Consciousness and Cognition 41 (2016) 119–134

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Review article

Self-deception as affective coping. An empirical perspective on philosophical issues

Federico Lauria [a,b,c,*], Delphine Preissmann [c,d,e], Fabrice Clément [a,c]

[a] Swiss Centre for Affective Sciences and Philosophy Department, University of Geneva, Campus Biotech, Chemin des Mines 9, 1202 Geneva, Switzerland
[b] Italian Academy for Advanced Studies in America, Columbia University in the City of New York, 1161 Amsterdam Avenue, New York, NY 10027, United States
[c] Cognitive Science Centre, University of Neuchâtel, Espace Louis-Agassiz 1, 2000 Neuchâtel, Switzerland
[d] Department of Psychiatry, Faculty of Medicine, University of Geneva, Rue Gabrielle-Perret-Gentil 4, 1211 Geneva 14, Switzerland
[e] Centre for Psychiatric Neurosciences, Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 13 July 2015
Revised 22 December 2015
Accepted 6 February 2016
Available online 23 February 2016

Keywords:
Self-deception
Affect
Appraisal
Somatic marker
Dopamine
Coping mechanism
Bias
Happiness
Cognitive science

ABSTRACT

In the philosophical literature, self-deception is mainly approached through the analysis of paradoxes. Yet, it is agreed that self-deception is motivated by protection from distress. In this paper, we argue, with the help of findings from cognitive neuroscience and psychology, that self-deception is a type of affective coping.

First, we criticize the main solutions to the paradoxes of self-deception. We then present a new approach to self-deception. Self-deception, we argue, involves three appraisals of the distressing evidence: (a) appraisal of the strength of evidence as uncertain, (b) low coping potential and (c) negative anticipation along the lines of Damasio's somatic marker hypothesis. At the same time, desire impacts the treatment of flattering evidence via dopamine. Our main proposal is that self-deception involves emotional mechanisms provoking a preference for immediate reward despite possible long-term negative repercussions. In the last part, we use this emotional model to revisit the philosophical paradoxes.

Published by Elsevier Inc.

Contents

* Corresponding author at: Swiss Centre for Affective Sciences and Philosophy Department, University of Geneva, Campus Biotech, Chemin des Mines 9, 1202 Geneva, Switzerland.

E-mail addresses: federico.lauria@unige.ch (F. Lauria), delphine.preissmann@unine.ch (D. Preissmann), fabrice.clement@unine.ch (F. Clément).

http://dx.doi.org/10.1016/j.concog.2016.02.001
1053-8100/Published by Elsevier Inc.

That evening, Sam was late again. He entered the room, kissed Mary, and apologized for the delay. It was like any other day, except for one thing. Unbeknownst to Sam, there was a red lipstick stain on the collar of his bright white shirt, a very salient mark. The truth is that he had just been kissing Sally, his secret lover of several months. It was almost impossible for Mary not to have seen the stain, this strong evidence of Sam's infidelity, but she didn't draw the obvious conclusion that her husband had been unfaithful. Rather, she surprisingly still believed that Sam was faithful. She trusted Sam's explanation that his being late was due to the usual unreliableness of the Brooklyn line, which was of course a lie. When asked by her friends the next day how things were going with Sam, she replied that everything was fine and that she loved him more and more every day. And the striking fact is that she was sincere in saying so and was not lying – except to herself. This example has been used as one of the self-deception paradigms in philosophy since it was first employed by Meiland (Meiland, 1980).

Experiments have revealed that self-deception is widespread (Kunda, 1990). People usually believe that they are good drivers (Lajunen, Hakkarainen, & Summala, 1996), professors typically believe that they are well above average (Mele, 2001), and seriously ill patients often believe that they will recover (Goldbeck, 1997), among other examples. As reality is less flattering, it appears that we deceive ourselves and that our desires significantly bias our cognition. How do we manage to avoid facing the facts when evidence speaks for itself? What is self-deception?

In the philosophical literature, it is agreed that self-deception involves at least three elements: (i) a cognitive state (ii) formed despite sufficient evidence to the contrary (iii) because of a subject's desire. Self-deception is thus a type of motivated cognition. Nonetheless, there are debates about self-deception: one is about the very process of deceiving oneself and the other is about the result of this process. Both controversies emerge from the traditional picture of self-deception, which conceives of self-deception as structurally analogous to interpersonal deception (Davidson, 1985). People deceive other persons by intentionally leading them to believe what they do not believe themselves. In deceiving Mary, Sam intends to induce in Mary the belief that the train is the cause of his lateness, while not believing this himself. Deception is thus an intentional process involving two simultaneous beliefs with contradictory content. If self-deception is deceiving oneself, it should involve an intention and two simultaneous conflicting beliefs (Davidson, 1985). As people deceive others by *intentionally* inducing in others a belief that they themselves do not hold, people deceive themselves by *intentionally* forming a belief that conflicts with another of theirs. To return to our example, Mary intends to believe that Sam is faithful and, as a result, believes so, while also believing that he is not faithful to her. However, as intuitive as it seems, this picture gives rise to two paradoxes (see Mele, 2001).

The *dynamic* paradox arises as soon as it is observed that deceiving someone implies that the deceiver intentionally hides the truth from the deceived person and knows that she is doing so. Unlike the case of interpersonal deception, it is notably difficult to understand how one can both know (or believe) the truth and intentionally hide it *from oneself* so as to "lie" to oneself. Indeed, knowing that one is hiding the truth from oneself makes this endeavor self-defeating. As a consequence, self-deception seems impossible, which collides with the fact that it is widespread. This puzzle is called the 'dynamic paradox' because it concerns the process of deceiving oneself. Several solutions have been proposed. To mention just two, the traditional picture appeals to an unconscious intention in understanding self-deception (Bermudez, 2000; Davidson, 1985), while aficionados of alternative views have renounced the appeal to intention altogether (Bach, 1981; Mele, 1997).

The traditional picture also gives rise to the *static* paradox, which concerns the state one is in when being self-deceived (the so-called "product" of self-deception). If self-deception is structurally analogous to interpersonal deception, self-deceived subjects should be in a contradictory state of mind. But it seems impossible to hold two simultaneous beliefs with contradictory content in a conscious manner. How can I hold the beliefs that it is raining in Central Park and that it is not raining in Central Park, at the same time and in a conscious way? This has motivated some authors to think of self-deception as involving an *unconscious* belief (Davidson, 1982; Davidson, 1985; Pears, 1984). Alternatively, some have argued that self-deception consists in avoiding thinking about certain distressing things (Bach, 1981). Others understand the product to be a mental state other than believing, such as pretense (Gendler, 2007). Finally, some have defended that deceiving oneself results in the deluded belief only: Mary merely believes that Sam is faithful *period* (Bermudez, 2000; Mele, 1997).

The philosophical controversy mainly revolves around the solutions to these paradoxes. As each solution comes with a price, the state of the philosophical debate seems to have reached an impasse, at least in the absence of further empirical evidence (more on this in part 1). Yet, despite these controversies, there is an important feature of self-deception that is accepted by all sides of the debate without controversy, and that immediately comes to mind when one thinks about self-deception: suffering and, more generally, affect. Indeed, it is plausible to think that people deceive themselves in order to avoid suffering from distressing truths. Some truths are difficult to live with and significantly impair one's well-being. As Price (1954) put it, there are beliefs we *cannot afford*: The fact that her husband is having an affair is too distressing for Mary to face, as it frustrates one of her strong desires. This natural explanation of why we deceive ourselves is taken for granted by the main models of self-deception (see part 1) and has been empirically tested (Erez, Johnson, & Judge, 1995). Moreover, the literature on the biological adaptiveness of self-deception (Hippel & Trivers, 2011; Van Leeuwen, 2007), its cognitive function (Johnston, 1988; Barnes, 2007) and ethical puzzles, such as whether self-deception makes us happy (Van Leeuwen, 2009), mainly makes sense on the assumption that the conflict between truth and happiness lies at the heart of self-deception. This being said, it is striking that, to our knowledge, very few accounts revolve around this idea and investigate its affective

dimension in detail. Among the very few affective models that have been proposed, none of them recruits affect to approach the paradoxes of self-deception with minutiae. The aim of this paper is to redress this imbalance by providing a picture of self-deception centered on the conflict between truth and happiness and putting emphasis on the importance of affect and emotions in biasing the treatment of information in light of one's desire.

Since accounts of self-deception describe the mechanisms by which we deceive ourselves, we shall leave the armchair and appeal to findings in cognitive neuroscience and psychology (see for instance Mele, 2001; Bayne & Fernández, 2009 for a similar approach), using this experimental evidence to revisit the philosophical paradoxes. In a nutshell, we shall argue that the key to understanding self-deception lies in cognitive and affective filters that account for how desire influences attention and biases cognition. At the psychological level, self-deception involves three attentional filters, or appraisals, of the evidence that one's desire is frustrated – the "distressing" evidence: (a) appraisal of the evidence as being *ambiguous*, (b) *negative* evaluation of the evidence in the light of one's well-being, (c) *low* coping potential. In addition, it involves a *positive* evaluation of the evidence speaking in favor of the satisfaction of desire – the "flattering" evidence. We propose in turn two neurobiological mechanisms grounding these appraisals and explaining how they can result in the subject's favoring the flattering evidence over the distressing evidence. The first is Damasio's somatic marker hypothesis about decision-making. The second is the regulation of dopamine, the neurotransmitter of desire, which has been expansively studied in the neuroscience of motivation and self-control. These appraisals and mechanisms involve an affective dimension: They constitute affective filters on our thoughts that, we think, lie at the heart of self-deception.

In this sense, our picture echoes the affective revolution of decision-making. Our main proposal is that self-deception involves emotional mechanisms leading to a preference for immediate reward similar to the ones involved in decision-making. It thus constitutes a kind of affective coping with reality: In deceiving oneself, one carefully avoids short-term despair. This project thus aims to offer a new piece to the affective revolution of the mind by describing how affect is an integral part of self-deception.[1] Similar approaches have been proposed. Sahdra and Thagard (2003) appeal to hot cognition and emotional coherence (the computational model HOTCO) to approach self-deception. As far as the neurobiological mechanisms are concerned, Thagard (2007) alludes to somatic markers, and Balcetis (2008) recruits specialized neural pathways that are inherently affective, in particular the amygdala and basal ganglia, to illuminate motivated cognition.[2] This project shares the same spirit. Yet, our picture partly relies on other empirical tools and aims to offer a more systematic account of self-deception by specifying several appraisals and by proposing which neurobiological mechanisms ground them. It therefore aims to both develop the existing affective approaches more fully and use affect to disentangle the paradoxes with more detail. We shall indeed argue that this picture invites us to dissolve the static paradox and to adopt a non-intentionalist account of self-deception.

We shall first motivate our view by examining the main solutions to the paradoxes of self-deception. This examination will call for a new approach to self-deception, which we propose in the second section: the affective filter view. In the third part, we use the view to revisit the paradoxes. We conclude by showing how the affective filter view can contribute to a better understanding of motivated biases.

1. Motivating an affective approach: the impasse of the solutions to the paradoxes of self-deception

In this section, we present the challenges one is faced with as far as the present state of the literature on the philosophical paradoxes is concerned. This will set the agenda for our investigation and motivate our affective approach to self-deception. If we are right, the literature has reached an impasse that can be overcome by adopting an affective approach to self-deception.

1.1. The static paradox: the action and anxiety challenge

According to the standard picture, self-deception involves two beliefs with contradictory content, the belief about the distressing fact being unconscious (Davidson, 1982, 1985; Quattrone & Tversky, 1984; Sackeim & Gur, 1978; Sackeim & Gur, 1979; Sackeim & Gur, 1985).[3] Coming back to our example, Mary believes that Sam is not having an affair, while *unconsciously* believing that he has a secret lover. The unconsciousness of this latter belief is understood in terms of the subject's awareness or attention: The subject does not attend to the content of her belief, i.e. the subject's belief is not activated. This has led Davidson to think of self-deception as involving a "division of the self."

One might be skeptical about this appeal to the unconscious (see Mele, 2001). In the absence of further motivation, the appeal to unconscious belief appears *ad hoc* and mysterious: It seems to be motivated only by the need to solve the paradox, and one should explain how subjects form unconscious beliefs when deceiving themselves.

These suspicions have motivated several authors to propose alternative descriptions of the product of self-deception. For some, people deceive themselves when they believe some distressing fact and *pretend* that the flattering fact is true (Gendler,

[1] Some have approached self-deception with the help of emotions by discussing how emotions – rather than desire – can elicit self-deception (Mele, 2001) or how emotions themselves can be self-deceptive (e.g. de Sousa, 1978). Our approach differs in introducing emotions at the very heart of self-deception, rather than as a possible cause or as a possible candidate for self-deception. The volume edited by Bayne and Fernández (2009) is dedicated to affective and motivational influences on belief formation. However, no account in the volume approaches self-deception in detail by means of affect.

[2] Balcetis (2008) provides an excellent synthesis of empirical studies on motivated cognition.

[3] We focus here on the standard picture in which the subject simultaneously holds the two beliefs and we ignore accounts that involve two beliefs at different times (for instance, Pears (1984)). Our arguments apply to these accounts as well.

2007). On this description, Mary believes that Sam is unfaithful, yet pretends that he is not having an affair. Unlike belief, pretense does not constitute a commitment to the truth: I can pretend to be a tiger without believing it. Hence, in this respect, the subject's mental states involved in self-deception are not inconsistent. This option is one way of avoiding the costs of the traditional view, since it does not recruit the unconscious. For others, self-deception need not involve unconscious belief either: The trick is done by the subject's *not thinking* about the distressing thought, in the presence or absence of the distressing belief (Bach, 1981). Mary may believe that Sam is unfaithful, or she may not form this belief; what is crucial is that she avoids *thinking* this thought. The monitoring of attention can capture the absence of awareness that is at the heart of the appeal to unconscious belief. Since we do not think about the content of each of our beliefs, this picture is not *ad hoc*. In addition, Bach proposes three mechanisms by which thought avoidance takes place (Bach, 1981: 357–362): *rationalization* (subjective assessment of the distressing evidence as being inconclusive), *evasion* (turning one's attention away from the distressing evidence) and *jamming* (for instance, by vividly imagining the flattering fact anytime the thought of the distressing fact occurs). The charge of appealing to mysterious mechanisms would thus be misplaced. A final influential way to solve the static paradox is to restrict the product of self-deception to the belief about the flattering fact only (the deluded belief; Mele, 1997, 2001). Empirical evidence, Mele argues, is compatible with this deflationary picture. If self-deception involves only the deluded belief, it is neither paradoxical nor impossible. The widespread and well-known mechanism of motivated bias accounts for the absence of the distressing belief, which avoids the charge of being *ad hoc*, unlike the traditional view. It appears that these alternatives to the appeal to unconscious belief do not suffer from the charge of being *ad hoc* or of appealing to mysterious mechanisms.

There is a problem for both the alternative and the traditional views, however. Imagine that Sam deceives himself into believing that his son is good at mathematics, yet hires a mathematics teacher for his son (Mele, 1997). This action suggests that, at some point, the thought that his son is not faring well in mathematics has crossed Sam's mind. How can he act in such a way if his belief is unconscious (*pace* Davidson), if he does not think about its content (*pace* Bach) or if he does not have it altogether (*pace* Mele)? Although some actions, in particular automatic ones, seem to involve unconscious beliefs, taking the steps toward his son's mathematical "virtuosity" does not seem to be of this kind. It seems that beliefs can motivate actions of this type only if they are conscious or activated. If so, most accounts of self-deception cannot accommodate the behavior that may come with self-deception, in particular some actions. Similar worries arise about anxiety: Although Sam believes that his son is faring well in mathematics, he might still be anxious about his son's mathematical abilities. Does not this provide evidence that he suspects that things are not going that well and that he is aware of the truth? *Prima facie*, the answer to this question seems affirmative, in which case the main accounts cannot be all there is to self-deception. Likewise, it is not straightforward that pretending that his son is good at mathematics suffices to capture Sam's anxiety, because pretense and belief are not in tension, as just outlined. More can be said (see part 3), but this observation suffices to show the potential costs of the main proposed solutions to the static paradox.

One way to accommodate the cases of self-deception accompanied with the actions and anxiety mentioned suggests itself. Holding two conscious beliefs with contradictory content at the same time is impossible only for *full* beliefs, i.e. one cannot both be certain that p and certain that not p. However, lots of our beliefs come with uncertainty. It might thus be that some self-deceived subjects believe the flattering fact at some degree of credence (say 0.7) and also believe the distressing fact at another degree of credence (say 0.3). This does not involve any contradiction, despite the beliefs being in conflict, and both beliefs can be conscious. The distressing belief might then motivate actions like the one described, and this conflict between the beliefs might account for the anxiety of self-deceived subjects. Going back to our example, Sam might believe that his son is faring well in mathematics, yet still be uncertain about this and suspect that he has difficulties. This is why he hires a mathematics teacher and feels anxious. This proposal can correspond to the cases of self-deception *that come with the actions and anxiety described,* but not all instances of self-deception involve such behavior. As Mele (2001) put it, one can really succeed in deceiving oneself and have no beliefs about the distressing fact, in which case no action or anxiety of the type described would ensue. For these cases, the main accounts discussed remain a live option.

What are we to conclude from this exploration of the main answers to the static paradox? Our discussion was not meant to settle the debate, but to point to the challenges each option has to face, at least at first glance, in order to offer a model of self-deception that will have the resources to meet them. Given what we have said, it appears that aficionados of the traditional view need to provide a convincing answer to the suspicions surrounding the unconscious. More generally, proponents of any main account should rebut the action and anxiety challenges. In the absence of convincing answers and of further empirical evidence, it is tempting to adopt a picture of the product of self-deception that integrates the various proposals. What if self-deception comes with a variety of cognitive products? It might involve an unconscious belief in some instances, not thinking about some things in others, absence of belief for some cases and the presence of two conscious beliefs with distinct degrees of credence when self-deception is accompanied by the action and anxiety described. A picture of self-deception that admits to such variety in its product might capture the grain of truth in the accounts proposed without suffering from their flaws. From an empirical perspective, this would not be surprising given the complexity of cognitive processes. This integrative picture does not commit one to denying that self-deception is a unitary phenomenon: All cases of self-deception involve cognitive states that are biased by one's desire in the presence of sufficient evidence to form another cognitive state. If this is correct, the impasse of the literature might be due to unfortunate generalizations on some cases of self-deception, and the static paradox should rather be dissolved. The task would thus be to motivate such a variety. In the third part, we will explain how the affective filter view can help to rebut each challenge presented by means of affective processing, although the actual state of empirical research favors the integrative picture sketched.

1.2. The dynamic paradox: the challenge of cognitive mechanisms

According to the traditional picture, when we deceive ourselves, we *intentionally* bring about some false belief within us (Bermudez, 2000; Davidson, 1985). For this project to go through, one should be aware of what one is doing. The problem is that this awareness precludes one from succeeding. Self-deception thus appears impossible, hence the paradox.

The standard solution to this puzzle resorts to an *unconscious* intention. In being hidden, unconscious intentions do not defeat the endeavor of deceiving oneself.[4] It should be clear to the reader that this picture will inherit the same flaws as the appeal to unconscious belief discussed in the previous section: As it stands, it seems to be *ad hoc* and explains the obscure by the more obscure. Only by explaining how unconscious intentions operate can this account illuminate self-deception (Mele, 1997).

The main alternative solution to the paradox is to drop intentions (Bach, 1981; Mele, 1997). For self-deception is a type of bias: a bias due to a desire, i.e. a motivated bias. Now, biases do not need to be intentional to be successful. So why should self-deception be intentional? Since the empirical evidence is compatible with this deflationary picture, a non-intentionalist account of self-deception seems promising.

This view is by definition theoretically economical, as it does not commit one to the presence of intentions and *a fortiori* unconscious ones. However, it is not without problems. To mention two, it has been argued that it is prone to the *selectivity problem*: It cannot explain why only some desires lead to self-deceptive beliefs while others don't (Bermudez, 1997; Bermudez, 2000). After all, nothing in the picture prevents any desire from exerting such a distorting influence on one's beliefs. Why wouldn't any desire lead to self-deception? The intentionalist has a ready answer to this question: Some desires come with the intention to deceive oneself and hence lead to self-deception; other desires are not accompanied with such an intention and hence do not result in self-deception.[5] Second, in dropping intentions, the non-intentionalist account cannot secure a strict analogy between interpersonal and intrapersonal deception. Given the force of the intuition that self-deception is a type of deception, this seems problematic. Consequently, the defender of non-intentionalism should rebut the *selectivity problem* and justify the partial dis-analogy between interpersonal deception and self-deception.[6] Ideally, this should provide us with a mechanism involved in self-deception, which will enlighten us on both issues. In the last part of this paper, we shall argue that the affective filter view can provide us with such a mechanism and thus favors a non-intentionalist account.

Despite these static and dynamic controversies, it is worth observing that all accounts of self-deception implicitly rely on the conflict between truth and happiness. Why would one *intend* to deceive oneself in the absence of an awareness of a threat? Why would one *avoid* thinking about some things or be unaware of them without having sensed their devastating effect on one's well-being? How could desires bias cognition when evidence speaks for itself in the absence of signals indicating potential suffering? The conflict between truth and happiness is thus implicitly present in all accounts of self-deception, despite important differences between them. As Sackeim & Gur put it, in order for your attention to be diverted from an unwelcome piece of information, the latter should be somehow recognized as such (Sackeim & Gur, 1997). This motivates starting our inquiry there, by spelling out the mechanisms that constitute such a conflict. It is time now to present our model, before turning with a new eye to the paradoxes presented.

2. The affective filter view: a new approach to self-deception

Any view of self-deception should account for how subjects respond to some evidence at the expense of others. Now, not many systematic accounts of the cognitive mechanisms involved in self-deception have been proposed. The empirical studies generally focus on just one aspect of self-deception: vagueness of the evidence (e.g., Sloman, Fernbach, & Hagmayer, 2010), selective attention and impaired categorization of anomaly (Peterson, Driver-Linn, & Deyoung, 2002; Peterson et al., 2003), optimism bias and self-assessment (Baxter & Norman, 2011; Lopez & Fuxjager, 2012; Chance, Norton, Gino, & Ariely, 2011) or social desirability tested by Impression Management or Self-Deceptive Enhancement scores in personality traits questionnaires (Ashley & Holtgraves, 2003; Norem, 2002; Uziel, 2013). Yet, few systematic pictures of self-deception integrating the existing findings have been proposed in the empirical literature. Likewise, the philosophical debate is dominated by the attempts to solve the paradoxes presented, at the expense of detailed descriptions of the mechanisms involved in self-deception. In this section, we aim to fill this gap: We propose a systematic picture of self-deception involving distinct cognitive and neurobiological mechanisms that illuminate in more detail how desires bias cognition (as Mele has it) or how we avoid thinking about some distressing things (as Bach puts it).

The crux of the view relies on affective filters, so it is worth saying a few words about this notion. It appears that subjects spontaneously evaluate the epistemic status of information. As the metaphor goes, they filter information. One type of filter is cognitive: It contributes to the subject's epistemic vigilance (Sperber, 2000; Sperber et al., 2010). But some filters are

[4] We ignore here Pears's (1984) account of the dynamic paradox that involves forgetting one's intention to believe the flattering fact. We believe that this account partly suffers from the same problems as the traditional one. See Bach (1981) and Mele (2001) for criticism of Pears's account.

[5] It is worth noting that Mele argued that non-intentionalist accounts suffer from a selectivity problem of their own. Indeed, only some intentions to deceive oneself seem to succeed. Why is this so? We shall not discuss this controversy here (see Mele, 2001; Bermudez, 2000). Still, if the reader thinks that the selectivity problem also applies to the intentionalist picture of self-deception, this does not impact our model. Rather this reveals that the solution to the selectivity problem that we propose avoids this worry for both intentionalist and non-intentionalist pictures to be found in the literature (see part 3).

[6] We leave aside other potential worries for the non-intentionalist account, notably the distinction between self-deception and wishful thinking, and issues surrounding responsibility. We think indeed that these issues are partly orthogonal to the controversy between intentionalist and non-intentionalist accounts.

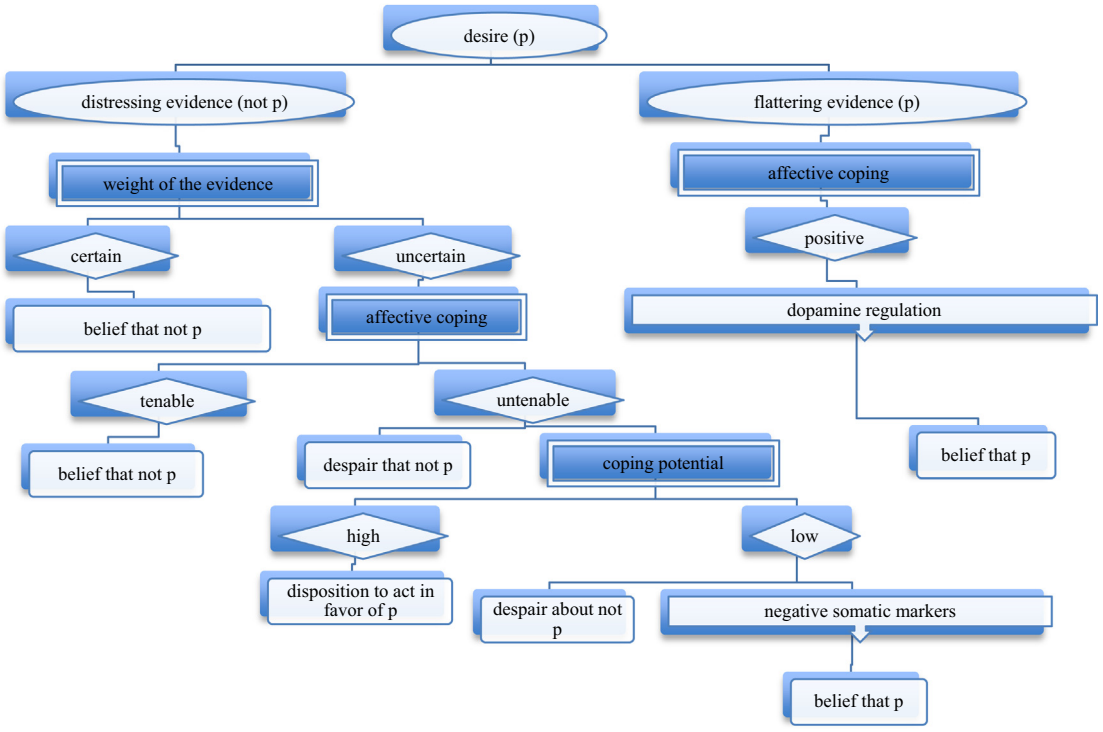

Fig. 1. Summarizes the various cognitive and affective appraisals involved in self-deception: the evaluation of the weight of evidence, the evaluation in the light of one's well-being and coping potential appraisal. The outputs of these appraisals can lead to the belief in the distressing fact, disposition to removing the threat or despair, by contrast to self-deception or belief in the flattering fact. The joint neurobiological mechanisms of somatic markers and dopamine elicited by some appraisals lead to self-deception. These mechanisms flesh out in more detail the motivated bias at stake in self-deception.

affective: They are regulated not by truth but rather by well-being. On the view we propose, self-deception crucially involves *affective* filters in addition to cognitive ones (Clément, 2006).

Our model is divided in two parts.[7] On the one hand, desire modulates attention by inhibiting the treatment of the evidence indicating that some desire is frustrated, i.e. the *distressing* evidence. This involves three appraisals at the psychological level, and is made possible by somatic markers at the neurobiological level (2.1). On the other hand, in addition to this inhibiting effect of desire on cognition, desire favors the treatment of *flattering* evidence or evidence for desire satisfaction (2.2). This involves a positive appraisal and is mediated, at the neurobiological level, by dopamine regulation. The view thus relies on two opposing yet complementary psychological and neurobiological processes induced by desire. Let us present each in more detail (see Fig. 1).

2.1. Desire and distressing evidence

Which cognitive and neural mechanisms enlighten the neglect of the distressing evidence by the self-deceived subject? In this section, we argue that three types of appraisals or evaluations of the distressing evidence are involved in self-deception: (i) appraisal of the evidence as *ambiguous*, (ii) *negative* anticipation of one's affective state and (iii) *low* coping potential appraisal. At the neurobiological level, these affective appraisals activate negative somatic markers, leading to the neglect of the distressing evidence by the subject. In other words, when Mary is presented with the stain, she discards this evidence speaking in favor of Sam's unfaithfulness in the following way. (i) She first evaluates it as *ambiguous*. This means that alternative explanations of the distressing evidence are available to her – explanations that point toward a happier conclusion than the one it suggests. (ii) Moreover, she anticipates the *negative impact* the evidence has on her well-being. (iii) Finally, she assesses that there is *nothing she can do* to neutralize this threat (or at least, that there is no easy way to do so). From a neurobiological perspective, these appraisals activate negative somatic markers in Mary's brain: The latter inhibit the treatment of the distressing evidence in the same way as they have been considered to discard some courses of action from the range of deliberation (Damasio, 1994).

We shall present each component in turn by adopting a dialectical order, which does not commit us to any sequence of appraisals. The order is methodological, not natural. Rather, we assume that there is no fixed order between these

[7] We focus here on straight self-deception. Our model can accommodate cases of twisted self-deception (the subject believes something that she does *not* desire to be true, because of a desire of hers), see Footnotes 10 and 13.

mechanisms and that they might even operate simultaneously (Balcetis, 2008: 374). Thus, this picture is far from adopting the view of various *homunculi* thinking about the situation in one's mind.

2.1.1. The strength of evidence

Evidence comes with strength. For instance, seeing Sam kiss Sally is strong evidence that he is unfaithful to Mary. By contrast, Sam's being late is weak evidence that he is having an affair, as lots of alternative explanations are available. But some situations are less clear than these extremes and involve ambiguity. For instance, seeing a photograph of the same kiss is still a good piece of evidence, but a less good one, as the picture might ultimately be a photomontage.

The subjective assessment of the strength of evidence is a function of the availability of alternative explanations: The more alternative explanations of some evidence are available to one, the less weight one will give to it. Far from being a mere conscious and rational endeavor, this evaluation is more often akin to a quick and automatic task (Huang & Rao, 2013). Of course, it can go wrong and is often biased. For instance, postdecisional counterfactual thinking studies have revealed how being an actor in an event increases the availability and salience of alternative explanations, in contrast with being an external observer of the same event (Girotto, Ferrante, Pighin, & Gonzalez, 2007). Consider a person who chooses an envelope containing a difficult multiplication task and fails to solve it, whereas she could have chosen another envelope containing a simple one. People reading this story typically think that the subject could have solved the problem if she had chosen the other envelope. By contrast, the actor of the event sees more alternative explanations: She might think that the failure is due to lack of time or the unavailability of a writing support (Girotto et al., 2007). As this example suggests, alternative explanations that are salient to participants tend to disculpate them (Gilbert, Morewedge, Risen, & Wilson, 2004). Since subjects are mostly actors or are personally involved in the event they deceive themselves about (vs. external observers), alternative explanations of the distressing evidence will be more salient to them than to external observers. For instance, alternative explanations of the lipstick stain might be more salient for Mary, as she is involved in the romantic relationship with Sam, while external observers are tempted to explain the same fact by Sam's infidelity only. She might think, say, that a woman has accidentally brushed his shirt with her lips in the subway. Since the availability of alternative explanations contributes to the decrease of the weight of evidence, being an actor increases the possibility of downgrading the weight of some evidence. Our evaluation of the weight of evidence is also biased by affect when the evidence presented is ambiguous. For instance, people in a negative state are inclined to form pessimistic beliefs, whereas people in a good mood tend to form optimistic ones, in the presence of ambiguous evidence (Eysenck, Mogg, May, Richards, & Al, 1991; Macleod & Byrne, 1996; Nygren, Isen, Taylor, & Dulin, 1996; Wright & Bower, 1992).

The first component of self-deception is that self-deceived subjects assess the distressing evidence as being ambiguous. How could Mary fail to believe that Sam is unfaithful when seeing clearly that he is kissing Sally and declaring to her his love? The only way she could fail to form this belief is a pathological one, one which does not count as irrational anymore and falls outside the realm of (ir)rationality. Empirical findings on self-deception support the presence of the evaluation of evidence as ambiguous: People deceive themselves in a dot-tracking task when told that their results reflect their intelligence, but only when the feedback on their performance was vague (Sloman et al., 2010). The authors of this experiment interpret their results considering other studies. For instance, they mention the famous experiment showing that people deceive themselves in leaving their hands longer in cold water when they are told that this indicates that they have a good cardiovascular system (Quattrone & Tversky, 1984). A plausible interpretation appeals to the uncertainty of the participants about their own threshold for tolerating pain. In another elegant study, Chance et al. (2011) have shown that people who had the opportunity to cheat on a first test (the answers were available and the subjects read them) overestimate their results on a second one, even if they have seen that the answers will not be available anymore for the second test. The authors propose that this overestimation might be due to a doubt about the evidence: The subjects did not interpret their performance as cheating because of ambiguous evidence (they might have thought that they already knew the answers when reading them even if it was not the case).

One might doubt that self-deception involves the evaluation of the distressing evidence as ambiguous. Indeed, the latter might *constitute* self-deception or *result* from it, rather than being a *condition for* it. As Bach and Mele observed, one way to deceive oneself precisely consists in modifying one's epistemic standards in the light of one's desires, for instance by decreasing the weight of distressing evidence. In the grip of self-deception, Sally might consider that being alcoholic requires drinking much more than the criterion adopted by non-alcoholic people and thus does not believe that she is alcoholic. We agree that in some cases downgrading of the evidence *results* from self-deception. This, however, is possible because the subject recruits some vagueness in her epistemic standard. Sally might, for instance, believe that health policies state more stringent standards than they really have. If so, some vagueness (about general epistemic standards) makes possible general self-deceived beliefs (such as "being alcoholic requires eight glasses of wine each day") that, in turn, will lead to more specific ones (such as "I am not alcoholic"). This reveals that the assessment of evidence is dynamic and that ambiguity can linger at the roots of belief-formation, i.e. affect one's epistemic standards.

So far, we have focused on the evaluation of the weight of distressing evidence. Yet, some cases of self-deception seem to involve the absence of recognition of the evidence altogether. Consider the well-studied phenomenon of change blindness: Large changes in the environment might be totally ignored if they are unexpected. For instance, people have been shown to fail to notice that one person has been replaced with another despite the fact that they were talking with the first person (Partyka, 2011; Simons & Levin, 1998). This might be explained by the fact that such a change is too discordant with one's expectations. Treating such information by means of attention would therefore be uneconomical for the brain. Still, change

blindness is possible only when changes are subtle (e.g. replacing a man with a woman rather than another man will be immediately detected). Self-deception might thus involve a similar blindness about the distressing evidence: If the event is unexpected yet subtle enough, the conditions for change blindness are met. If so, Mary might fail to notice the stain on Sam's collar. If there are such cases, the evidence is not assessed as ambiguous, as it is not attended at all. Yet, it is not the case that these subjects are certain that their desire is frustrated, which speaks in favor of our component of ambiguity.

Although ambiguity is required, it is not sufficient for self-deception to arise, and the reason why is instructive. Consider that Sally only slightly desires Sam to be faithful. Sam arrives at the restaurant very late. His hair is a real mess, he is unusually distracted and his shirt smells of a female perfume. This is quite strong evidence that he is having an affair, although there still is room for alternative explanations. Facing this situation, Sally believes that Sam is having an affair. One relevant difference between Sally's case and our initial example is that Sally only slightly cares about Sam being faithful, while Mary deeply desires Sam's faithfulness. As a consequence, the evidence Sally is presented with is not as hard to face as the evidence Mary is presented with. This brings us to the second component.

2.1.2. Anticipation of well-being

Not all negative events are on par. There are some we can affectively handle: Although they bring about some suffering, they do not impact our well-being in an important way. At most, they come with disappointment or temporary sadness. But there are more tragic events – some that are more difficult to live with. Depending on what one cares about most, some heartaches are more painful than others, some deaths can be profound wounds and some disappointments can be harsh. It is thus adaptive that people assess events by anticipating the impact they have on their well-being. People are notably bad at predicting how well they will fare when confronted with negative events (Gilbert & Wilson, 2009). Call this evaluation of events in the light of one's well-being "affective coping." It has been extensively studied in the context of the appraisal theory of emotion (Scherer, Schorr, & Johnstone, 2001).[8] Can we say more about the neurobiological mechanisms it might involve?

Damasio's work on decision-making provides some insights to answer this question, as he proposed a neurobiological mechanism devoted to the task of evaluating the consequences of courses of actions one considers (Damasio, 1994). Damasio proposed that decision-making importantly involves emotions by investigating subjects suffering from emotional deficits who had great difficulty making optimal decisions. For instance, his patient "Elliott" whom he considered to be a modern Phineas Gage, suffered from a ventromedial prefrontal (VMPC) lesion without manifesting any deficit in his ability to imagine and evaluate the consequences of his actions, as revealed by neuropsychological tests of decision-making. Yet, he was unable to make the right decisions. According to Damasio, this impairment in decision-making is explained by the emotional hyporeactivity due to his lesion (see Clark, Cools, & Robbins, 2004). He proposed that this patient could not eliminate options that would have negative emotional repercussions, as bad choices were not provoking somatic emotional responses. The moral Damasio draws from this case is that successful decision-making is far from being a cold process taking into account all the possibilities of action. Rather, making the right decision requires emotional anticipation signaling the bad consequences of some options and automatically eliminating them from the range of deliberation. Emotions do this via the somatic states they come with, which are automatically reactivated when we imagine future actions. This process can occur unconsciously and has been considered as part of *gut feelings* unconscious intelligence (Gigerenzer, 2007; Gigerenzer, 2008; Isenman, 2013). For instance, experiments on the Iowa gambling task reveal how subjects begin to choose the good decks of cards after the bad choices are accompanied by an increase of skin conductance and before the conscious understanding of the behavior modification. The hunch leading to the right decision was thus unconscious (Bechara, 1997). Conversely, experiments on addictive behavior have revealed that the consumption decision is underlain by anomalies in somatic states when addicts think of a risky decision that is not beneficial in the long term (Verdejo-García & Bechara, 2009; Naqvi & Bechara, 2009; Volkow, Fowler, Wang, Baler, & Telang, 2009). Thus, as in Elliott's case, addicts are impaired in their decision-making because of emotional deficits.

We suggest that self-deception involves an assessment of the distressing evidence in the light of one's well-being: People assess the evidence as being untenable or difficult to handle, which activates a mechanism similar to the somatic markers involved in decision-making. To come back to our example, when Mary sees the stain, somatic markers indicate to her how she would be devastated if Sam was having an affair. Thagard (2007) already proposed that somatic markers are involved in self-deception, although he did not pursue the analogy in full detail.[9] As the traditional somatic markers have the function of excluding some courses of action from the range of the ones considered seriously by the subject, we suggest that affect might automatically discard distressing evidence in the case of self-deception. If the analogy stands, self-deception and decision-making crucially involve affect in an automatic manner, i.e. non-intentionally and subconsciously, since this is what makes somatic markers economical.

This appraisal is one way of understanding the conflict between truth and happiness that is at the heart of self-deception. The paradigmatic content and population of self-deception speak in favor of such negative affective anticipation. Self-deception typically concerns things we deeply care about, such as our intelligence, romantic relationships, and health (Goldbeck, 1997; Mele, 2001). Being presented with evidence that our desires for these things are frustrated is likely to come

[8] See the *adjustment capacity dimension* of coping appraisal, i.e. the evaluation of one's capacity to adapt to a change in the environment (Scherer et al., 2001).

[9] Thagard (2007) is concerned with conflicts of interest and incidentally touches on self-deception.

with negative affect. Indeed, negative emotions have the function of informing us about how well we fare with regard to our concerns, in particular of signaling threats and obstacles to desire satisfaction. Similarly, the populations especially prone to self-deception, i.e. addicts, seriously sick people and paranoid persons, point our inquiry in the same direction. For instance, facing one's addiction, one's impending death or one's loss of control might be too devastating for one and thus might be accompanied by negative affect.

Moreover, experiments support the view that self-deception involves somatic markers. People scoring high on self-deception questionnaires like the Balanced Inventory of Desirable Responding (Paulhus, 1998) seem to have more difficulties in integrating information which is incongruent with their desire, and they show perseverative errors even if their strategy has negative consequences for them (Peterson et al., 2002, 2003). Peterson and colleagues interpret these findings by means of the somatic markers hypothesis: As for the traditional Iowa gambling tasks experiments, it is plausible that incongruent information is affectively marked. For rational subjects, this mark comes with the motivation to investigate and adapt one's behavior accordingly. In contrast, self-deceived subjects seem to ignore such signals in their consequent behavior. The authors conclude that self-deception is thus a failure of accommodating incongruent information despite the affective signal that comes with it.

At this stage, one might think that there is a step from the negative evaluation of the distressing evidence to somatic markers. One might agree that the negative evaluation is a type of affective appraisal. Yet, this in itself does not require adopting the somatic markers hypothesis, since the latter specifies a mechanism by which information can be discarded and not only negatively assessed. We agree that there is a step from affective appraisal to somatic markers that needs clarification. However, restricting at the psychological level of appraisal does not suffice to explain how subjects – or subjects' brains – discard some piece of evidence. Distressing information is all too relevant for subjects to be ignored and is therefore likely to be treated automatically. The appeal to somatic markers is meant to provide us with a mechanism by which negative affect can lead to silencing evidence, as this seems to be the case in self-deception. Moreover, an experiment about the way political preferences bias our treatment of incongruent information suggests that somatic markers are involved in motivated cognition (Westen, Blagov, Harenski, Kilts, & Hamann, 2006). Participants presented with incongruent information (say, a Republican presented with unflattering evidence about Republican candidates in the 2004 US election) exhibit greater activation in the medial prefrontal cortex and, as a result, do not believe the incongruent information. Now, this cerebral region corresponds to somatic markers, a conclusion to which the authors allude.[10]

We have argued that somatic markers contribute to the subject's discarding of the distressing evidence. But how exactly is this to be understood? Does the subject's brain attend to the distressing evidence and, in a second step, discard it, or can affective filters prevent one from noticing the evidence? This touches on controversies surrounding selective attention.[11] We will not settle this issue here, as this goes far beyond the scope of this paper. It suffices to observe that the affective filter described might be an early filter of information: Rather than filtering the distressing evidence before rejecting it, it might distract subjects from consciously attending to the unwelcome piece of evidence in the first place. Mary's distress in the presence of Sam's stain could distract her from noticing it. Indeed, affect can influence pre-conscious attention (e.g., Mathews & Macleod, 1994). Thus, self-deception might involve low levels of information processing exacerbated by emotional processes.

Since the picture of self-deception sketched so far relies on the recognition of an important threat to one's well-being, this requires explaining why self-deceived subjects do not act so as to face adversity. Negative emotions typically come with behavioral tendencies – a fight-or-flight response – that help us face bad events (Frijda, 1987). It is thus mysterious that subjects do not act to neutralize the threat provided that they have evaluated the situation as unbearable. Somatic markers are only the beginning of an explanation, as they are in principle compatible with behavioral tendencies. In the absence of a further explanation, self-deception will appear strongly maladaptive as it prevents one from finding a solution. Although it is an open question whether self-deception is on the whole adaptive, our model so far makes it strikingly maladaptive. Yet the persistence of self-deception suggests that it is at least partly adaptive. This is where our third component enters the picture.

2.1.3. Coping potential appraisal

In addition to assessing events in the light of our well-being, we also appraise events through our ability to deal with them by *acting* on the situation (coping potential appraisal). Lazarus, in his seminal study on stress (1966), proposed that stressful situations involve an appraisal of the situation as personally significant and exceeding one's resources for coping, an appraisal that is essentially affective. More generally, the appraisal theory of emotions has emphasized that coping potential is one of the appraisal dimensions that determine the elicitation and differentiation of emotions (Scherer et al., 2001). For instance, anger typically comes with the tendency to reinstate justice and thus with an assessment that one can act to modify the world (Frijda, 1987), as opposed to sadness, which involves the awareness that there is nothing to do to improve the situation. The evaluation of our coping potential is more fine-grained than this: We assess whether *we* can act on the

[10] We think that this component can make sense of twisted self-deception. Imagine that Sam is paranoid or very anxious. In the absence of sufficient evidence that Mary is having an affair, he believes that she is betraying him, because he strongly desires her not to do so. One plausible explanation of such cases appeals to the idea of costs and benefits of beliefs for one's well-being (Mele, 2001): The subject is afraid of being mistaken and thus prefers believing a distressing fact rather than being wrong about a flattering one. It should be clear to the reader that this explanation clearly shares the spirit of the affective filter view.

[11] For instance, the "cocktail party" effect shows that self-pertinent information (i.e. one hears one's own name) is treated predominantly even if it falls outside the scope of focused attention (Arons, 1992). Yet, scholars disagree on whether the information has been treated and then consciously attended or whether the filter operates earlier on before treating the relevant information (not processed by higher and semantic analyses).

situation (vs. other agents) – the *control* dimension of coping (Scherer et al., 2001) – and evaluate the resources available to cope with a situation – the *power* dimension of coping. Among situations one can act on, some involve more effort, pain, luck and so on than others. The coping potential appraisal is thus far from being an all-or-nothing matter.

The last type of evaluation of the distressing evidence required for self-deception consists in the subject assessing her coping potential as being low. The event suggested by the evidence is either assessed as falling outside one's control or as controllable only with difficulty. In other words, people do not deceive themselves when they assess the situation as one that can be easily changed by their own action, even if they appraise it as difficult to handle affectively. Consider a variation of Mary's case: Imagine that she thinks that she can easily convince Sam to stop having an affair in the case that he has one. Seeing the lipstick stain, it is likely that she will not deceive herself despite assessing the affair as being devastating for her; rather, she will try to avoid this suffering by acting on the situation, say, by talking to Sam. By contrast, we hypothesize that self-deceived Mary assesses that she cannot handle Sam's infidelity (or can only handle it with difficulty) *and assesses that there is nothing she can do to avoid this suffering (or that acting to improve the situation comes with a cost)*.

Paying attention to the content of self-deception and the populations especially prone to it suggests that this component is on the right track. Consider matters on which people deceive themselves, such as intelligence, driving skills, health, attractiveness, and relationships. Although we have some control over these things, this control can be importantly low or difficult to exercise – a fact people seem to be aware of. If our third component is correct, it is not astonishing that people deceive themselves about these matters much more often than they do about their ability to take a bus, to talk, to wake up and so on.[12] Similarly, the people especially prone to self-deception (addicts, people suffering from an incurable sickness, paranoid people) are in conditions where a significant loss or impairment of control has taken place – or at least subjects may believe so. The very striking cases of genocide denial come to mind as well: One way to explain this independently of emotions such as hatred or racism is the affective unbearableness of one's responsibility and the impossibility of properly reinstating one's dignity. This distribution of self-deception is no longer mysterious if one adopts our component.[13]

Some experiments suggest that self-deception involves appraisal of impaired control. Studies reveal that motivated biases in information gathering depend on the control one thinks one has on the topic at stake. For instance, people are less likely to require a medical diagnostic for a disease or to ask for more information about it when they are told that the disease is untreatable (vs. treatable; Dawson, Savitsky, & Dunning, 2006). Controllability of the disease appears more determinative than its severity in information seeking. Similar results have been found about academic ability: Even when academic ability is important for the participants, the controllability is the key to the subjects' treatment of information (Dunning, 1995). Although motivated information seeking strictly speaking differs from self-deception, both might involve similar mechanisms to some extent, such as low coping appraisal.

Some cases of self-deception might result in a reassessment of one's coping potential. For instance, alcoholic people tend to believe that they have control over their drinking consumption despite evidence to the contrary (Strom & Barone, 1993). Still, we think that such cases are compatible with our condition of low coping potential: It is likely that, at some point, addicts realize how difficult stopping their consumption is and deceive themselves into believing the contrary. This underlines that self-deception involves the dynamic appraisal of one's coping abilities.

So far we have restricted our attention to the way desire impacts the treatment of the distressing evidence. Yet subjects may be uncertain about some evidence, assess it as disastrous and falling outside their control, but not deceive themselves. Rather, they might be anxious or even form a distressing belief and fall into despair. Interestingly, depressed people deceive themselves much less than non-depressed people do (Surbey, 2011) – a phenomenon called "depressive realism." Since our picture is so far compatible with despair, i.e. the opposite of self-deception, it needs an important supplementation. This, we think, is provided by the appeal to dopamine.

2.2. Dopamine, desire and flattering evidence

Any plausible account of self-deception should explain how subjects form the deluded belief and respond to flattering evidence – not merely how they neglect the distressing information. We propose that the role of desire in monitoring one's attention is the key to understanding how flattering evidence takes precedence over distressing evidence in self-deception. As dopamine is the neurotransmitter of desire, we shall account for this influence of desire on cognition by means of the neurobiological mechanism of dopamine regulation.

Dopamine is a neurotransmitter known to be central in reward and reinforcement circuits, and it is more related to desire than pleasure. In some famous experiments, Schultz (1997) and Schultz, Tremblay, and Hollerman (1998) showed that dopaminergic activity increases when a monkey sees the cue predicting a reward and decreases if the reward does not follow the anticipating cue. This suggests that dopamine might be implicated in coding for prediction errors and is related to the anticipation of pleasure, in particular proximal expected pleasure, since dopaminergic neurons' activity is decreased when reward delivery is delayed and increased when the reward is delivered rapidly (Bermudez & Schultz, 2014). Conversely, lack of dopamine is associated with apathy, lack of motivation and frequently with depression and anxiety, as observed in patients with Parkinson's disease (Sagna, Gallo, & Pontone, 2014).

[12] Of course, when people believe that they cannot easily act on the things just mentioned, they might as well deceive themselves about them.

[13] As cases of twisted self-deception involve paranoia and anxiety, conditions in which subjects think that their control is significantly impaired, our component fits well with such cases as well.

This dopaminergic reward system is modulated by prefrontal cortex top-down control. In fact, self-control (regulation between reason and desire) seems to strongly rely on the balance between prefrontal cortex and dopaminergic neurotransmission (Heatherton & Wagner, 2011). An increase in dopamine is thus responsible for what is typically considered to be practically irrational behavior. Hyperdopaminergic symptoms that occurred in patients treated with L-Dopa show that too much dopamine increases stereotypic behaviors, hypersexuality, pathological gambling and compulsive shopping (Ardouin et al., 2009). Likewise, addictions (to drugs or behaviors) provoke strong changes in dopaminergic circuits and an increase in dopaminergic activity. For instance, gambling disorder, a form of maladaptive behavior considered to be a behavioral addiction, is strongly linked with dopaminergic dysfunctions (Linnet, 2014). Dopaminergic neurons hijacked by drugs respond automatically to cues associated with the drug. These cues become hypersalient and quite impossible for an addicted person to ignore even if they do not provoke pleasure anymore. According to Robinson & Berridge, the increase in dopaminergic activity that is involved in addiction provokes a shift from liking to wanting (Berridge, Robinson, & Aldridge, 2009). In the beginning, people take a drug to obtain a pleasurable positive reinforcement; later they continue taking it despite absence of pleasure and because of the intense craving induced by stopping. This hypersensitivity might explain relapses induced by reminder cues, even for addicts who stop taking drugs for several years. This increase in the dopaminergic system is accompanied with a decrease in frontal activation, in particular in the ability of the frontal cortex to inhibit compulsory behavior in the presence of reminder cues (Crews & Boettiger, 2009). Like people with a ventromedial prefrontal lesion, people suffering from addictions show decision-making deficits in choosing immediate rewards even if they have negative future repercussions.

Dopamine also impacts one's cognitive states. The previously mentioned effect of dopamine on attention, i.e. that dopamine comes with hypersalience of cues, seems also present in non-addicted people (Dill & Holton, 2014). Similarly, hallucinations and delusions in psychiatric diseases like schizophrenia strongly involve an increase in dopaminergic transmission (Os & Kapur, 2009). We suggest that dopamine also influences cognition by easing us into deceiving ourselves.

Empirical findings on self-deception provide further justification for this speculation. It appears that self-deception involves a frame of mind in which dopamine is widespread. People with an increased number of dopamine receptors are more prone to self-deception, and dopamine enhances the optimism bias (Sharot, Guitart-Masip, Korn, Chowdhury, & Dolan, 2012). Experiments on motivated cognition point in the same direction: Participants being presented with congruent information about their favorite politician manifest increased activation in the ventral striatum of the right caudate, i.e. a dopaminergic region of the brain (Delgado, Miller, Inati, & Phelps, 2005), which explains why they believe congruent rather than incongruent information. We propose that dopamine contributes to self-deception by means of the equilibrium between reason and desire explained earlier.

On the one hand, dopamine comes with the positive evaluation of desire satisfaction and privileged attention directed to positive cues or evidence for desire satisfaction. Turning to our example, Mary is anticipating how pleased she would be by Sam's faithfulness and attends primarily to evidence for it, such as Sam's caring for her. This process is partly affective and is the positive counterpart of the negative evaluation and somatic markers presented earlier. On the other hand, as was just explained with the case of addicts, this positive anticipation can come with a drop in prefrontal cortex activation and can thus take precedence over rational cognition. As noted, addicts are more prone to deceive themselves (Ferrari, Groh, Rulka, Jason, & Davis, 2008; Walker, 2010). The decrease of frontal lobe activation in conjunction with the increase in the dopamine system should lead to a neurobiological situation favoring self-deception to its maximum. Indeed, increased frontal lobe activation decreases unrealistic optimism and anosognosia, which can be considered forms of self-deception (Mckay et al., 2013). In other words, the more one's prefrontal cortex is activated, the less one deceives oneself; and the more dopamine transmission, the more one deceives oneself. The role of dopamine might explain why addicts are so prone to self-deception and how *flattering* evidence is predominantly treated by self-deceived subjects at the expense of rational cognition. If there is an analogy between behavioral addiction and self-deception, as suggested by our proposal, one might speculate that people deceive themselves at the beginning to obtain a pleasurable sensation, while having the possibility to stop deceiving themselves. But, as time goes by, self-deception might become like a habit (typically encouraged by dopaminergic transmission) to pay attention to possible rewarding cues instead of negative ones. Dopamine might thus partly explain the persistence of deluded beliefs.

Finally, dopamine is released in large quantities especially in cases of uncertainty (Anselme & Robinson, 2013; Preuschoff, Bossaerts, & Quartz, 2006). For instance, the placebo effect, which relies on dopaminergic transmissions, is successful only in case of vague evidence about one's healing mechanisms (Benedetti, 2014). Since we have argued that the appraisal of distressing evidence as ambiguous is a component of self-deception, it can increase dopamine transmission, which, in turn, will favor self-deception. Conversely, dopamine might contribute to the subject decreasing the weight of the distressing evidence, given that the neurotransmitter orients attention toward pleasure. The stereotype of the businessman, for example, inundated with all the tasks implied by his job, could bring some support to Mary's image of a good partner and explain away the distressing evidence. Dopamine thus provides a fully consoling story, from positive anticipation of some event to an appeasing explanation of incongruent information.

To summarize, people form self-deceptive beliefs because they prefer immediate rewards, even if this could have long-term negative consequences, in the same manner as addicts choose immediate rewards despite negative repercussions. This preference involved in self-deception can be explained by the cognitive and affective appraisals described and the joint mechanisms of somatic markers and dopamine. Fig. 1 summarizes the affective filter view. We think that self-deception emerges as a pattern out of the components described, a hypothesis that merits being empirically tested. Future empirical research based on this model might provide more detail about the way these features interact.

If this picture is right, self-deception is a type of affective coping. When there is space for doubt, when we think that suffering is inescapable and too hard to handle, self-deception might be our last chance to avoid distress, at least momentarily (see Rorty, 1972; Rorty, 1994). Self-deception is thus a protection from depression, like other biases such as the immune neglect in future emotional forecasting (subjects underestimate positive events that will occur in the future, a mechanism that is meant to protect them; Gilbert & Wilson, 2009). Our picture is in line with the view that self-deception has the function of reducing anxiety (Johnston, 1988; Barnes, 2007), although we do not take a stance on the efficacy of self-deception in reducing anxiety, let alone the issue of the function of self-deception and appeal to avoidance of suffering rather than anxiety (see Scott-Kakures, 2000). More generally, our model aligns self-deception with the various protective mechanisms of the mind, or what has been called the psychological immune system. These mechanisms lie at the heart of psychoanalysis and have been recently investigated beyond this literature by studies on cerebral underpinnings of psychoanalytic mechanisms (Freud, 1955/1920; Alberini, 2013; Ansermet & Magistretti, 2007; Ansermet & Magistretti, 2010; Bazan & Detandt, 2013). As we shall see now, this can illuminate the paradoxes of self-deception. Let us close this paper by revisiting them with the help of the affective filter view.

3. Revisiting the paradoxes of self-deception

In this section, we argue that the affective filter view meets the challenges we have presented in the first part: It invites us to defuse the static issue and to adopt a non-intentionalist picture of self-deception.

3.1. The affective filter view and the static paradox

As mentioned, describing the product of self-deception is a controversial issue. In what follows, we defend that the affective filter view can illuminate this controversy by providing aficionados of each solution with the further motivation they need and by motivating an integrative account of self-deception.

The affective filter view relies on appraisals and mechanisms that inhibit the treatment of distressing evidence. This is, in principle, compatible with several products of self-deception. Indeed, the emotional "scanning" that is part of our picture is fast, automatic and largely implicit. In a situation involving a threat or salient emotional cues (reinforcements, such as food or an erotic partner), the brain evaluates the situation very quickly and reacts to these emotional stimuli *before* subjects become aware of this evaluation. For Scherer, appraisal checks such as the ones described can be carried out at different levels of complexity and do not imply conscious procedures – in fact, far from it (Leventhal & Scherer, 1987). Self-deception might thus be a form of automatic emotional evaluation with minor top modulation and might then result in an unconscious belief (the standard account), not thinking about some things (Bach's proposal) or even the absence of the distressing belief altogether (Mele's account). On the basis of empirical literature on attention, we have even ventured that subjects might fail to notice the distressing evidence. Our picture thus does not commit us to a single product of self-deception, but rather favors an integrative picture. Yet, as observed in the first part, this variety of cognitive products of self-deception should be motivated. In what follows, we shall explain how the affective filter view has the resources to do so.[14]

Somatic markers are crucial in decision-making in virtue of automatically discarding some courses of action from the range of deliberation and preventing the formation of some intentions (see part 2). Analogously, if self-deception involves somatic markers, subjects might automatically discard the distressing evidence and believe the flattering fact only, as in Mele's picture. As speculative as this hypothesis is, it is based on the function of somatic markers and follows from our analogy. This is compatible with the subject assessing the distressing evidence as described, since information can be stored in one's brain without one activating the relevant belief, as it appears in *aha* experiences (Jung-Beeman et al., 2004).

Moreover, the same analogy can help rebutting the action challenge to the main accounts, i.e. explaining why self-deceived subjects may act in a way that suggests that they believe the distressing fact (see part 1). If self-deception involves similar mechanisms to decision-making, self-deceived subjects may discard some evidence and then provide *post hoc* rationalizations of their actions. For instance, Sam's hiring a mathematics teacher for his son can be explained by his belief that his son does not have difficulties in mathematics and that hiring a teacher will make him even better. This type of confabulation is common in decision-making: People tend to be more convinced about the soundness of their reasons to act after some decision has been made, which is explained by affective processing of the type mentioned (Haidt, 2001). In the case of self-deception, people will form other deluded beliefs (such as Sam's belief that his son can become even better in mathematics), which leads to self-deceptive actions (hiring a teacher). If the analogy stands, literature on decision-making can help rebut the action challenge.

As for the anxiety challenge, Mele argued that being presented with distressing evidence might suffice to capture the anxiety of some self-deceived subjects without them believing the distressing fact (Mele, 2001). Our picture can provide some further justification for this proposal, as it appeals to the assessment of the evidence as harmful – a piece of information that is stored in the subject's brain and might thus account for anxiety.

Our view can also provide the defender of the traditional picture with a mechanism accounting for unconscious belief. Recall that our model is in line with psychological protective mechanisms. Now, protective mechanisms leading to uncon-

[14] For reasons of space, we ignore Gendler's proposal here, as it constitutes a less standard solution to the static paradox.

scious states are plausibly activated when anticipation of suffering and low coping potential are involved: People unconsciously believe or forget things they assess as too difficult to handle, or form unconscious desires because the latter are too difficult to avow, and these unconscious phenomena all concern things people think they have little if any control over. Moreover, recent approaches bridging psychoanalysis and neuroscience have shed new light on drive by its link with somatic markers (Alberini, 2013; Ansermet & Magistretti, 2007; Ansermet & Magistretti, 2010) and dopaminergic transmission (Bazan & Detandt, 2013). It thus appears that our view is in line with recent studies on the mechanisms of unconscious states' formation, which helps remove the charge of mystery raised against the traditional picture of self-deception.

So far, our discussion suggests that the main accounts are on a par, since our model can provide the resources to meet the challenges raised. Still, more justification – in particular empirical support – is needed to warrant these answers to the challenges. Moreover, the action and anxiety challenges can be met by appealing to the presence of two conscious beliefs with distinct degrees of credence, as emphasized in the first part. Defenders of the main views should explain why this is not what is going on in some cases.

Our picture is compatible with the latter explanation. An integral part of our view is that self-deceived subjects assess their evidential situation as being ambiguous. It is thus no wonder that they might form two beliefs with distinct degrees of credence, the distressing belief having a weaker degree of credence than the flattering one, given the respective evaluations of the strength of evidence. Dopamine might even reinforce this effect, as suggested earlier. In the absence of further empirical evidence favoring one picture over another, it is plausible to think that the product of self-deception comes in a variety: Sometimes people deceive themselves by forming the deluded belief only; in other cases, they might form an unconscious belief or not think about the distressing fact; and they may sometimes form two beliefs with distinct degrees of credence.

This integrative account can be motivated further by appealing to biases in general, as Mele does. In our view, self-deception is a motivated bias, but also an affective one, since its main mechanisms are affective. This invites investigating self-deception alongside affective biases such as optimism, anxiety, phobias, or biases that are part of addiction. Now, affective biases do seem to admit different cognitive products. Consider that Sam is an anxious person and is presented with some ambiguous piece of evidence that, say, his performance of some task is going well. Studies reveal that he is likely to pay more attention to threatening stimuli than non-anxious people (Mathews & Macleod, 1994; Mineka, Watson, & Clark, 1998; Mogg & Bradley, 1998). He might not have noticed the positive cue and thus might not believe that the task is going well. He might also have noticed it and might then subconsciously believe that the task is going well, while consciously believing that a disaster is happening. Alternatively, he might believe that things will be fine to a lesser degree of credence than he believes the contrary. Similar observations apply to other types of affective biases. If so, a unified picture of affective bias invites us to adopt an integrative picture of their products, including those of self-deception. Although this observation mirroring Mele's approach should be empirically tested, it is sufficient to switch the burden of proof onto defenders of non-integrative pictures of the product of self-deception. In the absence of further empirical evidence, it appears that the static controversy should be dissolved in the light of the variety of the cognitive outputs of affective filters. This suggests we change perspective on self-deception by starting with intuitions about what self-deception involves and supporting them with empirical evidence, rather than by approaching it in a way that makes it *prima facie* impossible. This approach will also help solve to the *dynamic* paradox.

3.2. The affective filter view and the dynamic paradox

How does the picture presented contribute to the dynamic paradox, i.e. does it speak in favor of the presence of an intention in self-deception? Our view does not mention intentions. We have argued that the several appraisals and mechanisms described suffice to make sense of self-deception. Moreover, as observed, they can be automatic and therefore do not necessarily involve intentions. The view thus favors a non-intentionalist account. Yet, any non-intentionalist picture needs to rebut the challenges of the selectivity problem and of the analogy between self-deception and interpersonal deception. How does the proposed view fare in this respect?

In presenting our model, we emphasized that not all desires are on par as far as anticipation of well-being, coping potential and ambiguity of evidence are concerned. This, we think, can meet the selectivity challenge. Our account does not restrict itself to claiming that self-deception is a motivated bias. Rather, we fleshed out distinct components constituting this bias. The several conditions for self-deception make it less and less likely that any desire will lead to self-deception. Our components can thus explain how only some desires lead to self-deception and thus solve the selectivity problem without recruiting intention.[15]

Of course, the affective filter view cannot secure a strict analogy between self-deception and interpersonal deception, as it does not appeal to intentions. Nonetheless, it guarantees a close enough analogy between the two types of deception. People deceive others when similar appraisals are involved. Deceivers think that telling the truth will be too harmful and expect some benefits from the deception. They assess the distressing situation as one that cannot be easily modified and evaluate the evidence in favor of the distressing fact as, at most, ambiguous for the deceived person. Isn't this enough similarity between self-deception and interpersonal deception? It seems so. Moreover, empirical findings on neural mechanisms involved in interpersonal deception suggest that there is a difference between interpersonal and intrapersonal deception in terms of intentions. Interpersonal deception requires conscious efforts (for instance, to imagine possible contradictions), which translates itself into

[15] As emphasized earlier, it has been argued that intentionalists suffer from a selectivity problem of their own (see Footnote 5). Since our model offers a solution to the selectivity problem, it does not suffer from the pitfalls of both intentionalist and non-intentionalist pictures that do not appeal to emotions.

strong frontal activation (Christ, Essen, Watson, Brubaker, & Mcdermott, 2008) and can be identified with the efforts to bring about a plan. In contrast, as observed, self-deception implies a decrease in frontal cortex activation due to dopamine regulation and relies on emotional structures that are less controlled and less inhibited by the prefrontal cortex than others. Both features suggest that no plan is involved. The difference between interpersonal deception and self-deception in terms of intentions is thus vindicated, while the partial analogy between the two types of deception is acknowledged.

Let us close this section with a final note on affective biases, one that echoes Mele's approach again. Affective biases like optimism, anxiety and others do not seem to involve intentions. Rather, they are paradigmatic instances of automatic cognition, i.e. subconscious and *non-intentional* processes. For instance, in being optimistic, Mary is inclined to believe that good events will happen in the face of equally strong evidence for good or bad scenarios, without any intention to do so. Likewise, in addicts, the amygdala and the dopaminergic *nucleus accumbens* are automatically activated by drug-related stimuli. The latter orient our attention even if they are presented subliminally (Childress et al., 2008) and particularly when the prefrontal cortex cannot play its inhibiting role. No intention is needed to describe this attentional bias. If self-deception is an affective bias and if affective biases need not be intentional, self-deception might not be intentional either. We thus agree with Mele: We can understand self-deception without appealing to intentions and by appealing to studies on biases. However, we emphasize that we can do without intentions because emotions play the role traditionally ascribed to intentions in this debate, i.e. easing us into deceiving ourselves. In this sense, we hope that the promise of the affective revolution of self-deception has been partly met.

4. Conclusions

What can cognitive science teach us about self-deception? In this paper, we have recruited findings from cognitive neuroscience and psychology to approach self-deception. We have argued that affective filters, in particular emotional appraisals, lie at the heart of self-deception and that the neurobiological mechanisms of somatic markers and dopamine regulation can enlighten how we deceive ourselves. This picture is inspired by empirical evidence but is also driven by armchair intuitions. After all, it is a common experience that some truths are difficult to handle, especially when there is no way out. And it is part of everyday life that desires impact our attention. The model we have proposed puts these intuitions together and confronts them with empirical evidence. Fortunately, our findings are consonant with the armchair in this case.

We have drawn two morals with regard to the paradoxes of self-deception. First, an empirically minded look at the static controversy favors an integrative account, as far as the available evidence is concerned. Second, we have further motivated a non-intentionalist account of self-deception. In this respect, we agree with Mele, except for one thing: the neglect of emotions.

In addition to its empirical plausibility and the contribution to the paradoxes mentioned, the affective filter view opens new paths for understanding motivated biases by means of emotions. For instance, it can easily accommodate the differences and similarities between self-deception and wishful thinking. If wishful thinking usually does not involve being presented with distressing evidence or at least not with the same salience, it might not involve the very same mechanisms with the same intensity. However, it is plausible that coping appraisals, somatic markers and dopamine are involved in wishful thinking. The view can also approach motivated information gathering along similar lines. It can as well be extended to cognitive dissonance, as far as *cognitive* well-being and *cognitive* coping are considered: Some thoughts are difficult to handle, since they conflict with too many of our core theoretical beliefs and choices we have made (well-being), and accommodating them requires drastic changes in our beliefs, a task that can extend our cognitive capacities (coping). More generally, the combined activation of cognitive and affective checking devices is susceptible to explain other cases of credulity. The notion of "cognitive lures", for example, is particularly useful to give an account of phenomena such as advertising, swindles or adherence to cults (Clément, 2006).

Finally, the affective filter view can enlighten the controversy about the adaptiveness of self-deception and, more generally, positive illusions. One might deem it mysterious that self-deception evolved, since it involves distortion of our beliefs and thus does not fit their purpose. As true beliefs are crucial to fare well, self-deception seems maladaptive. Wonder decreases, however, as soon as one recognizes the existence of two attentional filters that are both largely inherited by phylogenesis: the cognitive and the affective. The role of the cognitive evaluation process is to check for the truthfulness of the incoming information. It can be seen as a "cognitive filter" and proceeds in a largely implicit way. One of the functions of the affective evaluation process is to check the emotional impact of information on the well-being of the organism. It is thus not astonishing that self-deception has evolved: We speculate that it emerges from two systems that, in isolation, have proper and precious functions. It is then in a sense adaptive, to the extent that affective coping mechanisms are: After all, emotional well-being is one face of adaptiveness. Is this to say that self-deception is a spandrel, accident or by-product of other adaptive mechanisms (Van Leeuwen, 2009)? Is our speculation compatible with self-deception having a *sui generis* function? These questions open the path to the vexed issue of whether self-deception ultimately leads to happiness. Here, we have only argued that the promise of happiness is the crux around which self-deception revolves. Whether the promise is met or not is a matter for another investigation.[16]

[16] This article has been presented at the International Society for Research on Emotions (Geneva 2015), Thumos Seminar (Geneva 2015), Cognitive Science Centre Colloquium (Neuchâtel 2015) and the conference "Beliefs that Feel Good" (Basel 2015). We thank the audience of these talks for their comments. We also wish to thank Richard Dub and the anonymous referees for their constructive suggestions as well as the Centre for Cognitive Sciences (University of Neuchâtel) for having founded this research. The study was partially founded by a public grant overseen by the French National Research Agency (ANR) as part of the program "Licornes" (ANR-12-CULT-0002).

References

Alberini, C. M. (2013). *Memory reconsolidation*. London: Elsevier Academic Press.
Anselme, P., & Robinson, M. J. F. (2013). What motivates gambling behavior? Insight into dopamine's role. *Frontiers in Behavioral Neuroscience, 7*.
Ansermet, F., & Magistretti, P. J. (2007). *Biology of freedom: Neural plasticity, experience, and the unconscious*. New York: Other Press.
Ansermet, F., & Magistretti, P. (2010). *Les énigmes du plaisir*. Paris: O. Jacob.
Ardouin, C., Chéreau, I., Llorca, P.-M., Lhommée, E., Durif, F., Pollak, P., & Krack, P. (2009). Évaluation des troubles comportementaux hyper- et hypodopaminergiques dans la maladie de Parkinson. *Revue Neurologique, 165*(11), 845–856.
Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society, 12*(7), 35–50.
Ashley, A., & Holtgraves, T. (2003). Repressors and memory: Effects of self-deception, impression management, and mood. *Journal of Research in Personality, 37*(4), 284–296.
Bach, K. (1981). An analysis of self-deception. *Philosophy and Phenomenological Research, 41*(3), 351.
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.
Barnes, A. (2007). *Seeing through Self-Deception*. Cambridge: Cambridge University Press.
Baxter, P., & Norman, G. (2011). Self-assessment or self deception? A lack of association between nursing students' self-assessment and performance. *Journal of Advanced Nursing, 67*(11), 2406–2413.
Bayne, T., & Fernández, J. (2009). *Delusion and self-deception: Affective and motivational influences on belief formation*. New York: Psychology Press.
Bazan, A., & Detandt, S. (2013). On the physiology of jouissance: Interpreting the mesolimbic dopaminergic reward functions from a psychoanalytic perspective. *Frontiers in Human Neuroscience, 7*.
Bechara, A. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293–1295.
Benedetti, F. (2014). Drugs and placebos: What's the difference? Understanding the molecular basis of the placebo effect could help clinicians to better use it in clinical practice. *EMBO Reports, 15*(4), 329–332.
Bermudez, J. L. (1997). Defending intentionalist accounts of self-deception. *Behavioral and Brain Sciences, 20*(1), 107–108.
Bermudez, J. L. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis, 60*(4), 309–319.
Bermudez, M. A., & Schultz, W. (2014). Timing in reward and decision processes. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1637). 20120468–20120468.
Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'Liking', 'wanting', and learning. *Current Opinion in Pharmacology, 9*(1), 65–73.
Chance, Z., Norton, M. I., Gino, F., & Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. *Proceedings of the National Academy of Sciences, 108*(Supplement_3), 15655–15659.
Childress, A. R., Ehrman, R. N., Wang, Z., Li, Y., Sciortino, N., Hakun, J., ... O'brien, C. P. (2008). Prelude to passion: Limbic activation by "Unseen" drug and sexual cues. *PLoS ONE, 3*(1).
Christ, S. E., Essen, D. C. V., Watson, J. M., Brubaker, L. E., & Mcdermott, K. B. (2008). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex, 19*(7), 1557–1566.
Clark, L., Cools, R., & Robbins, T. (2004). The neuropsychology of ventral prefrontal cortex: Decision-making and reversal learning. *Brain and Cognition, 55*(1), 41–53.
Clément, F. (2006). *Les mécanismes de la crédulité*. Genève, Paris: Droz.
Crews, F. T., & Boettiger, C. A. (2009). Impulsivity, frontal lobes and risk for addiction. *Pharmacology Biochemistry and Behavior, 93*(3), 237–247.
Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.
Davidson, D. (1985). Deception and division. In E. Lepore & B. McLaughlin (Eds.), *Actions and events*. New York: Basil Blackwell.
Davidson, D. (1982). Paradoxes of irrationality. In R. Wollheim & J. Hopkins (Eds.), *Philosophical essays on freud*. Cambridge: Cambridge University Press.
Dawson, E., Savitsky, K., & Dunning, D. (2006). "Don't tell me, I don't want to know": Understanding people's reluctance to obtain medical diagnostic information. *Journal of Applied Social Psychology, 36*(3), 751–768.
de Sousa, R. (1978). Self-deceptive emotions. *Journal of Philosophy, 75*, 684–697.
Delgado, M., Miller, M., Inati, S., & Phelps, E. (2005). An fMRI study of reward-related probability learning. *NeuroImage, 24*(3), 862–873.
Dill, B., & Holton, R. (2014). The addict in us all. *Frontiers in Psychiatry, 5*.
Dunning, D. (1995). Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives. *Personality and Social Psychology Bulletin, 21*(12), 1297–1306.
Erez, A., Johnson, D. E., & Judge, T. A. (1995). Self-deception as a mediator of the relationship between dispositions and subjective well-being. *Personality and Individual Differences, 19*(5), 597–612.
Eysenck, M. W., Mogg, K., May, J., Richards, A., & Al, E. (1991). Bias in interpretation of ambiguous sentences related to threat in anxiety. *Journal of Abnormal Psychology, 100*(2), 144–150.
Ferrari, J. R., Groh, D. R., Rulka, G., Jason, L. A., & Davis, M. I. (2008). Coming to terms with reality: Predictors of self-deception within substance abuse recovery. *Addictive Disorders & Their Treatment, 7*(4), 210–218.
Freud, S. (1955/1920). Beyond the pleasure principle. In J. Strachey (Ed.), *The standard edition of the complete psychological works of sigmund freud, trans.* London: Hogarth Press.
Frijda, N. H. (1987). *The emotions*. Cambridge: Cambridge University Press.
Gendler, T. S. (2007). Self-deception as pretense. *Philosophical Perspectives, 21*(1), 231–258.
Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. New York: Viking.
Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20–29.
Gilbert, D. T., Morewedge, C. K., Risen, J. L., & Wilson, T. D. (2004). Looking forward to looking backward. The misprediction of regret. *Psychological Science, 15*(5), 346–350.
Gilbert, D. T., & Wilson, T. D. (2009). Why the brain talks to itself: Sources of error in emotional prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1335–1341.
Girotto, V., Ferrante, D., Pighin, S., & Gonzalez, M. (2007). Postdecisional counterfactual thinking by actors and readers. *Psychological Science, 18*(6), 510–515.
Goldbeck, R. (1997). Denial in physical illness. *Journal of Psychosomatic Research, 43*(6), 575–593.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.
Heatherton, T. F., & Wagner, D. D. (2011). Cognitive neuroscience of self-regulation failure. *Trends in Cognitive Sciences, 15*(3), 132–139.
Hippel, W. V., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34*(01), 1–16.
Huang, Y., & Rao, R. P. N. (2013). Reward optimization in the primate brain: A probabilistic model of decision making under uncertainty. *PLoS ONE, 8*(1).
Isenman, L. (2013). Understanding unconscious intelligence and intuition: "*Blink*" and beyond. *Perspectives in Biology and Medicine, 56*(1), 148–166.
Johnston, M. (1988). Self-deception and the nature of mind. In B. McLaughlin & A. O. Rorty (Eds.), *Perspectives on self-deception*. Berkeley: University of California Press.
Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., ... Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology, 2*(4), e97.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498.
Lajunen, T., Hakkarainen, P., & Summala, H. (1996). The ergonomics of road signs: Explicit and embedded speed limits. *Ergonomics, 39*(8), 1069–1083.
Lazarus, R. S. (1966). *Psychological stress and the coping process*. New York: McGraw-Hill.

Leventhal, H., & Scherer, K. (1987). The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition & Emotion, 1*(1), 3–28.
Linnet, J. (2014). Neurobiological underpinnings of reward anticipation and outcome evaluation in gambling disorder. *Frontiers in Behavioral Neuroscience, 8*.
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.
Macleod, A. K., & Byrne, A. (1996). Anxiety, depression, and the anticipation of future positive and negative experiences. *Journal of Abnormal Psychology, 105*(2), 286–289.
Mathews, A., & Macleod, C. (1994). Cognitive approaches to emotion and emotional disorders. *Annual Review of Psychology, 45*(1), 25–50.
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.
Meiland, J. W. (1980). What ought we to believe? Or the ethics of belief revisited. *American Philosophical Quarterly*, 15–24.
Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences., 20*(01).
Mele, A. R. (2001). *Self-deception unmasked*. Princeton, NJ: Princeton University Press.
Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology, 49*(1), 377–412.
Mogg, K., & Bradley, B. P. (1998). A cognitive-motivational analysis of anxiety. *Behaviour Research and Therapy, 36*(9), 809–848.
Naqvi, N. H., & Bechara, A. (2009). The hidden island of addiction: The insula. *Trends in Neurosciences, 32*(1), 56–67.
Norem, J. K. (2002). Defensive self-deception and social adaptation among optimists. *Journal of Research in Personality, 36*(6), 549–555.
Nygren, T. E., Isen, A. M., Taylor, P. J., & Dulin, J. (1996). The influence of positive affect on the decision rule in risk situations: Focus on outcome (and especially avoidance of loss) rather than probability. *Organizational Behavior and Human Decision Processes, 66*(1), 59–72.
Os, J. V., & Kapur, S. (2009). Schizophrenia. *The Lancet, 374*(9690), 635–645.
Partyka, J. C. (2011). *Change blindness: Predictors and effects of distraction*. Pensacola, FL: University of West Florida.
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.
Pears, D. (1984). *Motivated irrationality*. Oxford: Clarendon Press.
Peterson, J. B., Deyoung, C. G., Driver-Linn, E., Séguin, J. R., Higgins, D. M., Arseneault, L., & Tremblay, R. E. (2003). Self-deception and failure to modulate responses despite accruing evidence of error. *Journal of Research in Personality, 37*(3), 205–223.
Peterson, J. B., Driver-Linn, E., & Deyoung, C. G. (2002). Self-deception and impaired categorization of anomaly. *Personality and Individual Differences, 33*(2), 327–340.
Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron, 51*(3), 381–390.
Price, H. H. (1954). The inaugural address: Belief and will. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 1–26.
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.
Rorty, A. O. (1972). Belief and self-deception. *Inquiry, 15*(1–4), 387–410.
Rorty, A. O. (1994). User-friendly self-deception. *Philosophy, 69*(268), 211.
Sackeim, H. A., & Gur, R. C. (1978). Self-deception, self-confrontation, and consciousness. In *Consciousness and self-regulation* (pp. 139–197). US: Springer.
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. (1985). Voice recognition and the ontological status of self-deception. *Journal of Personality and Social Psychology, 48*(5), 1365–1368.
Sackeim, H. A., & Gur, R. C. (1997). Flavors of self-deception: Ontology and epidemiology. *Behavioral and Brain Sciences, 20*(1), 125–126.
Sagna, A., Gallo, J. J., & Pontone, G. M. (2014). Systematic review of factors associated with depression and anxiety disorders among older adults with Parkinson's disease. *Parkinsonism & Related Disorders, 20*(7), 708–715.
Sahdra, B., & Thagard, P. (2003). Self-deception and emotional coherence. *Minds and Machines, 13*, 213–231.
Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford, NY: Oxford University Press.
Schultz, W. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.
Schultz, W., Tremblay, L., & Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology, 37*(4–5), 421–429.
Scott-Kakures, D. (2000). Motivated believing: Wishful and unwelcome. *Nous, 34*(3), 348–375.
Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R., & Dolan, R. J. (2012). How dopamine enhances an optimism bias in humans. *Current Biology, 22*(16), 1477–1481.
Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*(4), 644–649.
Sloman, S. A., Fernbach, P. M., & Hagmayer, Y. (2010). Self-deception requires vagueness. *Cognition, 115*(2), 268–281.
Sperber, D. (2000). *Metarepresentations: A multidisciplinary perspective*. Oxford: Oxford University Press.
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359–393.
Strom, J., & Barone, D. F. (1993). Self-deception, self-esteem, and control over drinking at different stages of alcohol involvement. *Journal of Drug Issues, 23*. 705–705.
Surbey, M. K. (2011). Adaptive significance of low levels of self-deception and cooperation in depression. *Evolution and Human Behavior, 32*(1), 29–40.
Thagard, P. (2007). The moral psychology of conflicts of interest: Insights from affective neuroscience. *Journal of Applied Philosophy, 24*(4), 367–380.
Uziel, L. (2013). Impression management ("lie") scales are associated with interpersonally oriented self-control, not other-deception. *Journal of Personality, 82*(3), 200–212.
Van Leeuwen, D. S. N. (2007). The spandrels of self-deception: Prospects for a biological theory of a mental phenomenon. *Philosophical Psychology, 20*(3), 329–348.
Van Leeuwen, D. S. N. (2009). Self-deception won't make you happy. *Social Theory and Practice, 35*(1), 107–132.
Verdejo-García, A., & Bechara, A. (2009). A somatic marker theory of addiction. *Neuropharmacology, 56*, 48–62.
Volkow, N., Fowler, J., Wang, G., Baler, R., & Telang, F. (2009). Imaging dopamine's role in drug abuse and addiction. *Neuropharmacology, 56*, 3–8.
Walker, M. J. (2010). Addiction and self-deception: A method for self-control? *Journal of Applied Philosophy*.
Westen, D., Blagov, P. S., Harenski, K., Kilts, C., & Hamann, S. (2006). Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 US presidential election. *Journal of Cognitive Neuroscience, 18*(11), 1947–1958.
Wright, W. F., & Bower, G. H. (1992). Mood effects on subjective probability assessment. *Organizational Behavior and Human Decision Processes, 52*(2), 276–291.