increases* with age[348] and Moran's et al. (2014) argumentation lead to the conclusion that complexity reduction is to be interpreted as self-deception. On the one

[347] The reason for this is that the quantity that is to be minimized is actually free energy, which can be approximated by prediction error (see next section for free energy).

[348] One could explain age-related optimism bias in two ways: as *goal*-related optimism bias such that elderly people accomplish something by it (case of selectivity) or as *neutral* optimism bias such that elderly people simply do not care as much about their environment at all. To remind the reader, Scott-Kakures' cognitive dissonance account (section 3.1.1) argues that self-deception in general (and not only age-related optimism bias) is characterized by the striving to find out the truth, instead of pursuing other goal representations. Yet in that account, there were still certain kinds of expectations that self-deceivers wanted to fulfill. The question is now whether a fully neutral optimism bias could be a case of self-deception and my answer would be: probably not. This is because the phenomenological and behavioral characteristics, e.g. defend one's position in an argument with other people, would not be (or would be very weakly) fulfilled.

hand, it is plausible since complexity reduction can result in accuracy reduction. On the other hand, self-deception is not restricted to occur in elderly people. While elderly people can be assumed to have accumulated enough evidence, in order to start reducing the complexity, why young self-deceivers perform such complexity reduction has to be explained. This is where experts come into play. Interestingly, complexity reduction is argued to be the case not only in elderly people, but also in *experts*, e.g. trained ballerinas (Moran et al., 2014, p. 6). So, not age is the deciding factor, but training (which means the sufficient amount of data instances an agent is presented with). If self-deceivers were to be described by means of an accuracy-complexity trade-off, they would be described as experts that reduce complexity in a specific way that does not produces an accurate model. A tension between goal representations of discovering the truth and other-directed goal representations has been argued to exist in self-deception (see section 2.1.3 and 2.2.1): those two stand in conflict. If the balance between accuracy and complexity, which also stand in conflict to the degree that reduction of complexity reduces accuracy, were to depend on the respective balance between the two kinds of goal representations (discovering the truth for accuracy and other-directed for complexity), then other-directed goal representations might be argued to determine in which way complexity is reduced that results in self-deceivers being experts with inaccurate models.

So, depending on which of these cases (priors, goal congruency and the two trade-offs) one accepts to represent self-deception, one could “recycle” the explanations, already given for them in the predictive coding paradigm, also for self-deception. Such an explanation would be explanatory cheap and would not be able to explain all the behavioral and phenomenological characteristic satisfactorily, but for completeness reasons this possibility was useful to mention.

Before I review other predictive coding tools that might be helpful in the next section, I want to point to another complication regarding the personal level, namely, the question of how the model of an agent who *observes* self-deception actually looks like. Self-deception is, at least in some case, recognized by others and requires justification. In these cases there is a social interaction between the self-deceiver and the observer. The same is the case in a psychological study: the experimenter observes the self-deceiver and tries to reconstruct how the self-deceiver’s model would look like.

Daunizeau et al. (2010a,b) propose an approach to explain decision-making under uncertainty in terms of predictive coding. Since decision-making and belief-forming processes both are goal-directed processes of information acquisition and manipulation, I will shortly sketch Daunizeau’s et al. approach. It consists in taking the stance of the observer who explains in a Bayesian fashion the behavior of the participants on the assumption that their decision-making stems from Bayesian processes:

> Ideal Bayesian observer assumptions map experimental stimuli to perceptual representations, under a perceptual model; $m^{(p)} : x \rightarrow^{\vartheta} u$; while representations determine subject responses, under a response model; $m^{(r)} : \lambda \rightarrow^{\theta} y$. The central idea of our approach is to make inversion of the perceptual model (i.e. recognition: $u \rightarrow^{\vartheta} \lambda$) part of the response model. This provides a complete mapping $m^{(r)}: u \rightarrow^{\vartheta} \lambda \rightarrow^{\theta} y$ from experimental stimuli to observed responses.[349] (Daunizeau et al., 2010a, p. 7)

In other words, subjects construct posteriors or subjective representations of causes via perceptual inference and decide upon actions on the basis of those representations, loss-

[349] u – sensory inputs, x – hidden causes, ϑ and θ – perceptual parameter or prior precision (p. 7), λ encodes the approximate posterior density on the hidden state (Daunizeau et al. 2010a, p. 4). Interestingly, *representation* as posterior by arguing that “the posterior represents information about the hidden states given some sensory inputs” (Daunizeau et al., 2010a, p. 3).

and utility functions (Daunizeau et al., 2010b, p. 1). The perceptual model provides subjects with the causes of sensory input, while the response model predicts the decision of the subject for the observer on the basis of *loss- and utility functions*. The emphasis is thus that subjects construct the inverted perceptual model, while observers – a response model that includes the inverted perceptual model (see next section for more detail).[350] Daunizeau et al. (2010b) have tested their approach on a speeded discrimination task (participants had to decide whether they see a house or a face, each being presented after one of two sounds that indicate with different probability what the participants will see (p. 2).

The model inversion described above is interesting for two reasons: first, if ever a predictive coding experiments were made regarding self-deception, instead of a speeded discrimination task, then the question would be how one would invert the model there? Will a loss- and utility function be enough? This question depends on the design of the experiment that one would then have, so I will leave it open for the future experimenter to determine. There is a related question though, namely whether self-deceivers possess personal level errors. As I mentioned in the optimality-discussion, a reason one might give for why self-deceivers err on the personal level is that given the same environment that self-deceivers and observers are in, both come to possess different models. On the one hand, it is not surprising at all, because since it is unlikely that for two different individuals the environment will be exactly the same. On the other hand, regarding the specific topic that self-deceivers are self-deceiving on, two contradictory models require an explanation. An easiest way would be to say that the loss- and utility functions in self-deceivers and observers differ so that nobody errs and observers think that self-deceivers err merely because of the difference in those functions. Yet, two contradictory models cannot represent the causal structure of the environment equally well, because the latter is utility-independent, so self-deceivers possess personal level errors.

Second, though the experiment above concerned a speeded *discrimination* task, which suggests that there would possibly be an epistemic agent model involved, a *perceptual* model was constructed for the participants. Is there a need for an additional cognitive model to operate on top of the perceptual one in such a manner that it takes the perceptual model to provide the context? Or is higher-order cognition in some sense *embedded* into the construction of the perceptual model such that no clear-cut distinction between the two can be made? I think that so far it seems that predictive coding would embrace the second alternative, hence, no speaking of additional cognitive models. This choice may stem from the equivocal use of the term 'belief' for the personal, as well as subpersonal level of description. This has insofar a consequence for an explanation of self-deception, as I argued that two kinds of selection might bring it about – a world/self-model selection, as well as an epistemic agent model selection. If there is no additional cognitive model, then both selection processes would be modeled by a *perceptual* model. Then one would need to ask the question whether it would be the *same* perceptual model and if yes, why then the resulting attitude would be *transparent* in one case and in the other would not. I will leave this question open until section 4.5.

Interim conclusion: So far I have introduced four possibilities for epistemic agent models (phenomenal models of higher-order cognitive processes) to be explained by Bayesian

[350] An observer can also manipulate the perceptual inference of the subject: "Finally, it is worth highlighting the importance of experimental design for identifying (Bayesian decision theoretic) models. This is because perceptual inference results from interplay between the history of inputs and the subject's priors. This means that an experimenter can manipulate the *belief of the observer (e.g., creating illusions or placebo effects)* and ensure the model parameters can be quantified efficiently" (Daunizeau et al., 2010a, p. 9; my emphasis).

rules. Predictive coding falls under option four, because it is neither “as if”, nor implementation of *any* Bayesian scheme, but empirical Bayes. I exemplified in which way the terms ‘inference’ and ‘prediction error’ are also ambiguous regarding the personal/subpersonal level distinction. I then illucidated different ways (behavior, model, inference) by which Bayesian-optimality might be violated and in accordance to the literature stayed with the second option. Further, I clarified which parameters of the model might lead to a suboptimal self-deceptive model and, lastly, I noted that the difference between self-deceiver’s and observer’s models suggests that the former possess personal level errors.

4.4 Conceptual tools extracted from the predictive coding approach

> This assumption is false: using probabilistic models to provide a *computational-level* explanation does not require that hypothesis spaces or probability distributions be explicitly represented by the underlying[351] *psychological or neural processes*, or that people learn and reason by explicitly using Bayes’ rule.
> (Griffiths et al., 2010, p. 362; my emphasis)

So far in this chapter I criticized connectionist models of self-deception, as well as personal level error minimization accounts. The first ones possess only a very limited amount of conceptual tools that one could use, while the second do not do any justice to the phenomenology of our higher-order cognitive processes (e.g., see Metzinger, 2015 for the claim that our 'conscious thought' is subpersonal). Thereafter, I elaborated on different kinds of Bayesian explanations that one can construct. The aim was to show, first, how far personal and subpersonal level Bayesian accounts differ (phenomenology belongs only to the scope of the first kind) and, second, that one could implement them in several ways, of which predictive coding is, as an implementation of empirical Bayes, only one such possibility. I am going to present in this section its conceptual tools, that predictive coding offers, with the aim to apply them in the following section.

According to predictive coding theory, a hierarchical model of the causes of our sensory input is constructed by level-wise comparison of precision-weighted predictions (of those causes[352]) with precision-weighted prediction errors. Precision is, informally, uncertainty about the predictions or prediction errors. Sufficiently large errors are propagated upwards and might change the model, but the influence of those errors is modulated by (subpersonal) confidence in them. To remind the reader, several terms, such as prediction error, inference, confidence can be used personally and subpersonally (see previous section). In predictive coding, the primary usage is a subpersonal one so that an argument is required to conclude that subpersonal confidence is related to the personal one. I will come to this point later in this section. As an alternative to the update of the model, the environment may be further sampled in the search for further evidence for or against the model. To give a (more sophisticated) example,[353] a self-deceiver, upon facing certain kinds of evidence contradicting the hypothesis that she possesses at the moment, either changes the hypothesis or sticks to it and samples other kinds of (subpersonal) evidence. If the self-deceiver is uncertain about the implications of the given piece of evidence (e.g., whether it contradicts the hypothesis or not), then no change or sampling may be needed. In this section, I will go into details and present the tools, that predictive coding offers, for explaining perception, action and cognition, in order to apply them later to self-deception.

[351] As far as I understand, underlying means an identity relation here.

[352] On the first level, the causes are the events on the sensory sheet.

[353] This example is one of personal level hypothesis construction, not of a subpersonal one.

In short, the dependencies between tools, that are to be presented, are as follows: there is a certain quantity that agents (as self-organizing systems) minimize, namely *free energy*. States with low free energy are *attractors* and agents visit a *limited* set of them. Agents possess a *generative model* of the environment that consists of several layers of *precision-weighted prediction and error units* and that models the *hidden and causal states*. How the model looks like, depends on the free energy, because it is constructed such that it minimizes free energy. This model can be *updated* or agents can *act* on it. In particular, the generative model allows also for goal-directed (*model-based*) action in virtue of *policies* (*counterfactual* representations of future actions). Such a generative model is not only a neurobiological and functional model, but it can also account for the *phenomenology* of the agent, e.g. *interoceptive inference* has been proposed to model affective states, and for the interpersonal relations, e.g. *interpersonal inference* has been proposed to explain how agents acquire personality traits (that as priors become parts of the model). *Precision* accounts for context-sensitive action and has been used to explain delusions. Due to conceptual similarities between self-deception and delusion, I will at the end of the section in detail present different accounts, explaining how precision can be applied to explain delusions.

Free energy: In the previous sections of this thesis I have spoken about how one might describe a model of the environment that an agent constructs. At this point, before talking more about the (generative) model by which agents estimate the causes of their sensations and how higher-order cognition (and self-deception) might fit into that picture, I will answer the question of why there is a model estimating the environment in the first place. Biological self-organizing system resist the tendency of disorder, which is to say that they preserve their inner structure and states in the face of changing environment (Friston, 2010), e.g. maintain constant blood pressure. Free energy is an information theoretical quantity (Friston, 2009, p. 293) that bounds surprisal (Friston & Stephan, 2007) which is the entropy (measure of disorder) of external states (Friston, 2012a). The model of the causes of sensory data that agents construct is such that it minimizes free energy. In other words, a system expects to find itself in certain states (a restricted amount of them) and if it does not happen – experiences surprisal. Surprisal, which measures how probable an outcome is (e.g., agent being in a certain state such as a fish out of water), is not to be equated with personal level surprise (e.g., Clark, 2013b), although the two might coincide (Friston, 2010). Free energy is a measure of such surprise and relates the model of the environment a system constructs (hypothesized causes of sensory input) to the environment (actually sampled sensory inputs). Free-energy *minimization* is a principle that ensures that a biological system stays in its environmental niche, which is to say that it does not die because it explored a state it cannot survive in (Friston & Stephan, 2007). So, a system minimizes its free-energy to ensure that it finds itself in unsurprising states. For that, it has to actually estimate correctly what state it is in – construct a truthful model of the environment (causes of sensations, be it sensory or proprioceptive sensations).[354] While free energy (as an upper bound) is a proxy for surprise, prediction error is a proxy for free energy. This is because free-energy minimization, under certain assumptions (Friston & Kiebel, 2009), is equal to the suppression of hierarchical prediction error (Friston

[354] Prediction error minimization minimizes surprise by building the model of the environment: “This is a central part of the attraction of prediction error minimization. Minimizing prediction error (or free energy) does two things at the same time. First, it approximates surprise so that a given organism tends to sample the sensory evidence that defines its phenotype. Second, it obliges the phenotype to select plausible hypotheses about the causes of its sensory input” (Hohwy, 2013, p. 53).

& Stephan, 2007, p. 442). In short, the relationship between both procedures (free energy and prediction error minimization) is this: the system wants to minimize free energy, for which one has to solve a difficult inversion problem (to revert sensory input to causes), solution to which a particular implementation can achieve - predictive coding[355] (Friston, 2002, 2005). So, because *free energy* minimization is difficult, one minimizes something else instead – prediction error. Thus, one of the most appealing things about predictive coding is the possibility of the mathematical formalization in terms of free energy "that quantifies the amount of prediction error" (Friston, 2012b). This is to say that free energy minimization is primary to predictive coding minimization or that the second kind of minimization is an approximation of the first. Free energy minimization is, therefore, a principle that underlies predictive coding. I will go into more detail when I speak about the generative model and its structure below.

Since free-energy minimization is a principle commonly found in physics and biology (self-organizing systems minimize it), I will give a simpler example of how one might imagine such kind of minimization, before turning to the complex way the brain minimizes free energy. For example, one method for finding out the tertiary (3D) structure of proteins (in other words the positions in 3D of all the atom that a protein consists of) given their amino acid sequence (sequence of molecules) is finding the locations of atoms so that they minimize a certain function - global free energy (Klepeis & Floudas, 2003). In the case of proteins, interactions within the protein system, e.g. ionic, hydrogen, determine the free energy (Klepeis & Floudas, 2003). The aim is to find *one* set of atoms positions for the given protein that minimizes free energy. To account for all possible effects though, the determination of the global minimal free energy requires the calculation of an *ensemble* (probability distribution) of low energy states that the system might transition into (Klepeis & Floudas, 2003).

I made this short excurse into the tertiary structure determination of proteins because this is what humans supposedly also do, if they are free energy minimizers.[356] They also have such an ensemble of states that they visit, or in Friston & Stephan's (2007) words, "[b]iological systems act on the environment and sample it selectively to avoid phase-transitions that will irreversibly alter their structure" (p. 451). The generalizability of the free energy principle to all living organisms is not only a benefit for the predictive coding theory, but also it is in itself an explanatory tool, because it shapes the way we think of human beings and explain their actions: The extension of proteins in 3D (which is their tertiary structure) is governed by the amount of free energy they possess and, similarly, if predictive coding is true, the way humans use the 3D space[357] (how they act in what they experience is the outer world) also depends on the amount of free energy that the system possesses. Proteins prefer to visit states with low free energy, which is calculated by considering atom interactions. In case of humans, what one has to consider is much more complex, including, for example, goal representations that are also part of their model of the world. The models of proteins and humans differ, as well as the possible kinds of

[355] Difference between generative and inverse models: "Predictive, or more generally, generative, models turn the inverse problem on its head. Instead of trying to find functions of the inputs that predict the causes they find functions of causal estimates that predict the inputs" (Friston, 2002, p. 124).

[356] I will leave the comparison of free energy definitions for proteins and predictive coding out. Even if they are different, my general argumentative point still applies.

[357] For the sake of the analogy, I ignore the fourth temporal dimension here. On the other hand, since the extention of protein in space also changes with time, one could also add the fourth dimension to proteins in the analogy instead of taking one from the human space.

changes that minimize the free energy of the system, but the principle (free energy minimization) is the same. Let me now look at those visited states in more detail.

Set of attractors: As agents minimize free energy by visiting a limited set of states in order to avoid surprise, the second conceptual tool is that of a 'set of attractors' – which is the way to label those limited states. That systems, such as human beings, visit a *limited* amount of external states is true because they are *ergodic* (Friston, 2013b). Ergodicity describes the relation between the probability that a certain state is visited to the time that the system stays in this state: "intuitively, an ergodic system forgets its initial states, such that the probability a system is found in any state becomes—for almost every state—the proportion of time that state is occupied" (Friston, Breakspear & Deco, 2012, p. 2). A toy example: consider a dice that you throw at regular time points. The probability of getting each number (1-6), where those numbers are the states that a dice can visit, is 1/6. So at a particular point in time the probability of getting every number is 1/6. A way to verify this is to actually throw the dice many times and compute the average for every state – determine the time average. Due to ergodicity a self-organizing system will converge to an invariant set of global attractors – states that it visits (Friston, 2013b; Friston, Sengupta, & Auletta, 2014). Such a system will not fall prey to chaos and preserve its structure. A set of attractors, that a system visits, is a tool, because if it is unique for every system (which I think is possible on the assumption that there exist *phenotype*-dependent valuable states, see below), then differences between sets of attractors can be used to explain different properties of the system, e.g. in particular proneness to self-deception. The application of the limited set of attractors to self-deception might proceed along the following lines: one could analyze which parts of the generative model have which influence onto the determination of the set of attractors that self-deceivers possess. I will not explicitly speak about the set of attractors any further, but it is useful to keep in mind that any changes to the generative model determine the set of attractors of the agent such that, generally, the question why self-deceivers do not behave as observers expect might be answered by stating that their state of attractors deviates. This in itself is not very informative, so one needs to go on and ask which parts of the generative model are responsible for such a deviant set of attractors. So, let me explain what a generative model is and what it consists of, so that one knows which parts one could change in order to construe an explanation of self-deception.

Generative model: A generative model is a probabilistic model about the distribution and dependencies of the causes of our sensations. As mentioned above, the parameters of this model are fine-tuned in a way minimizing free energy. At this point it might be helpful to remind the reader about which kind of models I have talked before, in order to distinguish them from the generative model. When I spoke about the cognitive process of the self-deceiver, I discussed *phenomenal* models that she might possess (see section 2.2.2.3). Phenomenal kinds of models (world-, self- and epistemic agent models) describe the representational structure of our phenomenal experience. Generative models generate predictions of sensory input, for which the causal structure of our experience has to be captured. If it is phenomenally represented, both may be equivalent (Hobson, Hong, & Friston, 2014).

With respect to the to-be-predicted entity, one can distinguish two kinds of models: generative (forward) and inverse (recognition) models. The first ones, given the causes, approximate inputs and the second ones, given the inputs, approximate causes (Friston & Stephan, 2007, p. 444). This distinction resembles to a certain degree the top-down/bottom-up one: from above, a generative model generates sensory inputs that one would encounter if the causal structure were a certain way; from below, an inverse model computes those causes given the actual inputs. So, a generative model describes dependencies between causes and inputs in the form of likelihood of those inputs and prior probability (Friston, 2010). If the inversion of the generative model is possible, then no additional recognition model is needed, because one can get it via the inversion operation (Friston, 2003), else the agent has to approximate both the generative and the recognition model (in a way minimizing free energy). In the following, I will first describe the structure of the generative model and then discuss how it might relate to the phenomenology of the agent. I will not mention recognition models (counterparts of the generative model) for simplicity reasons.

Basic structure of the generative model: Which different kinds of states does the generative model need to represent the environment? If, for example, I wanted to write a simple program that would show a ball bouncing from a certain kind of surface realistically, I would need to know the *location* of the ball at every moment in time and how it *moves*, in other words, position and momentum (Friston & Stephan, 2007, p. 431). This is what *hidden states* are for – to model *motion* or change of states over time; additionally, another kind of variable is needed to model *causality* - *hidden causes* (Friston, 2012a). Now, it is seldom so simple that there is only one hidden state and a single hidden cause. A model may be complex to contain several levels of causes and states so that states and causes at the higher level are more abstract. For example, recognizing a certain melody entails, first, identifying the single notes and, second, connecting them. *Causal states* are computed *across* hierarchical level, while *dynamic states* use states at the same hierarchical level to estimate *motion* or how the states will evolve (Friston & Stephan, 2007, p. 439). In other words, causes are updated *between* hierarchical levels, while states – *within* one level (Shipp, Adams, & Friston, 2013, p. 708). This is just to say that there are different kinds of dependencies among hidden causes and hidden states that are to be modeled. This does not answer the question about *how* those *hidden* states and causes are actually modeled, what means that it does not answer the question about what the model looks like.

The system does not have access to hidden states or hidden causes, because they are beyond the evidentiary boundary (Markov blanket) and are called "hidden" precisely for that reason. A Markov blanket (sensory + active states) describes the separation of the model states (internal states) from hidden ones to which internal states do not have any access[358] (Friston, Sengupta et al., 2014). The generative model states may, accordingly, be divided into sensory, active and internal states so that through sensory states the system possessing this kind of model takes in new information and influences the environment by means of active states.

In contrast to hidden causes and states which are abstract entities, internal states are actually neural states (Friston, Sengupta et al., 2014). Internal states represent the hidden states and causes: they are hidden states' a posteriori estimates, conditioned on the generative model that one has (Friston, 2012a). There are two kinds of neurally realized states: *prediction* units and *error* units which represent two different functions of internal representations (Friston, 2010, p. 5). Prediction units encode the predicted state on the basis of predictions from the same level and the level above and error units encode the prediction error on the basis of errors from the same level and the level below (Friston, 2012a). Such hierarchical message-passing allows for prediction error minimization.

[358] I use 'access' here not in a personal sense, but just to emphasize that there is a cutting point between the system and everything else such that everything else has to be predicted, precisely because there is no other way to "know" or "access" it.

Precision: There are three kinds of expectations according to predictive coding: two of which I already elaborated, namely *causes* that model invariant aspects of the world and *states* that model the dynamics, e.g. motion of an object, as well as *precision* that models the reliability of causes and states (Shipp, Adams, & Friston, 2013, p. 707). *Precision* is precisely the inverse of the variance (Feldman & Friston, 2010). The generative model is a probability distribution (Friston, 2009): it specifies the probabilities for the values that certain variables can take. For such values one can calculate the mean (average value) and the variance (how much the value of a variable, in our case – sensory input, varies). Great variance indicates that it is difficult to establish on the basis of the mean of the distribution whether the sampled sensory input really belongs to the given distribution (what has caused it). For example, a beginner and a profi at tennis exercise by hitting balls at the wall. The profi would be able to hit at the same spot, while a beginner's hits would be scattered all over the place. One could calculate the mean spot a beginner hits, but would this help to predict the exact location on the wall where the next hit will go? Precision is important for explanations of misrepresentations, as I will demonstrate later by the example of delusions. It has the fine-tuning role: to determine how much influence its object will get, e.g. a certain prediction or error.

Two kinds of precision have been distinguished: *informational* or lack of knowledge, as well as *environmental* or volatility of the environment (Mathys, Daunizeau, Friston, & Stephan, 2011). In a fixed environment, when sensory input is ambiguous or noisy, there is informational uncertainty: it is unclear how the hidden states actually look like, e.g. in the dark.[359] Over time, the agent visits several environmental states, in which expected precision may vary, e.g. with the light being switched on expectations to recognize the objects clearly grow. Environmental uncertainty estimates how great that variance of precision is. So, there are two kinds of top-down predictions: not only predictions about the causes of our sensory input, but also predictions about precision[360] (Kanai, Komura, Shipp, & Friston, 2015, p. 4).

The distinction between these two kinds of precision is relevant for self-deception for two reasons. First, let me assume that precision about precision is lower for external information when the self-deceptive hypothesis is tested than when other hypotheses are tested. Then, the same kind of the operation – Bayesian inference – will lead to a different kind of the degree to which new information is incorporated into the model. Thus, a puzzling assumption that self-deceivers exhibit dual rationality, which would imply that the processes of the evaluation of information differ, can be described as the divergence in the precision of the precision about incoming information, when the self-deceptive hypothesis is to be tested. Here is an example: Suppose that you are arguing with a friend, Angie, because she, on the one hand, has no problem detecting that Sam, another friend, is deceiving himself about his attitude towards his current job position, but on the other hand, does not recognize the similarity of her own position to Sam. Sam argues to like his job, although in reality he is just anxious about looking for a new one. Suppose that Angie works somewhere else, but she, like Sam, is self-deceiving about how much she values her job position. It seems here as though Angie's evaluation of her own situation and Sam's situation differs – different criteria are applies. Instead of arguing that the kind of evaluation

[359] The input layer in virtue of the boundary gives signs about what the environment behind it looks like. When in the dark, or when the signal is ambiguous, there is much space for alternative interpretation and, thus, less precision.

[360] If predictions of content (causes) and context (precision) are seen as first- and second-order statistics (Kanai, Komura, Shipp, and Friston, 2015), then predictions of precision are third-order statistics.

differs (different operations are applies – one of them optimal and one suboptimal), it could be argued that precision of the quite similar information when testing the hypothesis "I am not fond of my job" and "Sam is not fond of his job" is different. That precision expectations differ regarding the testing of certain hypotheses is a core description of delusions (see the end of this section). Importantly, precision of precision is the parameter that determines how precision differs from situation to situation. In the example I have given Angie evaluated her situation and Sam's, but an analogical case could be constructed where Angies evaluates information in two different situation regarding only herself, e.g. her liking her job on the one hand and her being good in some kind of sports. To sum up, the first reason why the distinction between precision and precision of precision is important for self-deception is that self-deception is *dynamic*: the variation of precision across situations has to be analyzed. The second reason is that the optimal degree of instability, that will be central to my predictive coding analysis of self-deception in the following section, is dependent on the precision of precision. In short, the idea will be that the formation of the hypothesis regarding the incoming information allows for a certain level of freedom – exploration of alternative hypotheses that might not fit the evidence most tightly (static aspect). The degree of instability varies across situations that the agent encounters (dynamic aspect). The distinction between precision and precision of precision demonstrates the static and the dynamic aspect respectively.

In the following, I will focus on the static aspect of precision as enabling context-sensitivity, which is important for self-deception, because it, too, is context-sensitive. Above I have mentioned that predictive coding models *motion* by hidden states and *causality* by hidden causes. Both concern the *content*, which means in which state the world. Precision,[361] on the other hand, has been argued to model the *context* which is a property of content – its reliability (Hohwy, 2012; Kanai et al., 2015). It influences whether we perceive something at all, e.g. cues pointing at the location of an object make an agent expect to find that object there in virtue of enhanced precision expectations. Feldman & Friston (2010) use the metaphor of a searchlight for precision: it illuminates a part of the environment from the darkness of absent knowledge about the environment. In addition, in predictive coding this searchlight also makes the agent expect to find something there;[362] context consistent with the existence of certain stimuli makes an agent acknowledge that existence.

Context-sensitivity of self-deception has been explained by such personal level notions as direction of 'attention,' 'salience' and 'confidence.' Those three have been argued to be facets of precision (Kanai, Komura, Shipp, and Friston, 2015). So, I will first describe those three in predictive coding terms and then remind the reader in the next paragraph how the three facets play a role in self-deception. *Saliency* is a term used to denote the attention grabbing property of a stimulus (Feldman & Friston, 2010), e.g. a black swan in a group of white ones. Salience mediates an affordance (a property of being desirable for the agent and inducing a certain action), e.g. an agent pointing at the location of salient stimuli fulfills a touching affordance (Friston, Shiner et al., 2012). *Attention* is the upshot of precision optimization in virtue of the functional role that both possess (Hohwy, 2012, p. 12). Here,

[361] Precision is the optimization of *lateral* interactions (Friston & Stephan, 2007, pp. 440-441): "More specifically, precision sets the synaptic gain of error-units to their top-down and lateral inputs." (Feldman & Friston, 2010, p. 7)

[362] The analogy to the searchlight holds only to the degree that both precision and the searchlight is restricted to a certain limited area and everything else beyond that area is dark/out of focus/imprecise. The expectation to find something does not mean to find a particular object, but anything at all beyond darkness.

attention is used to denote not only bottom-up (saliency-driven, exogenous) attention, but also top-down (agent-driven, endogenous) one. Bottom-up attention is due to saliency: something grabbing the attention so that the agent feels attention being driven independently of her will. Endogenous attention, on the other hand, is accompanied by the feeling of volitionally having directed the attention towards something. According to predictive coding, endogenous attention has been argued to result from the increase in baseline activity of precision units prior to stimulus onset (Feldman & Friston, 2010; Hohwy, 2012). One can imagine this as follows: all units are waiting for the stimulus, but some are more eager than others, being made more sensitive. Then, the more sensitive units will detect the signal first. Lastly, *confidence* denotes sensitivity to the reliability of signals in a perceptual decision task, e.g. monkeys get a reward for completing such a task and a certain subset of their neurons fired not in response to the task content itself, but varied with the difficulty and amount of the reward (Kanai et al., 2015). Summing up, precision fulfills a functional role of selection of certain content, in virtue of providing the context for such a selection and as such is a mighty tool that has been used to explain mental and developmental disorder, e.g. delusions and autism, as well as the property of content becoming conscious (Hohwy, 2015). Regarding the latter, the idea is that in accordance with global workspace theories content becomes conscious in order to be acted upon and a *precise* hypothesis is preferred for being tested in action.

There are two ways in which precision as a selecting tool might be used for a conceptual analysis of self-deception. First, one of the main ways mentioned in the self-deception literature used to self-deceive is to misallocate attention and attention is precision optimization. Salience and confidence (two other facets of precision) also play a role in self-deception in that self-deceivers in virtue of their motivation find specific aspects of the situation more salient than others and are more often than not confident in front of observers in defending their self-deception. But redescribing certain phenomena in predictive coding terms – salience, attention and confidence – does not per se yield a *new* perspective on self-deception. The second way is to use properties of precision as a selection tool that attention, as a personal level selection tool, does not possess, in order to make the analysis more fine-grained. And this property coming in handy is the *dispersed* nature of precision which has, as a consequence, that a variety of other to-be-presented tools can be ascribed to possess precision, e.g. policies, counterfactuals, models and even precision itself.

Ways to change the generative model: What I have said so far is that free-energy minimizers occupy a limited set of states, which build a generative model at every point in time which infers the causes of sensations that those agents experience. Free-energy/prediction error minimization leads to the maximization of the evidence for the model of the causal structure of the environment. This happens either by perception[363] (change the model in accordance with the sensory input) or by action (sample more input to confirm the given model). In accordance, perceptual and active inference has been distinguished. The causal structure of the world is inferred via *perceptual inference* by means of the hierarchical subpersonal Bayesian updating (e.g., Clark, 2013b). Inference is characteristic for perception (inferring the *causes* of our sensations), but there are at least three[364] other ways to change the generative model an agent might possess, namely learning and model comparison (FitzGerald et al., 2014, p. 2; Friston, 2010):

- *Inference:* inferring causes of sensations

[363] Perception is “selection of a single hypothesis from competing alternatives that could explain sensations” (Friston, Breakspear et al., 2012, p. 1).

[364] Active inference has another direction of fit, namely not to change the model, so it has been excluded from the list.

- *Learning:* inferring/adjusting model parameters, e.g. acquisition of certain priors/hyperpriors
- *Model comparison:* averaging/selecting models for inference and learning
- *Precision optimization*[365]

The first two differ in the way that model evidence is maximized – by altering either causes or model parameters, e.g. the prior that only *one* object can be at the same spatial location. The latter prior plays a role in binocular rivalry (showing different images to each eye) in that only *one* image is perceived at a time (Hohwy, Roepstorff, & Friston, 2008). Model comparison enables to change the model all together, which would mean that one relinquishes one causal explanation and accepts another one, or one constructs a new explanation from parts of the old ones. The difference between inference, learning and model comparison can also be described as follows (Friston, Adams, & Montague, 2012): Suppose that you want to know the causes of you sensations. You might try to infer them, but for this it would be good to know the *constraints* set on this kind of inference, e.g. the laws of physics would help for example. Those are the priors that you learn on a slower temporal time scale. But even if you have learned the constraints and inferred the causes, still you might be wrong, because your model is wrong all together so that changing its states and parameters will still produce prediction error. You need to establish new causal relationships – new connections between states and parameters. So you compare other available (already optimized) models and choose another one. Those models that you compare against each other may be hierarchically deep (complex with a lot of levels describing causal relationship) or shallow (simple). *Habit formation* has been explained as averaging between hierarchically deep and shallow models – a trade-off between complexity and accuracy (FitzGerald et al., 2014, p. 5). *Bounded rationality* can also be explained by such a trade-off: an accurate model would allow for rational behavior on a task, yet a complex model has a computational cost in constructing and using them. In such a way, e.g. in abstract reasoning tasks, more simple models are preferred which lead to suboptimal task performance (Ibid., p. 7). In some cases though, instead of models being weighed against each other, they can be integrated into *one* model, if such a joint model is realistic, e.g. as in a ventriloquist illusion (p. 7).

Since there are four ways of updating the generative model, each of them can be theoretically used as a conceptual tool in the analysis of what the generative model of the self-deceiver looks like. I, in addition, showed how bounded rationality can be explained in predictive coding terms, since, as you already know (see e.g. sections 3.1.2.2 and 3.2.3.1), self-deception has been argued to be a case of bounded rationality. But as in the case of precision, just saying that predictive coding can explain bounded rationality and via

[365] As for precision, though there is something that unifies states and parameters that precision is not – it does not describe "deterministic dynamics in the world" (Friston, 2009) but random fluctuations of states, nevertheless it is still part of the generative model (Feldman & Friston, 2010). Below is a neurobiological description of inference, learning and attention:

"According to the free-energy principle, the sufficient statistics representing all three sorts of quantities will change to minimise free-energy. This provides a principled explanation for perception, memory and attention; it accounts for perceptual inference (optimisation of synaptic *activity* to encode the states of the environment); perceptual learning and memory (optimisation of synaptic *connections* that encode contingencies and causal regularities) and attention (neuromodulatory optimisation of synaptic *gain* that encodes the precision of states)." (Friston, 2009, p. 295)

So, optimization of activity, efficacy and gain (weight) corresponds to inference, learning and attention respectively, plus model comparison, which optimizes something else, namely the *complexity* of the model: complex models generate prediction error in the long run, though they may be accurate for a short period of time and, thus, have to be simplified (Hobson et al., 2014); Hohwy, 2015).

this it can explain self-deception is not to offer new conceptual insights. I offered this piece of information to the reader for reasons of completeness. In the next section, I will use model comparison in the attempt to gain new conceptual insights regarding self-deception.

Active inference: Active inference is the counterpart to updating the model of the world because during active inference the model is kept unchanged and new information/evidence is searched for the model (e.g., Hohwy 2012). Switching between perception and action depends on the precision of prediction errors: *precise* prediction errors change hypotheses, while *imprecise* ones lead to action (Brown et al., 2013). The reason for this precision relationship is intuitive: if the causal model is certainly wrong, then it has to be changed, otherwise one would better get more evidence. In a case of a physical action (in difference to mental action, e.g. remembering something, or to interoceptive action, e.g. controlling one's heartbeat or calming down, about which I will talk later in this section), proprioceptive prediction errors can be suppressed by engaging classical reflex arcs. Since both predictions and prediction errors are precision-weighted, action ensues only if the *ratio* between both these precisions fits. Predictions are sent downwards, at some point translated into proprioceptive predictions and then, via the workings of classical reflex arcs, change sensory input and in this way suppress proprioceptive prediction error (Brown et al., 2013). Whether proprioceptive predictions are corrected (and the model changed) or whether classical reflex arcs are engaged to change sensory input depends on the precision of both. In fact, precision of sensory prediction errors has even to be *attenuated* in healthy subjects and this is why one can't tickle oneself (Ibid.). Interestingly, if the precision of predictions is increased instead of the precision of prediction errors being attenuated, this results in delusions, e.g. an external force moving one's limbs. The reason is that precise proprioceptive prediction errors propagated *upwards* need to be explained in any case, independently of the level of precision of predictions (Brown et al., 2013).

Active inference per se will not feature in my conceptual analysis of self-deception in the next section, but model-based active inference (as a subset of active inference) will be part of it in order to explain the goal-directedness of self-deceivers, as well as other important phenomenological features such as tension and transparency of attitudes because of counterfactual representations that take part in model-based active inference. Further, the short excurse into how proprioceptive prediction error minimization depends on the ratio of precision estimates may sensitize the reader to the *fragility* of precision estimates that makes precision a useful tool for explanations of delusions, about which I will talk at the end of this section.

Model-based active inference: How to incorporate the goal-directed aspect into active inference (the following elaboration is based on Friston, Adams, & Montague, 2012; Friston, Samothrakis, & Montague, 2012; Friston, Adams, Perrinet et al., 2012)? One way is to assume the existence of a *value* function, as in optimal control or reinforcement learning models. The idea is that an agent in a certain state can complete an action to get into another state that will be more beneficial, e.g. find a way in the labyrinth to a treasure. Given that one has to complete several actions in order to get the desired state, the optimal sequence of such actions is the one that in every step brings an agent nearer to the goal. A value function is a mapping from state to an amount of reward for visiting that state. A sequence of optimal actions is a *policy*. The policy is found either by *model-free* or *model-based* learning. If there are not so many states that one can visit, one can learn the full model: reward *and* transition probabilities. The latter are the probabilities that if one is in a certain state and computes a certain action, then one will get to a specific other state (with a certain value). As an alternative, one can only learn the value function – a model-*free* variant, because the generative model of transition probabilities is not learned, for example because one does not know all the states that one can visit. One can then use the value function to determine the next action (choose an action that brings about the most valuable state). Interestingly, free energy minimizers do not need a value or cost function. In analogy to model-based and model-free learning one makes a distinction between *agency free* and *agency based* schemes (in the latter the agent represents its own actions, see Friston, Adams, & Montague, 2012, p. 5), but in both value functions have been substituted by *prior beliefs* about certain parameters of the model, namely transition probabilities between hidden states. The reason is that this substitution allows to solve the optimal control problem by already available means – Bayesian inference:

> The key distinction between optimal control and active inference is that in optimal control, action optimises the expected cost associated with the hidden states a system or agent visits. In contrast, active inference requires action to optimise the marginal likelihood (Bayesian model evidence) of observed states, under a generative model. (Friston, Samothrakis et al., 2012, p. 524)

Free energy minimizers always choose the least surprising actions and the amount of surprise depends on the (optimized) model of the environment that consists of states and parameters (priors). So actions depend on priors and there is no need for an additional value function: value is the evidence for the model. In *agency based* schemes, an additional subset is added to hidden states – *hidden control states*. Control states are beliefs about future actions (or *decisions* that an agent is to take; Moutoussis et al., 2014b, p. 4) and a policy becomes a sequence of control states that are connected by transition probabilities (Friston, Samothrakis et al., 2012). Control states can then be optimized (inferred) to find a policy that minimizes free energy (reduces uncertainty about our predictions) the most. For example, in the context of face categorization hidden control states may be the spatial locations that attract visual fixation, e.g. ear, eye, mouth of a face (Friston, Adams, & Montague, 2012, p. 15, 18). A model has been constructed how free energy minimizers perform saccadic eye movements (reduce uncertainty about which kind of a face they see) and it corresponds to the empirically available data about salient features of faces. Note that *hidden* control states are abstract entities that are neurologically realized through prior expectations about hidden controls and they are counterfactual in that they encode "what we would infer about the world, if we sample it in a particular way" (Friston, Adams, Perrinet et al., 2012, p. 2).

I think that goal-directed higher-order cognition can be constructed on the basis of agency based active inference, but as Friston, Samothrakis et al. (2012) note, the inference of a policy per se does not mean that it is also owned by the agent.[366] This is just a word of caution meaning that to say that in addition to the analysis of higher-order cognition along the lines of agency based inference (which I will do in the next section) one needs an additional argument to couple that analysis with the phenomenology of the agent, e.g. an inference to the best explanation. In the next section I will argue that an epistemic agent model can be described as trajectory merging. Each of the trajectories is a certain pursued policy – a strand of thoughts. This is because policies in virtue of their functional role – representing the agent as acting in a certain way in the future – fit the functional role of trajectories that can be potentially integrated into an epistemic agent model because of the modelled *agent* component. One can imagine that there is a pool of policies of which only

[366] Note that though Friston et al. (2012) define the "sense of agency as probabilistic representation of control" (p. 524), they also warn that "crucially, the agency implied by inference on control is not necessary owned by the agent" (p. 539).

one on the phenomenal level is being pursued, namely the one embedded into the epistemic agent model. How they are merged, depends on interoceptive counterfactuals, which offer means to explain the non-rational behavior of the self-deceiver. Another is that policies change our perception – this has been offered as a predictive coding explanation of the optimism bias.

There is a mutual dependency between transition probabilities of control states and transitions between hidden states, by which optimism bias has been explained (Friston et al., 2013). The explanation goes as follows: States have *expected values*, averaged over policies. These values can lead to *optimism bias*[367] "in the sense that, when precision is high, perception is biased toward the state that has the greatest potential to realize the agent's goal" (p. 10). This is the case because there is an interdependence between transition probabilities of control and hidden states. In other words, what we perceive depends on what goal representations we possess and the other way around: the way we perceive the world determines what we desire. The interdependence between hidden and control states enables *context-sensitive action* (p. 5), insofar as they influence each other's transition probabilities (p. 4) by means of influencing each other's values: "symmetry between the expected value over states – that provides the value of a choice, and the expected value over choices – that provides the value of a state" (p. 10). If precision of policies (policies also have precision, because they are inferred too; Moutoussis et al. 2014b, p. 10) is too high, there is even an augmentation of optimism bias, possibly the case in schizophrenia (Friston et al., 2013, p. 11; Schwartenbeck et al., 2014, p. 10). It is the *precision weighted value* of a policy that determines the belief that this policy will be performed (Schwartenbeck et al., 2014, p. 4). Precision optimization ensures that the prediction error with the most precision is selected. Such selection can lead to *biasing* as in case of optimism bias: biasing perception toward goal states and enhancing confidence in action choices (Friston et al., 2013).

Optimism bias is an example of how misrepresentations can be explained in predictive coding terms. Since self-deceptive attitude is a personal level misrepresentation (at least in the eyes of the observers), the reader could now explain the inability of self-deceivers to look at the personal level evidence with the eyes of observers: they do not *see* the same thing (e.g. think about the relativist account of Clegg & Moissinac, 2005 presented in section 1.3.4 who state that self-deceivers' and observers' evaluations are different in virtue of different perspectives they are taking onto the matter).

In the last section I mentioned two trade-offs that predictive coding can incorporate: speed-accuracy and accuracy-complexity. Accuracy-complexity trade-off is achieved by model comparison: to prevent overfitting, a generative model is simplified (Hobson et al., 2014). Speed-accuracy trade-off has been explained by the assumption of a posterior risk minimization: in conditions of uncertainty an agent would choose a policy that minimizes risk, e.g. being eaten by a predator (Daunizeau et al., 2010a). Another kind of a trade-off has been explained in terms of free energy minimization, namely that of exploration-exploitation. The value of a policy (divergence between states the agents can reach and those that the agent would like to reach) has been argued to be decomposable into *entropy* (intrinsic reward) and *utility* (extrinsic reward) (Friston et al., 2013; Friston, Schwartenbeck

[367] Moutoussis et al. (2014b) see *cooperation bias* in an interated Trust Game as an interpersonal analogue of the optimism bias, because interpersonal representations can influence beliefs about the desirability of outcomes (p. 10); e.g. "a fair person will not exploit me" (p. 2) and my co-player is fair, thus I will invest more money. In a Trust Game one player gets certain amount of money and can decide which part of it to invest with the other (given a certain gain by which this amount will be multiplied). The other can the return a part of the investment.

et al., 2014). Entropy corresponds also to the *novelty bonus* for exploring new states (Schwartenbeck et al., 2014, p. 4). So, intrinsic reward is that of exploration and extrinsic – that of exploitation. The general idea is that if the agent is certain in the policy, she can go for it, if not (if at least several states possess the same value) – it is better to explore more (Friston et al., 2013, p. 7).

The exploitation-exploration trade-off will become important for an analysis of self-deception in the next section, because I will argue that an optimal degree of instability (degree of uncertainty about the current hypothesis that ensures the novelty bonus), explains the fragility of self-deception, namely the policy and even the world-model changes in which self-deceptive attitudes either acquire a transparent signature of knowledge or become transparent themselves.

Interoceptive inference: Classical reflex arcs are not the only means that an agent has to control her behavior. Those control only motor behavior. Autonomic reflexes, in analogy to motor reflexes can also exhibit control. The difference is that the direction of this control is not outwards (external), but inwards – towards controlling the physiological condition of the body via *interoceptive* inference (Seth, 2015b):

> Drawing a parallel with models of perception, predictive interoception would involve hierarchically cascading top-down interoceptive predictions counter flowing with bottom-up inte- roceptive prediction errors, with subjective feeling states being determined by the joint content of the top-down predictions across multiple hierarchical levels. (Seth, Suzuki, & Critchley, 2012, p. 7)

If there is interoceptive inference, then there must be physiological states that are beyond the evidentiary boundary which means that they are not part of the model and have to be predicted. In accordance, the part of the body that enforces an evidentiary boundary and predicts external states is only the *brain* (Hohwy, 2014). This is so because our body is a means of interaction with the environment and in virtue of this tight relationship bodily states should also be modeled (Hohwy, 2014, p. 11). Even more so, there has been an argument that inference serves, in the first place, the aim of homeostasis (keeping essential internal variables stable), e.g. maintaining constant sugar level in the body, then the aim of acquiring an accurate model of the environment (Seth, 2015b). In addition, if the two factor theory of emotions is accepted, according to which emotional content arises as a result of the cognitive appraisal of physiological changes in the body, then there is a link between affective states and prediction error minimization (Seth, 2013; 2015a; 2015b; Seth, Suzuki, & Critchley, 2012). In accordance, Anil Seth (et al. 2012, 2013) has recently introduced an interoceptive model of emotions that includes that of feelings, given their definition as consciously experienced emotional states (Seth, 2013, p. 565). It states that emotions are the result of interoceptive inference that involves *interoceptive* prediction errors about *bodily* states.

How affective states (feelings and emotions) are explained in predictive coding terms – by means of interoceptive inference – is important for the conceptual analysis of at least two features of self-deception: first, I will argue in the next section that *interoceptive* counterfactuals (potential affective consequences of an action) make mental actions causally efficacious, second, I will explain tension by means of interoceptive counterfactuals.

Two factor theories of affective states on which interoceptive inference is based have been criticized. Proust (2015a; 2015b), for example, argues with respect to feelings (a subset of affective states) that they fulfill an indicator-function in a non-conceptual way (see section

2.1.3). This relates to a discussion about the possibilities of cognitive penetrability[368] in predictive coding. ‘Concept’ is a personal-level term and one might define cognitive penetrability in a narrow sense as influence of priors that are personal level beliefs (Vance, 2014), or one might opt for a broad understanding of cognitive penetrability as the influence of priors/expectations in general (Goldstone et al., 2015; Vetter & Newen, 2014). Since I provided a characterization of tension along the lines of Proust’s theory of metacognitive feelings in section 2.1.3, but am going to combine it with the predictive coding characterization in terms of counterfactuals in the next section, I will now briefly comment on the seeming incompatibility between interoceptive inference based on two factor theories and tension as non-conceptual indicators. My answer is analogical to the one that I will give later in this section to the seeming incompatibility between terror-management theory (which I favored in section 3), on the one hand, and sociometer theory as the basis for *interpersonal* inference on the other. Terror management theory and sociometer theory provide different view on the function of self-esteem. There, as here in the case of feelings, I think that one can use interoceptive inference without accepting all implications of the two-factor theory, because the latter can be treated as an *inspiration* for the former.

Motivational power of free energy reduction: There is a more general account of affective states as the rate of reduction of free energy (Joffily & Coricelli, 2013). The premise for such an account is that there is no need for a value function in free energy minimization so that free energy itself is actually the only ultimate value that exists and what else descriptions of valuable states one might give, if they do not minimize free energy, they would not be valuable. One can define a cost function as the “rate of change of value” (Friston, 2010, p. 8). This just means that the bigger the positive difference in value between two states (the current one and the state to be visited), the better, e.g. if money were the value then winning a lottery for a beggar would be a huge change of value. An *evolutionary* value of a phenotype has been argued to depend on the amount of time that is spent by the phenotype in valuable states (Friston, 2010, p. 7). If value is substituted by free energy reduction, then one would get Joffily & Coricelli’s (2013) explanation of emotional valence as rate of change of free energy, e.g. an agent that could reduce free energy the most would be very happy. To remind the reader, there are two kinds of precision estimates: informational uncertainty that estimates how unreliable our senses are and environmental uncertainty about how informational uncertainty changes over time. [369] Rate of change of free energy possesses the functional role of volatility, because it also assesses changes in the environment (Joffily & Coricelli, 2013).

This account demonstrates one important implication of free energy reduction: since it is valuable in itself, it possesses *motivational power*. There may be multiple ways to achieve free energy minimization (at least three kinds of changes of the generative model: inference, learning, model selection, as well as at least three kinds of inferences: perceptual,

[368] Cognitive penetrability can be interpreted as a question about the modularity: if something is impenetrable, then it is informationally encapsulated (Farennikova, 2014). Proust (2014), for example, argues for two different kinds of *representational systems*, a non-conceptual and a conceptual one.[368] Proust (2014) then argues that integration of procedural and analytical metacognition is an instance of a general integration of control structures “by a hierarchy of controls that have mainly a top-down effect, but also propagate error signals in a bottom-up way” (p. 743).

[369] Here Joffily & Coricelli (2013) accept Yu & Dayan’s (2005) distinction between expected and unexpected uncertainty: Expected uncertainty is the one about known unreliability of predicting relationships *within* a context and unexpected uncertainty is the one about the appropriateness of the context itself such that when unexpected uncertainty is high, it is a signal that a *context switch* should be made.

active (with several subtypes, e.g. model-based, model-free, interoceptive), so a variety of changes of states, parameters and precision might be responsible for what observers describe as the agent's motivation.

This implication of the free energy principle – that changes in the generative model, which reduce its free energy, possess motivational power – is important for a model of self-deception, because it allows to see motivation, similarly to dispersed precision, not as a certain centralized instance that radiates its influence somewhere (that would be a personal level understanding of motivation so far favored in accounts of self-deception), but motivation would then be an emergent property resulting in some changes of the generative model whose origin and workings are yet to be found out.

Transfer problem: On the one hand, predictive coding is a subpersonal kind of hypothesis testing (a causal hypothesis about the structure of the world is inferred or acted upon) – it prima facie does not relate to consciousness and reasoning. On the other hand, it uses personal level terminology and can be actually used to explain the personal level counterparts of those termini too. As I noted in the last section, terms 'prediction error' and 'inference' can be used either personally or subpersonally (see previous section). Now I can put the term 'uncertainty' on my list along with '(Bayesian) inference' and 'prediction error' as one which has personal, as well as subpersonal connotations and is used both ways in predictive coding. Precision (uncertainty) is not the one we experience, but uncertainty (low precision) about the rate of change of environment has also been argued to cause *anxiety* (Mathys et al., 2011). If in our construction of predictive coding models we use personal level notions, we might transfer the personal level relations and attributes of such notions to the subpersonal level, yet the rightfulness of such transfer requires justification (see section 1.3.4 for the same criticism of psychological experiments of self-deception; see also Drayson, 2012, for the general discussion of the problem). Let me call this the *transfer problem*. In other words, the relations between prediction errors and inferences as objects of our epistemic agent models (conscious models of oneself as a reasoning being) need not be the same as the relation between the prediction errors in the generative model underlying it. On the other hand, different facets of our experience, e.g. consciousness, delusions, illusions (for a concise summary see Hohwy, 2015), phenomenal self-model (Limanowski & Blankenburg, 2013) are being successively explained by predictive coding. The procedure hereby is usually to explain key characteristics of the phenomenon and offer predictive coding as the best guess:

> In this view [Limanowski & Blankenburg, 2013], the experience of embodied selfhood is specified by the brain's "best guess" of those signals most likely to be "me" across exteroceptive and interoceptive domains. From the perspective of cybernetics the embodied self is both that which needs to be homeostatically maintained and also the medium through which allostatic interactions are expressed. (Seth, 2015b)

According to Seth et al. (2012) *sense of agency* depends on how good exteroceptive prediction errors could be explained away, while *sense of presence* depends on the explanation of the interoceptive prediction errors (p. 3-4). A toy example: I am jogging one sunny morning on the beach. The model of my body and its interactions with the environment predict that I move my arms and legs a certain way, that my blood pressure may rise a little in comparison to sitting around, that I may feel a light breeze on my skin. In other words, there are several kinematic, proprioceptive, but also interoceptive predictions and if the senses confirm those predictions, I feel that I caused those movements and that I am immersed into the beach experience I am having. So, a system may not only construct a model of itself and its environment, but also make it *phenomenally* available.

Thus, you as a reader are now trying to understand how a prediction error model of self-deception can be constructed. You possess an epistemic agent model – transparent conscious self-representation of testing the hypothesis that self-deception is also explainable by prediction error minimization – which is, if the latter account is correct, already a case of subpersonal prediction error minimization that acquired the property of consciousness for some reason, e.g. to execute some actions (criticize my account for example).

For such phenomenally represented, but not necessarily controlled phenomena as dreaming, mind wandering and imagination, a predictive coding explanation has also been given. Those are phenomena in which state transitions happen in the absence of external input (Sadaghiani et al., 2010). A free energy minimizer visits a *limited* set of states, but from the perspective of brain dynamics there is also a process of continuous *exploration of hypotheses*[370] so that the existing generative model is changed, e.g. to minimize complexity as in dreaming (Hobson et al., 2014, p. 7). Imagination has been described as a case in which a *generative* model of causes, once formed, generates patterns on its own (Clark, 2013a). Internal brain activity has been described as "brain's internal context for processing external information and generating behavior" (Sadaghiani et al., p. 12). Result of itinerant fluctuations have been interpreted as searches over the hypothesis space (Sadaghiani et al., 2010).

With the transfer-problem subsection I meant to achieve three aims: first, I reminded the readers of the caution given in the previous section (about personal and subpersonal uses of 'inference' and 'prediction error' and now also uncertainty); second, I introduced the argument that the transfer from the generative model to the phenomenology as an inference to the best explanation (which is the foundation for conceptually analyzing phenomenological characteristics of self-deception in predictive coding terms) and, third, reviewed the predictive coding characterization of mind wandering as a constant exploration of hypotheses that I will apply to self-deception in the next section.

Counterfactuals: The idea that counterfactuals are needed for the agent to exhibit control over its environment is developed in Friston, Adams, Perrinet et al. (2012). There the authors argue that since there is a subset of hidden states that agents have control over (*hidden controls*) and which are represented as prior beliefs about future actions, the latter, before the action is executed, are actually counterfactuals – encoding the possible outcomes conditioned on the kind of action that would be taken (p. 2). To remind the reader, I have used counterfactuals in two senses so far in this thesis. First, I noted during the tension-discussion that according to Dokic's modal understanding of metacognitive feelings, those feelings indicate the extent to which a certain process would be successful in nearby possible worlds (see section 2.1.3). One can understand a possible worlds in different ways (e.g., see Menzel, 2015), but simplistically one could imagine that a world contains objects over which one can form relations (predicates). Since a possible world might differ in objects and relations from the actual world, it is a counterfactual world. Those might differ so that, for example, there might be no bird named Albert in a possible world p_i, or that bird would belong to a kind of birds that do not fly, e.g. an ostrich. Second, during the discussion of goal-directedness (see section 2.2.1), I mentioned studies that via the induction of personal level counterfactual thoughts (imagining counterfactual outcomes of a certain situation, or generating thoughts of the form "If only I had done x in situation y"),

[370] It is in question whether the predictive coding explanations about how mind-wandering and dreaming are unconstrained by external stimuli can be applied to self-deception, see Windt et al. (2014) for the discussion of sensory attenuation and agency in dreams and wakefulness in predictive coding.

indicate the presence of certain personal level traits (e.g. kind of counterfactuals generated by the subjects indicates creativity), or improve the results on a certain task (e.g. building counterfactuals in a negative mood improves anagram solving). So, one could use the term 'counterfactuals' either personally or subpersonally. Now I will explain which role subpersonal counterfactuals play in predictive coding and how one could enrich their role. According to Friston, Adams, Perrinet et al. (2012), a probabilistic representation of control is needed to infer hidden controls and this leads to the sense of agency: "[t]he agent may infer that these control states are produced by its own movements and thereby infer agency" (p. 4). Anil Seth (2014) has proposed a PPSCM (Predictive Perception of Sensory-Motor Contingencies) account to explain the absence of the phenomenology of *perceptual presence* in synesthesia. The main idea is that "PPSCM requires that counterfactual predictions be *explicitly* incorporated as part of the priors in a HGM [hierarchical generative model]" (Seth, 2014, p. 104). *Subjective veridicality* then is dependent on counterfactual richness[371] or the "range of conditional sensorimotor relations that are counterfactually encoded" (p. 105). For example, we perceive a red tomato (and not its picture) lying before us on the table, because there are a lot of counterfactual relations between our actions, e.g. going in a circle around the tomato and the part of the tomato that we expect to see. Those unseen but predicted parts that are conditioned on action complement each other so that we see a tomato as an object in the world lying there. According to Seth (2014), it is *intermediate*-level counterfactual probability densities (= secondary sensory cortical processing) that decide for high-level prior of image-hood against the high-level prior of object-hood (p. 105). Note also that Seth's (2012, 2014) accounts of the sense of presence (explaining away *interoceptive* prediction error vs. counterfactually-rich model so that counterfactuals encode *sensorimotor* contingencies) differ, but that they can be reconciled on a premise that I will elaborate in the next section, namely that sensorimotor contingencies include interoceptive predictions as well.

If there are two kinds of ways control can be executed – allostatic by motor reflexes and autonomic – so that there is *perceptual* and *interoceptive* active inference, then Friston's argument for counterfactuals can be extended to the interoceptive domain as well (Seth, 2015b). Further, counterfactuals may not only aid in exhibiting control, but also contribute to certain kinds of experience. This is important for an explanation of self-deception, because tension (an affective state that I will relate to interoceptive counterfactuals in the next section), as well as transparency of self-deceptive attitudes (that has been explained by exteroceptive counterfactuals by PPSCM) belongs to the phenomenological profile of the self-deceiver.

I first introduced counterfactuals in this section when I spoke about goal-directed action and then at the end of the discussion of interoceptive inference. I promised that I will argue in the next section that *interoceptive* counterfactuals (potential affective consequences of an action) make mental actions causally efficacious. I will explain tension by means of interoceptive counterfactuals. Those exteroceptive counterfactuals explain transparency will, in addition, offer another perspective on how self-deceptive attitudes acquire a transparent signature of knowing or how they themselves become transparent. For that, I will use as a premise that *interoceptive* counterfactuals go along with *exteroceptive* counterfactuals, which is to say that along with sensory consequences of an action affective consequences are modelled as well. My reason for such a hypothesis is that perception and interoception cannot be modularized, as I will argue in the next section.

[371] Perceptual presence and transparency may also express *epistemic reliability* which correlates with degrees of counterfactual richness (Metzinger, 2014).

Interpersonal inference: As interoceptive states have been argued to be subject of prediction by the model, the characteristics of the transparent self-model which agents take to be themselves can also be argued to be predicted. Moutoussis et al. (2014a,b) suggest that self-representations are inferred via *interpersonal inference.*[372] As Seth embraces two factor theory for convenience reasons, Moutoussis et al. (2014a,b) embrace the *sociometer theory of self-esteem* (which is a rival of the terror management theory, see section 3.1.4) and extend it to self-representations in general (Moutoussis et al., 2014a, p. 71). For sociometer theory, self-esteem is a means of social inclusion and for terror-management theory – a buffer against death anxiety. Moutoussis et al. (2014a,b) accept sociometer theory in order to be able to claim that "[t]ype-based interpersonal representations, of which self-esteem is only a subset, serve to optimize context-dependent social computations" (p. 71). A type is a trait that is context-relevant (Moutoussis et al., 2014a). Social interaction (goal-directed active inference) and self-representations are argued to be interdependent so that agents infer their own self-representations on the basis of the results of social interaction and, on the other hand, self-representations determine which control states (goal representations) are pursued:

> In an interpersonal context, traits such as 'talented', 'harsh' etc. can feed directly into the outcomes that a person wants to reach (or indeed to avoid). In terms of active inference, agents can *inform goals* by the desirable self-representations that they are likely to reach as outcomes. An example is: "If the outcome of this exchange is that I cooperate and my partner runs off with the money, this would be evidence that *I am a fool*. This is highly undesirable – I'll attach very low probabilities to outcomes implying that I'm a fool". (Moutoussis et al., 2014a, p. 71)

There are two implications of the interpersonal inference view that are worth mentioning in the light of the discussion of self-deception. First, the mutual influence between control states and self-representations suggests an explanation of cognitive and affective biases as *prior beliefs* about "noxious states in self and others" (Moutoussis et al., 2014a, p. 74). As a result, incongruency of the agent's decisions with the evidence can be explained by the influence of the agent's *type* instead of self-deception, as the authors argue (Ibid, p. 75). Regarding the difference in updating one's own self-representations and representations about others, Moutoussis et al. (2014b) argue that *self*-representations are more resistant to change than those about others given that the evidence basis for the former is much bigger than for the latter (p. 11). What does this mean for a theory of self-deception? The claim that 'self-deception' is an unnecessary term to use in the context of interpersonal inference reminds me of Borge's claim that if what we call 'self-deception' might be explained by the influence of emotional microtakings, we should not use the term (see section 1.2.5). Yet since the behavioral and phenomenological profile of self-deception is more complex than just goal-directed action, I think that interpersonal inference may serve as *one* of the explanatory tools that aid explaining the social aspect of self-deception.

Second, if interpersonal inference were to take part in analyzing self-deception, the claim, with which one has to deal, is that interpersonal inference is not defensive, as self-deception has been often claimed to be in the psychological literature (see e.g. section 3.1). I will argue now that one can reconcile both interpersonal inference and the claim that self-

[372] It is underquestion whether it is useful to distinguish *interpersonal inference* as a special kind of inference. *Perceptual* and *interoceptive inference* as inferring causes of external and internal signals, as well as *active inference* as acting on the present model in order to *change* the signals (see Fotopoulou, 2013a, p. 35 for a short discussion) may be plausible as different kinds of inference, but I am not sure about the distinctive characteristic of interpersonal inference.

deception may be to a degree defensive. For that let me first reflect on the difference between homeostasis, which may be seen as a defense of the system against outer influences (striving for balance despite forces acting to disrupt that balance), and social exchange as a dependence on outer influences. On the one hand, self-representations has been argued to be not a result of *homeostasis* (maintaining desirable representations for their desirability, psychological defense theories), but of *social exchange* by Moutoussis et al. (2014a) and they are updated on the basis of one's own actions and actions of one's interaction partners (p. 72) in a way that minimizes interpersonal surprise (p. 75). On the other hand, homeostasis – understood as maintenance of the internal variables in the expected range – has been argued to be a *fundamental imperative* such that perception serves rather this imperative than world-model construction (Seth, 2015b). Anil Seth (2015b) emphasizes the value of homeostasis as the state that not only perception, but also cognition and action strive for, comparable in its importance to the free-energy principle.[373] As Parvizi & Damasio (2001) put it, "[i]n fact, these functions - wakefulness, basic attention, and emotion - are interrelated and all aim, in one way or another, at achieving homeostatic balance" (p. 152). I would say that rather than asking what came first: homeostasis or world-model construction, the question should be: why homeostasis is achieved by world-model construction rather than by anything else?[374] So, instead of seeing homeostasis and world-model construction as rivals, one could also see them as complementing each other: world-model construction as a way to achieve homeostasis and maintaining homeostasis as a way by which world-model construction can proceed (I do not say the *only* way to leave the options open). In analogy to homeostasis – world-model construction, the defensive-offensive dichotomy might be reconciled in a similar way as I show below.

Answering the question about whether defense (terror management theory can also be viewed as a defensive theory) or social inclusion best describe interpersonal representations, this question actually mirrors Trivers' extension of the function of self-deception (see section 3.2) from a defensive to an offensive one – deceiving others, which implies social interaction, as a means of doing so. Actually, the interpersonal inference view answers the question about the usefulness of the defensive-offensive dichotomy itself. It has namely been argued that self-representations serve the aim of reducing the complexity of the generative model, since their *heuristic* nature allows for omitting "a complex tree of possible future states and outcomes" (Moutoussis et al., 2014a, p. 72). Dichotomies are also heuristics, so they reduce complexity, but do not necessarily point out the truthful state of affairs. So, the intermediate conclusion is that interpersonal inference, as other explanatory tools mentioned here, might aid in explaining self-deception. The fact that it makes accents different to those made by self-deceptive theories does not constitute a hindrance in its use. One need not accept sociometer theory (main claim: self-esteem reflects the level of social inclusion) as a whole, but it is only the claim that an agent's personality traits and behavior are predicted on the basis of its actions in a social environment that underlies interpersonal inference. And the latter can be reconciled with terror-management theory, if one, for example, argues that social interactions primarily reduce anxiety of death and that social inclusion serves this aim.

[373] Daydreaming and mind-wandering might be interpreted as activities that aid in maintaining homeostasis: "People strive for a balance between active goal pursuit and inner life such as ordinary mind-wandering and imaginative daydreaming, which are themselves goal-related." (Klinger, 2013, p. 14)

[374] Why in humans world-model construction has been preferred, has probably evolutionary reasons.

Against sociometer theory which argues that the function of self-esteem is to "reflect the extent of one's inclusion or fitness for inclusion in social groups," Pyszczynski et al. (2004a) point to instances of self-deception in which participants view themselves more favorably than others do; in this case sociometer theory would be violated, because self-esteem does not serve as an indicator for social inclusion anymore, because others recognize that it is not accurate (pp. 461-462). This is the case in which constraints (which theory about acquisition of self-esteem to prefer) would need to be shifted from the social inclusion view. Independently of the explanation for which aim self-representations serve, the interdependence between them and precision-weighted policies, which are sequences of control states linked by transition probabilities, is for me a more important point.

Interpersonal inference can become useful in two ways for conceptual analyses of self-deception. First, according to interpersonal inference personality traits are priors. Together with the motivational power implication discussed above, the acquisition of self-deception can be analyzed as the motivational workings of personality priors (because changes to the generative model that reduce free energy have motivational power, e.g. changes in the personality priors). Second, interpersonal inference can be used as a tool to describe the *behavioral* characteristics of self-deception, namely why and how self-deceivers *justify* their self-deception. For example, it has been argued that agent's attitudes about oneself are more resistant to change than attitudes about observers, because in time one samples more evidence for the former (and, thus, is more certain in their truth). Yet since self-deceivers' attitudes are actually often inconsistent with their behavior, inferences about attitudes cannot be drawn from behavior, so self-deception is more of a challenge for interpersonal inference than it is a tool for its conceptual analysis. I included interpersonal inference mostly for completeness reasons rather than for use in the next section.

Different facets of precision explaining delusions: I will now exemplify the direction and object of influence of precision for delusions, because they are conceptually tightly related to self-deception (see sections 1.2.6 and 1.2.7). According to doxastic accounts of delusion, an aberrant belief-formation is *selective* in updating its contents. In the cases in which such a selection is motivated, delusion has been argued to be equated with self-deception (see section 1.2.7). In predictive coding inferences are made by means of Bayesian updating (posterior = prior * likelihood) where each element can have its own *precision*: priors, likelihood and posteriors. Precision regulates the amount of belief updating (Frith & Friston, 2013, p. 8) and it is an attribute of different kinds of internal states, e.g. predictions, prediction errors, policies. Self-deception has also been argued to be a case of aberrant updating of the self-concept (section 3.1.2). If one wants to specify in which way precision changes the generative model to explain delusions, one should ask the following questions: How does precision operate? What is the direction of precision (high or low)? What is the object the precision weighs/selects? What is the level at which precision is changed? How is precision updated, or how does it change with time? Let me look closer at the answers to each of these question for delusions now. The reason for doing so is the following.

Dispersed nature of precision: How this kind of selection operates and what it actually selects are, then, two questions that one needs to answer in order to be able to apply this kind of selection to an explanation of some phenomena, e.g. self-deception. The answers are that selection is *dispersed* throughout the model and that the objects of selection, as well as the direction (more or less precision) depends on the nature of phenomenon one is about to explain. I will below demonstrate it on the example of delusions. First, I will say a few words about the dispersed nature of selection. The above description of precision as context and attention as precision optimization, points, according to me, mostly into the direction of representations of precision expectations being *dispersed* throughout the brain.

The alternative to dispersion would be an attentional system that can be encapsulated in some way. Precision is dispersed because of the way the generative model is like: several objects, e.g. priors, errors, policies, precision of precision itself and are connected among each other in multiple ways. Thus, there is no central locus of precision. As a direct consequence, the kind of *control* that precision in virtue of its functional role of selection provides is also decentralized. Interestingly, it is not only the case that precision provides a means for decentralized control, but also that one can argue that the attentional system, which one usually takes to be the control instance, is itself in need of a control instance, which is the default mode network (DMN).

Let me now discuss the latter view, not only because it broadens the understanding of how control could be executed according to predictive coding, which is useful for explanations of control/sense of control in self-deception, but also because the latter view about DMN as a controlling instance has been exemplified on repression, which is conceptually connected to self-deception (see for example section 1.3). If *precision* is dispersed across the hierarchical *network*, and attention is an emergent property of precision optimization, why should be there an additional attentional *system*? Carhart-Harris & Friston (2010) argue namely for a separate attentional system (see

figure 31). According to them, delusional thinking arises in the case of not properly functioning DMN so that it cannot control the attentional system[375] (p. 1276).

> The picture that emerges is of a hierarchy of brain systems with the DMN at the top and the salience and dorsal attention systems at intermediate levels, above thalamic and unimodal sensory cortex. Under a Helmholtzian model, each system is trying to suppress the free-energy of its subordinates, through a process of optimizing predictions to reduce prediction-errors. (Carhart-Harris & Friston, 2010, p. 1270)

One possible answer is that speaking in terms of 'systems' is in this case a useful simplification if the *function* – that of selection - is to be emphasized. The speculation that DMN activity is comparable with repression (Carhart-Harris & Friston, 2010, p. 1271) is unusual. In section 1.2.6 I have reviewed Gerrans' view that distinguishes the following levels: reflexive, DMN (affect-laden) and decontextualized. The difference between Carhart-Harris & Friston's and Gerrans' characterization is that the first ascribes a certain kind of *control* to the DMN whereas for Gerrans decontextualized processing is necessary to control the DMN.

I am reluctant to use the personal level heavily theoretically-laden term 'repression' to the DMN, but, what I agree with, is that the simplification of the parameters of the generative model, that might result from DMN activity, changes the contents of experience in virtue of the (causal) relationship between the generative model and the phenomenology of the agent. This is a two-step claim: first, that DMN activity changes the generative model in a certain way and second, that the generative model determines the phenomenology of the agent. The first step reflects the fact that predictive coding, as an *implementation*, offers a neurobiological model. Thus, changes in brain activity reflect changes in the generative model in general and changes in the DMN activity in particular (that have been related to mind wandering and dreaming). They have been argued to change a model in a certain way, namely to *simplify* it (Hobson et al., 2014). The arguments formulated for the second step so far, as noted above, have the form of an inference to the best explanation: if predictive

[375] Neural oscillations are argued to index the free-energy (Carhart-Harris & Friston, 2010). Interestingly, Shepherd (2014) argues that beta and theta band activity can be *controlled* by subject if feedback about it is provided (p. 12).

coding is true, and if phenomenology is necessary for certain kinds of actions, e.g. complex volitional (mental) actions, then some model or level of the model becomes conscious or acquires a certain phenomenology, e.g. presence, agency etc.

The central point about DMN as a control instance is that control need not be connected to *effort*: repression is often thought of as effortful, while mind wandering (which DMN activity might underlie, e.g. Schooler et al., 2011) is not. On the premise that repression and self-deception are conceptually overlapping (see section 1.3.1), I would refrain from the assumption that self-deception needs to involve effortful control, at least if DMN is involved. Wandering, on the other hand, exhibits *instability* (Friston et al., 2012) which seems to fit for the description of the brittle nature of self-deception and repression, e.g. that self-deceivers possess *fragile* self-esteem (see section 3.1.3.2). More on instability in self-deception in the next section.

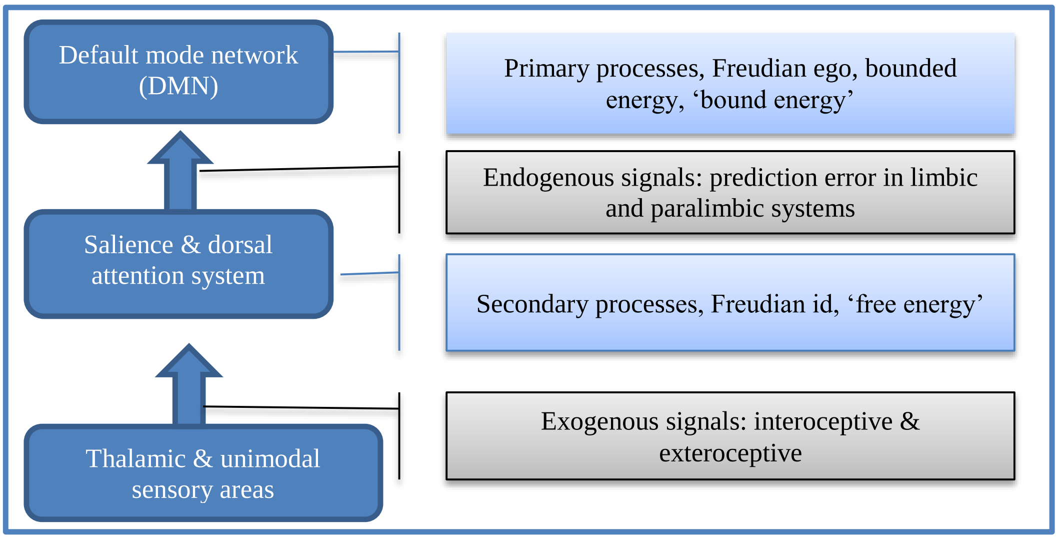

Default mode network (DMN)
Salience & dorsal attention system
Thalamic & unimodal sensory areas
Primary processes, Freudian ego, bounded energy, 'bound energy'
Endogenous signals: prediction error in limbic and paralimbic systems
Secondary processes, Freudian id, 'free energy'
Exogenous signals: interoceptive & exteroceptive

Figure 31. Carhart-Harris & Friston: Relationship between DMN, Attention and Sensory areas. Distinctions from Carhart-Harris & Friston (2010).

Precision is a mighty tool as I elaborated above. In order to use it, the first step is to ask when a conceptual analysis in terms of different precision expectations will be complete. I think, that it is the case when one has to clarify which object (priors, policies, models) in which direction (high or low) at which level (e.g., low/sensory, intermediate, high/cognitive), as well as in which way those precision expectations have been acquired in the first place. How quick they have been acquired could, for example, be a sign of how long those delusion symptoms will stay and, thus, say something about the stability of the phenomenon in question that is being explained by a certain kind of precision expectations. So, I will present accounts of delusion each of which highlights one or several of these aspects in order to introduce every of those completeness conditions.

Direction of precision (high vs. low): Hohwy (2013a) attempts to bypass the difficulties of the doxastic view of delusions[376] by providing a predictive coding account of how evidence

[376] Referring to Radden, Hohwy (2013a) points out that there are three major challenges to a unifying framework of delusions:

1. the categorization is difficult because of the heterogeneity of delusions (doxastic view vs. non-doxastic view);
2. challenge of not ignoring polythematic delusions (p. 58 - 61);
3. private nature of delusions – they oppose interpersonal standards of belief evaluation (p. 60).

is handled in delusions (p. 59). Delusions are beliefs (Hohwy, 2013a), and are subject to biases when they are updated (p. 64). For example, those experiencing *delusions of (alien) control* have the feeling that the movements that they have caused were not caused by them. One possible explanation is that sensory attenuation needed for active inference to take places does not happen so that, as a result, precise prediction errors – one's own executed movements – have to be explained (Frith & Friston, 2013). Delusion of control is a monothematic delusion, because it concerns a certain topic – the control of one's own movements. Davies & Egan (2013) have described monothematic delusions as "islands of delusion in a sea of apparent normality" . Polythematic delusions provide the contrast in that they encompass a wide variety of topics. A unifying notion of *perceptual and cognitive uncertainty* is argued to account for well-specified anomalous experiences of monothematic and ambiguous experiences of polythematic delusions (p. 59-60) in the following way: both monothematic and polythematic delusions are responses to uncertainty, but the *kind* of uncertainty, that they are subject to, differs. Monothematic delusions are driven by unusual sensory evidence which has *high precision* such that belief formation is determined bottom-up by sensory evidence. Polythematic delusions, on the other hand, are driven by ambiguous/hard to interpret evidence that has *low precision* so that belief formation is determined top-down by priors one already had (p. 65-68). Precision expectations, thus, determine the extent to which new evidence is incorporated:

> Thus, by wielding the different ways in which a Bayesian inferential system processes and responds to levels of uncertainty it is possible to identify different ways this system can be pathological. Polythematic delusions arise when *low precision expectations* cause one to be driven by poorly shaped prior expectations, and monothematic delusions arise when unexpectedly *high precision content* in evidentially insulated areas cause one to be in the thrall of spurious sensory input. (Hohwy, 2013, p. 68; my emphasis)

Self-deception has been argued to be similar to delusion that allows for a motivational explanation (see section 1.2.7). Polythematic delusions might be argued to be unrelated to self-deception: First, one can assume that motivational explanations are easier for monothematic than polythematic delusions, e.g. anosognosia being a paradigmatic self-deceptive delusion (see section 1.2.7). Second, self-deceivers exhibit, apart from the topic on which they are self-deceiving, rationality. This is why observers ascribe self-deception to them in the first place. Yet, the difference in the explanation of mono- vs. polythematic delusions is important, because self-deceivers' evidence may be of different kind – not only unusual, but also ambiguous. According to Büchel et al.'s (2014) model, the greatest amount of placebo analgesia should occur if both *top-down expectations of analgesia are high and precise*[377] (1); *bottom-up sensory signals (painful stimulus) are highly variable (low precision)* (2). Placebo has been argued to be a kind of self-deception by Trivers (see section 3.2.2.1). Büchel's et al account suggests an explanation of placebo as resulting from a certain balance of precisions (high top-down and low bottom-up). So, Büchel's et al account serves as an example for how a *combined* (both low and high precision, but at different levels) direction account for a conceptual description of a certain phenomenon, related to self-deception, might look like.

Interestingly, Hohwy (2013a) also argues that probabilistically independent sources of evidence might create evidential insulation, because they produce *contingent partitions in the flow of information* (p. 60). Good probabilistic inference gains more from having

[377] Opioids and the dopaminergic system change according to Büchel et al. (2014) the precision in virtue of being modulatory neurotransmitters (p. 1229).

independent sources of evidence than repeated sampling from the unreliable ones (p. 65). This is interesting because evidential insulation is Davidson's explanation of self-deception (1.1.1.1), which, if Hohwy were correct, Bayesian model of updating can incorporate.

Object of precision: Similar to Hohwy's (2013) is Frith & Friston's (2013) explanation of false beliefs which distinguish between two kinds of ways in which false beliefs might result:

- Too much weight on *prediction error* (evidence) which results in over-updating of beliefs
- Too much weight on *prior expectations* which results in under-updating of beliefs (p. 6).

Fotopoulou (2013b) also mentions, as possible explanations of anosognosia, difficulties in precision optimization, as well as the resistance of *priors about one's self-schema* to change in the face of missing prediction errors.[378] It has also been argued that delusion might develop because of *imprecise predictions* at lower levels leading to imprecision across hierarchical *interoceptive* levels:

> To the extent that imprecise predictions at low levels of (interoceptive) hierarchies are unable to suppress interoceptive prediction error signals, imprecise predictions will percolate upward, eventually leading not only to generalized imprecision across cortical hierarchical levels but also to re-sculpting of abstract predictive models underlying delusional beliefs. (Seth et al., 2012, p. 9)

Summing up, priors (about self-schemata or otherwise) and/or (interoceptive or otherwise) prediction errors, policies, even precision itself can be objects of precision. Since one way of changing the generative model is model comparison, those must possess weights too. This is a very broad basis on which one could construct a precision-based explanation for some phenomenon, e.g. self-deception. Given those different kinds of objects of precision and its dispersed nature in a sufficiently complex network, if some objects were part of others, e.g. prediction errors part of a self-schema, because the latter object is at a more general level of abstraction, then precisions for the former might contribute to the precision of the latter (this is what follows from the *dispersion* of precision). This is a word of caution: experimentally determined precision deviations of certain objects (e.g., self-schemata) need not be the *lowest* ones (e.g. more low level priors).

Level of the change of precision: In cases of *anosognosia* and *Capgras syndrome* prediction errors are precise, but false, both being a result of brain damage: right parietal damage and damage to amygdala and its connections respectively (Friston & Frith, 2013). Delusions are characterized by the feature that content is believed despite contradictory evidence. This feature is explained by the authors due to the beliefs at the *top hierarchy* being the least likely to change because overly precise bottom-up prediction errors are propagated upwards in the hierarchy (p. 10-11). Not only for anosognosia and Capgras syndrome, but also for *schizophrenia*, abnormal levels of precision, as well as distortions of precision updates, have been hypothesized as the cause (Fogelson et al., 2014, p. 8). This shows that a precision-based explanation of difficulties in Bayesian updating, in case of a certain

[378] Fotopoulou (2013b) argues for the following functional disruptions to play a role in anosognosia and explain the adherence to past expectations (Fotopoulou, 2013b, pp. 13-14):

- Lack of *active inference*, because affected limbs cannot move;
- Aberrant *perceptual inference* due to weak/absent signals about prediction errors;
- Compromised perceptual learning;
- Difficulties in precision optimization;
- Resistance to change of *premorbid priors* about e.g. one's self-schema in the face of missing prediction errors.

Since they are argued to follow from some sort of neurological damage, which is not the case in self-deception, I will not go into detail about them.

phenomenon, can be given. This explanation is so general though that *several* phenomena could be explained in that way. Such an explanation, thus, would fail to meet the demarcation criterion – how to distinguish this phenomenon from others? Surely, the *mechanism* need not be a demarcation criterion (the behavioral and phenomenological profile would be such as one in case of self-deception), but one still has to be aware of this caveat.

How is precision updated: Griffiths et al. (2014) argue in accordance with Fletcher & Frith (2009) that aberrant prediction error might be the cause of delusions. They emphasize that it is *associative* prediction error that controls attention (Griffiths et al., 2014, p. 9) such that "it predicts that participants' recent experiences with a cue will be the basis for future estimates of the consequences of that cue" (p. 19). According to associative learning theory, two events become mentally associated if one reliably follows the other, e.g. for a deluded person a bench may acquire a special meaning by being associated to the person sitting on it (Ibid., p. 4). I would like to note that here the direction of the causal relation is from prediction error (associations formed about causal relations in the world) to attention, which is the upshot of precision optimization, although it is the precision that determines the weight of prediction errors. Thus, updating is determined by precision and, in turn, determines the precision in the next step. Another example for such a circular role of precision is Mathys' et al. model. (Mathys et al., 2011) distinguish two different kinds of variables: *states* that change quickly and *parameters* (priors) that change slowly (p. 4) and argue for a Bayesian model that is equivalent to reinforcement models such that certain parameters "these parameters influence the precision of the prediction on the *next trial*, the precision of the posterior belief, and the learning rate" (p. 18; my emphasis). The learning rate is in this model is a weighing factor of prediction error as precision is (p. 8).

Summing up, I presented accounts of delusion that differ in the object, direction, level of change and the way precision is updated. In the next section I will focus on policies and models as objects of precision. Since I will argue that different kinds of self-deception depend on the idiosyncratic optimal degree of instability, the level of precision may vary and the same goes for how quick precision estimates are updated. I will not say anything about the level of change explicitly, though I assume that cognitive policies belong to a sufficiently high level of the hierarchy.

In this section I introduced different tools that predictive coding offers, e.g. set of attractors, generative model and the ways to change it, active, interoceptive and interpersonal inference. In the next and final section I will apply those tools to conceptually analyze a toy example of self-deception.

4.5 Application of conceptual tools extracted from predictive coding to self-deception

This section will attempt to conceptually analyze self-deceiver's behavior and phenomenology from the predictive coding perspective. First, I will describe the explanandum and the analysandum that correlates with it. Second, I will proceed to offering a new conceptual model of what is to be analyzed. For this purpose I will step by step re-introduce some of the new conceptual tools presented in the previous section. Last, I will apply those tools to enrich the explanation of self-deception. I propose to consider the following toy examples of self-deception:[379]

> A is motivated to believe X. She starts out the hypothesis testing process with the aim to find out the truth. Somehow, despite the unbiased evidence pointing into one direction, she ends up believing another conclusion because of *motivation* Y [the goal to believe this false conclusion/believe otherwise/relieve anxiety etc]. For that, she has to somehow misallocate *attention* to crucial pieces of evidence and experiences *tension* because of such misallocation, yet is driven to misallocate it nevertheless (*counterfactual goal-directed pull*). Upon relinquishing self-deception, A experiences insight: she has the feeling to have known the truth all along. When being questioned about believing X by observers, A would *justify* her belief, but at different time slices her behavior would be inconsistent from the point of view of the observer regarding the belief that X. It would be *inconsistent*. A central philosophical question then is whether the self-deceiver has *control* and/or experiences a sense of control over such misallocation, in order to be able to ascribe responsibility. The reader is encouraged to feel free to fill out X, Y and the concrete arguments with cases of her personal life.

> A soldier Ben has been lethally wounded. In the last minutes of his life he is making plans for an upcoming vacation [construction of an epistemic agent model with planning content]. He feels extreme distress upon trying to inspect his wounds or even think about them [tension condition]. Instead, he is immersed into the trail of thoughts about the vacation and sun and sea [counterfactual goal-directed pull] such that, from time to time, he feels like smelling the sea and feeling the sand under his feet [the goal component of the epistemic agent model becomes transparent – it is a world model now].

The application of the predictive coding tools will proceed in such a way that I first sketch how those can be applied to explain traditional examples of self-deception in which there are only certain changes in the epistemic agent model and then suggest how the second toy example might be analyzed when considering those tools. The first example is actually a schema (stands for a set of examples) along which one usually construes standard examples of self-deception. The second example is very simplistic: there is a planning epistemic agent model, not a reasoning one and the interpersonal aspect is weak (no observer would in this context argue with Ben over the fact that his planning is unrealistic). The focus is, thus, on the oscillation aspect between the two kinds of selection. To remind the reader, some authors in the self-deception literature (in virtue of the difficulty of ascribing a clearcut belief to self-deceivers) have favored the view that self-deceptive attitudes oscillate in the degree of certainty (see section 1.2.2). Michel (2014) has even argued that the constancy of our beliefs is an accident property thereof: at each moment in time attitudes are constructed from the evidence available at that time and, luckily, they often stay constant (see section 1.2.3 for more on his view). I have argued thereafter that it is not

[379] The before-death example was suggested by a friend who argued to have seen it in an unknown movie.

only certainty that might change as a property of self-deceptive beliefs, but also transparency (Pliushch & Metzinger, 2015). Here I will argue for an *overtone model of self-deception*. Let me explain in which way the overtone metaphor applies to self-deception. I borrowed the term 'overtone' from music where it denotes an additional frequency of a tone beyond the base frequency. Each tone that you hear, e.g. a music instrument playing or a cup breaking, has an intensity, as well as one or several frequencies (the lowest being called fundamental). [380] Those frequencies determine how the tone sounds (its timbre) and can be modelled as sine or cosine waves. Such wavefunctions (several added sine and cosine waves) describe an *oscillation* (how much it oscillates in each direction is the amplitude and how long a wave is – how long one has to wait until the same value will be repeated – is the frequency). The analogy to self-deception holds as follows: I am going to argue in this section that self-deceiver's optimal degree of instability (which determines how much exploration is pursued, in order to preclude overfitting of a model) is heightened so that constant exploration (of a certain number of hypotheses) is pursued at the cost of disambiguation in favor of a certain hypothesis.[381] These hypotheses can be explored in alternation, or in parallel and can be also recombined or overlayed in any manner one over the other. These hypotheses are like overtones of the currently active self-deceptive hypothesis (the base frequency) so that what we as self-deceivers or also observers perceive as one tone (self-deception) is actually a merge of different frequencies. In application to the second toy example, an epistemic agent model of an agent directed at certain contents, e.g. vacation planning, has to be overlayed by the model where that goal component is modelled as already being achieved and part of the active world- and self-model. At some point the base frequencies switch (this means that overtones also change) and the agent models itself as on vacation. One could argue that in this case there is only a switch and no overlay, but consider the traditional cases of describing self-deception as "knowing deep down that truth". For example, if a friend were to say: "I know that we broke up but I don't want to call him, because it really will be over (I would know for sure)." The tools that a rational analysis would provide us with would not suffice to analyze such an example, because of the logical impossibility of a contradicting (knowing and not knowing of any sort). Yet an agent might even reflect on several models of reality that she is possessing at the same time – they are somehow connected with each other, overlayed by each other such that the constant in both is the agent, but with changed thoughts and feelings depending on the accepted reality model. The overtone metaphor might be particularly well suited for this set of self-deceptive examples, but also the consideration of the simpler cases, where there is an alternation, might profit from the overtone metaphor when the question is considered, why self-deceivers do not actually notice the change in their models. My answer would be that those are gradually overlayed, instead of switched from one to the other. Furthermore, overtones are different depending on the *kind* of instrument being played, hence the richness and shrillness of the sounds also change from

[380] For more on overtones see http://www.britannica.com/science/overtone, accessed on 6.11.2015.

[381] My idea bears a certain resemblance to the one tension between goal representations can be creative and be upheld, instead of resolved: "Often, one does not (and perhaps cannot) seamlessly meld the two original conflicting goals into a unified higher-level goal, but rather holds them in creative tension with each other – both goals remain, and continue to pull in opposite directions in many instances, and one is simply required to decide one way or the other in any particular circumstance. Our view of cognitive coherence as a *defeasible* rational requirement (McIntyre 1990) enables us to permit this approach, and we argue further that in some cases this creative tension presents a preferable solution to integration." (Saunders & Over, 2009, p. 328)

instrument to instrument (and from human voice to human voice). In analogy, self-deception is idiosyncratic so that whether and how it develops depends on several factors, such as the phenotype of the agent and his personality traits. Furthermore, I argued that self-deceptive representations might possess several properties: degree of consciousness, certainty and transparency (see section 2.2.3). Those do not need to oscillate the same way, so the overtone models allows to describe those as different kinds of oscillations where one of them takes the lead.

What is to come in this section is that, after I have summarized the results of my conceptual analysis from the previous chapters and the tools, I will first focus on exploration as an alternative to disambiguation and then go through the alternatives that one could explore and whose contents could switch: epistemic agent models and/or world/self-models. With respect to epistemic agent models, I will argue that mental action becomes causally efficacious in virtue of interoceptive counterfactuals and will describe it in analogy to perception as a trajectory, as well as argue for traditional cases of self-deception being a case of trajectory *merging* with a post-hoc sense of control. With respect to world/self-models, I will apply the idea that exteroceptive counterfactuals influence the degree of transparency (Seth, 2014) to self-deceptive attitudes that vary in transparency. For that I will argue that *exteroceptive* counterfactuals include also modelling of *interoceptive* counterfactuals - the controlling element, as for epistemic agent models. Then, I will proceed to describe *tension* as depending on the interoceptive counterfactuals richness, more precisely as depending on the *kind* of counterfactuals (goal-conducive or not) being generated and not only on their *amount*, as it is the case for transparency.

The following conceptual analysis has been worked out in the previous sections: Self-deceivers behave inconsistently, but justify their behavior, while their phenomenology is characterized by tension and insight (section 2.1.2). Tension might be characterized as a kind of metacognitive feeling that indicates that certain rules of the process have been violated, given the goal representation guiding that process (section 2.1.3). Inconsistent goal representations (find out the truth, but also preserve certain views) might also lead to a similar tension. Thus, self-deceptive tension might be a complex feeling generated by multiple causes. When self-deception is relinquished self-deceivers experience insight. From the fact that self-deceiver's behavior is inconsistent it can be followed that some changes either in the epistemic agent model or in the world/self-model have to take place (section 2.2.1; see also figure 32). This is the case on the premise that those models determine the behavior of agents. This premise can be followed from predictive coding by the two-step argument. The first step is that the generative model determines action, what it does via active inference as testing of the model. The second step is that the generative model or one of its higher levels are causally related to the phenomenological models that an agent possesses, which is an inference to the best explanation. It is furthermore conceivable that self-deception might be sustained by different kinds of selection. If this were the case, then the alternation may be not only from one epistemic agent model to another, or from one world/self-model to another, but also crosswise.

These two different kinds of selection generate different kinds of attitudes (section 2.2.3). Selection is a means of achieving control over something. Whether control is experienced or not is then the question worth exploring. The latter question is particularly important for the first kind of selection, since I assume that agents seldom feel control over the generation of the transparent self/world-model. Self-deceiver's epistemic agent model is best described by means of the dolphin model of cognition as cognitive binding of different strands of thoughts (section 2.2.2.3). This claim is based on an important premise that conscious thinking is a subpersonal process in that it is the result of some subglobal physical dynamics that may be integrated into the epistemic agent model (Metzinger, 2015). An epistemic agent model is the conscious representation of the system as an agent executing epistemic actions, e.g. direction of attention or controlling strands of thought. So, there may be several subpersonal processes going on and several ways in which they are integrated into the epistemic agent model. Thus, the epistemic agent model may contain gaps, switch its object and, particularly in the last case, calling it a unified self-deceptive process might be too farfetched. For those readers wondering how the dolphin and the overtone metaphor relate to each other, I note that the dolphin metaphor is best used to imagine the embedding of thought processes into the EAM, which for the agent follow serially in succession on the phenomenal level. The overtone metaphor, on the other hand, is best to be used when imagining the subpersonal workings of those processes that, when embedded, we perceive as strands of thoughts.

There are following possibilities for self-deceivers to experience control over their epistemic agent models. First, even if self-deceptive attitudes only popped-out in the head of the self-deceiver without the self-deceptive process being available on the personal level at all, when those attitudes (that are not transparent but may possess the transparent signature of self-knowledge) are justified, self-deceivers might still experience a post-hoc feeling of control over a process they were not even aware of. My argument for such a post-hoc feeling of control is the following: there are three options – that self-deceiver's attitudes are accompanied by a feeling of control (1), that they are accompanied by a feeling of an attitude being externally generated and implanted into one's head (2) or that there is neither the feeling of control nor its opposite. The second case is a case of delusion that is not characteristic of self-deception, because else self-deceivers would be unable to defend their self-deception and argue for it and, thus, would not be self-deceived in the first place. I prefer the first case to the third one, because all attitudes integrated into the epistemic agent model are accompanied by the feeling of control. It is the central feature of the epistemic agent model that it represents the agent as being in control over performing certain mental actions. If such an attitude has been generated by some possibly subpersonal processes so that not the process, but *only* the resulting attitude has been integrated into the EAM, then the agent would still experience control over it for the same reason that sliding into mind wandering goes unnoticed most of the time: it is not only the case that there is such a thing as EAM, but also that an agent experiences the *continuity* of one EAM being followed by another. The case EAM-gap-EAM is not part of our phenomenology. If only a limited amount of our attitudes has been generated explicitly via an EAM, then if we would not experience control over all the others attitudes when they are embedded into the current EAM in order to be acted upon, then there would be agents who acknowledge that they are acting upon something that they do not have an exact idea where it came from, or would be forced to rethink each attitude over again once it popped into their mind. In the latter case they would never come to act, so it would be also evolutionary not a very efficient strategy. The central philosophical point here is that the feeling of control does not require explicit argument construction. As Proust (2013) holds, metacognitive feelings might lead to the sense of agency too, along with explicit argument construction (see section 2.1.3).

Second, suppose that the self-deceiver experiences counterfactual goal-directed pull (section 2.2.1) upon consideration of certain pieces of evidence. This pull can have either the standard phenomenological description in terms of tension (sense of uneasiness and distress) upon changing focus from/to certain kinds of information or this pull can also be described as a feeling of being attracted to something, or compelled to dwell upon something (remember Friston's claim that a policy as a sequence of control states implicitly leads to the sense of agency and the sense of agency is a certain kind of control). Enriching the phenomenological profile of the self-deceiver this way may then provide demarcation criteria for distinguishing different kinds of self-deception.

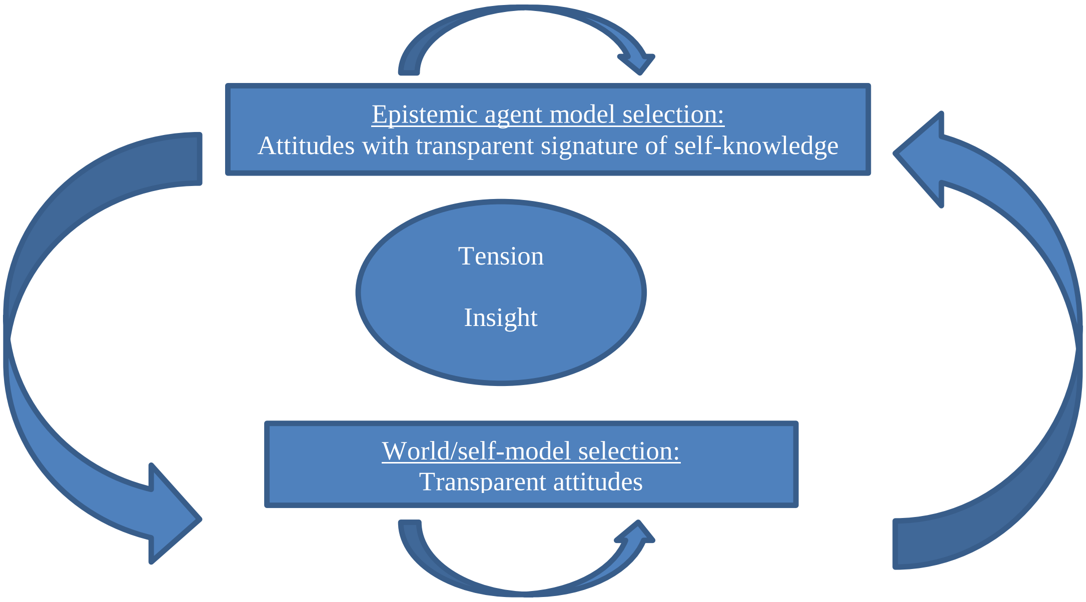

Epistemic agent model selection:
Attitudes with transparent signature of self-knowledge
Tension
Insight
World/self-model selection:
Transparent attitudes

Figure 32. Phenomenology of the self-deceiver
The arrows denote changes of the model due to changes in various aspects of selection.

Second, the conceptual tools of predictive coding introduced in the last section can be integrated as follows:

- Set of attractors (limited states an agent visits) such that free energy level determines which states are valuable, with value being just free energy reduction
- States of the generative model: control states, counterfactuals, precision (direction, object, level, updating procedure),
- Procedure that changes the generative model: inference (priors*likelihood=posterior), (parameter) learning, model comparison, precision estimation
- Inference: perceptual, active (goal-directed), interoceptive, interpersonal

Idiosyncratic set of attractors: Before I explore self-deception as continued exploration hypothesis, there is a metatheoretical comment to make. In previous chapters I have introduced a lot of different theories characterizing self-deception, the main flaw of which was their generality claim (to explain self-deception as a whole and not its specific types). Such an endeavor is difficult, because one counterexample suffices to destroy a theory with such an explanatory scope. There were those who, on the other hand, were arguing for the idiosyncrasy of the view of each agent with respect to another (Clegg & Moissinac, 2005). What comes for granted here is the explanation of the difference between self-deceiver and observer on a certain matter. I do not want to pursue neither one nor the other extreme, but offer a more fragmented take: several behavioral and phenomenal constraints can be set on self-deception and each of them can have its own description. Depending on the preferred subset of the reader, the description of the case of self-deception that the reader would like to have is then achieved by her putting together the puzzle pieces that she thinks belong together and whose description is offered in this section. Here the motivational power of free energy reduction, which I described in the previous section, comes in handy. First, a self-deceiver, like any agent, minimizes free energy by occupying a limited set of states[382] (attractors). The choice of those states is, at least partly, a matter of priors that might have been acquired in the course of the evolution (see previous section) and, partly, reflect the idiosyncratic motivational drive of the self-deceiver. The motivational drive of the self-deceiver – to minimize free energy in one way and not in another – is the result of the way her generative model looks like, e.g. changes in states, parameters (e.g. priors) and precision. According to interpersonal inference, personality traits are priors that influence the chosen action policy. One the assumption that self-deceiver differ from observers in the possession of certain personality traits and that those are priors (ergo, those are parts of the generative model), those also influence the way self-deceivers minimize free energy. This is to say that in every single case different parts of the generative model of the self-deceiver might lead to self-deception, because it is free energy reduction per se that is motivation, not the *way* it is achieved. So there is a lot of freedom in explaining self-deception.

Nested phenomenal models: Not only might self-deceivers' models look different, but they are also nested – they are connected in a certain, probably hierarchical way with each other. As a result, complicated patterns of how and which kind of precision expectations are possible. Think about the dispersed nature of precision argued for in the previous section, as well as the *balance* of precision expectations being important, not only a certain degree of it. The latter has been an issue not only in the explanations of delusion introduced in the previous section, but also the description of active inference in general: one could compensate bottom-up error imprecision with top-down prior precision, but at a cost. Whether self-deception is such a cost and whether it involves such a compensation, or just a certain pattern of precision expectations and if yes, at which level – these question cannot be determined a priori and, thus, I will not explore them in this thesis. But the general pictures of nested models and distributed precision is important for the reader to retain, in order to broaden the horizon on future (more specific) hypothesis about self-deception that might be tested. So I would like to introduce my take on how the epistemic agent model and the world-model might be hierarchically connected, so that I can start describing the exploration of hypotheses/models and the phenomenological characteristics of self-deception by means of different relationships in those models.

According to predictive coding, an agent possesses a *generative* model of the causal structure of its environment (Friston, 2009). Generative models are part of the computational level description. On the phenomenal level though, another kinds of models have been argued for: phenomenal mental models or "those mental models which are, functionally speaking, globally available for cognition, attention, and the immediate control of behavior" (Metzinger, 2003, p. 2010). It is possible that those two are formally equivalent (Hobson et al., 2014) and also that phenomenal models are also Bayes-optimal,[383] yet as such they belong to different levels of description. Having said that, a phenomenal self-model (what we experience as the self) includes a world-model such that this self-model is embedded into it (see figure 33). This is because from the observer perspective, a certain agent is only one small part of the world, but from the perspective of that agent, the world-model is being part of the self-model in virtue of the agent modelling the relation of being situated in the world. This is to say that an agent models itself as part of the environment with which he interacts. An epistemic agent model is a phenomenal

[382] "Crucially, the emergent dynamics are consistent with active inference, in which internal states couple back to hidden external states to (apparently) preserve the attracting set." (Friston, Sengupta et al., 2014, p. 439).

[383] Minimal phenomenal self has been argued to be a Bayes-optimal region of maximal invariance (Metzinger, 2013b, p. 3) and intuitions to be Bayes-optimal (Metzinger & Windt, 2014).

model of the intentionality relation and it, in its turn, is embedded into the self-model in the same sense that the self-model is embedded into the world-model: without the distinction between self- and world-model there would be no self experience (see 3.1.4) and the same goes for the epistemic agent model, namely that as it is an *agent* model there is an agent which is a transparent self-model that is being experienced in control of something, e.g. epistemic actions one executes.

Listing the properties of each of these kinds of models and their relation to each other is decisive in explaining the claim how epistemic agent model content can switch to become world-model content. This is because the latter is transparent and because the relations to the self-model changes: the possibilities and ways to act on that content change. I offer to call this the *transparency oscillation* claim. To change a thought, one employs other strategies than those of changing the environment. One experiences further more control with respect to one's own thoughts (with notable exceptions such as thoughts about the problems one is having).

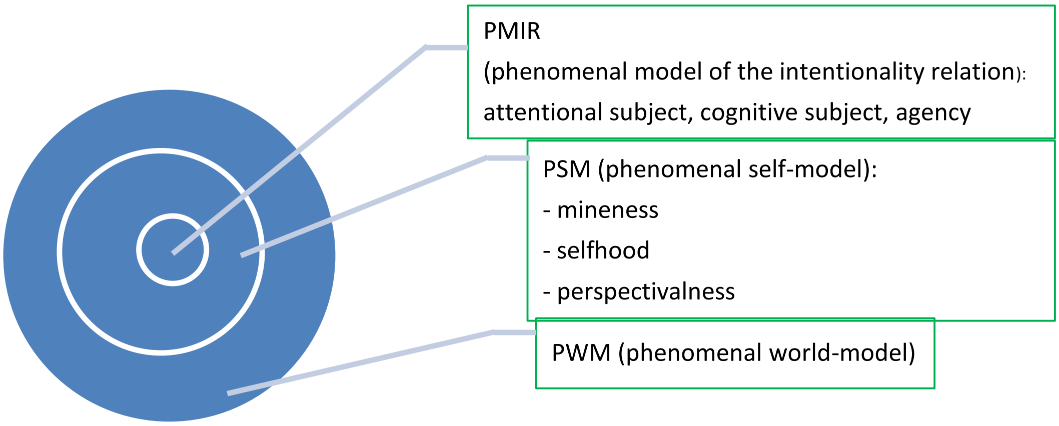

PMIR
(phenomenal model of the intentionality relation):
attentional subject, cognitive subject, agency
PSM (phenomenal self-model):
- mineness
- selfhood
- perspectivalness
PWM (phenomenal world-model)

Figure 33. Relationships between phenomenal models.
Distinctions from Metzinger (2003).

Phenomenal mental model is a mental model that is globally available for cognition, attention and control of behavior (Metzinger, 2003, p. 210). It is characterized by *mineness* (are experienced as owned by the agent), *selfhood* (are experienced as diachronically prevailing) and *perspectivalness* (experienced from the first-person perspective).
Below is an example for the first two characteristics:

Mineness: e.g., "I always experience *my* thoughts, *my* focal attention, and *my* emotions as part of *my own* stream of consciousness." (Ibid., p. 302)
Selfhood: e.g., "I experience myself as being *identical* through time." (Ibid., p. 302)
A *first-person perspective* is characterized by a transparent phenomenal self-model (PSM) and a transparent PMIR (Metzinger, 2003, p. 572).

Phenomenal model of the intentionality relation (PMIR) is a conscious mental model whose content are subject-object relations, e.g. "I am someone currently grasping the content of the sentence I am reading." (Metzinger, 2003, p. 411) PMIR includes *agency*, which is an experience of making a possible self-state part of the phenomenal world-model (Ibid., p. 408). There are two types of the PMIR: the attentional and the cognitive subject. Attentional and cognitive agency are two ways in which an agent controls (*mental autonomy*) one's own mental function (Metzinger, 2013a).

Attentional subject: a conscious self-model that experiences itself as causing shifts of attention (Ibid., p. 391)
Cognitive subject: a conscious self-model that experiences itself as selecting cognitive contents for subsequent processing (Ibid., p. 406).

Constant exploration as alternative to disambiguation: My take on the explanation of content shifts between the epistemic agent model and the world model is to use an analogy to a phenomenon from perceptual experience - binocular rivalry. For that, I first need to introduce another kind of description for the models being constructed by the systems, namely in terms of hypotheses.[384] Predictions about the causes of sensory input can also be characterized as hypotheses, in analogy to scientific hypotheses – both need to be tested (e.g., Seth, 2015b). Thus, if self-deception can be explained by predictive coding, it is trivially a kind of hypothesis testing, yet not a personal level one like the FTL model (see section 4.2). (Seth, 2015b) argues in accordance to the hypothesis testing analogy that active inference can have three aims with respect to those hypotheses: confirm, disconfirm or disambiguate between alternative hypotheses.

Binocular rivalry is characterized as rivalry between competing, alternating perceptual models from which one is experienced at a time. One image is shown to one eye and another image – to another eye. Due to such a procedure, a condition is evoked when for an agent it looks like both objects occur at the same spatiotemporal location (Hohwy et al., 2008, p. 688). Because of the prior that there is only *one* object per spatiotemporal location (1), as well as the presence of prediction errors when either model is chosen, e.g. that one sees a house or a face (2), there is alternation of models about the causes of visual input, e.g. that there is a face or a house in the vicinity (Hohwy et al., 2008). The alternation is achieved by model *selection*.

There are two reasons why I do not dismiss the idea that the same way binocular rivalry is explained in terms of predictive coding can be applied to self-deception. First, distinctiveness in *experience* does not entail distinctiveness in *process* (Duncan & Barrett, 2007). From the fact that epistemic agent models lead to a different kind of experience it cannot be followed that the process by which those can be explained is different from the perceptual one. Second, though self-deception is a higher-order cognitive process and binocular rivalry a perceptual process, both can be seen as construction processes and both may be *motivated* (Zadra & Clore, 2011 cite evidence that in binocular rivalry a more emotionally significant image is perceived longer). Third, the effects of motivation may even be *similar* on perception and cognition, e.g. Ebbinghaus illusion and semantic priming are reduced by sad mood (Zadra & Clore, 2011).

The application of the binocular rivalry analogy to self-deception looks as follows: Both are characterized by an alternating pattern of hypotheses that are provided by the system as an explanation of the sensory input. Both hypotheses cannot be true at the same time (for logical reasons in the case of self-deception and empirical or evolutionary priors in case of binocular rivalry). In case of binocular rivalry, different models explain different sensory prediction errors. In case of self-deception, sensory prediction errors seem to point towards one hypothesis and the motivation towards another.[385] To return to the self-deception example presented above, proprioceptive and interoceptive input suggests that one is dying, but strong motivation to stay alive leads to the construction of plans for the future. A model with a minimal free energy is chosen, but, as described in the last section, there are several ways to minimize free energy. To remind the reader, there is a circular dependency of goal representations and perception in predictive coding by which optimism bias has been explained.

In the way that binocular rivalry and self-deception minimize free energy both may deviate. This is where the comparison with another phenomenon that is characterized by an exploration of hypotheses might be helpful, namely mind wandering. This is because mind wandering has been argued to result from the *optimal degree of instability*. Dreaming and

[384] Since it is another kind of *description*, I did not put it onto the list of explanatory tools in the previous section.

[385] In an extreme case, one can construct a moderate version where there is no such unidirectional discrepancy between sensory input and motivation.

mind-wandering are produced by the changes in internal dynamics of brain activity in order to reduce the complexity of the model or promote exploration:

> In brief, if neuronal activity represents the causes of sensory input, then it should represent *uncertainty* about those causes in a way that precludes overly confident representations. This means that neuronal responses to stimuli should retain an *optimal degree of instability* that allows them to explore *alternative hypotheses* about the causes of those stimuli. (Friston, Breakspear & Deco, 2012, p. 3; my emphasis)

In other words, in order not to miss something important (not to be too fixed on a certain hypothesis), uncertainty has to be preserved to a certain degree. Here I will compare what optimal degree of instability leads to optimism, binocular rivalry, mind wandering and self-deception: in optimism the degree of instability is very low (there is almost no exploration of alternative hypotheses), in binocular rivalry the degree of instability leads to a switch between a limited amount of hypotheses (two), in mind wandering there is a flow of hypotheses that is not restricted in this manner. I think that at this point one need not restrict self-deception to a certain pattern of how the optimal degree of instability operates there, but be aware that depending on the idiosyncratic self-deception one is having, at least one of these three cases is possible.

In case of binocular rivalry, there is uncertainty because of contradictory sensory prediction *errors* (bottom-up). As for exploration, there is uncertainty of *predictions* (top-down). Simplistically, one might consider that the first is outer and the second – inner uncertainty. In case of self-deception, outer uncertainty (ambiguity) has been postulated (see section 1.3.2.3), but there seems to be less of an inner uncertainty, e.g. in the case of optimism. To remind the reader, optimism has been explained in terms of self-deception as a case in which control states (beliefs about future actions) are overly precise and influence the transitions between hidden states (assumed states of the world). Thus, in case of optimism it is argued to be *high* precision (low uncertainty) of control states. On the other hand, optimism can be seen as occupying one end of the *optimal degree of instability* scale by which self-deceivers may be characterized. In case of alternation of competing hypotheses, there is more instability than in the case of optimism. A certain degree of instability may have as a result that instead of hypotheses being *disambiguated*, ambiguity is *preserved*, to enable the switching characteristic of at least some cases of self-deception. Instability per se only says that there will be changes, but in difference to dreaming or mind wandering, in case of binocular rivalry and self-deception there is *constancy*: the exploration does not drift too far away, but several models keep returning. Thus, to explain self-deception one needs not only to explain what leads to an enhanced degree of instability, but also what keeps the way that instability is being resolved over time constant. A change in estimates of uncertainty over time is *volatility*. In the previous section I reviewed that volatility can be applied to different objects, e.g. priors, policies, models themselves. I think that from the arm chair it would be premature to commit oneself to one certain object at this point, but let me show which effects periodic changes in volatility might have on the example of optimism: when precision of control states is high, perception is less veridical, so if precision of control states alternates from high to low, then constructed hypotheses will be more or less veridical depending on the amount of precision of control states. The main insight from predictive coding on self-deception so far is that it need not be external evidence that leads to the alternation of explanatory models, but that a high level of exploration is kept in order to ensure that ambiguity prevails. This is the *optimal degree of instability*. This is incongruent with psychological findings that self-deception supposedly correlates with openness to new experiences (one of the five personality factors; see section

1.3.1). To pin down the results so far, I have argued that in self-deception the optimal degree of instability is such that there is more exploration than disambiguation. Self-deceivers, thus, seem to be open, but not open enough – they are not open to acknowledge their self-deception and experience insight upon relinquishing it.

Insight: In terms of the overtone theory of self-deception, the prior-post insight difference in the cognitive situation can be described as a change of the base frequency. As long as the self-deceiver's epistemic agent models are determined by the overtones of a certain self-deceptive hypothesis (base frequency), there is exploration, but it is bounded by those overtones. Insight experience is the experience of the change of the base frequency. One can note that the implicit premise in describing self-deception as repetitive hypothesis exploration is that at the phenomenal level the subtle changes of the hypotheses being explored are *not* experienced. I will talk about different kinds of blinks – descriptions of phenomena in which certain things are not experienced – later on in this section. So the standard case is experiencing our thoughts and models of reality as *one* coherent stable entities, despite the subtle and not so subtle changes in trajectories that might ensue. Insight is a case when the change is experienced. It is this *openness* that I think precedes insight – the second phenomenological characteristic of self-deception that follows when self-deception is abandoned. Conveniently, the value of a policy can be decomposed into *extrinsic* (utility) and *intrinsic* (exploration) reward (see previous section) so that when utility of a policy drops, exploration kicks in, an alternative might be explored that contradicts the policy pursued before and insight might occur. The central question is now why exactly such a policy is explored instead of uncertainty of a policy (that underlies exploration) being used to uphold the self-deceptive cycle, analogue to binocular rivalry. To remind the reader, I have above argued that a certain degree of instability responsible for exploration enables self-deception in the first place. My hypothesis is that both acquiring and relinquishing depend on changes in precision expectation, while determining the kind of these changes is left for future research. Worth mentioning is that such relinquishing need not be any effortful activity. Let me draw an analogy to intuitions. I have mentioned the view before that intuitions involve a transparent signature of self-knowledge (see Metzinger & Windt, 2014 and section 2.1.1). My point at that time was to introduce the possibility that certain self-deceptive attitudes might possess a transparent signature of self-knowledge – one would be sure about them, but would not be able to explain step by step how one came to such a conclusion. Feeling of *certainty* is a connecting link between intuitions and insight (that insight involves a strong feeling of certainty see Picard, 2013). One could say that uncertainty leads to openness to new experiences and interpretation. The current point is that to change someone's intuitions one needs to distract attention from the context about which intuitions are to be changed (Weatherson, 2014, p. 526). The personal level control is not involved. Quite the opposite, one needs distraction. Insight acquired after self-deception might also be the result of some subpersonal frequency changing. For the reader to keep in mind is the antagonistic role of the transparent signature of knowing in upholding self-deception on the one hand and insight in abolishing self-deception on the other.

Epistemic agent model selection: An epistemic agent model is a transparent representation of the agent as establishing knowledge relations (Metzinger, 2013a). An epistemic agent model is characterized by attentional (control of attention) and cognitive (control of deliberative thought) agency (Metzinger, 2013a) and is a representationalist description of mental agency. A mental action as finding a selected (intended) path in the behavioral space with multiple inputs - memory contents - and outputs - possible kinds of behavior (Wu, 2013). Personal level hypothesis testing, e.g. weighing the evidence in favor of one or the other hypothesis is a case of a construction of an epistemic agent model. Planning, as in the self-deception case presented above is another such case. Most traditional examples of self-deception present it as a skewed epistemic agent model construction, e.g. weighting the evidence falsely in coming to a conclusion that one's child is innocent of a crime, one's spouse is not cheating or one's own personality characteristics are more favorable than they are so that one starts with the aim to find out the truth and ends up with an improperly constructed argumentation that for observers looks like it has the aim to come to a desired conclusion, instead of an impartial one (see previous review chapters). I think that human reasoning is at best full of gaps: thoughts popping out, skipping (e.g., think about trying to understand a proof of a theorem given by a skillful mathematician – she will do three steps at once that need to be chewed through for you without noticing it) and merging. As an example of the latter, think of an argumentative fallacy of using a certain concept in one way for the first half of an argument and then switching to another one of its connotations in the second half. Here, the ambiguity of a concept allowed to misuse it. Likewise, our argumentations are dolphinlike trajectories in the argumentative space that at certain points can merge and change into one another. This happens because of the ever changing precisions of pursued policies.

The development of my own positive model of traditional self-deception (the one concerning only changes in the epistemic agent model and not the switch between it and the world model) will proceed as follows: First, I will argue that mental action becomes causally efficacious in virtue of interoceptive counterfactuals. Second, I will describe it in analogy to perception as a trajectory and argue for traditional cases of self-deception being a case of trajectory *merging*. Third, I will emphasize that that motivated attention is a result of volatility and precision influencing both our policies and our model of the world, as well as the circular relationship between both. Fourth, I will argue that neither the presence nor absence of the sense of control is characteristic for self-deception.

The first argument goes along the next lines. In predictive coding there has been no additional *cognitive* inference postulated. Is then cognition a kind of interoceptive or perceptual inference? This question makes sense only if the two operate independently of each other. In the following I will argue that it is not the case. If there are two kinds of inferences in predictive coding, a perceptual and an interoceptive one and if in self-deception both are at work, then one important question is about how the two interact. I presume that postulations of different (perceptual and interoceptive) systems would be a result of *modularity* intuitions: either information being processed in a different way in those systems or being encapsulated. This question is also analogical to the one that I posed with respect to precision in the previous section: if *precision* is dispersed across the hierarchical network, and attention is an emergent property of precision, why should be there an additional attentional *system*? I think that perceptual and interoceptive inference are interdependent and cannot be modularized. On the one hand, as with the attentional system, one could postulate the presence of a certain system because there are (circumscribed) neural areas focused on a certain kind of processing (e.g., limbic areas in case of emotions), but as in the case of precision, when one looks at how the different states are connected in the (generative) model, which is a network containing several levels in which units are connected in a certain way where each of these units computes hidden states and causes that are connected yet in another way (see the basic structure of the generative model in the last section), then the most probable hypothesis is not to try to modularize, but to determine the degree of influence for each phenomenon separately. One consequence of this hypothesis is the close link between affective and cognitive states, for which I have already given some evidence in previous chapters.

Actions generated by active inference have an effect on the environment via classical reflex arcs. What about mental acts?[386] My answer is that since there are exactly two ways in which control can be executed – allostatic and autonomic (both operating by means of counterfactuals representing the consequences of possible actions; see previous section), it is autonomic control by means of *interoceptive* counterfactuals that makes mental acts causally effective, because it cannot be allostatic control, since the agent is not action on the environment. I take interoceptive counterfactuals to be those encoding the affective consequences of an action, in analogy to perceptual consequences encoded in exteroceptive counterfactuals.

A short digression into what my results concerning affective states in self-deception have been so far: I have distinguished four functions that tension (and affective states more generally) could play in self-deception (see introduction into chapter 4). It could:

- cause self-deception and result out of self-deception because of inconsistent attitudes, e.g. believing that my arguments in this thesis are sound and that they are not quite up to the standard would be two dissonant attitudes that would stay in the way of me believing that I have written a nice thesis;
- arise out of a self-deceptive process and influence that said process, e.g. while in possession of a certain epistemic agent model (thinking through an argument) I could have the feeling that some premises do not validate the conclusion that I make, which might then lead me to change the conclusion, or the way I actually validate the conclusion might be already the result of the influence of those affective states.

I did not discard any of these functions, but argued, relying on Proust (2013), that in the third case (tension arising out of a self-deceptive process) it is a metacognitive feeling fulfilling an indicator function. I further extended Dokic's (2012) view that one could describe the *degree* of a metacognitive feeling modally as depending on the amount of possible worlds in which a certain mental action would be successful, e.g. being certain that one would remember something, to the hypothesis that the *kind* of counterfactuals might determine the metacognitive feeling (tension, intuitivity or counter-intuitivity) that one might be having (see section 2.1.3). I will elaborate more on this below, but now I will present an argument for the claim that epistemic agent model selection is trajectory switching or merging that is heavily relied on interoceptive states and counterfactuals. Let me call this the *interoceptive trajectory merging* claim. As the first step, I will argue that mental acts become causally effective in virtue of interoceptive counterfactuals and then sketch how trajectory switching comes into this picture. Note that this claim is different from the one about tension in that it is more general: it is not about how and which counterfactuals determine the experience of a certain feeling during a certain process – self-deception, but more fundamental about how all kinds of mental actions might become causally effective. So, right now I am laying the foundation for the tension-hypothesis. Further, in difference to tension as metacognitive feeling, the function of affective states that is being elaborated is not the third, but the fourth one, namely the influence of affective states onto the process. To remind the reader, that affective states can control how the acquisition of cognitive attitudes is accomplished, is exactly what I appreciated in Sahdra & Thagard's connectionist model of self-deception (see section 4.1). To recapitulate, a concise version of the argument for the hypothesis that mental actions become causally effective in virtue of interoceptive counterfactuals (let me call this the *causal efficacy* claim) looks as follows:

1. There are exactly two ways in which control can be executed – allostatic and autonomic.

[386] This question was posed by Wanja Wiese during our discussion of our poster about mental agency and predictive coding we presented at KogWis2014.

2. Since the agent acts on its own mental states, instead of the environment, it is not allostatic control by means of which mental actions can become causally efficacious.
3. Autonomic control operates by means of *interoceptive* counterfactuals, in analogy to allostatic control operating by means of exteroceptive counterfactuals. For allostatic control this means that proprioceptive and kinematic consequences of actions are represented and then fulfilled via active inference. I extend this by arguing that during autonomic control then counterfactual affective consequences must be represented and fulfilled.

Now I will on the basis of the causal efficacy claim argue for the second part of the interoceptive trajectory merging claim, namely that epistemic agent model selection can be understood as trajectory merging. This includes answering the question why such merging goes unnoticed. My answer will be that on the premise that trajectories possess valence and intensities (the two dimensions of affective states), they acquire a property of resolution through it which not only explains the unobtrusive way of merging, but also the counterfactual goal-directed pull (another phenomenological characteristic of self-deception: being pulled towards consideration of certain mental contents).

If dolphin model of cognition is correct (see section 2.2.2.3), namely that the right kind of a metaphor for cognitive processes is not a serial or dual one, but the binding of different selected trajectories of a cognitive (sea) landscape on its surface (Pliushch & Metzinger, 2015), then there is not only *one* policy that might be involved. The change from linguistic to spatial metaphors has been marked as a sign of a Graphical era we are in (Anderson, 2003, p. 122). In line with this claim, Furl et al. (2010) argue that facial expressions are represented as anticipated *trajectories* of change of those expressions: Pictures of neutral and fearful faces were morphed to different degrees so that participants got to see trajectories from a neutral to a fearful face and vice versa. After seeing such a sequences of pictures, participants had to rate another picture for fearfulness. The results indicated that predictable sequences in which the degree of being morphed rose or fell monotonously, thus forming a trajectory, biased perception (Furl et al., 2010, p. 696). I think that an analogical case can be argued for cognition: In Furl et al.'s experiments participants learned perceptual trajectories, in case of mental action policies can be described as already learned cognitive trajectories.

An open question is how those trajectories switch or merge (which metaphor is correct is open to debate), in order for "cognitive binding," or continuous integration of cognitive processes into the epistemic agent model, to be able to take place (Pliushch & Metzinger, 2015). I argue that cognitive binding is made possible by the fact that our thoughts, due to varying affective intensity, possess limited resolution which enables such binding. Resolution has been argued for in the case of perception: "[w]hatever the metarepresentational instruments used by a system to inspect its own active perceptual states, they can only generate a certain resolution" (Metzinger, 2003, p. 193). It is the result of the intensity constraint such that perceptual states vary in three intensity dimensions: hue, saturation and brightness (p. 185). Perceptual states having these dimensions are such that transitions between them are smooth and the state themselves may become more or less fine-grained (the resolution might increase or decrease). Phenomenal presentations vary in intensity and are "ultra-smooth" (Metzinger, 2003, p. 189). I argue for an analogy in the domain of higher-order cognition where intensity is an affective one and resolution is the degree of clarity with which we experience entertaining a certain thought. The less clarity, the more possibilities for cognitive binding and changing trajectories. This hypothesis rests on premises that every thought is affective to a certain degree (Duncan & Barrett 2007, p. 1202). If thoughts further possess an affective component (intensity + valence), then it may further restrain our possibilities to focus attention on these thoughts, in analogy to the case of pain: “[i]t is the globally available *intensity* of a pain experience that makes it harder and harder to guide our attention anywhere else and that eventually *locks* our attentional focus onto the damaged body part” (Metzinger, 2003, p. 187). Notice the difference to goal-directed pull that Irving argued for: Resolution is not about anxiety experienced when leaving the dolphin trail, metaphorically speaking, but about sticking to the trail and being immersed into it, for example think about the thought crossing your mind that you might have left your credit card unattended somewhere in the supermarket and the clarity with which this thought will hit you and the subsequent sequence of thoughts you might be entertaining. My argument for the premise that attitudes vary in intensity is as follows:

1. *Realness* constraint can be decomposed into the *intensity* constraint and the *transparency* constraint (Metzinger, 2011, p. 290).
2. What is experienced as real is *intuitive*, for the basis of both is a *phenomenological possibility* (for the notion of a phenomenologically possible world see Metzinger, 2011, p. 288). Metzinger & Windt (2014) argue, for example, that in intuitions of certainty the phenomenal signature of knowing has become transparent.
3. Cognitive representations possess degrees of intuitiveness.

---

4. Cognitive representations vary in degrees of *transparency* and *intensity*.

In addition to resolution, cognitive blink is another element enabling cognitive binding. Here, I draw the analogy from *attentional blink* – a phenomenon of not noticing certain visual stimuli upon attending to too many other distracting stimuli. Van Vugt & Slagter (2014) have tested attentional blink to be reduced in experienced open monitoring (OM) meditators, which was not the case for focused-attention (FA) meditators. Thus, attentional processes can be manipulated to reduce the attentional blink. Given that attentional processes play the same role in cognition, as they do in perception, it might not only be the case that there is such a phenomenon as *cognitive blink* which is a brief loss of self-awareness, followed by a shift of the object of PMIR. To distinguish, a self-representational blink “is characterized by a brief loss of self-awareness, followed by a shift in the unit of identification.” (Metzinger, 2013a, p. 9) The point is that attentional shifts (a generative model changing the part of the model that possesses high precision level) tend to go unnoticed and there might be different kinds of them, apart from the attentional blink and the self-representational blink.

Cognitive trajectories can be described as precision-weighted policies at the computational level. I think that their differential weighting allows for switching of merging of such policies. In the previous section I argued that understanding emotions as rate of reduction of free energy leaves it open to what feature is actually responsible for such minimization. Policies are also chosen to minimize free energy, because it is the only ultimate value that the agent strives for. But what happens when in the middle of pursuing a policy another policy minimizes free energy more than the current one? I think that it is a plausible explanation for why cognitive binding takes place. The question how exactly this happens I will leave open for those capable of constructing such a model and testing my hypothesis. Summing up, epistemic agent model selection is trajectory switching or merging that is heavily relied on interoceptive states and counterfactuals.

As cited above, mental action possesses three main characteristics: attention (1) and sense of control (2) over a sequence of epistemic aims that are accomplished (3). *Attention* is one facet of precision (see previous section). I argue that attention in mental action is to be explained the same way that attention in perceptual inference has been. There is “[t]he proposal that we might re-deploy the *same attentional mechanisms* with regard to inner (which are really just off-line) representations is thus entirely plausible” (Waskan 2006, p.

141; my emphasis). It has been also argued that an "'imagery network' is composed of a general attentional mechanism arising in parietal cortex and a context-sensitive mechanism originated in prefrontal cortex" (Mechelli et al., 2004, p. 1262). Here, the similarity of the role of attention in perception and cognition is to be emphasized, namely acquisition of knowledge: Attention has been argued to be a property of cognitively substantional self-knowledge, or self-knowledge based on evidence such that attention to one's thoughts determines the extent of self-knowledge (Boghossian, 2000, p. 493-494). Thus, there is no prima facie reason to suppose that inward attention is to be explained differently from outward attention.

Note that both outward and inward attention is a case of *nested* selection. In perceptual inference precision plays the role of selection: the more precise the prediction error, the more it will change the hypothesis about causes of input, which would cause bigger changes in the world model. Thus, the world model is itself already the result of subpersonal selection, but the agent can also experience attentive agency[387] which is the control of one's attention onto different aspects of the (transparent) world model (Metzinger, 2013a). So, personal-level attention to an object in the world is nested so that the world is a case of subpersonal selection from which an agent performs another kind of selection. The analogue case can be argued for inward attention. As a result, attention of self-deceivers being distributed in another way than that of observers is a result of a different volatility and precision distribution across different levels (and/or models). From this follows that there might be nested cases of self-deception, e.g. if optimism via precision of control states influences perception in a certain way, but on the other hand other kind of precision estimation (attention) determines what we focus on in the perceived world.

A central question for accounts of self-deception is whether self-deceivers experience agency over their self-deception (see the motivational debate in section 1.1). Related to this question, Dana Nelkin has argued that self-deceivers are responsible for their self-deception since if they were to recognize the causal connection between their motivation and the acquired self-deceptive attitude, they would relinquish the latter (section 1.1.2.5). One can explain agency in a *post-action* way: one does something and after that if there is a match between what one wanted to do and what one has done, then one has caused that action. As I mentioned in the last section, according to Seth et al. (2012) the *sense of agency* depends on how good exteroceptive prediction errors could be explained away (matched the predictions). Another post-action agency account is, for example, to equate intention to perform a certain action with attention (for this account see Shepherd, 2014).[388] Decision making is here understood as a process guided by a general intention to find the right decision such that attention is focused on the performed mental operations and leads to the recognition of the indication that the decision is appropriate (pp. 14-15). Mind wandering would be a counterexample to equating attention with intention, because there is attention but no guidance (see Irving, 2015). Finally, explaining agency as inference on the causes of one's actions, namely that one's own control states have caused it (Friston, Adams, Perrinet et al., 2012), would also be a post-action account of agency.

[387] Only the selection process that were part of the PSM could lead to the experience of agency (Metzinger, 2003).

[388] Intention as attention: "When deliberating, an agent is trying to solve a question about what to do. Various plans for action are available to her in thought, and at the moment of decision she performs the mental operation of intention formation in response to recognition of an indication that doing so is appropriate. How is it that she recognizes the indication? It must be because the plan she adopts (or a reason(s) that favours it) is, at the time, the object of attention." (Shepherd, 2014, p. 13)

Such a post-action account of agency is at best incomplete, because it does not explain when exactly an agent would make such an inference and what for. I think that as in the case of the property of consciousness in general arising for precise models that an agent can act upon (Hohwy, 2015), the sense of agency is an indicator that an action can be performed, so it arises prior to action. As such, it could be subject to self-deception, e.g. as in the case of addition the sense of agency may uphold self-deception. This is because "I could if I would but I need not to" way of thinking (false sense of control) actually *precludes* action. So, sense of control arises on precise policies in case when action is to be initiated or precluded. Both its presence and absence can be a means of self-deception.

As for responsibility because of the sense of agency that one could have acquired over one's mental operations, I think that a transparent signature of self-knowledge is a circuit breaker that precludes the acquisition of a sense of agency in certain cases. To remind the reader, the phenomenal signature of knowing becomes transparent through repeated explanation giving. Khemlani & Johnson-Laird (2012) conducted experiments in which participants had to detect logical inconsistencies. Their results suggest that the additional fact that participants actively construct explanations of the inconsistencies makes it harder for participants to detect those afterwards (section 2.2.3). Further, ecstatic seizures also evoke a transparent feeling of knowing (subjective feeling of certainty) and have been explained as a case when an interoceptive mismatch is precluded so that no prediction errors are explained and as a result certainty in one's prediction arises (Picard, 2013). When a feeling of knowing is transparent, one would be certain in knowing something, but not knowing why. Since personal level control does not underlie the acquisition of the property of transparency, self-deceivers are not responsible in these cases (see end of section 1.1.2.4 for more on responsibility).

World/self-model selection: Changes in transparency that determine the change from the epistemic agent model to the world-model, as well as tension are to be explained by the workings of different counterfactuals.[389] Towards the end of section 2.2.3 I mentioned Moritz et al.'s (2014) experiment who have shown that putting schizophrenics in a certain kind of virtual environment attenuates overconfidence in errors. Aymerich-Franch et al. (2014) have tested whether embodiment of a dissimilar avatar decreases anxiety in public speaking (giving a talk in front of a virtual audience), on the premise that anxious individuals preferred dissimilar avatars. The results are not statistically significant, but the trend points to the confirmation of this hypothesis. What this shows is that the world/self-model one possesses affects one's cognitive and affective processes. But I went further and argued that certain cognitive attitudes may acquire the property of realness (transparency) so that they become part of the world/self-model. Note that the fact that thoughts may acquire the property of realness presupposes that we do not experience the *change* in realness, as e.g. we do not experience the change from explicit deliberation to mind wandering (Metzinger, 2013a). It is in question whether we experience the *change* of the degree to which we are certain of something or, with other words, the change in

[389] Referring to the need to distinguish the personal and subpersonal use of counterfactuals (see section 4.3), the determination of the degree of realness by counterfactuals leads to a possibly counter-intuitive conclusion that subpersonal counterfactual selection cannot determine personal counterfactual selection as the representation of possible phenomenal world- and self-models one could possess. If transparency correlates with counterfactual richness (Seth, 2014, Metzinger, 2014b), then counterfactual richness would play a different role than phenomenal possible worlds we might represent – it might aid in our experiencing those phenomenal possible worlds as real in the first place due a change in the degree of transparency. This conclusion came up in our discussion of mental agency with Wanja Wiese for the poster at KogWis14.

transparency during this change itself, thus not in retrospection. Not experiencing such a change might allow for content to switch between opaque object component of the PMIR (subject being directed to it) and the world-level (that content being experienced as real). Thus, one kind of selection might switch to another. Something that was experienced as mental content might now gain the robust phenomenal quality of mind-independence.

How could one explain this in terms of predictive coding? First, transparency may be acquired endogenously or exogenously. This means that either it is possible for mental content to become transparent without the change of the external state, or one would need to put oneself in a certain external state to acquire (or loose) transparency. For example, you may have experienced that some worry you had that was real after discussion with friends loses this quality. Or that in vacation all the problems become surreal, because the attractor has been changed. So, which mental content is experienced as real may depend on the currently active attractor. As for the feature of the generative model that accounts for it, counterfactual richness of generative models have been argued by Anil Seth to evoke the feeling of presence (see previous section). Feeling of presence, on the other hand, has the same functional role as transparency – to indicate epistemic reliability/veridicality (Metzinger, 2014). What we experience as present in the world (mind-independent) has to have acquired the property of transparency (not to be represented as being represented by the model and, thus not to be model-dependent and subject to error). This property is determined by how rich the action-outcome repertoire for the given content is. If some content belongs to the epistemic agent model, e.g. one's future-directed plans, then for it to become real would mean that the fact that those plans depend on the condition of the body that is unsatisfactory would not be available anymore. So, an action repertoire for a healthy body that is counterfactually rich but nonveridical is modelled as veridical. On the one hand, one could hypothesize that this is because the alternative case that one is dying is actually not only counterfactually poor (not a lot of actions can be executed). Yet not everybody who is dying would be susceptible to this kind of self-deception and for some case of self-deception, e.g. the ubiquitious (spouse) cheating example, the distribution of counterfactual richness is unclear. So, I hypothesize that the reason why in the latter case content may also become transparent is the simultaneous modelling of interoceptive consequences of our actions along with exteroceptive ones. To remind the reader, in the previous section I presented Joffily & Coricelli's (2013) account of modelling affective states in predictive coding who argued that valence as a rate of change of free energy possesses the functional role of volatility. So one might argued that valence might also influence volatility (and not only substitute it). Thus, interoceptive counterfactuals might change the precision of the body model in the soldier example (and some other aspects of the generative model in other examples of self-deception).

Tension: In the following I will argue that *interoceptive*[390] counterfactual richness determines the degree of tension. In the previous section I already introduced Seth's (2014) concept of 'counterfactual richness' as a range of counterfactual sensorimotor contingencies that determine the degree or realness. In previous publications, Seth has already argued that "presence is the result of successful suppression by top-down predictions of informative interoceptive signals evoked (directly) by autonomic control signals and (indirectly) by bodily responses to afferent sensory signals" (Seth et al., 2012,

[390] An evidentiary boundary between hidden causes of sensory input and hypotheses generated to explain them is also present both in perceptual and interoceptive inference (Hohwy, 2014), since bodily states have to be inferred too (p. 18): "parts of our own bodies that are not functionally sensory organs are beyond the boundary, so cognitive states are not extended into the body – there is not embodied extension." (Hohwy, 2014, p. 11)

p. 2). I hypothesize that if *perceptual* inference involves counterfactually rich representations, then the same applies to *interoceptive* inference or with other words, I think that there might be interoceptive counterfactuals.[391] One of the roles that I ascribe to such counterfactuals is that of providing agents with affective responses to their cognitive processes, self-deceptive feelings of uneasiness being among these affective responses. According to Dokic (2012), noetic feelings are directed at one's own cognitive competence and can be framed in terms of "I can do this:" "For instance, the feeling of knowing is the feeling that one's performance is or will be successful in possible worlds close to the actual world." (p. 316) If it is accepted that intensity of feelings is the result of *interoceptive* counterfactual richness (range of counterfactual relations between the outcome of the cognitive process and goal representations that this process is conducive and not conducive to), then counterfactual interoceptive richness might explain tension in self-deception. Self-deceptive processess violate the constraint of being truth-conducive in virtue of being *belief* forming processess (truth-conduciveness has been emphasized by Michel, 2014). As such, the valence of generated feelings would be negative and the intensity of these feelings would grow in dependence of the range of counterfactual outcomes that violate goal representations of the agent. I argued in section 2.1.3 that intuitivity, counter-intuitivity and anxiety are different kinds of phenomenological feelings that might arise out of belief forming processes where intuitivity and counter-intuivity are neutral with respect to agent's goals, whereas anxiety is negatively related to achieving agent's goals (other than truth-conduciveness). Intuitivity is feeling of certainty about something being the case or about interpreting things in a certain way. Note that in predictive coding, uncertainty (a feeling in between intuitivity as certainty that something *is* the case and counter-intuitivity as certainty that something *is not* the case) might be beneficial to creativity as an exploration of alternatives. In short, if interoceptive counterfactual richness of a cognitive action is poor (most counterfactuals indicate that an outcome would not be goal-conducive), then there is tension that self-deceivers experience. Given the assumption that there may be several policies from which *one* is executed at a time, this outcome may still be executed, depending on the precision of the policy. Imagine there are two policies: acquire the truth or come to a certain motivated conclusion. The cognitive action one performs (as a result of there being a corresponsing control state so that this action minimizes free energy the most) will evoke tension because its counterfactual richness either to acquire truthful or motivated representation will be poor. The more ways there will be to acquire the truth (when there is a lot of evidence that one might acquire or consider), the poorer the counterfactual richness for the motivated conclusion and the other way around. If there are more policies, then this might get more complicated, because of the dependencies between counterfactual richness of different actions. Note that counterfactual richness is a property of the model (Seth, 2014) and when I use it as a property of an action I imply that once an action is executed, there is a new model, because upon an execution of an action different dependencies have to be updated.

In short, a predictive coding explanation of self-deception is that it is a constant exploration of a restricted set of models with high precision, akin to the case of binocular rivalry. *For the dying soldier, the two alternating models would be that his future-directed plans are only a distracting strategy or that those are really going to happen in the future.* An optimal degree of instability is determined by the volatility which depends on the environment and

[391] But note that Friston's (2014) argue that there are also "purely sensory expectations that do not inform future or counterfactual outcomes – and have no sensorimotor contingency" (p. 120).

the psychological characteristics of the self-deceiver. *For the soldier, the fatal consequences of the situation restrict the instability to be low so that the model implying a future is chosen.* Self-deceiver's epistemic agent model is constructed out of different precision-weighted policies. *In the example, future-directed plans are generated.* The content of this model may become transparent and switch to the world model, if exteroceptive counterfactual richness (amount of potential actions that could be executed) of this content is sufficiently high. *Given a dying body, there is not many kinds of actions one could do, but on the premise of a false (healthy) body-model, one could execute a lot of future-directed actions*. If the interoceptive counterfactual richness of the currently pursued policy is low, then tension is experienced. *A policy of bandaging a lethally wounded body generates tension, because it contradicts all the possible goal representations one could have, e.g. go on vacation or simply stay alive. Other policies would be equally aversive in this example, except that one changes the content of the world model.*

Interim conclusion: In this section I introduced the *overtone* model of self-deception. I argued that several (subpersonal) hypotheses might be explored at once (only one of them having the benefit of becoming a part of some phenomenal model). Those are like overtones to the one being currently phenomenally available. What those overtones look like (or how they develop) depends on the idiosyncratic characteristics of the self-deceiver. Self-deception is a kind of continued exploration in which disambiguation is precluded. Thereafter, I focused on the explanation of two kinds of selection responsible for self-deception – epistemic agent model and world/self-model – in terms of predictive coding. From the behavioral profile I provided an explanation of *inconsistency* as switch between two kinds of selections and *tension* as interoceptive counterfactual richness of cognitive actions that are violating goal representations. Crucially, I argued that mental action has an effect in virtue of interoceptive counterfactuals. Further, to the degree that sensorimotor contingencies employ interoceptive counterfactuals, those might also regulate the acquisition by attitudes of the property of realness. It is transparency of self-deceptive content that attributes to a greater degree to the fact that self-deceivers more often than not *justify* their attitudes instead of abandoning them. An intermediate case is one in which a self-deceptive attitude acquires a transparent signature of self-knowledge (one becomes very certain of it without knowing why). An opposite case would be a self-deceiver gaining *insight* on its own self-deception. I left for future research to determine on which kind of precision estimation this depends. In accordance I did not distinguish whether bottom-up or top-down processing is more responsible for evoking self-deception: both precision of priors and prediction errors might interact in a unique way.

All in all, interoception plays a very important role for self-deception in particular and cognition in general: being immersed into the dolphin's track, feeling counterfactual goal-directed pull if leaving the track, feeling uneasiness when following the track in case of epistemic agent models, but also acquiring transparency in case of world/self-models. In virtue of the definition of emotional valence as rate of reduction of free energy, free energy being the only ultimate value and policies reducing free energy (see previous section), one not only could say that everything is about free energy reduction, but also the other way round: everything is about affect.

5 Conclusion

In this thesis I developed an overtone theory of self-deception. I argued that a satisfying theory of self-deception has to be parsimonious, demarcate self-deception in a proper way, explain which kind of disunity violates which kind of unity (this is the enigmatic part about self-deception) and be truthful to the phenomenology of the self-deceiver (chapter 1). For parsimony reasons I came to the conclusion that the behavioral and phenomenological profile is the explanandum. I kept the behavioral (inconsistency + justification) and phenomenological (tension + insight) profile general so that this description might not demarcate self-deception from other phenomena nor between different kinds of self-deception (section 2.1), but I subsequently refined the phenomenological profile: tension might not only result as a metacognitive feeling that certain rules of a belief-acquisition process have been violated, but also as a counterfactual goal-directed pull (section 2.2). It is the phenomenological profile, I argued, that can distinguish between the kinds of selection that has brought about self-deception: epistemic agent model selection or world/self-model selection. In the first case only the signature of knowledge has become transparent, while in the latter – the whole attitude (section 2.2). The construction of the epistemic agent model has been described by the *dolphin* model of cognition: subpersonal cognitive processing is like a group of dolphins traveling below the surface. On the surface (our epistemic agent model) the agent usually follows only one dolphin's trail and often does not notice changes to the trail.

As for the function that self-deception has, both deception of others and denial of death are plausible, but even more the combination of both: the distinction between self and other is central to both and the combination can account for why often we are not deceived by the self-deception of others (and why self-deceivers have to justify their self-deception in front of observers) on the one hand and why self-deception seems so important for our self-esteem. The combination would look as follows: self-deception, by which self-esteem is upheld, evolved as a means against anxiety of death and has then developed an addition function – that of other deception. Thus, it is not always the case the others do not notice one's self-deception (and are successfully fooled), so that self-deceivers have to justify their self-deceptive attitudes (chapter 3). Lastly, I applied predictive coding tools to for conceptual analysis of the two kinds of selection by which self-deception can be acquired (chapter 4). Here I presented the *overtone* theory of self-deception: self-deceivers might develop several hypotheses to test in parallel, which are like overtones to the basic frequency of a (music) tone. They do not notice the switches in epistemic agent or world/self-model selection that happen and, thus, do not realize when they behave inconsistently. I argued that self-deceivers are caught in constant exploration as an alternative to disambiguation.

With respect to epistemic agent models, I argued that mental action becomes causally efficacious in virtue of interoceptive counterfactuals and described it in analogy to perception as a trajectory, as well as argued for traditional cases of self-deception being a case of trajectory *merging* with a post-hoc sense of control. With respect to world/self-models, I have applied Anil Seth's idea, that exteroceptive counterfactuals influence the degree of transparency, to self-deceptive attitudes that vary in transparency. I have argued that *exteroceptive* counterfactuals include also modelling of *interoceptive* counterfactuals - the controlling element, as for epistemic agent models. Thus, the interaction of exteroceptive and interoceptive counterfactuals explains transparency which itself explains why self-deceivers are so hard to persuade into relinquishing self-deception. Then I proceeded to describe *tension* as depending on the interoceptive counterfactuals *richness*, more precisely as depending on the *kind* of counterfactuals (goal-conducive or not) being generated and not only on the *amount* of it, as it is the case for transparency.

For the future, new experiments with clear objects of measurements and refined questionnaires that fit the behavioral and phenomenological profile of the self-deceiver are to be conducted (summary of my proposed changes to the experimental procedures see in section 2.2.3), as well as the profile itself more refined to fit the demarcation constraint.

6 References

Aarts, H., & Custers, R. (2012). Unconscious goal pursuit: Nonconsicuos goal regulation and motivation. In R. M. Ryan (Ed.), *Oxford library of psychology. The Oxford handbook of human motivation* (pp. 232–247). Oxford: Oxford Univ. Press.

Adams, R. A., Aponte, E., Marshall, L., & Friston, K. J. (2015). Active inference and oculomotor pursuit: the dynamic causal modelling of eye movements. *Journal of neuroscience methods*, *242*, 1–14. doi:10.1016/j.jneumeth.2015.01.003

Ainslie, G. (2014). Selfish goals must compete for the common currency of reward. *The Behavioral and brain sciences*, *37*(2), 135–136. doi:10.1017/S0140525X13001933

Andersen, S. M., Moskowitz, G. B., Blair, I. V., & Nosek, B. A. (2007). Automatic Thought. In A. Kruglanski & E. Higgins (Eds.), *Social Psychology: Handbook of Basic Principles* (pp. 138–175). New York: Guilford Press.

Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences*, *18*(6), 279–292. doi:10.1016/j.tics.2014.03.002

Anderson, M. L. (2003). Embodied Cognition: A field guide. *Artificial Intelligence*, *149*(1), 91–130. doi:10.1016/S0004-3702(03)00054-7

Anderson, M. L. (2015). Mining the Brain for a New Taxonomy of the Mind. *Philosophy Compass*, *10*(1), 68–77. doi:10.1111/phc3.12155

Audi, R. (1997). Self-deception vs. self-caused deception: A comment on Professor Mele. *Behavioral and Brain Sciences*, *20*(1), 104.

Aymerich-Franch, L., Kizilcec, R. F., & Bailenson, J. N. (2014). The Relationship between Virtual Self Similarity and Social Anxiety. *Frontiers in human neuroscience*, *8*, 944. doi:10.3389/fnhum.2014.00944

Bach, K. (1981). An Analysis of Self-Deception. *Philosophy and Phenomenological Research*, *41*(3), 351–370. Retrieved from http://www.jstor.org/stable/2107457

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.-P. Dupuy (Ed.), *CSLI publications: Vol. 69. Self-deception and paradoxes of rationality* (pp. 163–189). Stanford, Calif: Center for the Study of Language and Information.

Baghramian, M., & Nicholson, A. (2013). The puzzle of self-deception. *Philosophy Compass*, *8*(11), 1018–1029.

Bagnoli, C. (2012). Self-Deception and Agential Authority. A Constitutivist Account. *Humana.Mente Journal of Philosophical Studies*, *20*, 99–116.

Balcetis, E. (2008). Where the Motivation Resides and Self-Deception Hides: How Motivated Cognition Accomplishes Self-Deception. *Social and Personality Psychology Compass*, *2*(1), 361–381. doi:10.1111/j.1751-9004.2007.00042.x

Bandura, A. (2011). Self-deception: A paradox revisited. *Behavioral and Brain Sciences*, *34*(01), 16–17. doi:10.1017/S0140525X10002499

Barnes, A. (1997). *Seeing through self-deception*. *Cambridge studies in philosophy*. Cambridge, New York: Cambridge University Press.

Baron, R. (2011). Certainty. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2011st ed.). Retrieved from http://plato.stanford.edu/archives/win2011/entries/certainty/

Barrett, H. C. (2007). Modularity and Design Reincarnation Thanks to Peter Carruthers, Brad Duchaine, and Greg Bryant for helpful comments on this chapter. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Volume 2: Culture and Cognition* (pp. 199–217). Oxford: Oxford University Press.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. doi:10.1146/annurev.psych.59.103006.093639

Barttfeld, P., Wicker, B., McAleer, P., Belin, P., Cojan, Y., Graziano, M.,. . . Sigman, M. (2013). Distinct patterns of functional brain connectivity correlate with objective performance and subjective beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(28), 11577–11582. doi:10.1073/pnas.1301353110

Baumeister, R. F. (1996). Self-Regulation and Ego Threat: Motivated Cognition, Self-Deception, and Destructive Goal Setting. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action. Linking cognition and motivation to behavior* (pp. 27–47). New York, NY: Guilford Press.

Baumeister, R. F., & Cairns, K. J. (1992). Repression and Self-Presentation: When Audiences Interfere With Self-Deceptive Strategies. *Journal of Personality and Social Psychology*, *62*(5), 851–862.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the public interest*, *4*(1), 1–44.

Baumeister, R. F., & Newman, L. S. (1994). Self-Regulation of Cognitive Inference and Decision Processes. *Personality and Social Psychology Bulletin*, *20*(1), 3–19. doi:10.1177/0146167294201001

Baumeister, R. F., & Winegard, B. M. (2014). Fashioning a selfish self amid selfish goals. *The Behavioral and brain sciences*, *37*(2), 136–137. doi:10.1017/S0140525X13001945

Bayne, T., & Fernández, J. (2009). Delusion and Self-Deception. Mapping the Terrain. In T. Bayne & J. Fernández (Eds.), *Macquarie monographs in cognitive science. Delusion and self-deception. Affective and motivational influences on belief formation* (pp. 1–21). New York, NY: Psychology Press.

Beer, J. S., Chester, D. S., & Hughes, B. L. (2013). Social threat and cognitive load magnify self-enhancement and attenuate self-deprecation. *Journal of Experimental Social Psychology*, *49*(4), 706–711. doi:10.1016/j.jesp.2013.02.017

Bermúdez, J. L. (2000a). Personal and sub-personal; A difference without a distinction. *Philosophical Explorations*, *3*(1), 63–82. doi:10.1080/13869790008520981

Bermúdez, J. L. (2000b). Self-deception, intentions, and contradictory beliefs. *Analysis*, *60*(4), 309–319.

Bermúdez, J. L. (1997). Defending intentionalist accounts of self-deception. *Behavioral and Brain Sciences*, *20*(1), 107–108.

Billon, A. (2011). Have We Vindicated the Motivational Unconscious Yet? A Conceptual Review. *Frontiers in Psychology*, *2*, 224. doi:10.3389/fpsyg.2011.00224

Bliss-Moreau, E., & Williams, L. A. (2014). Tag, you're it: affect tagging promotes goal formation and selection. *The Behavioral and brain sciences*, *37*(2), 138–139. doi:10.1017/S0140525X13001969

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. doi:10.1017/S0140525X00038188

Blokpoel, M., Kwisthout, J., & van Rooij, I. (2012). When Can Predictive Brains be Truly Bayesian? *Frontiers in psychology, 3,* 406. doi:10.3389/fpsyg.2012.00406

Boghossian, P. A. (2000). Content and self-knowledge. In S. Bernecker & F. I. Dretske (Eds.), *Knowledge. Readings in contemporary epistemology* (pp. 480–498). Oxford, New York: Oxford University Press.

Borge, S. (2003). The Myth of Self-Deception. *The Southern Journal of Philosophy, 41*(1), 1–28. doi:10.1111/j.2041-6962.2003.tb00939.x

Bornstein, R. F. (1997). Varieties of self-deception. *Behavioral and Brain Sciences, 20*(1), 108–109.

Bortolotti, L. (2010). *Delusions and other irrational beliefs*. Oxford, New York: Oxford University Press.

Bortolotti, L., & Broome, M. R. (2008). Delusional Beliefs and Reason Giving. *Philosophical Psychology, 21*(6), 821–841. doi:10.1080/09515080802516212

Bortolotti, L., & Mameli, M. (2012). Self-Deception, Delusion and the Boundaries of Folk Psychology. *Humana.Mente Journal of Philosophical Studies, 20*, 203–221.

Boudry, M., & Braeckman, J. (2012). How convenient! The epistemic rationale of self-validating belief systems. *Philosophical Psychology, 25*(3), 341–364. doi:10.1080/09515089.2011.579420

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389–414. doi:10.1037/a0026450

Braude, S. (2009). The Conceptual Unity of Dissociation: A Philosophical Argument. In J. A. O'Neil & P. F. Dell (Eds.), *Dissociation and the dissociative disorders. DSM-V and beyond* (pp. 27–36). New York: Routledge.

Brentano, F. (1971[1911]). *Psychologie vom empirischen Standpunkt. Zweiter Band: Von der Klassifikation der psychischen Phänomene.* Hamburg: Meiner.

Brigard, F. de, & Giovanello, K. S. (2012). Influence of outcome valence in the subjective experience of episodic past, future, and counterfactual thinking. *Consciousness and Cognition, 21*(3), 1085–1096. doi:10.1016/j.concog.2012.06.007

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing, 14*(4), 411–427. doi:10.1007/s10339-013-0571-3

Brown, R. (2003). The Emplotted Self: Self-Deception and Self-Knowledge. *Philosophical Papers, 32*(3), 279–300. doi:10.1080/05568640309485128

Brown, S. L., & Kenrick, D. T. (1997). Paradoxical self-deception: Maybe not so paradoxical after all. *Behavioral and Brain Sciences, 20*(1), 109–110.

Bryant, R. A., & Kourch, M. (2001). Hypnotically induced emotional numbing. *Clinical and Experimental Hypnosis, 49*(3), 220–230.

Büchel, C., Geuter, S., Sprenger, C., & Eippert, F. (2014). Placebo Analgesia: A Predictive Coding Perspective. *Neuron, 81*(6), 1223–1239. doi:10.1016/j.neuron.2014.02.042

Burke, B. L., Martens, A., & Faucher, E. H. (2010). Two Decades of Terror Management Theory: A Meta-Analysis of Mortality Salience Research. *Personality and Social Psychology Review, 14*(2), 155–195. doi:10.1177/1088868309352321

Byrne, C. C., & Kurland, J. A. (2001). Self-deception in an Evolutionary Game. *Journal of Theoretical Biology, 212*(4), 457–480. doi:10.1006/jtbi.2001.2390

Carhart-Harris, R. L., & Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain*, *133*(4), 1265–1283. doi:10.1093/brain/awq010

Carruthers, P. (2006). The Case for Massively Modular Models of Mind. In R. Stainton (Ed.), *Contemporary debates in philosophy: Vol. 7. Contemporary debates in cognitive science* (pp. 3–21). Malden, MA, Oxford: Blackwell Pub.

Chin, J., Mrazek, M., & Schooler, J. W. (2012). Blind spots to the self. Limits in knowledge of mental contents and personal predispositions. In S. Vazire & T. Wilson (Eds.), *Handbook of Self-Knowledge* (pp. 77–89). New York: Guilford Publications.

Clark, A. (2013a). Dreaming the Whole Cat: Generative Models, Predictive Processing, and the Enactivist Conception of Perceptual Experience. *Mind*, *121*(483), 753–771. doi:10.1093/mind/fzs106

Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. doi:10.1017/S0140525X12000477

Clegg, J. W., & Moissinac, L. (2005). A relational theory of self-deception. *New Ideas in Psychology*, *23*(2), 96–110. doi:10.1016/j.newideapsych.2006.03.001

Clos, M., Langner, R., Meyer, M., Oechslin, M. S., Zilles, K., & Eickhoff, S. B. (2014). Effects of prior information on decoding degraded speech: An fMRI study. *Human Brain Mapping*, *35*(1), 61–74. doi:10.1002/hbm.22151

Colombo, M. (2013). Constitutive relevance and the personal/subpersonal distinction. *Philosophical Psychology*, *26*(4), 547–570. doi:10.1080/09515089.2012.667623

Cooper, G. (2007a). *The Science of the Struggle for Existence: On the Foundations of Ecology*. Cambridge: Cambridge University Press. Retrieved from http://books.google.de/books?id=qzf-WpEdmXIC

Cooper, J. (2007b). *Cognitive dissonance: Fifty years of a classic theory* (1st ed.). Los Angeles: SAGE Publ.

Cosmelli, D., & Preiss, D. D. (2014). On the temporality of creative insight: a psychological and phenomenological perspective. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.01184

Craver, C. F. (2015). Levels. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (8T). Frankfurt am Main: MIND Group.

Crawford, C., & Salmon, C. (2004). The Essence of Evolutionary Psychology. An Introduction. In C. B. Crawford & C. Salmon (Eds.), *Evolutionary psychology, public policy, and personal decisions* (pp. 23–50). Mahwah, N.J: Lawrence Erlbaum Associates.

Crawford, C. B. (2004). Public Policy and Personal Decisions: The Evolutionary Context. In C. B. Crawford & C. Salmon (Eds.), *Evolutionary psychology, public policy, and personal decisions* (pp. 3–22). Mahwah, N.J: Lawrence Erlbaum Associates.

Crocker, J., & Nuer, N. (2004). Do People Need Self-Esteem? Comment on Pyszczynski et al. (2004). *Psychological Bulletin*, *130*(3), 469–472. doi:10.1037/0033-2909.130.3.469

Crocker, J., & Park, L. E. (2004). The Costly Pursuit of Self-Esteem. *Psychological Bulletin*, *130*(3), 392–414. doi:10.1037/0033-2909.130.3.392

Cummins, R. A., & Nistico, H. (2002). Maintaining life satisfaction: The role of positive cognitive bias. *Journal of Happiness Studies*, *3*, 37–69.

Damm, L. (2011). Self-deception about emotion. *The Southern Journal of Philosophy*, *49*(3), 254–270.

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Friston, K. J., Stephan, K. E., & Sporns, O. (2010b). Observing the Observer (II): Deciding When to Decide. *PLoS ONE*, *5*(12), e15555. doi:10.1371/journal.pone.0015555

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., Friston, K. J., & Sporns, O. (2010a). Observing the Observer (I): Meta-Bayesian Models of Learning and Decision-Making. *PLoS ONE*, *5*(12), e15554. doi:10.1371/journal.pone.0015554

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2012). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1379–1390. doi:10.1098/rstb.2012.0002

Davidson, D. (1986). Deception and division. In J. Elster (Ed.), *Studies in rationality and social change. The multiple self* (1st ed., pp. 79–92). Cambridge: Cambridge Univ. Press.

Davidson, D. (1998). Who is fooled? In J.-P. Dupuy (Ed.), *CSLI publications: Vol. 69. Self-deception and paradoxes of rationality* (pp. 1–18). Stanford, Calif: Center for the Study of Language and Information.

Davies, M. (2009). Delusion and motivationally biased belief. Self-deception in the two-factor framework. In T. Bayne & J. Fernández (Eds.), *Macquarie monographs in cognitive science. Delusion and self-deception. Affective and motivational influences on belief formation* (pp. 71–86). New York, NY: Psychology Press.

Davies, M., & Egan, A. (2013). Delusion: Cognitive approaches - Bayesian inference and compartmentalization. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, & J. Z. Sadler (Eds.), *International perspectives in philosophy and psychiatry. The Oxford handbook of philosophy and psychiatry*. Oxford: Oxford Univ. Press.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 208–216. doi:10.3758/CABN.10.2.208

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter Than We Think: When Our Brains Detect That We Are Biased. *Psychological Science*, *19*(5), 483–489. doi:10.1111/j.1467-9280.2008.02113.x

de Sousa, R. (2014). Emotion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2014th ed.). Retrieved from http://plato.stanford.edu/archives/spr2014/entries/emotion/

Demos, R. (1960). Lying to Oneself. *The Journal of Philosophy*, *57*(18), 588–595. Retrieved from http://www.jstor.org/stable/2023611.

Dennett, D. C. (1991). *Consciousness explained* (1st ed.). Boston: Little.

Dennis, P. A., & Halberstadt, A. G. (2013). Is believing seeing? The role of emotion-related beliefs in selective attention to affective cues. *Cognition & Emotion*, *27*(1), 3–20. doi:10.1080/02699931.2012.680578

Deweese-Boyd, I. (2012). Self-deception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2012nd ed.). Retrieved from http://plato.stanford.edu/archives/spr2012/entries/self-deception/

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences*, *16*(6), 317–321. doi:10.1016/j.tics.2012.04.007

Ditto, P. H. (2009). Passion, Reason, and Necessity. A quantity-of-Processing View of Motivated Reasoning. In T. Bayne & J. Fernández (Eds.), *Macquarie monographs in cognitive science. Delusion and self-deception. Affective and motivational influences on belief formation* (pp. 23–53). New York, NY: Psychology Press.

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology*, *75*(1), 53–69. doi:10.1037/0022-3514.75.1.53

Dokic, J. (2012). Seeds of self-knowledge: noetic feelings and metacognition. In M. J. Beran, J. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 302–321). Oxford: Oxford University Press.

Domschke, K., Stevens, S., Pfleiderer, B., & Gerlach, A. L. (2010). Interoceptive sensitivity in anxiety and anxiety disorders: an overview and integration of neurobiological findings. *Clinical psychology review*, *30*(1), 1–11. doi:10.1016/j.cpr.2009.08.008

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, *26*(1), 1–18. doi:10.1111/phpe.12014

Drayson, Z. (2014). The Personal/Subpersonal Distinction. *Philosophy Compass*, *9*(5), 338–346. doi:10.1111/phc3.12124

Dretske, F. I. (1982). *Knowledge & the flow of information* (2nd ed.). *A Bradford book*. Cambridge, Mass: MIT Press.

DuBois, D. L., & Flay, B. R. (2004). The Healthy Pursuit of Self-Esteem: Comment on and Alternative to the Crocker and Park (2004) Formulation. *Psychological Bulletin*, *130*(3), 415–420. doi:10.1037/0033-2909.130.3.415

Duncan, S., & Barrett, L. F. (2007). Affect is a form of cognition: A neurobiological analysis. *Cognition & emotion*, *21*(6), 1184–1211. doi:10.1080/02699930701437931

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M.,. . . Dalgleish, T. (2010). Listening to your heart. How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21*(12), 1835–1844. doi:10.1177/0956797610389191

Dunning, D. (2011). Get thee to a laboratory. *Behavioral and Brain Sciences*, *34*(01), 18–19. doi:10.1017/S0140525X10002530

Egan, A. (2009). Imagination, Delusion, and Self-Deception. In T. Bayne & J. Fernández (Eds.), *Macquarie monographs in cognitive science. Delusion and self-deception. Affective and motivational influences on belief formation* (pp. 263–280). New York, NY: Psychology Press.

Evans, J. S. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds. Dual processes and beyond* (pp. 33–54). Oxford: Oxford University Press.

Farennikova, A. (2014). Perception of Absence and Penetration from Expectation. *Review of Philosophy and Psychology*. doi:10.1007/s13164-014-0188-1

Farrow, Tom F D, Burgess, J., Wilkinson, I. D., & Hunter, M. D. (2015). Neural correlates of self-deception and impression-management. *Neuropsychologia*, *67*, 159–174. doi:10.1016/j.neuropsychologia.2014.12.016

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, *4*, 215. doi:10.3389/fnhum.2010.00215

Fernández, J. (2013). Self-deception and self-knowledge. *Philosophical Studies, 162*, 379–400. doi:10.1007/s11098-011-9771-9

Fernbach, P. M., Hagmayer, Y., & Sloman, S. A. (2014). Effort denial in self-deception. *Organizational Behavior and Human Decision Processes*, *123*(1), 1–8. doi:10.1016/j.obhdp.2013.10.013

Ferrè, E. R., Bottini, G., Iannetti, G. D., & Haggard, P. (2013). The balance of feelings: Vestibular modulation of bodily sensations. *Cortex*, *49*(3), 748–758. doi:10.1016/j.cortex.2012.01.012

Festinger, L. (1957). *A theory of cognitive dissonance* (Repr 1968). Stanford, Calif: Stanford Univ. Press.

Fingarette, H. (2000[1969]). *Self-deception*. California: Univ. of California Press. Retrieved from http://www.loc.gov/catdir/description/ucal042/99016358.html

Fishbach, A. (2014). The motivational self is more than the sum of its goals. *The Behavioral and brain sciences*, *37*(2), 143–144. doi:10.1017/S0140525X13002021

Fishbach, A., & Ferguson, M. (2007). The goal construct in social psychology. In A. Kruglanski & E. Higgins (Eds.), *Social Psychology: Handbook of Basic Principles* (pp. 490–515). New York: Guilford Press.

FitzGerald, Thomas H B, Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in human neuroscience*, *8*, 457. doi:10.3389/fnhum.2014.00457

FitzGerald, Thomas H B, Moran, R. J., Friston, K. J., & Dolan, R. J. (2015). Precision and neuronal dynamics in the human posterior parietal cortex during evidence accumulation. *NeuroImage*, *107*, 219–228. doi:10.1016/j.neuroimage.2014.12.015

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature reviews. Neuroscience*, *10*(1), 48–58. doi:10.1038/nrn2536

Fogelson, N., Litvak, V., Peled, A., Fernandez-del-Olmo, M., & Friston, K. J. (2014). The functional anatomy of schizophrenia: A dynamic causal modeling study of predictive coding. *Schizophrenia Research*, *158*(1-3), 204–212. doi:10.1016/j.schres.2014.06.011

Fotopoulou, A. (2013a). Beyond the reward principle: Consciousness as precision seeking. *Neuropsychoanalysis*, *15*(1), 33–38.

Fotopoulou, A. (2013b). Time to get rid of the 'Modular' in neuropsychology: A unified theory of anosognosia as aberrant predictive coding. *Journal of Neuropsychology*, *8*(1), 1–19. doi:10.1111/jnp.12010

Frankish, K. (1998). A matter of opinion. *Philosophical Psychology*, *11*(4), 423–442. doi:10.1080/09515089808573271

Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds. Dual processes and beyond* (pp. 89–107). Oxford: Oxford University Press.

Frankish, K. (2012). Delusions, Levels of Belief, and Non-doxastic Acceptances. *Neuroethics*, *5*(1), 23–27. doi:10.1007/s12152-011-9123-7

Frankish, K., & Evans, J. S. B. T. (2009). The duality of mind: a historical perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds. Dual processes and beyond* (pp. 1–29). Oxford: Oxford University Press.

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. doi:10.1037/0033-295X.100.2.298

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. doi:10.1098/rstb.2005.1622

Friston, K. J. (2002). Functional integration and inference in the brain. *Progress in Neurobiology*, *68*, 113–143.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352. doi:10.1016/j.neunet.2003.06.005

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. doi:10.1016/j.tics.2009.04.005

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience*, *11*(2), 127–138. doi:10.1038/nrn2787

Friston, K. J. (2012a). A Free Energy Principle for Biological Systems. *Entropy (Basel, Switzerland)*, *14*(11), 2100–2121. doi:10.3390/e14112100

Friston, K. J. (2012b). Prediction, perception and agency. *International journal of psychophysiology: official journal of the International Organization of Psychophysiology*, *83*(2), 248–252. doi:10.1016/j.ijpsycho.2011.11.014

Friston, K. J. (2013a). Consciousness and hierarchical inference. *Neuropsychoanalysis*, *15*(1), 38–42.

Friston, K. J. (2013b). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475. doi:10.1098/rsif.2013.0475

Friston, K. J. (2014). Active inference and agency. *Cognitive Neuroscience*, *5*(2), 119–121. doi:10.1080/17588928.2014.905517

Friston, K. J., Adams, R., & Montague, R. (2012). What is value-accumulated reward or evidence? *Frontiers in neurorobotics*, *6*, 11. doi:10.3389/fnbot.2012.00011

Friston, K. J., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, *3*, 151. doi:10.3389/fpsyg.2012.00151

Friston, K. J., Breakspear, M., & Deco, G. (2012). Perception and self-organized instability. *Frontiers in Computational Neuroscience*, *6*. doi:10.3389/fncom.2012.00044

Friston, K. J., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1521), 1211–1221. doi:10.1098/rstb.2008.0300

Friston, K. J., Samothrakis, S., & Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biological Cybernetics*, *106*(8-9), 523–541. doi:10.1007/s00422-012-0512-8

Friston, K. J., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, *7*, 598. doi:10.3389/fnhum.2013.00598

Friston, K. J., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1655). doi:10.1098/rstb.2013.0481

Friston, K. J., Sengupta, B., & Auletta, G. (2014). Cognitive Dynamics: From Attractors to Active Inference. *Proceedings of the IEEE, 102*(4), 427–445. doi:10.1109/JPROC.2014.2306251

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H.,. . . Sporns, O. (2012). Dopamine, Affordance and Active Inference. *PLoS Computational Biology*, *8*(1), e1002327. doi:10.1371/journal.pcbi.1002327

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417–458. doi:10.1007/s11229-007-9237-y

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158. doi:10.1016/S2215-0366(14)70275-5

Frith, C. D., & Friston, K. J. (2013). False perceptions & false beliefs: Understanding schizophrenia

. *Pontifical Academy of Sciences*.

Funkhouser, E. (2005). Do the self-deceived get whay they want? *Pacific Philosophical Quarterly*, *86*(3), 295–312. doi:10.1111/j.1468-0114.2005.00228.x

Funkhouser, E. (2009). Self-Deception and Limits of Folk Psychology. *Social theory and praxis*, *35*(1), 1–16.

Furl, N., van Rijsbergen, N J, Kiebel, S. J., Friston, K. J., Treves, A., & Dolan, R. J. (2010). Modulation of perception and brain activity by predictable trajectories of facial expressions. *Cerebral Cortex*, *20*(3), 694–703. doi:10.1093/cercor/bhp140

Füstös, J., Gramann, K., Herbert, B. M., & Pollatos, O. (2013). On the embodiment of emotion regulation: interoceptive awareness facilitates reappraisal. *Social Cognitive and Affective Neuroscience*, *8*(8), 911–917. doi:10.1093/scan/nss089

Galeotti, A. E. (2012). Self-Deception: Intentional Plan or Mental Event? *Humana.Mente Journal of Philosophical Studies*, *20*, 41–66.

Galeotti, A. E. (2014). Liars or Self-Deceived? Reflections on Political Deception. *Political Studies*, 1–16. doi:10.1111/1467-9248.12122

Gangestad, S. W. (2011). Understanding self-deception demands a co-evolutionary framework. *Behavioral and Brain Sciences*, *34*(01), 23–24. doi:10.1017/S0140525X10002578

Gärdenfors, P. (2004). Conceptual Spaces as a Framework for Knowledge Representation. *Mind and Matter*, *2*(2), 9–27.

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. doi:10.1016/j.biopsycho.2014.11.004

Gawronski, B., & Bodenhausen, G. (2012). Self-insight from a dual-process perspective. In S. Vazire & T. Wilson (Eds.), *Handbook of Self-Knowledge* (pp. 22–37). New York: Guilford Publications.

Gendler, T. S. (2007). Self-deception as pretense. *Philosophical Perspectives*, *21*(1), 231–258. doi:10.1111/j.1520-8583.2007.00127.x

Gerlach, K. D., Dornblaser, D. W., & Schacter, D. L. (2014). Adaptive constructive processes and memory accuracy: Consequences of counterfactual simulations in young and older adults. *Memory*, *22*(1), 145–162. doi:10.1080/09658211.2013.779381

Gerrans, P. (2013). Delusional Attitudes and Default Thinking. *Mind & Language*, *28*(1), 83–102.

Gerrans, P. (2014). Pathologies of hyperfamiliarity in dreams, delusions and déjà vu. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.00097

Gerrans, P. (2015). All the Self We Need. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (15T). Frankfurt am Main: MIND Group.

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. doi:10.1016/j.neuropsychologia.2013.11.010

Gigerenzer, G. (2006). Bounded and Rational. In R. Stainton (Ed.), *Contemporary debates in philosophy: Vol. 7. Contemporary debates in cognitive science.* Malden, MA, Oxford: Blackwell Pub.

Girotto, V., Ferrante, D., Pighin, S., & Gonzalez, M. (2007). Postdecisional Counterfactual Thinking by Actors and Readers. *Psychological Science*, *18*(6), 510–515. doi:10.1111/j.1467-9280.2007.01931.x

Gleicher, F., Kost, K. A., Baker, S. M., Strathman, A. J., Richman, S. A., & Sherman, S. J. (1990). The Role of Counterfactual Thinking in Judgments of Affect. *Personality and Social Psychology Bulletin*, *16*(2), 284–295. doi:10.1177/0146167290162009

Goel, V., Tierney, M., Sheesley, L., Bartolo, A., Vartanian, O., & Grafman, J. (2007). Hemispheric Specialization in Human Prefrontal Cortex for Resolving Certain and Uncertain Inferences. *Cerebral Cortex*, *17*(10), 2245–2250. doi:10.1093/cercor/bhl132

Goldstone, R. L., de Leeuw, Joshua R., & Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, *135*, 24–29. doi:10.1016/j.cognition.2014.11.027

Goleman, D. (1985). *Vital lies, simple truths: The psychology of self-deception*. New York: Simon and Schuster.

Greenwald, A. G. (1988). Self-knowledge and self-deception. In J. S. Lockard & D. L. Paulhus (Eds.), *Century psychology series. Self-deception. An adaptive mechanism?* (pp. 111–129). Englewood Cliffs, NJ: Prentice-Hall.

Greenwald, A. G. (1997). Self-Knowledge and Self-Deception: Further consideration. In M. Myslobodsky (Ed.), *The mythomanias. The nature of deception and self-deception* (pp. 51–71). Mahwah, N.J: L. Erlbaum Associates.

Greve, W., & Wentura, D. (2003). Immunizing the Self: Self-Concept Stabilization Through Reality-Adaptive Self-Definitions. *Personality and Social Psychology Bulletin*, *29*(1), 39–50. doi:10.1177/0146167202238370

Greve, W., & Wentura, D. (2010). True lies: Self-stabilization without self-deception☆. *Consciousness and Cognition*, *19*(3), 721–730. doi:10.1016/j.concog.2010.05.016

Griffiths, O., Langdon, R., Le Pelley, Mike E., & Coltheart, M. (2014). Delusions and prediction error: re-examining the behavioural evidence for disrupted error signalling in delusion formation. *Cognitive Neuropsychiatry*, 1–29. doi:10.1080/13546805.2014.897601

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. doi:10.1016/j.tics.2010.05.004

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, *37*(2), 147–169. doi:10.1037/0022-3514.37.2.147

Hagedorn, J. W. (1996). Happiness and self-deception: An old question examined by a new measure of subjective well-being. *Social Indicators Research*, *38*(2), 139–160.

Haight, M. R. (1980). *A study of self-deception*. Brighton, Sussex: The Harvester Press.

Harmon-Jones, E. (2012). Cognitive Dissonance Theory. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior* (2nd ed., Vol. 1, pp. 543–549). London: Acad. Press.

Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2015). The special case of self-perspective inhibition in mental, but not non-mental, representation. *Neuropsychologia*, *67*, 183–192. doi:10.1016/j.neuropsychologia.2014.12.015

Hegedüs, G., Darnai, G., Szolcsányi, T., Feldmann, Á., Janszky, J., & Kállai, J. (2014). The rubber hand illusion increases heat pain threshold. *European Journal of Pain*, *18*(8), 1173–1181. doi:10.1002/j.1532-2149.2014.00466.x

Helzer, E., & Dunning, D. (2012). On motivated reasoning and self-belief. In S. Vazire & T. Wilson (Eds.), *Handbook of Self-Knowledge* (pp. 379–396). New York: Guilford Publications.

Herwig, U., Kaffenberger, T., Jäncke, L., & Brühl, A. B. (2010). Self-related awareness and emotion regulation. *NeuroImage*, *50*(2), 734–741. doi:10.1016/j.neuroimage.2009.12.089

Hesselmann, G., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2010). Predictive Coding or Evidence Accumulation? False Inference and Neuronal Fluctuations. *PLoS ONE*, *5*(3).

Hirsh, J. B. (2014). Mapping the goal space: personality integration and higher-order goals. *The Behavioral and brain sciences*, *37*(2), 144–145. doi:10.1017/S0140525X13002033

Hirstein, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. *Disorders in mind*. Cambridge, Mass: MIT Press.

Hobson, J. A., Hong, C. C.-H., & Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.01133

Hohwy, J. (2013a). Delusions, Illusions and Inference under Uncertainty. *Mind & Language*, *28*(1), 57–71.

Hohwy, J. (2013b). *Predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in psychology*, *3*, 96. doi:10.3389/fpsyg.2012.00096

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J. (2014). The Self-Evidencing Brain. *Noûs*, *48*(1), 1–27. doi:10.1111/nous.12062

Hohwy, J. (2015). Prediction Error Minimization, Mental and Developmental Disorder, and Statistical Theories of Consciousness. In R. Gennaro (Ed.), *Disturbed consciousness* (pp. 293–324). Cambridge, MA: MIT Press.

Hohwy, J., Roepstorff, A., & Friston, K. J. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, *108*(3), 687–701. doi:10.1016/j.cognition.2008.05.010

Holcomb, H. (2004). Darwin and Evolutionary Moral Psychology. In C. B. Crawford & C. Salmon (Eds.), *Evolutionary psychology, public policy, and personal decisions* (pp. 73–95). Mahwah, N.J: Lawrence Erlbaum Associates.

Holton, R. (2001). What is the Role of the Self in Self-Deception? *Proceedings of the Aristotelian Society (Hardback), 101*(1), 53–69. doi:10.1111/j.0066-7372.2003.00021.x

Huang, J. Y., & Bargh, J. A. (2014b). The Selfish Goal: Autonomously Operating Motivational Structures as the Proximate Cause of Human Judgment and Behavior. *Behavioral and Brain Sciences*, *37*(2), 121–135.

Huebner, B., & Rupert, R. D. (2014). Massively representational minds are not always driven by goals, conscious or otherwise. *The Behavioral and brain sciences*, *37*(2), 145–146. doi:10.1017/S0140525X13002045

Irving, Zachary C. (2015). Mind-wandering is unguided attention: accounting for the "purposeful" wanderer. *Philosophical Studies*, 1–25. doi:10.1007/s11098-015-0506-1

Jackendoff, R. (1975). On Belief-Contexts. *Linguistic Inquiry*, *6*(1), 53–93.

Jackendoff, R. (2012). *A user's guide to thought and meaning*. New York: Oxford University Press.

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: a critique. *Trends in Cognitive Sciences*, *9*(1), 21–25. doi:10.1016/j.tics.2004.11.003

Janicki, M. G. (2004). Beyond Sociobiology: A Kinder and Gentler Evolutionary View of Human Nature. In C. B. Crawford & C. Salmon (Eds.), *Evolutionary psychology, public policy, and personal decisions* (pp. 51–72). Mahwah, N.J: Lawrence Erlbaum Associates.

Jennings, C. D., & Nanay, B. (2014). Action without attention. *Analysis*, 1–7. doi:10.1093/analys/anu096

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS computational biology*, *9*(6), e1003094. doi:10.1371/journal.pcbi.1003094

Johnson, D. D. P., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317–320. doi:10.1038/nature10384

Johnson, K. J., & Fredrickson, B. L. (2005). "We All Look the Same to Me:" Positive Emotions Eliminate the Own-Race Bias in Face Recognition. *Psychological Science*, *16*(11), 875–881.

Johnston, M. (1988). Self-deception and the nature of the mind. In B. P. McLaughlin & A. O. Rorty (Eds.), *Topics in philosophy: Vol. 6. Perspectives on self-deception* (pp. 63–91). Berkeley Cal.: Univ. of California Press.

Jopling, D. A. (2013). Placebo effects in psychiatry and psychotherapy. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, & J. Z. Sadler (Eds.), *International perspectives in philosophy and psychiatry. The Oxford handbook of philosophy and psychiatry* (pp. 1202–1226). Oxford: Oxford Univ. Press.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. J. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *370*(1668). doi:10.1098/rstb.2014.0169

Khemlani, S., & Johnson-Laird, P. N. (2013). Cognitive changes from explanations. *Journal of Cognitive Psychology*, *25*(2), 139–146. doi:10.1080/20445911.2012.720968

Khemlani, S. S., & Johnson-Laird, P. (2012). Hidden conflict: Explanations make inconsistencies harder to detect. *Acta Psychologica, 139*, 486–491.

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex, 48*(7), 805–825. doi:10.1016/j.cortex.2011.04.006

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166. doi:10.1007/s10339-007-0170-2

Klein, T. A., Ullsperger, M., & Danielmeier, C. (2013). Error awareness and the insula: links to neurological and psychiatric diseases. *Frontiers in Human Neuroscience, 7*(14), 1–14.

Klepeis, J. L., & Floudas, C. A. (2003). ASTRO-FOLD: A Combinatorial and Global Optimization Framework for Ab Initio Prediction of Three-Dimensional Structures of Proteins from the Amino Acid Sequence. *Biophysical Journal*, *85*, 2119–2146.

Klinger, E. (2013). Goal Commitments and the content of thoughts and dreams: basic principles. *Frontiers in psychology*, *4*, 415. doi:10.3389/fpsyg.2013.00415

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences*, *27*(12), 712–719. doi:10.1016/j.tins.2004.10.007

Knobe, J. (2015). Philosophers are doing something different now: quantitative data. *Cognition*, *135*, 36–38. doi:10.1016/j.cognition.2014.11.011

Krebs, D., Ward, J. W., & Racine, T. (1997). The many faces of self-deception. *Behavioral and Brain Sciences*, *20*(1), 119.

Krebs, D. L., & Denton, K. (1997). Social Illusions and Self-Deception:. The Evolution of Biases in Person Perception. In J. A. Simpson & D. T. Kenrick (Eds.), *Evolutionary social psychology* (pp. 21–47). Mahwah, NJ: Erlbaum.

Krizan, Z., & Windschitl, P. D. (2009). Wishful Thinking about the Future: Does Desire Impact Optimism? *Social and Personality Psychology Compass*, *3*(3), 227–243. doi:10.1111/j.1751-9004.2009.00169.x

Kruglanski, A. W., & Webster, D. W. (1996). Motivated Closing of the Mind: "Seizing" and "Freezing". *Psychological Review*, *103*(2), 263–283.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, *53*(4), 636–647. doi:10.1037/0022-3514.53.4.636

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. doi:10.1037/0033-2909.108.3.480

Kunda, Z. (1992). Can Dissonance Theory Do It All? *Psychological Inquiry*, *3*(4), 337–339.

Kurt, A., & Paulhus, D. L. (2008). Moderators of the adaptiveness of self-enhancement: Operationalization, motivational domain, adjustment facet, and evaluator. *Journal of Research in Personality*, *42*(4), 839–853. doi:10.1016/j.jrp.2007.11.005

Lamba, S., & Nityananda, V. (2014). Self-deceived individuals are better at deceiving others. *PloS one*, *9*(8), e104562. doi:10.1371/journal.pone.0104562

Langdon, R., & Bayne, T. (2010). Delusion and confabulation: Mistakes of perceiving, remembering and believing. *Cognitive Neuropsychiatry*, *15*(1-3), 319–345. doi:10.1080/13546800903000229

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

Leary, M. R. (2004). The Function of Self-Esteem in Terror Management Theory and Sociometer Theory: Comment on Pyszczynski et al. (2004). *Psychological Bulletin*, *130*(3), 478–482. doi:10.1037/0033-2909.130.3.478

Lenggenhager, B., & Lopez, C. (2015). Vestibular Contributions to the Sense of Body, Self, and Others. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (23T). Frankfurt am Main: MIND Group.

Levesque, C., & Brown, K. W. (2007). Mindfulness as a moderator of the effect of implicit motivational self-concept on day-to-day behavioral motivation. *Motivation and Emotion*, *31*(4), 284–299. doi:10.1007/s11031-007-9075-8

Levy, N. (2009). Self-Deception Without Thought Experiments. In T. Bayne & J. Fernández (Eds.), *Macquarie monographs in cognitive science. Delusion and self-deception. Affective and motivational influences on belief formation* (pp. 227–242). New York, NY: Psychology Press.

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in human neuroscience, 7,* 547. doi:10.3389/fnhum.2013.00547

Lin, Y.-T. (2015). Memory for Prediction Error Minimization: From Depersonalization to the Delusion of Non-Existence. A Commentary on Philip Gerrans. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (15C). Frankfurt am Main: MIND Group.

Lockard, J. A. (1997). Distal versus proximal mechanisms of "real" self-deception. *Behavioral and Brain Sciences*, *20*(1), 120–121.

Lockie, R. (2003). Depth psychology and self-deception. *Philosophical Psychology*, *16*(1), 127–148.

Lopez, C., Schreyer, H.-M., Preuss, N., & Mast, F. W. (2012). Vestibular stimulation modifies the body schema. *Neuropsychologia*, *50*(8), 1830–1837. doi:10.1016/j.neuropsychologia.2012.04.008

Lopez, J. K., & Fuxjager, M. J. (2012). Self-deception’s adaptive value: Effects of positive thinking and the winner effect. *Consciousness and Cognition*, *21*(1), 315–324. doi:10.1016/j.concog.2011.10.008

Lynch, K. (2012). On the “tension” inherent in self-deception. *Philosophical Psychology, 25*(3), 433–450. doi:10.1080/09515089.2011.622364

Lynch, K. (2013). Self-Deception and Stubborn Belief. *Erkenntnis*, *78*(6), 1337–1345. doi:10.1007/s10670-012-9425-0

Lynch, K. (2014). Self-deception and shifts of attention. *Philosophical Explorations, 17*(1), 63–75. doi:10.1080/13869795.2013.824109

Lynch, R. F., & Trivers, R. L. (2012). Self-deception inhibits laughter. *Personality and Individual Differences, 53*(4), 491–495. doi:10.1016/j.paid.2012.02.017

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, *102*(1-3), 59–70. doi:10.1016/j.jphysparis.2008.03.004

Marcus, G. F., & Davis, E. (2013). How Robust Are Probabilistic Models of Higher-Level Cognition? *Psychological Science*, *24*(12), 2351–2360. doi:10.1177/0956797613495418

Marraffa, M. (2012). Remnants of Psychoanalysis. Rethinking the Psychodynamic Approach to Self-Deception. *Humana.Mente Journal of Philosophical Studies*, *20*, 223–243.

Mashour, G. A. (2013). Cognitive unbinding: a neuroscientific paradigm of general anesthesia and related states of unconsciousness. *Neuroscience and biobehavioral reviews*, *37*(10 Pt 2), 2751–2759. doi:10.1016/j.neubiorev.2013.09.009

Massoni, S. (2014). Emotion as a boost to metacognition: how worry enhances the quality of confidence. *Consciousness and Cognition*, *29*, 189–198. doi:10.1016/j.concog.2014.08.006

Mast, F. W., Preuss, N., Hartmann, M., & Grabherr, L. (2014). Spatial cognition, body representation and affective processes: the role of vestibular information beyond ocular reflexes and control of posture. *Frontiers in integrative neuroscience*, *8*, 44. doi:10.3389/fnint.2014.00044

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*(39). doi:10.3389/fnhum.2011.00039

McCulloch, K. C., & Smallman, R. (2014). The implications of counterfactual mind-sets for the functioning of implementation intentions. *Motivation and Emotion*. doi:10.1007/s11031-014-9408-3

McGuire, M. T., Marks, I., Nesse, R. M., & Troisi, A. (1992). Evolutionary biology: a basic science for psychiatry? *Acta Psychiatrica Scandinavica*, *86*(2), 89–96. doi:10.1111/j.1600-0447.1992.tb03234.x

McHugo, M., Olatunji, B. O., & Zald, D. H. (2013). The emotional attentional blink: what we know so far. *Frontiers in human neuroscience*, *7*, 151. doi:10.3389/fnhum.2013.00151

McIntosh, A. R., Fitzpatrick, S. M., & Friston, K. J. (2001). On the marriage of cognition and neuroscience. *NeuroImage*, *14*(6), 1231–1237. doi:10.1006/nimg.2001.0941

McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior*, *31*(5), 309–319. doi:10.1016/j.evolhumbehav.2010.04.005

McKay, R., Langdon, R., & Coltheart, M. (2007). Models of misbelief: Integrating motivational and deficit theories of delusion. *Consciousness and Cognition*, *16*, 932–941. doi:10.1016/j.concog.2007.01.003

McKay, R., Langdon, R., & Coltheart, M. (2009). "Sleights of Mind". Delusions and Self-Deception. In T. Bayne & J. Fernández (Eds.), *Macquarie monographs in cognitive science. Delusion and self-deception. Affective and motivational influences on belief formation* (pp. 165–186). New York, NY: Psychology Press.

McKay, R., Tamagni, C., Palla, A., Krummenacher, P., Hegemann, S. C., Straumann, D., & Brugger, P. (2013). Vestibular stimulation attenuates unrealistic optimism. *Cortex*, *49*(8), 2272–2275. doi:10.1016/j.cortex.2013.04.005

McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, *32*(06), 493. doi:10.1017/S0140525X09990975

Mechelli, A., Price, C. J., Friston, K. J., & Ishai, A. (2004). Where bottom-up meets top-down: neuronal interactions during perception and imagery. *Cerebral cortex (New York, N.Y. : 1991)*, *14*(11), 1256–1265. doi:10.1093/cercor/bhh087

Mele, A. (2000). Self-deception and emotion. *Consciousness & Emotion*, *1*(1), 115–137. doi:10.1075/ce.1.1.07mel

Mele, A. (2009). Self-Deception and Delusion. In T. Bayne & J. Fernández (Eds.), *Macquarie monographs in cognitive science. Delusion and self-deception. Affective and motivational influences on belief formation* (pp. 55–86). New York, NY: Psychology Press.

Mele, A. (2010). Approaching self-deception: How Robert Audi and I part company☆. *Consciousness and Cognition*, *19*(3), 745–750. doi:10.1016/j.concog.2010.06.009

Mele, A. R. (2001). *Self-deception unmasked*. *Princeton monographs in philosophy*. Princeton, NJ: Princeton Univ. Press. Retrieved from http://www.loc.gov/catdir/samples/prin031/00032626.html

Mele, A. R. (2012). When are we self-deceived? *Humana.Mente Journal of Philosophical Studies*, *20*, 1–15.

Menzel, C. (2015). Possible Worlds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2015th ed.). Retrieved from http://plato.stanford.edu/archives/sum2015/entries/possible-worlds/

Metcalfe, J. (1998). Cognitive Optimism: Self-Deception or Memory-Based Processing Heuristics? *Personality and Social Psychology Review*, *2*(2), 100–110. doi:10.1207/s15327957pspr0202_3

Metcalfe, J., Cottrell, G. W., & Mencl, W. E. (1992). Cognitive Binding: A Computational-Modeling Analysis of a Distinction between Implicit and Explicit Memory. *Journal of Cognitive Neuroscience*, *4*(3), 289–298.

Metzinger, T. (2013a). The myth of cognitive agency: subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in psychology*, *4*, 931. doi:10.3389/fpsyg.2013.00931

Metzinger, T. (2013b). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research1. *Frontiers in Psychology*, *4*, 746. doi:10.3389/fpsyg.2013.00746

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. *A Bradford book*. Cambridge, Mass: MIT Press.

Metzinger, T. (2011). The No-Self Alternative. In S. Gallagher (Ed.), *Oxford handbooks in philosophy. The Oxford handbook of the self* (pp. 279–296). Oxford: Oxford University Press.

Metzinger, T. (2014). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience*, *5*(2), 122–124. doi:10.1080/17588928.2014.905519

Metzinger, T. (2015). M-Autonomy. *Journal of Consciousness Studies*, *22*(11-12).

Metzinger, T., & Windt, J. (2014). Die Phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath, & J. Kipper (Eds.), *Suhrkamp Taschenbücher Wissenschaft: Vol. 2094. Die experimentelle Philosophie in der Diskussion* (1st ed., pp. 279–321). Berlin: Suhrkamp.

Michel, C. (2014). *Self-knowledge and Self-deception: The Role of Transparency in First Personal Knowledge*. Münster, Westf: mentis.

Michel, C., & Newen, A. (2010). Self-deception as pseudo-rational regulation of belief☆. *Consciousness and Cognition*, *19*(3), 731–744. doi:10.1016/j.concog.2010.06.019

Mihov, K. M., Denzler, M., & Förster, J. (2010). Hemispheric specialization and creative thinking: A meta-analytic review of lateralization of creativity. *Brain and Cognition*, *72*, 442–448.

Mijovic-Prelec, D., & Prelec, D. (2009). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1538), 227–240. doi:10.1098/rstb.2009.0218

Miłkowski, M. (2015). Satisfaction conditions in anticipatory mechanisms. *Biology & Philosophy*, *30*(5), 709–728. doi:10.1007/s10539-015-9481-3

Millham, J., & Kellogg, R. W. (1980). Need for Social Approval: Impression Management or Self-Deception? *Journal of research in personality*, *14*, 445–457.

Mograbi, D. C., & Morris, R. G. (2013). Implicit awareness in anosognosia: Clinical observations, experimental evidence, and theoretical implications. *Cognitive Neuroscience*, *4*(3-4), 181–197. doi:10.1080/17588928.2013.833899

Moran, R. J., Symmonds, M., Dolan, R. J., Friston, K. J., & Sporns, O. (2014). The Brain Ages Optimally to Model Its Environment: Evidence from Sensory Learning over the Adult Lifespan. *PLoS Computational Biology*, *10*(1), e1003422. doi:10.1371/journal.pcbi.1003422

Moritz, S., Voigt, M., Köther, U., Leighton, L., Kjahili, B., Babur, Z.,. . . Grzella, K. (2014). Can virtual reality reduce reality distortion? Impact of performance feedback on symptom change in schizophrenia patients. *Journal of Behavior Therapy and Experimental Psychiatry*, *45*(2), 267–271. doi:10.1016/j.jbtep.2013.11.005

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014a). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition*, *25*, 67–76. doi:10.1016/j.concog.2014.01.009

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014b). A formal model of interpersonal inference. *Frontiers in human neuroscience*, *8*, 160. doi:10.3389/fnhum.2014.00160

Müller, P. A., & Stahlberg, D. (2007). The role of surprise in hindsight bias: A metacognitive model of reduced and reversed hindsight bias. *Social Cognition*, *25*(1), 165–184.

Myers, A. L., McCrea, S. M., & Tyser, M. P. (2014). The role of thought-content and mood in the preparative benefits of upward counterfactual thinking. *Motivation and Emotion*, *38*(1), 166–182. doi:10.1007/s11031-013-9362-5

Nachson, I. (1997). Neuropsychology of Self-Deception: The Case of Prosopagnosia. In M. Myslobodsky (Ed.), *The mythomanias. The nature of deception and self-deception* (pp. 277–305). Mahwah, N.J: L. Erlbaum Associates.

Nagel, J. (2014). The Meanings of Metacognition. *Philosophy and Phenomenological Research*, *89*(3), 710–718. doi:10.1111/phpr.12145

Nelkin, D. K. (2002). Self-deception, motivation, and the desire to believe. *Pacific Philosophical Quarterly*, *83*, 384–406.

Nelkin, D. K. (2012). Responsibility and Self-Deception: A Framework. *Humana.Mente Journal of Philosophical Studies*, *20*, 117–139.

Newen, A. (2015). Understanding Others - The Person Model Theory. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (26T). Frankfurt am Main: MIND Group.

Nicholson, A. (2007). Cognitive Bias, Intentionality and Self-Deception. *teorema*, *XXVI*(3), 45–58.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Noordhof, P. (2003). Self-Deception, Interpretation and Consciousness. *Philosophy and Phenomenological Research*, *67*(1), 75–100. Retrieved from http://www.jstor.org/stable/20140582

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

Oaksford, M., & Chater, N. (2009). Précis of bayesian rationality: The probabilistic approach to human reasoning. *The Behavioral and brain sciences, 32*(1), 69-84; discussion 85-120. doi:10.1017/S0140525X09000284

Olsen, R. K., Moses, S. N., Riggs, L., & Ryan, J. D. (2012). The hippocampus supports multiple cognitive processes through relational binding and comparison. *Frontiers in human neuroscience*, *6*, 146. doi:10.3389/fnhum.2012.00146

Panksepp, J., & Panksepp Jules B. (2000). The Seven Sins of Evolutionary Psychology. *Evolution and Cognition*, *6*(2), 108–131.

Park, J., & Sloman, S. A. (2014). Causal explanation in the face of contradiction. *Memory & Cognition.* doi:10.3758/s13421-013-0389-3

Parvizi, J., & Damasio, A. (2001). Consciousness and the brainstem. *Cognition, 79*, 135–139.

Patten, D. (2003). How do we deceive ourselves? *Philosophical Psychology, 16*(2), 229–246.

Paulhus, D., & Buckels, E. (2012). Classic Self-Deception Revisited. In S. Vazire & T. Wilson (Eds.), *Handbook of Self-Knowledge* (pp. 363–378). New York: Guilford Publications.

Paulhus, D. L. (1984). Two-component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609.

Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology*, *74*(5), 1197–1208. doi:10.1037/0022-3514.74.5.1197

Paulhus, D. L., & John, O. P. (1998). Egoistic and Moralistic Biases in Self-Perception:: The Interplay of Self-Decceptive Styles With Basic Traits and Motives. *Journal of Personality and Social Psychology, 66*(6).

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and Denial in Socially Desirable Responding. *Journal of Personality and Social Psychology*, *60*(2), 307–317.

Pears, D. (1986). The goals and strategies of self-deception. In J. Elster (Ed.), *Studies in rationality and social change. The multiple self* (1st ed., pp. 59–77). Cambridge: Cambridge Univ. Press.

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

Pearson, J. M., Watson, K. K., & Platt, M. L. (2014). Decision Making: The Neuroethological Turn. *Neuron, 82*(5), 950–965. doi:10.1016/j.neuron.2014.04.037

Pedrini, P. (2012). What Does the Self-Deceiver Want? *Humana.Mente Journal of Philosophical Studies*, *20*, 141–157.

Peetz, J., Jordan, C. H., & Wilson, A. E. (2014). Implicit Attitudes Toward the Self Over Time. *Self and Identity*, *13*(1), 100–127. doi:10.1080/15298868.2012.762619

Perring, C. (1997). Direct, fully intentional self-deception is also real. *Behavioral and Brain Sciences, 20*(1), 123–124.

Pessoa, L. (2015a). Précis of The Cognitive-Emotional Brain. *Behavioral and Brain Sciences*, *38*(e71), 1–18.

Pessoa, L. (2015b). The cognitive-emotional amalgam. *Behavioral and Brain Sciences, 38*(e91), 47–66. doi:10.1017/S0140525X14001083

Pezzulo, G., Barca, L., & Friston, K. J. (2015). Active inference and cognitive-emotional interactions in the brain. *Behavioral and Brain Sciences*, *38*(e85), 37–39. doi:10.1017/S0140525X14001009

Picard, F. (2013). State of belief, subjective certainty and bliss as a product of cortical dysfunction. *Cortex; a journal devoted to the study of the nervous system and behavior, 49*(9), 2494–2500. doi:10.1016/j.cortex.2013.01.006

Picard, F., & Kurth, F. (2014). Ictal alterations of consciousness during ecstatic seizures. *Epilepsy & Behavior, 30*, 58–61. doi:10.1016/j.yebeh.2013.09.036

Pinker, S. (2011). Representations and decision rules in the theory of self-deception. *Behavioral and Brain Sciences*, *34*(01), 35–37. doi:10.1017/S0140525X1000261X

Pitt, D. (2012). Mental representation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2012nd ed.). Retrieved from http://plato.stanford.edu/archives/win2012/entries/mental-representation/

Pliushch, I. (2015). The Extension of the Indicator-Function of Feelings - A Commentary on Joëlle Proust. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (31C). Frankfurt am Main: MIND Group.

Pliushch, I., & Metzinger, T. (2015). Self-Deception and the Dolphin Model of Cognition. In R. Gennaro (Ed.), *Disturbed consciousness* (pp. 167–207). Cambridge, MA: MIT Press.

Pompili, M., Iliceto, P., Luciano, D., Innamorati, M., Serafini, G., del Casale, A.,. . . Lester, D. (2011). Higher hopelessness and suicide risk predict lower self-deception among psychiatric patients and non-clinical individuals. *Rivista di psichiatria*, *46*(1), 24–30.

Porcher, J. (2012). Against the Deflationary Account of Self-Deception. *Humana.Mente Journal of Philosophical Studies*, *20*, 67–84.

Powell, R., & Clarke, S. (2012). Religion as an Evolutionary Byproduct: A Critique of the Standard Model. *The British Journal for the Philosophy of Science*, *63*(3), 457–486. doi:10.1093/bjps/axr035

Preuss, N., Hasler, G., & Mast, F. W. (2014). Caloric Vestibular Stimulation Modulates Affective Control and Mood. *Brain Stimulation*, *7*, 133–140.

Price, M. C., & Norman, E. (2008). Intuitive decisions on the fringe of consciousness: Are they conscious and does it matter? *Judgment and Decision Making*, *3*(1), 28–41.

Proulx, T., Inzlicht, M., & Harmon-Jones, E. (2012). Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in Cognitive Sciences*, *16*(5), 285–291. doi:10.1016/j.tics.2012.04.002

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford: Oxford Univ. Press.

Proust, J. (2014). Précis of The Philosophy of Metacognition. *Philosophy and Phenomenological Research*, *89*(3), 703–709. doi:10.1111/phpr.12152

Proust, J. (2015a). Feelings as Evaluative Indicators: A Reply to Iuliia Pliushch. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (31R). Frankfurt am Main: MIND Group.

Proust, J. (2015b). The Representational Structure of Feelings. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (31T). Frankfurt am Main: MIND Group.

Pyszczynski, T., & Cox, C. (2004). Can We Really Do Without Self-Esteem? Comment on Crocker and Park (2004). *Psychological Bulletin*, *130*(3), 425–429. doi:10.1037/0033-2909.130.3.425

Pyszczynski, T., Greenberg, J., & Solomon, S. (1999). A dual-process model of defense against conscious and unconscious death-related thoughts: An extension of terror management theory. *Psychological Review*, *106*(4), 835–845.

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J., & Schimel, J. (2004a). Converging Toward an Integrated Theory of Self-Esteem: Reply to Crocker and Nuer (2004), Ryan and Deci (2004), and Leary (2004). *Psychological Bulletin*, *130*(3), 483–488. doi:10.1037/0033-2909.130.3.483

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J., & Schimel, J. (2004b). Why do people need self-esteem? A theoretical and empirical review. *Psychological Bulletin*, *130*(3), 435–468.

Quadt, L. (2015). Multiplicity Needs Coherence - Towards a Unifying Framework for Social Understanding. A commentary on Albert Newen. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (26C). Frankfurt am Main: MIND Group.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, *46*(2), 237–248. doi:10.1037/0022-3514.46.2.237

Rachlin, H., & Frankel, M. (1997). The uses of self-deception. *Behavioral and Brain Sciences*, *20*(1), 124–125.

Radden, J. (2013). Delusions Redux. *Mind & Language*, *28*(1), 125–139.

Ramachandran, V. (1996). The Evolutionary Biology of Self-Deception, Laughter, Dreaming and Depression: Some Clues from Anosognosia. *Medical Hypotheses*, *47*, 347–362.

Reuter, R. A. (1999). *On the function of consciousness: An adaptationist perspective*. Retrieved from http://www.academia.edu/648160/On_the_function_of_consciousness_-_an_adaptationist_perspective

Revonsuo, A. (1999). Binding and the Phenomenal Unity of Consciousness. *Consciousness and Cognition*, *8*, 173–185.

Richardson, R. C. (2007). *Evolutionary psychology as maladapted psychology*. *Life and mind A Bradford book*. Cambridge, Mass: MIT Press.

Rorty, A. O. (1986). Self-deception, akrasia and irrationality. In J. Elster (Ed.), *Studies in rationality and social change. The multiple self* (1st ed., pp. 115–131). Cambridge: Cambridge Univ. Press.

Rorty, A. O. (1988). The Deceptive Self: Liars, Layers, and Lairs. In B. P. McLaughlin & A. O. Rorty (Eds.), *Topics in philosophy: Vol. 6. Perspectives on self-deception* (pp. 11–28). Berkeley Cal.: Univ. of California Press.

Rorty, A. O. (1994). User-Friendly Self-Deception. *Philosophy*, *69*(268), 211–228. Retrieved from http://www.jstor.org/stable/3751346

Rorty, A. O. (2009). User-friendly Self-Deception. A Traveler´s Manual. In C. W. Martin (Ed.), *The philosophy of deception* (pp. 244–259). Oxford: Oxford Univ. Press.

Rosenberg, M., Schooler, C., Schoenbach, C., & Rosenberg, F. (1995). Globar self-esteem and specific self-esteem: Different concepts, different outcomes. *American Sociological Review*, *60*(1), 141–156.

Ryan, R. M., & Deci, E. L. (2004). Avoiding Death or Engaging Life as Accounts of Meaning and Culture: Comment on Pyszczynski et al. (2004). *Psychological Bulletin*, *130*(3), 473–477. doi:10.1037/0033-2909.130.3.473

Sackeim, H. A., & Gur, R. C. (1978). Self-Deception, Self-Confrontation, and Consciousness. In G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and Self-Regulation. Advances in Research and Theory* (Vol. 2, pp. 139–197). New York: Plenum.

Sackeim, H. A., & Gur, R. C. (1979). Self-Deception, Other-Deception, and Self-Reported Psychopathology. *Journal of Consulting and Clinical Psychology*, *47*(1), 213–215.

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

Sadaghiani, S., Hesselmann, G., Friston, K. J., & Kleinschmidt, A. (2010). The relation of ongoing brain activity, evoked neural responses, and cognition. *Frontiers in Systems Neuroscience*, *4, article 20.* doi:10.3389/fnsys.2010.00020

Sahdra, B., & Thagard, P. (2003). Self-deception and emotional coherence. *Minds and machines*, *13*(2), 213–231.

Samuels, R. (2009). The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds. Dual processes and beyond* (pp. 129–146). Oxford: Oxford University Press.

Sandoz, P. (2011). Reactive-homeostasis as a cybernetic model of the silhouette effect of denial of pregnancy. *Medical Hypotheses*, *77*(5), 782–785. doi:10.1016/j.mehy.2011.07.036

Sanford, D. H. (1988). Self-Deception as Rationalization. In B. P. McLaughlin & A. O. Rorty (Eds.), *Topics in philosophy: Vol. 6. Perspectives on self-deception* (pp. 157–169). Berkeley Cal.: Univ. of California Press.

Sass, L. A., & Pienkos, E. (2013). Delusion: The phenomenological approach. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, & J. Z. Sadler (Eds.), *International perspectives in philosophy and psychiatry. The Oxford handbook of philosophy and psychiatry* (pp. 632–657). Oxford: Oxford Univ. Press.

Saunders, C., & Over, D. E. (2009). In two minds about rationality? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds. Dual processes and beyond* (pp. 317–334). Oxford: Oxford University Press.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399.

Schmitt, F. (1988). Epistemic Dimensions of Self-Deception. In B. P. McLaughlin & A. O. Rorty (Eds.), *Topics in philosophy: Vol. 6. Perspectives on self-deception* (pp. 183–204). Berkeley Cal.: Univ. of California Press.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences.* doi:10.1016/j.tics.2011.05.006

Schulte, J. (Ed.). (2011[1953, postum]). *Bibliothek Suhrkamp: Vol. 3010. Philosophische Untersuchungen*. Berlin: Suhrkamp.

Schwabe, L., Merz, C. J., Walter, B., Vaitl, D., Wolf, O. T., & Stark, R. (2011). Emotional modulation of the attentional blink: the neural structures involved in capturing and holding attention. *Neuropsychologia*, *49*(3), 416–425. doi:10.1016/j.neuropsychologia.2010.12.037

Schwartenbeck, P., FitzGerald, Thomas H B, Mathys, C., Dolan, R., & Friston, K. J. (2014). The Dopaminergic Midbrain Encodes the Expected Certainty about Desired Outcomes. *Cerebral cortex (New York, N.Y. : 1991).* doi:10.1093/cercor/bhu159

Schwartenbeck, P., FitzGerald, Thomas H B, Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. J. (2015). Optimal inference with suboptimal models: addiction and active Bayesian inference. *Medical Hypotheses*, *84*(2), 109–117. doi:10.1016/j.mehy.2014.12.007

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

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

Scott-Kakures, D. (2009). Unsettling questions: Cognitive Dissonance in Self-Deception. *Social Theory and Practice*, *35*(1), 73–106.

Scott-Kakures, D. (2012). Can you succeed in intentionally deceiving yourself? *Humana.Mente Journal of Philosophical Studies, 20*, 17–39.

Sedikides, C. (2007). Self-enhancement and self-protection: Powerful, pancultural, and functional. *Hellenic Journal of Psychology, 4,* 1–13.

Sedikides, C., & Alicke, M. D. (2012). Self-enhancement and self-protection motives. In R. M. Ryan (Ed.), *Oxford library of psychology. The Oxford handbook of human motivation* (pp. 303–322). Oxford: Oxford Univ. Press.

Sedikides, C., Herbst, K. C., Hardin, D. P., & Dardis, G. J. (2002). Accountability as a deterrent to self-enhancement: The search for mechanisms. *Journal of Personality and Social Psychology*, *83*(3), 592–605. doi:10.1037//0022-3514.83.3.592

Sedikides, C., & Skowronski, J. (2012). Construct accessibility and interpretation of self-behaviors. Tracing and recuding the signatures of self-protection and self-enhancement. In J. P. Forgas, K. Fiedler, & C. Sedikides (Eds.), *The sydney symposium of social psychology series: Vol. 14. Social thinking and interpersonal behavior* (pp. 239–257). New York, NY: Psychology Press.

Sedikides, C., Skowronski, J., & Gaertner, L. (2004). Self-enhancement and self-protection motivation: From the laboratory to an evolutionary context. *Journal of Cultural and Evolutionary Psychology, 2*, 61–79.

Serrien, D. J., Ivry, R. B., & Swinnen, S. P. (2006). Dynamics of hemispheric specialization and integration in the context of motor control. *Nature Neuroscience, 7,* 160–167.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. doi:10.1016/j.tics.2013.09.007

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience*, *5*(2), 97–118. doi:10.1080/17588928.2013.877880

Seth, A. K. (2015a). Inference to the Best Prediction: A Reply to Wanja Wiese. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (35R). Frankfurt am Main: MIND Group.

Seth, A. K. (2015b). The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (35T). Frankfurt am Main: MIND Group.

Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An Interoceptive Predictive Coding Model of Conscious Presence. *Frontiers in Psychology*, *2*, 395. doi:10.3389/fpsyg.2011.00395

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941-R945. doi:10.1016/j.cub.2011.10.030

Shea, N. (2013). Neural mechanisms of decision making and the personal level. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, & J. Z. Sadler (Eds.), *International perspectives in philosophy and psychiatry. The Oxford handbook of philosophy and psychiatry.* Oxford: Oxford Univ. Press.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193. doi:10.1016/j.tics.2014.01.006

Sheldon, K. M. (2004). The Benefits of a "Sidelong" Approach to Self-Esteem Need Satisfaction: Comment on Crocker and Park (2004). *Psychological Bulletin*, *130*(3), 421–424. doi:10.1037/0033-2909.130.3.421

Shepherd, J. (2014). Deciding as Intentional Action: Control over Decisions. *Australasian Journal of Philosophy*, 1–17. doi:10.1080/00048402.2014.971035

Sheridan, Z., Boman, P., Mergler, A., Furlong, M. J., & Elmer, S. (2015). Examining well-being, anxiety, and self-deception in university students. *Cogent Psychology, 2*(1), 993850. doi:10.1080/23311908.2014.993850

Sherman, D. K., & Cohen, G. L. (2002). Accepting Threatening Information: Self-Affirmation and the Reduction of Defensive Biases. *Current Directions in Psychological Science*, *11*(4), 119–123. doi:10.1111/1467-8721.00182

Shipp, S., Adams, R. A., & Friston, K. J. (2013). Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in neurosciences*, *36*(12), 706–716. doi:10.1016/j.tins.2013.09.004

Siewert, C. P. (1998). *The significance of consciousness*. Princeton, N.J: Princeton University Press.

Sloman, S. A., Fernbach, P. M., & Hagmayer, Y. (2010). Self-deception requires vagueness. *Cognition*, *115*, 268–281. doi:10.1016/j.cognition.2009.12.017

Slotter, E. B., Gardner, W. L., & Finkel, E. J. (2010). Who Am I Without You? The Influence of Romantic Breakup on the Self-Concept. *Personality and Social Psychology Bulletin*, *36*, 147–160.

Smith, D. (2002). The evolution of the unconscious. *Psychoanalytische Perspectieven*, *20*(4), 525–548.

Smith, D. L. (2014). Self-Deception: A Teleofunctional Approach. *Philosophia*, *42*(1), 181–199. doi:10.1007/s11406-013-9464-z

Stankov, L., & Lee, J. (2014). Overconfidence Across World Regions. *Journal of Cross-Cultural Psychology*. doi:10.1177/0022022114527345

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current opinion in neurobiology, 25,* 85–92. doi:10.1016/j.conb.2013.12.007

Sterelny, K. (2007). Cognitive Load and Human Decision, or, Three Ways of Rolling the Rock Uphill. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Volume 2: Culture and Cognition* (pp. 218–234). Oxford: Oxford University Press.

Sturm, T. (2009). Selbsttaeschung: Wer is hier (ir)rational und warum? *Studia Philosophica*, *68*, 229–254.

Sun, R. (2008). Introduction to computational cognitive modeling. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 3–19). Cambridge: Cambridge Univ. Press.

Surbey, M. K. (2004). Self-Deception: Helping and Hindering Public and Personal Decision Making. In C. B. Crawford & C. Salmon (Eds.), *Evolutionary psychology, public policy, and personal decisions* (pp. 117–144). Mahwah, N.J: Lawrence Erlbaum Associates.

Surbey, M. K., & McNally, J. J. (1997). Self-Deception as a Mediator of Cooperation and Defecting in Varying Social Contexts Described in the Iterated Prisoner's Dilemma. *Evolution and Human Behavior*, *18*, 417–435.

Szabados, B. (1973). Wishful Thinking and Self-Deception. *Analysis*, *33*(6), 201–205. Retrieved from http://www.jstor.org/stable/3327197

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

Talbott, W. (1997). Does self-deception involve intentional biasing? *Behavioral and Brain Sciences*, *20*(1), 127.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind*. New York: Basic Books.

Taylor, S. E., & Brown, J. D. (1994). Positive Illusions and Well-Being Revisited. Sepadating Fact From Fiction. *Psychological Bulletin*, *116*(1), 21–27.

Taylor, S. E., Lerner, J. S., Sherman, D. K., Sage, R. M., & McDowell, N. K. (2003). Portrait of the self-enhancer: Well adjusted and well liked or maladjusted and friendless? *Journal of Personality and Social Psychology*, *84*(1), 165–176. doi:10.1037/0022-3514.84.1.165

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some Evidence for Unconscious Lie Detection. *Psychological Science*, *25*(5), 1098–1105. doi:10.1177/0956797614524421

Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical Fading: The Role of Self-Deception in Unethical Behavior. *Social Justice Research*, *17*(2), 223–236.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. doi:10.1016/j.tics.2006.05.009

Tesser, A., Martin, L. L., & Cornell, D. P. (1996). On the Substitutability of Self-Protective Mechanisms. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action. Linking cognition and motivation to behavior* (pp. 48–68). New York, NY: Guilford Press.

Thagard, P. (2000). *Coherence in thought and action*. *Life and mind*. Cambridge, Mass: MIT Press.

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. *A Bradford book*. Cambridge, Mass: MIT Press.

Thagard, P., & Nerb, J. (2002). Emotional Gestalts:: Appraisal, Change, and the Dynamics of Affect. *Personality and Social Psychology Review*, *6*(4), 274–282.

Thagard, P., & Nussbaum, A. D. (2014). Fear-Driven Inference: Mechanisms of Gut Overreaction. In L. Magnani (Ed.), *Studies in Applied Philosophy, Epistemology and Rational Ethics. Model-Based Reasoning in Science and Technology* (Vol. 8, pp. 43–53). Berlin, Heidelberg: Springer Berlin Heidelberg.

Thomas, M. S., & McClelland, J. L. (2008). Connectionist Models of Cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 23–58). Cambridge: Cambridge Univ. Press.

Thomason, R. H. (2010). *Belief, Intention, and Practicality: Loosening up Agents and Their Propositional Attitudes*. Retrieved from http://web.eecs.umich.edu/~rthomaso/documents/pr-reas/ep.pdf

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds. Dual processes and beyond* (pp. 171–195). Oxford: Oxford University Press.

Tiernan, B., Tracey, R., & Shannon, C. (2014). Paranoia and self-concepts in psychosis: A systematic review of the literature. *Psychiatry Research*, *216*(3), 303–313. doi:10.1016/j.psychres.2014.02.003

Treisman, A. (1996). The binding problem. *Current opinion in neurobiology*, *6*, 171–178.

Triandis, H. C. (2009). *Fooling ourselves: Self-deception in politics, religion, and terrorism*. *Contributions in psychology: Vol. 52*. Westport, Conn: Praeger Publishers. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10356289

Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Trivers, R. (1985). *Social evolution*. Menlo Park, Calif: Benjamin/Cummings.

Trivers, R. (1991). Deceit and self-deception. The relationship between communication and consciuosness. In M. H. Robinson & L. Tiger (Eds.), *Man & beast revisited* (pp. 175–191). Washington: Smithsonian Institution Press.

Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, *907*, 114–131.

Trivers, R. (2010). Deceit and Self-Deception. In P. M. Kappeler & J. Silk (Eds.), *Mind the Gap: Tracing the origins of human universals* (pp. 373–393). Berlin, Heidelberg: Springer Berlin Heidelberg.

Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to fool others*. London: Allen Lane.

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* (pp. 105–131). Mahwah, N.J: L. Erlbaum Associates.

Trope, Y., & Liberman, A. (1996). Social Hypothesis Testing: Cognitive and Motivational Mechanisms. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology. Handbook of basic principles* (pp. 239–270). New York: Guilford Press.

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2008). Do Today's Young People Really Think They Are So Extraordinary? An Examination of Secular Trends in

Narcissism and Self-Enhancement. *Psychological Science*, *19*(2), 181–188. doi:10.1111/j.1467-9280.2008.02065.x

Tschacher, W., & Bergomi, C. (2011). Cognitive binding in schizophrenia: weakened integration of temporal intersensory information. *Schizophrenia Bulletin, 37 Suppl 2*, S13-22. doi:10.1093/schbul/sbr074

Turnbull, O. H., Fotopoulou, A., & Solms, M. (2014). Anosognosia as motivated unawareness: the 'defence' hypothesis revisited. *Cortex*, *61*, 18–29. doi:10.1016/j.cortex.2014.10.008

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1993). Probabilistic reasoning. In A. I. Goldman (Ed.), *Readings in philosophy and cognitive science* (pp. 43–68). Cambridge, Mass.: MIT Press.

Uziel, L. (2014). Impression Management (“Lie”) Scales Are Associated With Interpersonally Oriented Self-Control, Not Other-Deception. *Journal of Personality*, *82*(3), 200–212. doi:10.1111/jopy.12045

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology*, *44*(5), 1334–1338. doi:10.1016/j.jesp.2008.03.010

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

Van Leeuwen, N. D. S. (2007b). The Spandrels of Self-Deception: Prospects for a Biological Theory of a Mental Phenomenon. *Philosophical Psychology*, *20*(3), 329–348. doi:10.1080/09515080701197148

Van Leeuwen, N. D. S. (2008). Finite rational self-deceivers. *Philosophical Studies*, *139*(2), 191–208. doi:10.1007/s11098-007-9112-1

Van Leeuwen, N. D. S. (2009). Self-Deception Won´t Make You Happy. *Social Theory and Practice*, *35*(1), 107–132.

Van Leeuwen, N. D. S. (2013a). Self-Deception. In H. LaFollette (Ed.), *International Encyclopedia of Ethics* (p. 486). Oxford: Wiley-Blackwell.

Van Leeuwen, N. D. S. (2013b). The folly of fools: The logic of deceit and self-deception in human life. *Cognitive Neuropsychiatry*, *18*(1-2), 146–151. doi:10.1080/13546805.2012.753201

van Vugt, Marieke K., & Slagter, H. A. (2014). Control over experience? Magnitude of the attentional blink depends on meditative state. *Consciousness and Cognition*, *23*, 32–39. doi:10.1016/j.concog.2013.11.001

Vance, J. (2014). Cognitive Penetration and the Tribunal of Experience. *Review of Philosophy and Psychology*. doi:10.1007/s13164-014-0197-0

Varki, A., & Brower, D. L. (2013). *Denial: Self-deception, false beliefs, and the origins of the human mind* (1st ed.). New York: Twelve.

Vauclair, J., & Donnot, J. (2005). Infant holding biases and their relations to hemispheric specializations for perceiving facial emotions. *Neuropsychologia*, *43*, 564–571.

Verguts, T., & Notebaert, W. (2009). Adaptation by binding: a learning account of cognitive control. *Trends in Cognitive Sciences*, *13*(6), 252–257. doi:10.1016/j.tics.2009.02.007

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, *27*, 62–75. doi:10.1016/j.concog.2014.04.007

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(45), 16214–16218. doi:10.1073/pnas.1403619111

von Hippel, W., & Trivers, R. (2011a). Reflections on self-deception. *Behavioral and Brain Sciences*, *34*(01), 41–56. doi:10.1017/S0140525X10003018

von Hippel, W., & Trivers, R. (2011b). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(01), 1–16. doi:10.1017/S0140525X10001354

Waskan, J. A. (2006). *Models and cognition: Prediction and explanation in everyday life and in science*. *A Bradford book*. Cambridge, Mass: MIT Press.

Weisberg, J. (2005). Consciousness Constrained: A commentary on Being No One. *Psyche*, *11*(5), 1–22.

Why, Y. P., & Huang, R. Z. (2011). Positive illusions and its association with cardiovascular functions. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, *81*(3), 305–311. doi:10.1016/j.ijpsycho.2011.07.016

Windt, J. M., Harkness, D. L., & Lenggenhager, B. (2014). Tickle me, I think I might be dreaming! Sensory attenuation, self-other distinction, and predictive processing in lucid dreams. *Frontiers in human neuroscience*, *8*, 717. doi:10.3389/fnhum.2014.00717

Wright, Gordon R T, Berry, C. J., Catmur, C., & Bird, G. (2015). Good Liars Are Neither 'Dark' Nor Self-Deceptive. *PLoS ONE*, *10*(6), e0127315. doi:10.1371/journal.pone.0127315

Wu, W. (2011). Confronting Many-Many Problems: Attention and Agentive Control. *Noûs*, *45*(1), 50–76.

Wu, W. (2013). Mental Action and the Threat of Automaticity. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the Will* (pp. 244–261). Oxford University Press.

Yamada, M., Uddin, L. Q., Takahashi, H., Kimura, Y., Takahata, K., Kousa, R.,. . . Suhara, T. (2013). Superiority illusion arises from resting-state brain networks modulated by dopamine. *Proceedings of the National Academy of Sciences.* doi:10.1073/pnas.1221681110

Yanal, R. J. (2007). Self-deception and the experience of fiction. *Ratio*, *20*(1), 108–121. doi:10.1111/j.1467-9329.2007.00350.x

Yee, N., Bailenson, J. N., & Ducheneaut, N. (2009). The Proteus Effect: Implications of Transformed Digital Self-Representation on Online and Offline Behavior. *Communication Research*, *36*(2), 285–312. doi:10.1177/0093650208330254

Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural Mechanisms of Belief Inference during Cooperative Games. *Journal of Neuroscience*, *30*(32), 10744–10751. doi:10.1523/JNEUROSCI.5895-09.2010

Yoshie, M., & Haggard, P. (2013). Negative Emotional Outcomes Attenuate Sense of Agency over Voluntary Actions. *Current Biology*, *23*(20), 2028–2032. doi:10.1016/j.cub.2013.08.034

Young, M. C. (2013). Do positive illusions contribute to human well-being? *Philosophical Psychology*, *27*(4), 536–552. doi:10.1080/09515089.2013.764860

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. doi:10.1016/j.neuron.2005.04.026

Zadra, J. R., & Clore, G. L. (2011). Emotion and Perception: The Role of Affective Information. *Wiley interdisciplinary reviews. Cognitive science*, *2*(6), 676–685. doi:10.1002/wcs.147

Zehetleitner, M., & Schönbrodt, F. (2014). When misrepresentation is successful. In T. Breyer (Ed.), *Epistemological foundations of evolutionary psychology*. New York: Springer.

Zeigler-Hill, V. (2006). Discrepancies Between Implicit and Explicit Self-Esteem: Implications for Narcissism and Self-Esteem Instability. *Journal of Personality*, *74*(1), 119–144. doi:10.1111/j.1467-6494.2005.00371.x