The Overtone Model of Self-Deception

Iuliia Pliushch

In this paper I will argue for what I call an 'overtone model of self-deception'. The analogy to overtones (higher-order frequencies of a tone) is as follows: a self-deceiver's optimal degree of instability (the term is borrowed from Friston et al. 2012a, and applied to self-deception) is elevated so that constant exploration (of a certain number of hypotheses) is pursued instead of disambiguation in favor of a certain hypothesis. These hypotheses are explored in parallel (for similar ideas with respect to higher-order cognition in general see Pezzulo and Cisek 2016, and Metzinger 2017) and are like overtones of the currently active self-deceptive hypothesis (the base frequency) so that what we as self-deceivers, as well as observers, perceive as one tone (self-deception) is actually a fusion of different frequencies. The term 'fusion' is relevant because the phenomenology of the self-deceiver is co-determined by overtones.

Keywords

Binocular rivalry | Counterfactuals | Interoception | Overtones | Predictive coding | Predictive processing | Self-deception

1 Introduction

Recently, I was sitting in a train when I heard the following snippet of a conversation "Well, you know how the saying goes: I have gone through many terrible things and some of them really *were* terrible". In other words, what one perceives as terrible may, on reflection, turn out not to be so. Thus, humans often misrepresent things, be it for the better (optimists) or for the worse (pessimists). One may say that misrepresentation already is self-deception, or one may require additional criteria to label something ´self-deception´. In the following, I will elaborate on the kinds of behavioral and phenomenological criteria those may be. In any case, the above example shows that our perception of the world and of ourselves is always filtered. This is the basis for several phenomena that may occur, for example delusions or illusions, and also self-deception. Though more often than not self-deception is described as over-optimism or denial, such that the aspects filtered out are negative, this is not necessarily the case (see, e.g., Mele 2001 for "twisted" self-deception). The basis for both is a certain kind of distortion, which, though elusive when philosophically analyzed (there is still no consensus on what self-deception is, how it is brought about, or whether one should relinquish — abandon — it), is a frequent visitor of everyday conversations. To err on the side of caution, I do not claim that the above conversation demonstrates a case of self-deception — there is not enough information for such a judgement — but rather that, if enriched with details (see the template below), it can become one.

On the one hand, we are able to successfully interact in the world we live in, which means that there has to be at least a kernel of truth in the way we perceive it. On the other hand, there is quite some divergence between the way the world is and how we perceive it, else there would be no research on self-deception. As I will argue later, self-deception is a cluster concept. Many different phenomena have been subsumed under it. For the reader to get a better feeling for this concept, I will first give a folk-psychological template for self-deception. This is how we get acquainted with it in our everyday life. A word of caution though: the template involves the simplistic folk-psychological assumption that self-deception is a personal-level process, which is to say that self-deceivers possess an epistemic agent model (EAM, see Metzinger 2017) — a conscious model of them as directing their attention and guiding their reasoning process in a certain manner. I will question this assumption in the next section.

Pliushch, I. (2017). The Overtone Model of Self-Deception.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 21. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573222

Self-Deception Template
A is motivated to believe X. She starts the hypothesis testing process with the aim of finding the truth. Somehow, despite the unbiased *evidence* (or at least unbiased information, because acknowledging something as evidence is already a step further in information processing[1]) pointing into one direction, she ends up believing the conclusion that is not supported by the evidence because of *motivation Y* [the goal representation to believe this false conclusion/believe otherwise/relieve anxiety etc]. While being self-deceived, she experiences (at least sometimes) *tension* because of the inconsistency between the acquired conclusion and the evidence. Upon relinquishing self-deception, *A* experiences *insight*: she has the feeling of having known the truth all along. When 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 with regard to her belief that *X*. The reader is encouraged to fill out *X*, *Y* and the concrete arguments with cases from her personal life.

Examples covered in the philosophical literature include denial of having a terminal illness such as cancer (Rorty 1988), denial of the unfaithfulness of one's spouse (Funkhouser 2009), and denial of one's child's criminal inclinations (Mele 2001). Among the psychological examples not only can one list unrealistic optimism and self-enhancement (Taylor 1989; Von Hippel and Trivers 2011), but also anosognosia, the denial of one's own paralysis (Levy 2009).

In this paper, inconsistency as characteristic of self-deception will take center stage. I employ the term 'inconsistency' here in a broad way as either first- or third-person inconsistency. First-person inconsistency is the possession of contradictory representations. Third-person inconsistency is behaving in a manner contradictory to one's verbal assertions, or to one's own behavioural dispositions. I call it 'third-person inconsistency,' because this kind of inconsistency is noticed and pointed out to us by others (friends, relatives, observers) and leads to another characteristic of self-deception, namely that self-deceivers *justify* themselves rather than relinquishing the self-deception. I employ this broad formulation of inconsistency because I do not want to commit myself to either an intentional or a deflationary account of self-deception. One point of disagreement between these two accounts is whether self-deception involves contradictory mental states. According to Alfred Mele's (Mele 2012) deflationary account, self-deception results from motivationally biased belief acquisition. In this vein, Erik Helzer and David Dunning (Helzer and Dunning 2012) hold that "motivated reasoning emerges as a paradigmatic case of self-deception". Yet this account is not without its problems, one of which is the 'slippery slope' problem that has been brought by Neil Van Leeuwen (Van Leeuwen 2013) against Robert Trivers (Trivers 2011) examples of self-deception. Van Leeuwen's (Van Leeuwen 2013) worry is that if the category of self-deception encompasses too large a range of phenomena, then the concept of self-deception would lose its scientific value. This criticism can be applied to the deflationary definition of self-deception as motivationally biased beliefs as well. If motivated biases are defined as those that "may be triggered and sustained by desires in the production of *motivationally* biased beliefs" (Mele 2012, p. 7), and if every cold (not motivated) bias can be triggered by motivation, then every bias is — potentially — self-deceptive. In this paper, to make the most philosophically compelling case, I sketch how one could analyze the strongest kind of inconsistency possible in self-deception. Weaker cases of inconsistency would be easier to analyze in a similar manner.

The main goal of this paper is to present a predictive processing inspired account of how to explain inconsistency in self-deception. First, I will present my own account of self-deception as a cluster concept with a certain behavioral and phenomenological profile. I will describe its profile, situate inconsistency as one of the behavioral characteristics of self-deception, and elaborate on how inconsistency has been previously incorporated into philosophical theories of self-deception. Thereafter, I will briefly introduce the predictive processing tools that I will need in order to, lastly, propose my

1 For discussion on acknowledging evidence see, for example, Michel and Newen 2010, and Bagnoli 2012.

'overtone theory of self-deception' for how inconsistency can be analyzed in cases of self-deception. I will argue that overtones enrich our phenomenal experience not only by generating tension, but also in enabling the experience of more nuanced affective consequences such as being glad that something is the case or anticipating that it is the case.

2 The Philosophical Problem: Self-Deception

In this section I will present my own account of self-deception and explain how inconsistency fits into the picture. Existing accounts of self-deception struggle in satisfying two important constraints for an adequate theory of self-deception, namely the *parsimony* and the *demarcation* constraints. The first constraint can be formulated as the requirement to keep the analysis of self-deception as simple as possible, e.g. not to postulate unnecessary internal states. The second requires identifying the criteria for distinguishing self-deception from other phenomena.

The necessity of the first constraint arises from the debate about the *nature of self-deceptive representations*, i.e. which kind of attitude self-deception is. Although the nature of a self-deceptive representation is standardly thought to be belief, alternatives exist. One is *pretense*, which is an attitude akin to imagining that fulfills a belief-like role (Gendler 2007). *Avowal* is another example of a belief-like attitude that has been argued to be produced by the process of self-deception (Audi 1997). A criticism brought against *avowal* and in favor of *belief* as a self-deceptive attitude is that there has to be independent motivation, apart from the wish to solve the paradox of self-deception, in order to postulate the attitude of avowal (Van Leeuwen 2007, p. 429-431). I think the parsimony constraint has to be applied not only to the nature of self-deceptive representations, but also to the question of the kind of self-deceptive motivation and the nature of the self-deceptive process.

The second constraint has its roots in the *intentionalist-deflationary debate*: it has been used as a point of critique for accounts of self-deception. The intentionalist-deflationary debate is about the nature of the motivation for self-deception. Intentionalists argue that self-deceivers are motivated by intentions, while proponents of deflationary theory argue for specific kinds of desires. The so-called selectivity problem is an example of the failure of the demarcation constraint. Curiously, both intentionalist and deflationary accounts have been argued to suffer from it. For example, it has been argued by Mele (Mele 2001) that it is possible to imagine cases in which one would not be able to self-deceive despite the intention to do so (p. 66). This is the so-called dynamic paradox of self-deception: if one has the intention to self-deceive, then in order to carry out this intention, one would need make oneself believe something one does not currently believe. On the other hand, José Luis Bermúdez (Bermúdez 2000) argues that it is not always the case that non-intentional motivation, such as desire, leads to the acceptance of certain self-deceptive hypotheses (p. 317). Another example is the criticism voiced by Neil Van Leeuwen against Robert Trivers that I mentioned in the previous section.

My solution to the parsimony and demarcation constraints is that self-deception is a cluster concept with a certain phenomenological and behavioral profile. Self-deception is a cluster concept in virtue of the slippery slope that has been enabled by the understanding of self-deception as a motivationally biased belief. The more behavioral and phenomenological properties characteristic of self-deception a phenomenon possesses, the stronger a case of self-deception it is. For example, inconsistency and justification belong to the behavioral properties. In other words, self-deceivers are prone to behave inconsistently while justifying their behavior. The self-deceiver's phenomenology is characterized by tension (feelings of uneasiness or distress), and insight when self-deception is relinquished. I will leave open the question of which and how many properties are required for a minimal case of self-deception. I would, however, like to frame self-deceptive motivation in terms of goal representations, because they are folk-psychologically neutral, unlike to intentions or desires.

From the fact that the self-deceiver's behavior is inconsistent, it might be deduced that there must be an inconsistency in their belief set. This is an inference to the best explanation given the following

three assumptions. First, beliefs are stored entities.[2] Second, beliefs (more often than other attitudes) determine our actions (Van Leeuwen 2007). Third, self-deceivers behave inconsistently (Funkhouser 2005; Funkhouser 2009). One possibility for framing this inconsistency is to argue that self-deceivers possess inconsistent *beliefs* (Davidson 1986).

Since, on the assumption that self-deceivers are generally rational beings, it is difficult to explain how such an inconsistency is possible, the inconsistency requirement has been weakened in many recent accounts to involve, for example, a belief and a suspicion that the belief is false (Mele 2001; Mele 2012), or attitudes other than beliefs are argued to result from self-deception, such as avowal (Audi 1997) or pretense (Gendler 2007). The implicit assumption in these accounts is that consistency is only required among attitudes of the same kind, whereas different kinds of attitudes can contradict each other. Psychological experiments testing self-deceptionsuppose the presence of inconsistency not among propositional attitudes, but between a propositional attitude and skin conductance response (Gur and Sackeim 1979) or between different types of processing: conscious/unconscious, implicit/explicit, automatic/controlled (Von Hippel and Trivers 2011).

In the remainder of this section I will address the process of self-deception and the properties of self-deceptive representations, like transparency or its affective component. In doing this, I will use the terminology of Thomas Metzinger's self-model theory of subjectivity. According to this theory, phenomenal mental models are those that incorporate consciously experienced content (Metzinger 2003). What an agent thinks of as real, is part of the transparent *world-model*, like an apple on a table. Transparency means that earlier processing stages are inaccessible. Thus, if something is transparent, it is experienced as real and not as a representation. This is true with respect to our experience of the world and of ourselves. We experience the world as real, not as the result of the transparent phenomenal world-model that has been constructed. The same applies to the phenomenal *self-model*. It is not the case, though, that the world- and self-models are the only models that an agent possesses. From time to time, humans use their cognitive capacities such that epistemic agent models (EAMs, see Metzinger 2017) are constructed. Epistemic agent models are transparent, conscious self-representations of the agent as possessing the capacity for epistemic agency (control of attention and control of goal-directed thoughts) and/or actually executing epistemic actions (Metzinger 2013; Pliushch and Metzinger 2015).

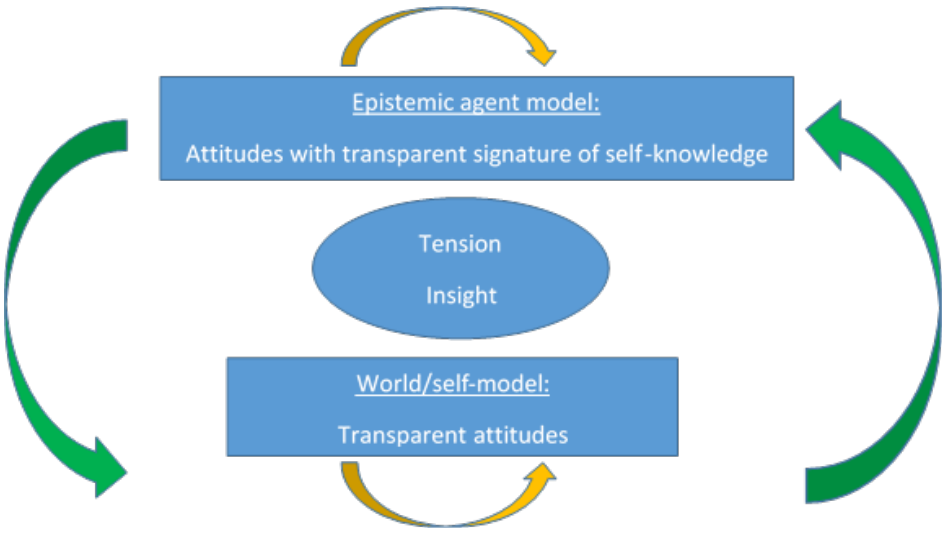

Figure 1. Phenomenology of the self-deceiver. Tension and insight are phenomenological characteristics of self-deception. The two kinds of model that can accommodate self-deceptive representations are the epistemic agent model and the world/self-model. The arrows denote changes in and between models due to changes in their various aspects, like transparency of certain attitudes.

2 For a criticism of this view and an argument in favour of a constructivist view which holds that beliefs are reevaluated and constructed at different time slices instead of being retrieved, see Michel 2014.

The application of these concepts to self-deception is as follows: the template I presented in the introduction suggests that a self-deceptive process involves what is folk-psychologically called 'conscious thoughts'. Mind wandering may also contain conscious thoughts, but since it lacks such characteristics as veto control (it cannot always be terminated at will) and epistemic agency, it is just a subpersonal process that has become conscious (Metzinger 2015, Metzinger 2017). Conscious thinking is a subpersonal process in that it is the result of specific patterns of neural activity 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., directing attention or controlling strands of thought. The unfolding of the cognitive landscape is an alternation between mind wandering episodes and episodes of epistemic agent model construction. The switches between the two often go unnoticed and have, thus, been called a 'self-representational blink' (Metzinger 2013). What about the switches in directing an agent's attention from one object to another? Or the control of the succession of thoughts? 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, without noticing it , do three steps at once that need to be explained to you in detail) and merging. As an example of merging, consider the argumentative fallacy of equivocation: using a certain concept in one way for the first part of an argument and then switching to another of its connotations in the second part. Here, the concept's ambiguity allowed its misuse. This gapful nature of reasoning may be described according to the dolphin model of cognition (Pliushch and Metzinger 2015) such that our argumentations are dolphin-like trajectories in the argumentative space that at certain points can merge and mutate into one another. In predictive processing terms, the reason for changes and fusions can be argued to consist in the ever changing precisions of pursued policies.

If self-deception is achieved by means of epistemic agent model construction, or reasoning in folk-psychological terms, then the result would be a certain attitude that is not experienced as real, but as a representation: as something that can be true or false. Another possibility is that a self-deceptive representation becomes transparent and thus becomes part of the agent's world- or self-model. Anosognosia, the denial of a disability, may be one example, such as when someone's misrepresentation that their arm is not paralyzed becomes transparent (Pliushch and Metzinger 2015). Another example is denial of pregnancy. Denial of pregnancy does not exhibit such characteristics of self-deception as inconsistency from the third-person perspective (in this case nobody would notice that the self-deceiver is self-deceiving) and, hence, there is no need for justification. No tension is present either, but there is something, below awareness, at the level of the autonomic system, which can be described as "knowledge" of the pregnancy (Sandoz 2011, p. 784). The reason for theassumption of such knowledge is a silhouette effect: the figure of someone denying pregnancy does not change; instead the fetus expands in a vertical direction (that of diaphragm; p. 783). When, upon medical examination, insight into their body condition comes, there are cases where almost instantly the position of the fetus changes and the pregnant woman's silhouette changes in front of the 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 by Sandoz 2011 (p. 784) for the silhouette effect is *reactive homeostasis:* reflexive immediate escape from the emergent situation. An emergent situation is one of denial, and is described as a case of the "persistence of paradoxical realities" (p. 784). The author wonders which kind of informational pathway between the semantic knowledge of pregnancy provided by the doctor and the autonomic system might have led to the immediate appearance of pregnancy signs. The tentative answer would be that the cognitive and interoceptive domain is more tightly connected than has been assumed thus far (see Pezzulo and Cisek 2016).

One should not overlook that a more thorough investigation of the phenomenology of the self-deceived is needed in order to establish whether in cases when the self-deceptive attitude is not itself transparent, it nevertheless contains another transparent element, namely the phenomenal signature of knowing. This may be a feature that self-deceptive attitudes share with intuitions. For example,

Thomas Metzinger and Jennifer Windt (Metzinger and Windt 2014) argue that in intuitions of certainty, the *phenomenal signature of knowing* has become transparent. 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). Self-deceivers may then also be susceptible to the epistemological fallacy (E-fallacy), which consists in ascribing epistemic status to phenomenal states due to the presence of the phenomenal signature of knowing.

It is possible that the phenomenal signature of knowing may become transparent through repeated explanation giving. Sangeet Khemlani and Philip Johnson-Laird (Khemlani and Johnson-Laird 2012) conducted experiments in which participants had to detect logical inconsistencies. Their results suggest that when participants actively construct explanations of inconsistencies, this makes it harder for participants to detect them afterwards. In an attempt to frame self-deception within predictive processing, consider the following findings. Ecstatic seizures evoke a transparent feeling of knowing (a subjective feeling of certainty) and have been explained as a case when an interoceptive mismatch is not computed, such that prediction errors remain unexplained 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, without knowing why.

Apart from transparency, another feature that attitudes may vary in are affective consequences. The affective dimension possesses many roles in self-deception:

- Cause/motivation (as desires) (e.g., Mele 2000)
- Phenomenological accompaniment or tension (e.g., Noordhof 2003)
- Functional role, e.g. to 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* according to which there are two sets of constraints — cognitive (activations) and emotional (valences) such that which attitude is accepted depends not only on cognitive, but also emotional constraints (Sahdra and Thagard 2003)

If Ray Jackendoff (Jackendoff 2012) is correct in arguing that every thought possesses an affective component, then, similarly to transparency, it is a feature of self-deceptive attitudes (and not only them, but of all attitudes in general) that they can vary in affective properties.

Summing up, how many attitudes self-deception involves; and which propositional content or other characteristics, such as transparency or affective consequences, self-deceptive attitudes possess, may vary. As a consequence, it is not only that propositional content of asserted self-deceptive attitudes may vary over time (the yellow arrows in figure 1), but also that the model to which the self-deceptive attitude belongs may change too (the green arrows in figure 1). An example of the change in the epistemic agent model over time is the claim that self-deceptive attitudes vary in their degree of conviction/certainty (see, e.g., Lynch 2012; Porcher 2012). An example of the change of the model itself over time would be a case in which self-deceptive attitudes become transparent.

3 New Conceptual Tools: The Predictive Processing Approach

In the previous section I suggested that self-deception can concern both the construction of epistemic agent models and world/self-models. Epistemic agent models — 'conscious thinking' — can be understood as personal level hypothesis testing, particularly given the assumption that human beings are mostly rational: a hypothesis is chosen, evidence is sought out, and conclusions are drawn depending on whether the hypothesis is supported or not. If predictive processing is true, then both the epistemic agent model and the world/self-model construction are cases of hypothesis testing, but of a subpersonal kind. This is the case because, first, predictive processing is a theory about subpersonal

level of information processing, according to which perception, action and cognition are the results of hierarchical prediction error minimization. Second, it is an inference to the best explanation that the "selection of mental representations into consciousness" can be accommodated by the predictive processing framework (Hohwy 2015, p. 321).

An obvious benefit for analyses of self-deception using predictive processing is that rationality is a personal-level property: hypotheses at the subpersonal level do not need to be consistent with each other. In this section I will, first, briefly introduce some useful aspects of predictive processing, then I will argue that self-deception is a case where ambiguity is preserved such that a range of hypotheses is present. The following section will be dedicated to describing these hypotheses and their relationships. For that I will employ the overtone metaphor.

Predictive processing is a subpersonal kind of hypothesis testing insofar as a causal hypothesis about the structure of the world is inferred or acted upon (Friston 2009). An agent possesses a *generative* model of the causal structure of its environment. It is generative in that it generates predictions of sensory input that is actually sampled. When sensory input deviates from predictions, there is a prediction error that leads to changes of predictions (or actions — to change the sensory input). The former process is called *perceptual inference*; the latter is called *active inference*. Predictions about the causes of sensory input can also be characterized as hypotheses, analogously to scientific hypotheses — both need to be tested (e.g. Gregory 1980, Seth 2015a).

Generative models are part of the computational level of description. They are to be distinguished from 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. 210). It is possible that those two are formally equivalent (Hobson et al. 2014), yet as such they belong to different levels of description. Nevertheless, if predictive coding can cast light onto consciousness (Hohwy 2015), then phenomenal models, which describe our experiences, and computational models, which describe brain processes, relate to each other in some way. It is not an aim of this paper to establish this relation (which would require much more empirical testing), but if predictive processing is to be a theory of cognition (Clark 2013), particularly higher-order cognition such as conscious thought, then, as a first step, a hypothesis about self-deception (a higher-order cognitive phenomenon) can be voiced that employs the tools of predictive processing. I will now turn my attention to the predictive processing explanation of binocular rivalry that shares at least two characteristics with self-deception: there is contradictory input (1) that leads models to alternate (2). The idea that self-deception involves alternating models (*alternation*, tool #1) will be extended by the suggestion that alternation is brought about by the optimal degree of instability (*instability*, tool #2), which then leads to the proposal that in cases of self-deception there is no sequential alternation, but rather a *parallel* exploration of hypotheses (*counterfactual processing*, tool #3).

Binocular rivalry is characterized as rivalry between competing, alternating perceptual models, only one of which is usually experienced at a time. The procedure by which binocular rivalry is brought about is such that one image is shown to one eye and a different image is shown to the second eye. This procedure can evoke a condition in which it appears to the agent that both objects occur at the same spatiotemporal location (Hohwy et al. 2008). The viewer experiences an alternation between models about the causes of visual input, for example, that there is a face or a house in the vicinity (Hohwy et al. 2008). This is the case 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, in this case, that one sees a house or a face (2).

The application of the binocular rivalry analogy to self-deception is 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 due to empirical or evolutionary priors in the case of binocular rivalry). In the case of binocular rivalry, different models explain different sensory prediction errors. In the case of self-de-

ception, it is argued that, at least sometimes, there is contradictory evidence (see previous section), analogously to the two different images presented to the eyes. Yet, and here the analogy to binocular rivalry breaks, self-deception extends over relatively long periods of time. To my best knowledge, study participants have been exposed to binocular rivalry only for short periods of time. For the description of longer-term contradiction, another predictive processing tool (#2) will be employed, namely the *optimal degree of instability*:

> 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 et al. 2012a, p. 3; my italics)

In other words, in order to not miss something important (not to become prematurely fixed on a certain hypothesis), uncertainty has to be preserved to a certain degree. The term 'uncertainty' can be used at both the personal (an agent being uncertain over something) and subpersonal (representations in the model having low degree of precision) levels. Further, regarding the subpersonal usage of this term, different kinds of representations can possess differing precision, i.e. be more or less uncertain. In the case of binocular rivalry, there is uncertainty because of contradictory (given higher-level assumptions) sensory prediction *errors* (bottom-up). Yet in the case of exploration (see the quotation above), there seems to be uncertainty of *predictions* (top-down). Simplistically, one might consider that the first is outer and the second is inner uncertainty. In the case of self-deception, outer uncertainty (ambiguity of the situation that enables its alternative interpretation) has been postulated (Sloman et al. 2010).[3] I will argue for the case of inner case of uncertainty. Robin Carhart-Harris et al. (Carhart-Harris et al. 2014) hold that wishful, imaginative and creative thinking is characterized by more entropy (more states are visited by the system) than rational or goal-directed thinking. The reasoning is that rational and/or goal-directed reasoning impose constraints on the system such that it is driven towards a more limited set of states than, for example, in the case of wishful inferences, which are biased by emotions. Though I agree that there is enhanced entropy in cases of self-deception, I think that constraints in the form of goal-representations and affective states constrain it so as to enable intricate patterns of oscillation with respect to different properties of self-deceptive representations.

Optimism can be seen as occupying one end of the *optimal degree of instability* scale by which self-deceivers may be characterized, because instead of alternating between different judgments on a certain topic, in the case of optimism the judgment is fixed — it is overly positive. Optimism (see Friston et al. 2013) has been explained in predictive processing terms 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). In other words, states of the world which are beneficial to the agent's

3 Steven Sloman et al. (Sloman et al. 2010) follow George Quattrone's and Amos Tversky's (Quattrone and Tversky 1984) pain endurance paradigm. In Quattrone and Tversky 1984, study participants endured cold water for longer when told that their ability to do so was indicative of their health. Similarly, Sloman et al. (Sloman et al. 2010), 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 which led to speeding up, but denied doing this deliberately. Philip Fernbach et al. (Fernbach et al. 2014), similarly, led participants to believe that the way in which they search for objects in a picture (detailed or holistic) is indicative of self-control. If told that detailed search is indicative of self-control, participants search longer for objects, but deny the effort of doing so.
These are, without a doubt, interesting results, but they demonstrate a fairly weak kind of self-deception, because it is devoid of such properties as first-person inconsistency, tension and need for justification. Weak kinds of self-deception can be explained in predictive processing terms in the same way as optimism (see Friston's idea referred to in the main text). First, there are two kinds of theories about how motivation influences cognition: *qualitative (e.g.,* Kunda 1990) and *quantitative (e.g.,* Ditto et al. 1998). According to the qualitative theory, bias influences the way information is processed. Mele's theory of self-deception employs this understanding of motivation. According to the quantitative theory, motivation changes only the time that we spend theorizing. The pain endurance paradigm is of this kind. Fernbach et al. (Fernbach et al. 2014) argue that "self-deception is enabled by people's tendency to adopt a mental representation of their own behavior that yields the most beneficial inference" (p. 6). The most parsimonious way of explaining participant behavior, though, is not by means of reasoning or conscious inference, but by the circular dependency between goal representations and perception (see main text). Goals representations are, on par with intentions and desires, one way of describing what motivates agents.

goal representations are assumed. Thus, in the case of optimism there is *high* precision (low uncertainty) of control states. Control states represent the goals of the system. Optimism occurs because there is a circular dependency between perception and goal representations (beliefs about future actions). Since control states influence the transitions between hidden states, and vice versa, in the case where goal representations are certain to be fulfilled, they influence perception.

In the case of alternation between competing hypotheses, there is more instability than in the case of optimism. A certain degree of instability may mean that instead of hypotheses being *disambiguated*, ambiguity is *preserved*, enabling the switching that is characteristic of at least some cases of self-deception. Instability, per se, says only that there will be changes in explored hypotheses, but in the case of binocular rivalry and self-deception there is also *constancy*: exploration does not cover too much ground but rather keeps returning to a select few models. Thus to explain self-deception, one needs to not only explain what leads to an enhanced degree of instability, but also why that instability is resolved in the same way over time. A change in estimates of uncertainty over time is *volatility* (Mathys et al. 2011). Volatility can be applied to different objects, including priors, policies, and models themselves. Further research needs to be done to establish how volatility can account for the constancy of model switches, but I think that the more complex the self-deception, the more kinds of volatility will be involved.

So far I have compared self-deception to binocular rivalry — where models also may alternate — and argued that a certain degree of instability may lead to the alternation observed in self-deception. What I want to argue for now is that, contrary to binocular rivalry where models alternate sequentially, in the case of self-deception two or more hypotheses are explored *concurrently*. In order show this, I first need to describe the role of counterfactuals in predictive processing. This role is fundamental in that action — accomplished by active inference — involves counterfactuals that encode "what we would infer about the world, if we sample it in a particular way" (Friston et al. 2012b, p. 2). Then, when sampling takes place, these counterfactuals are realized by active inference. There is a set of counterfactuals because the form of representation is probabilistic: several bets with different probabilities are encoded,[4] and replaced when another set of bets minimizes errors better.[5] A benefit of such probabilistic encoding is that perception and action are intertwined to a greater degree, enabling quicker action, for example, estimating the right trajectory of a flying ball and acting to catch it (Clark 2016), because multiple affordances are taken care of at the same time:

> It [pro-active readiness] must allow many possible responses to be simultaneously *partially prepared,* to degrees dependent upon the current balance of evidence — including estimations of our own uncertainty — as more and more information is acquired. (Clark 2016, p. 179; my italics)

Affordances (potential actions) have been argued to be computed during embodied decisions, for example, the affordance of reaching a berry or an apple (Cisek 2012, Cisek and Pastor-Bernier 2014). Recently, this claim has been extended to higher-order cognition: the same neural circuits and resources are used during a cognition process, as well as overt action (Pezzulo and Cisek 2016). Metzinger (Metzinger 2017) draws an analogy been drawn between affordances in the 'external' world (actually affordances in the internal, transparent world-model) and "affordances for cognitive agency", e.g. what to think or to calculate next: mind wandering creates a constant stream of the latter. Competing affordances are computed in parallel if both can still be realized. This is like fleeing from the predator along a road that forks at some future point; until the fork is reached, the two paths for fleeing can be

4 As Andy Clark (Clark 2016, p. 181) points out: "At every level, then, the underlying form of representation remains thoroughly probabilistic, encoding a series of deeply intertwined bets concerning what is 'out there' and (our current focus) how best to act".

5 For example, in case of speech processing it is argued that "we may rely upon stored knowledge to guide a set of guesses about the shape and content of the present sound stream: guesses that are constantly compared to the incoming signal, allowing residual errors to decide between competing guesses and (when necessary) reject one set of guesses and replace it with another" (Clark 2016, p. 194).

traversed together (Clark 2016).6 Parallel hypothesis exploration in self-deception is in this context an extreme case in which the parallel computation does not stop when the fork has been reached; options which, given the evidence, should have been abandoned, are instead kept available.

Interestingly, for action (executing movements) to take place instead of perception (changing the current hypothesis about the state of the world), precision expectations have to be down-played, and insofar as this occurs, they have been argued to be self-deceptive:

> In sum, action (under active inference) requires a kind of targeted dis-attention in which current sensory input is attenuated so as to allow predicted sensory (proprioceptive) states to entrain movement. At first sight, this is a rather baroque […] mechanism […] involving an implausible kind of self-deception. According to this story, it is only by downplaying genuine sensory information specifying how our bodily parts are *actually* currently arrayed in space that the brain can 'take seriously' the *predicted* proprioceptive information that determines movement, allowing those predictions to act […] directly as motor commands. (Clark 2016, p. 217)

This quotation is interesting for two reasons. First, it demonstrates the wide scope in which the term 'self-deception' is used. Second, it leads into a discussion of which phenomena are appropriately labelled 'self-deception' and which should be termed 'misrepresentation' in the predictive processing framework. I think that the above example is one of misrepresentation, not self-deception, because none of the behavioral or phenomenological characteristics of self-deception are present. Thus, the fact that the predictive processing framework is able to incorporate misrepresentations does not mean that each case of misrepresentation is a case of self-deception, or that if misrepresentations are produced in particular ways they are necessarily also kinds of self-deception. The fulfilment of behavioral and phenomenological characteristics is important to self-deception. This can be illustrated by two examples. First, Jean Daunizeau et al. (Daunizeau et al. 2010) argue that "categorization errors are optimal decisions if the risk of committing an error quickly is smaller than responding correctly after a longer period" (p. 6). In such cases of a speed-accuracy conflict, a speedy decision and not an accurate one, is optimal.[7] Second, Rosalyn Moran et al. (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 well the model predicts the data) is not the only quantity that is minimized during prediction error minimization, rather, accuracy minus complexity is minimized[8] (for an exact definition of accuracy and complexity see Friston 2010). Thus complexity is a penalty term. It is a penalty because a model is more useful if it is 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 instance perfectly. When it encounters a new piece of data and cannot predict it, the necessity of a perfect prediction would enforce a change to the model introducing new parameters, thereby making the model more complex. Thus, according to Moran et al. (Moran et al. 2014), Occam's razor, which is utilized here to avoid overfitting, leads to attenuated learning with age. This means that complexity reduction leads to the reduction of *short-term* Bayesian updating (according to which each time one encounters an instance the model cannot predict, one would be prone to change the model) and a shift to enhanced top-down processing (Moran et al. 2014, p. 6). The authors voice the optimistic conclusion that "as we age, we converge on an

6 Interestingly, competing affordances are resolved only when action is needed: "Instead, to minimize prediction error is to minimize failures to identify the affordances for action that the world presents. Here, a good strategy is to deliver (at every moment) a partial grip upon a number of competing affordances: an 'affordance competition' that is plausibly resolved only *as and when action requires*." (Clark 2016, p. 202; my italics)

7 Specifically, it is Bayes-optimal.

8 The reason for this is that the quantity that is to be minimized is actually free energy, which can be approximated by prediction error (Friston 2010).

accurate and parsimonious model of our particular world […] whose constancy we actively strive to maintain" (Moran et al. 2014, p. 1). Thus, the second example is one in which, given a complexity-accuracy conflict, the less complex model, not the more accurate one, wins. I doubt that these two ways *necessarily* lead to self-deception, nor are they the only ways in which self-deception in the predictive processing framework occurs. This is because it is not true that every misrepresentation is self-deceptive, or, for that matter, that it is every misrepresentation which has been acquired in a certain way. Rather, if some phenomenon fits the phenomenological and behavioral characteristics, then it is more or less self-deceptive, depending on how many characteristics it fulfils.

The main insight into self-deception thus far from the combination of the predictive processing approach and applying the tools of alternation (#1), optimal degree of instability (#2), and counterfactual processing (#3), is that there is no need for external evidence for the alternation of explanatory models, but that a high level of exploration can be kept in order to ensure that ambiguity prevails.[9] My main claim is that this high level of exploration, for the self-deceiver, is determined by the *optimal degree of instability*. This is congruent with psychological findings which show that self-deception supposedly correlates with openness to new experiences (Kurt and Paulhus 2008, p. 843; Paulhus and John 1998, p. 1030). To pin down the results so far, I have argued that in self-deception cases, the optimal degree of instability is such that there is more exploration than disambiguation. Self-deceivers, thus, seem to be open (on the personal level), but not open enough — they are not open to acknowledge their self-deception, and experience insight upon relinquishing their self-deception.

Not being open enough leads to a discussion of how intuitions contribute to self-deception, and how self-deception can be relinquished such that insight can kick in. In the previous section I hypothesized that self-deception and intuitions have a common phenomenological characteristic — a transparent signature of self-knowledge (Metzinger and Windt 2014). There is another feature that they might have in common, namely that one has to distract oneself in order to actually abandon both intuition and self-deception. Regarding intuitions, it has been argued 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). If in the case of self-deception there is not enough openness, then the same may be true regarding that phenomenon. Mere distraction is not enough, though. Changes in attention allocation have been argued to be a mechanism to *uphold* self-deception, not *relinquish* it (e.g., Noordhof 2009; for critique of the application of thought-suppression to self-deception see Lynch 2014). Thus, the effects of distraction can vary; an additional element is needed to make the acceptance of self-deceptive content possible.

My hypothesis is that this element is not of conceptual, but of interoceptive nature. Compelling arguments for this claim can be found when considering the following empirical evidence. *Anosognosia* (Turnbull et al. 2014) and *unrealistic optimism* (McKay et al. 2013) are (temporarily) attenuated by caloric vestibular stimulation (CVS). CVS consists in applying cold water to the ear canal of the patient, eliciting a nystagmus, i.e. rapid eye movements. In other words, for a short period of time after CVS, anosognosics acknowledge their paralysis, and optimistic people become less overly optimistic. Thus, something non-cognitive — the vestibular apparatus — influences higher-order cognition. Interestingly, CVS is also argued to "modulate the alternation rate in binocular rivalry" (Mast et al. 2014, p. 8). This, on the one hand, strengthens my analogy between binocular rivalry and self-deception. On the other hand, the fact that CVS influences binocular rivalry also makes the matter more complex: what is the underlying mechanism for such an inference? Oliver Turnbull et al. (Turnbull et al. 2014), on the premise that the vestibular sense influences both affective and spatial processing of information, view

9 My idea bears a certain resemblance to the idea that tension between goal representations can be creative and 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 and Over 2009, p. 328)

anosognosia as a "dynamic, emotional by-product of a cognitive deficit" (p. 24) in "veridical spatial cognition" (p. 21). In other words, they argue that anosognosics perceive the world (spatial cognition aspect) in an egocentric fashion, i.e. how they want the world to be (emotional by-product aspect). Given that both anosognosia and unrealistic optimism are reduced by CVS and the explanation of the former as an emotional by-product, this depiction might be applied to self-deception as well. The more cautious claim is, then, that emotions, by means of CVS as an intermediary, influence the process of self-deception. In this vein, Bigna Lenggenhager and Christophe Lopez (Lenggenhager and Lopez 2015) voice a hypothesis that vestibular stimulation might influence interoception (p. 14). On balance, the point, in predictive processing terms, is that perceptual, interoceptive and possibly other kinds of inferences, are connected (Clark 2016), and in the case of self-deception, this connection plays a more important role than has been acknowledged so far.

To sum up, while being self-deceived, there is exploration, which is bounded such that insight is precluded. In predictive processing, the sequences of beliefs about future actions are termed 'policy' and the value of a policy can be decomposed into *extrinsic* (utility) and *intrinsic* (exploration) reward (Friston et al. 2013). When the expected utility of a policy drops, exploration is engaged, such that alternatives are explored which might contradict the policy pursued before, allowing insight to occur. The central question now is why exactly such a policy is explored which leads to insight, given that previously, during the time of the self-deceptive episodue, uncertainty of the policy (which underlies exploration) was used to uphold the self-deceptive cycle (analogously to binocular rivalry), undermining insight. This is where motivation comes into the picture. Motivation determines when we start to self-deceive and when we relinquish self-deception. In the predictive processing framework, what motivation is and how it determines an agent's actions can be described in a very general manner. A standard view, because of its proximity to the folk-psychological level of description, is that motivation encompasses both goal representations and the affective consequences of our actions. A more general view would be one in terms of the reduction of free energy (of which prediction error is an approximation). Free energy is an upper bound on surprise regarding the sensory states of the agent. Minimization of surprise is crucial for survival of the agent (Friston 2010). There is no need for a value function in free energy minimization, as free energy itself is actually the only ultimate value that exists, and whatever other descriptions of purportedly valuable states one might give, if they do not minimize free energy, they would not actually be valuable. A cost function can be defined 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. For example, if money were the value in question, then for a beggar, winning a lottery would result in a huge change of value.[10] If value is substituted by free energy reduction, then one would get Mateus Joffily and Giorgio Coricelli's (Joffily and Coricelli 2013) explanation of emotional valence as the rate of change of free energy, such that, for example, an agent who could reduce free energy the most would be very happy. The consequence for self-deception of such a general account of value is that motivation should not be seen as an independent element that can be switched on and off. Rather, the sheer fact that agents reduce free energy, or pursue one policy and not another, is a demonstration of the ubiquity of motivation in the agent's cognition and action. In the following section, I will turn my attention to the set of hypotheses that are explored in self-deception.

4 The Overtone Metaphor

In this section, I will apply the overtone metaphor to self-deception. Recall that my own philosophical thesis is that during self-deception, uncertainty is preserved and several hypotheses are explored, but

10 The *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).

that they are at the same time bounded in order to preclude insight, and that affective consequences may play a role in overcoming the boundary. I will now describe how I think they relate to each other.

I borrowed the term 'overtone' from music theory where it denotes an additional frequency of a tone beyond the base frequency. Each tone that you hear, e.g. a musical instrument playing or a cup breaking, has an intensity, as well as one or several frequencies (the lowest being called fundamental). Those frequencies determine how the tone sounds (its timbre) and can be modelled as sine or cosine waves. Such wave functions (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 — i.e., how long one has to wait until the same value will be repeated — is the frequency). For example, a tone played on a violin will sound different to the same tone being sung. This is due to the difference in the expression of overtones. Imagine, for example, that you have computer software with which you can produce sound waves. Each wave will sound a certain way, and if you click to add more and more waves, then the sound produced will differ.

First and foremost, I want to mention that some authors in the self-deception literature (in virtue of the difficulty of ascribing a clear-cut belief to self-deceivers) have favored the view that self-deceptive attitudes oscillate in the degree of certainty (for a short summary see Pliushch and Metzinger 2015). Christoph Michel (Michel 2014) has even argued that the constancy of our beliefs is an accidental property thereof: at each moment in time attitudes are constructed from the evidence available at that time and, luckily, they often stay constant. Thus, what has been offered is the oscillation of one attitude over time.

What I want to propose instead is that, in analogy to physical action (see the previous section), there is a set of hypotheses that changes over time. A self-deceiver's optimal degree of instability (which determines how much exploration is pursued, in order to preclude overfitting a model) is heightened so that constant exploration (of a certain number of hypotheses) is pursued at the cost of disambiguation in favor of any particular hypothesis. These hypotheses are like overtones of the currently active self-deceptive hypothesis (the base frequency) so that what self-deceivers or observers perceive as one tone (self-deception) is actually a fusion of different frequencies. The hypotheses are those in which certain attributes — propositional content, transparency, and affective consequences — vary. It is a *fusion* insofar as the phenomenology is determined not only by the proposition that is currently verbally affirmed, but *by the entire set*. One of the reasons I favor the overtone metaphor is that overtones are different depending on the *kind* of instrument being played, hence the richness and shrillness of the sounds also changes from instrument to instrument (and from human voice to human voice). Analogously, self-deception is idiosyncratic such that whether and how it develops depends on several factors, such as the phenotype of the agent and his personality traits. Depending on the amount of tension one experiences during self-deception, it can be compared to a consonant (pleasant, absent tension) or a dissonant (unpleasant) sound.[11]

Let me consider two examples that concern sets of hypotheses in which the transparency attribute varies. First, consider the following statement: "I know that we broke up but I don't want to call him, because then it will really 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 believing a contradiction (knowing and not knowing the same proposition). Yet an agent might reflect on several models of reality that she is possessing at the same time — they are somehow connected with each other, overlaid upon each other so that in both cases only the agent remains constant, while thoughts and feelings change depending on the accepted reality model. In this example, I think that the base frequency is the knowledge about the break up, yet the other hypothesis — that it is not over — contributes to relieving the negative affective consequences of the first. The base frequency is transparent (or maybe also unconscious) to a certain degree.

11 I am grateful to Wanja Wiese for pointing this out to me.

Second, consider the following study. Laura Aymerich-Franch (Aymerich-Franch et al. 2014) tested the relationship between virtual self-similarity (the similarity between a person's virtual avatar and their real-world appearance) and social anxiety. Virtual self-similarity was varied in that participants were assigned avatars that were more or less similar to their appearance. They were then required to give a talk in front of a virtual audience, a task expected to produce social anxiety. The results were not statistically significant, but the trend indicated that embodying dissimilar avatars reduces public speaking-induced anxiety. One interpretation of the results is that embodying dissimilar avatars was associated with a weaker sense of presence. Thus, if the avatar is not experienced as being oneself, then embodying it would not lead to strong reactions. Alternatively, the results can be interpreted as showing that the world/self-model that one possesses affects one's cognitive and affective processes. Further, if one asked which self-model — one's own or that of the avatar — functioned as the subject's unit of identification during the public speaking task, my hypothetical answer would be: both, but each only to a certain degree.

These two examples are very simplistic. Both involve only two representations, which means there is only one overtone. Richer examples of self-deception with several overtones are also possible. Among other things, an agent's fears, hopes, and wishes, all of which are affective attitudes, can serve as overtones too.

My claim that overtones influence phenomenology rests on the hypothesis that counterfactuals possess such influence. In the previous section I used the claim that counterfactuals implement physical action to argue that several hypotheses might be explored at once. Notably, the role of counterfactuals in predictive processing is not restricted to that of implementing physical action (see also Metzinger 2017). They have also been hypothesized to lead to certain kinds of phenomenology. In this vein, the concept of 'counterfactual richness' has been introduced by Anil Seth (Seth 2014) as a range of counterfactual sensorimotor contingencies[12] and it is argued to influence objecthood and/or the phenomenology of how real[13] an object appears (Seth 2015b). The idea can be most intuitively described on the personal level as follows. We could use any given object in several ways. This means we can imagine how an object would change given our actions, which leads to a sense of objecthood. For example, if there is a potato in front of me, I see it as something I could peel, throw, or turn around its axis, instead of simply seeing a picture of a potato. What is invariant in all those counterfactual scenarios is the object — the potato. Counterfactual sensorimotor contingencies are the subpersonal equivalent of this idea. The implication of this idea is that insofar as our counterfactuals are not violated when we use objects, they appear real.

Generalizing this line of thinking, this means that counterfactuals that represent sensorimotor contingencies influence certain aspects of our phenomenology. In transferring this claim from perception to cognition, a problem arises: in cognition there are no sensorimotor contingencies. Luckily, *exteroceptive* (directed at the outer world) counterfactuals do not exhaust our possibilities, because given that there is autonomic control too, there may be interoceptive counterfactuals that also influence experience in certain ways. For example, while in allostatic control, proprioceptive and kinematic consequences of actions are represented and then fulfilled via active inference, in autonomic control, interoceptive counterfactuals may represent the affective consequences of actions. Interestingly, the idea that the affective dimension depends on the presence of counterfactuals of a certain kind does not depend on predictive processing. For example, Jérôme Dokic (Dokic 2012) hypothesizes that one

12 But note that Karl Friston (Friston 2014) argues that there are also "purely sensory expectations that do not inform future or counterfactual outcomes — and have no sensorimotor contingency" (p. 120).

13 Referring to the need to distinguish the personal and subpersonal use of counterfactuals, the determination of the degree of realness by counterfactuals leads to the possibly counter-intuitive conclusion that subpersonal counterfactual selection cannot determine personal counterfactual selection. Personal counterfactual selection is selection among the representations of possible *phenomenal* world- and self-models one could possess. If transparency correlates with counterfactual richness (Seth 2014; Metzinger 2014), then counterfactual richness would play a different role than the 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.

could describe the *degree* of a metacognitive feeling modally as depending on the number of possible worlds in which a certain mental action would be successful. Thus, for example, a strong feeling of knowing indicates that one's competence is robust and will not easily fail in nearby possible worlds. Drawing on these findings, I wish to extend the view found in the self-deception literature that inconsistent representations might produce tension (characterized by a feeling of uneasiness, see previous sections) and argue for a broader claim. Overtones enrich experience not only in degrees of transparency, but also through more nuanced affective consequences such as being glad that something is the case or anticipating that it is the case.

Thus far in this section, I have focused on and elaborated one possibility for the application of the overtone metaphor: that it may be applied to self-deceptive hypotheses which, if phenomenally available, are often characterized as beliefs that something is the case, such as the belief that this paper is insightful. There is an alternative way in which the metaphor might be applied, namely with respect to the different ways in which self-deceptive beliefs are *justified*. Justification of a beliefs requires an epistemic agent model and, so self-deception about justification is a special case. For example, consider a lazy person who is content not learning something new (such as the Italian language) in her spare time. Upon being questioned about why this is the case, several (possibly inconsistent) lines of argumentation may be generated, for example, too much to do at work, too expensive, or no conversation partners available to practice with. If one line of reasoning no longer fits, such as a pay rise making paying for lessons no longer too expensive, another justification is chosen. The actual underlying reason in this case, however, is just laziness, but this is a negative self-characteristic which not everyone is ready to accept.

In summary, I have shown that a predictive processing account of subpersonal hypothesis testing indicates that sets of hypotheses are tested in perception, cognition and action, but we also experience a more or less unified sequential phenomenology, even in cases of self-deception. To reconcile both these observations and offer a description of the self-deceivers' inconsistency, I propose applying the overtone metaphor to self-deception: in cases of self-deception, several hypotheses are explored in parallel such that there is a basic phenomenally available hypothesis and one or more overtones that to different degrees also are reflected in the phenomenal experience.

5 Conclusion

In this paper, I introduced the *overtone* model of self-deception. After introducing the topic, I argued in the second section that in classical philosophical discussion, "self-deception" is a cluster concept that is characterized by behavioral and phenomenological profiles. I focused on the behavioral characteristic of inconsistency. In the third section, I compared self-deception to binocular rivalry and came to the conclusion that self-deception is a kind of continued exploration in which disambiguation is precluded. My main claim in this paper was that several (subpersonal) hypotheses might be explored at once. In the final section, then, I applied the overtone metaphor to the hypotheses that are continually explored in self-deception. I argued that the phenomenology of the self-deceiver is determined by the fusion of the overtones that the self-deceiver possesses. What those overtones look like (or how they develop) depends on the idiosyncratic characteristics of the self-deceiver.[14]

14 I want to thank Thomas Metzinger, Wanja Wiese, Lisa Quadt and Paweł Gładziejewski for extremely helpful comments, as well as the Barbara-Wengeler foundation (BWS) for funding my PhD thesis, ideas from which can be found in this paper. I am also very grateful to Lucy Mayne, who has proofread the English language.

References

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. https://dx.doi.org/10.3389/fnhum.2014.00944.

Bagnoli, C. (2012). Self-deception and agential authority. A constitutivist account. *Humana.Mente Journal of Philosophical Studies, 20*, 99–116.

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

Bermúdez, J. L. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis, 60* (4), 309–319.

Borge, S. (2003). The myth of self-deception. *The Southern Journal of Philosophy, 41* (1), 1–28. https://dx.doi.org/10.1111/j.2041-6962.2003.tb00939.x.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R. & Nutt, D. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience, 8*, 20. https://dx.doi.org/10.3389/fnhum.2014.00020.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology, 22* (6), 927–936. https://dx.doi.org/10.1016/j.conb.2012.05.007.

Cisek, P. & Pastor-Bernier, A. (2014). On the challenges and mechanisms of embodied decisions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369* (1655). https://dx.doi.org/10.1098/rstb.2013.0479.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (3), 181–204. https://dx.doi.org/10.1017/S0140525X12000477.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Damm, L. (2011). Self-deception about emotion. *The Southern Journal of Philosophy, 49* (3), 254–270.

Daunizeau, J., den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Friston, K. J., Stephan, K. E. & Sporns, O. (2010). Observing the observer (ii): Deciding when to decide. *PLoS ONE, 5* (12), e15555. https://dx.doi.org/10.1371/journal.pone.0015555.

Davidson, D. (1986). Deception and division. In J. Elster (Ed.) *The multiple self* (pp. 79–92). Cambridge: Cambridge University 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. https://dx.doi.org/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.

Fernbach, P. M., Hagmayer, Y. & Sloman, S. A. (2014). Effort denial in self-deception. *Organizational Behavior and Human Decision Processes, 123* (1), 1–8. https://dx.doi.org/10.1016/j.obhdp.2013.10.013.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13* (7), 293–301. https://dx.doi.org/10.1016/j.tics.2009.04.005.

——— (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11* (2), 127–138. https://dx.doi.org/10.1038/nrn2787.

———(2014). Active inference and agency. *Cognitive Neuroscience, 5* (2), 119–121. https://dx.doi.org/10.1080/17588928.2014.905517.

Friston, K. J., Breakspear, M. & Deco, G. (2012a). Perception and self-organized instability. *Frontiers in Computational Neuroscience, 6*. https://dx.doi.org/10.3389/fncom.2012.00044.

Friston, K. J., Adams, R. A., Perrinet, L. & Breakspear, M. (2012b). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151. https://dx.doi.org/10.3389/fpsyg.2012.00151.

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. https://dx.doi.org/10.3389/fnhum.2013.00598.

Funkhouser, E. (2005). Do the self-deceived get whay they want? *Pacific Philosophical Quarterly, 86* (3), 295–312. https://dx.doi.org/10.1111/j.1468-0114.2005.00228.x.

———(2009). Self-deception and limits of folk psychology. *Social Theory and Praxis, 35* (1), 1–16.

Gendler, T. S. (2007). Self-deception as pretense. *Philosophical Perspectives, 21* (1), 231–258. https://dx.doi.org/10.1111/j.1520-8583.2007.00127.x.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 290* (1038), 181–197.

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. https://dx.doi.org/10.1037/0022-3514.37.2.147.

Helzer, E. & Dunning, D. (2012). On motivated reasoning and self-belief. In S. Vazire & T. D. Wilson (Eds.) *Handbook of self-knowledge* (pp. 379–396). New York: Guilford Publications.

Hobson, J. A., Hong, C. C.-H. & Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology, 5*. https://dx.doi.org/10.3389/fpsyg.2014.01133.

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. https://dx.doi.org/10.1016/j.cognition.2008.05.010.

Jackendoff, R. (2012). *A user's guide to thought and meaning.* New York: Oxford University Press.

Joffily, M. & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology, 9* (6), e1003094. https://dx.doi.org/10.1371/journal.pcbi.1003094.

Johnston, M. (1988). Self-deception and the nature of the mind. In B. P. McLaughlin & A. O. Rorty (Eds.) *Perspectives on self-deception* (pp. 63–91). Berkeley C.: University of California Press.

Khemlani, S. S. & Johnson-Laird, P. N. (2012). Hidden conflict: Explanations make inconsistencies harder to detect. *Acta Psychologica, 139*, 486–491.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108* (3), 480–498. https://dx.doi.org/10.1037/0033-2909.108.3.480.

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. https://dx.doi.org/10.1016/j.jrp.2007.11.005.

Lenggenhager, B. & Lopez, C. (2015). Vestibular contributions to the sense of body, self, and others. In T. Metzinger & J. M. Windt (Eds.) *Open MIND:* 23(T). Frankfurt am Main: MIND Group.

Levy, N. (2009). Self-deception without thought experiments. In T. Bayne & J. Fernández (Eds.) *Delusion and self-deception* (pp. 227–242). New York: Psychology Press.

Lynch, K. (2012). On the ‚tension" inherent in self-deception. *Philosophical Psychology, 25* (3), 433–450. https://dx.doi.org/10.1080/09515089.2011.622364.

——— (2014). Self-deception and shifts of attention. *Philosophical Explorations, 17* (1), 63–75. https://dx.doi.org/10.1080/13869795.2013.824109.

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. https://dx.doi.org/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). https://dx.doi.org/10.3389/fnhum.2011.00039.

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. https://dx.doi.org/10.1016/j.cortex.2013.04.005.

Mele, A.R. (2000). Self-deception and emotion. *Consciousness & Emotion, 1* (1), 115–137. https://dx.doi.org/10.1075/ce.1.1.07mel.

Mele, A. R. (2001). *Self-deception unmasked.* Princeton, NJ: Princeton University Press.

——— (2012). When are we self-deceived? *Humana.Mente Journal of Philosophical Studies, 20*, 1–15.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

——— (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology, 4*, 931. https://dx.doi.org/10.3389/fpsyg.2013.00931.

———(2014). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience, 5* (2), 122–124. https://dx.doi.org/10.1080/17588928.2014.905519.

———(2015). M-Autonomy. *Journal of Consciousness Studies, 22* (11-12).

———(2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.* Frankfurt am Main: MIND Group.

Metzinger, T. & Windt, J. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Hor-

vath & J. Kipper (Eds.) *Die experimentelle Philosophie in der Diskussion* (pp. 279–321). Berlin: Suhrkamp.

Michel, C. (2014). *Self-knowledge and self-deception: The role of transparency in first personal knowledge.* Münster: mentis.

Michel, C. & Newen, A. (2010). Self-deception as pseudo-rational regulation of belief. *Consciousness and Cognition, 19* (3), 731–744. https://dx.doi.org/10.1016/j.concog.2010.06.019.

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. https://dx.doi.org/10.1371/journal.pcbi.1003422.

Noordhof, P. (2003). Self-deception, interpretation and consciousness. *Philosophy and Phenomenological Research, 67* (1), 75–100. http://www.jstor.org/stable/20140582.

———(2009). The essential instability of self-deception. *Social Theory and Practice, 35* (1), 45–71.

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).

Pezzulo, G. & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences, 20* (6), 414–424. https://dx.doi.org/10.1016/j.tics.2016.03.013.

Picard, F. (2013). State of belief, subjective certainty and bliss as a product of cortical dysfunction. *Cortex, 49* (9), 2494–2500. https://dx.doi.org/10.1016/j.cortex.2013.01.006.

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.

Porcher, J. E. (2012). Against the deflationary account of self-deception. *Humana.Mente Journal of Philosophical Studies, 20*, 67–84.

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. https://dx.doi.org/10.1037/0022-3514.46.2.237.

Rorty, A. O. (1988). The deceptive self: Liars, layers, and lairs. In B. P. McLaughlin & A. O. Rorty (Eds.) *Perspectives on self-deception* (pp. 11–28). Berkeley: University of California Press.

Sahdra, B. & Thagard, P. (2003). Self-deception and emotional coherence. *Minds and Machines, 13* (2), 213–231.

Sandoz, P. (2011). Reactive-homeostasis as a cybernetic model of the silhouette effect of denial of pregnancy. *Medical Hypotheses, 77* (5), 782–785. https://dx.doi.org/10.1016/j.mehy.2011.07.036.

Saunders, C. & Over, D. E. (2009). In two minds about rationality? In J. Evans & K. Frankish (Eds.) *In two minds* (pp. 317–334). Oxford: Oxford University Press.

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. https://dx.doi.org/10.1080/17588928.2013.877880.

——— (2015a). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.) *Open MIND:* 35(T). Frankfurt am Main: MIND Group.

———(2015b). Presence, objecthood, and the phenomenology of predictive perception. *Cognitive Neuroscience, 6* (2-3), 111–117. https://dx.doi.org/10.1080/17588928.2015.1026888.

Sloman, S. A., Fernbach, P. M. & Hagmayer, Y. (2010). Self-deception requires vagueness. *Cognition, 115*, 268–281. https://dx.doi.org/10.1016/j.cognition.2009.12.017.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind.* New York: Basic Books.

Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to fool others.* London: Allen Lane.

Turnbull, O. H., Fotopoulou, A. & Solms, M. (2014). Anosognosia as motivated unawareness: The ‚defence' hypothesis revisited. *Cortex, 61*, 18–29. https://dx.doi.org/10.1016/j.cortex.2014.10.008.

Van Leeuwen, N. D. (2007). The product of self-deception. *Erkenntnis, 67* (3), 419–437.

——— (2013). The folly of fools: The logic of deceit and self-deception in human life. *Cognitive Neuropsychiatry, 18* (1-2), 146–151. https://dx.doi.org/10.1080/13546805.2012.753201.

Von Hippel, W. & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34* (01), 1–16. https://dx.doi.org/10.1017/S0140525X10001354.

Weatherson, B. (2014). Centrality and marginalisation. *Philosophical Studies, 171* (3), 517–533. https://dx.doi.org/10.1007/s11098-014-0289-9.