Les ateliers de l'éthique
The Ethics Forum

Les ateliers de l'éthique
The Ethics Forum

WHAT DOES EMOTION TEACH US ABOUT SELF-DECEPTION? AFFECTIVE NEUROSCIENCE IN SUPPORT OF NON-INTENTIONALISM

Federico Lauria et Delphine Preissmann
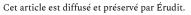
Citer cet article

Résumé de l'article

Intuitivement, l'affect joue un rôle indispensable dans la dynamique de l'autoduperie. Appelons cette conception « l'affectivisme ». Il importe d'examiner l'affectivisme, étant donné que les affectivistes soutiennent que cette conception favorise une approche non-intentionnaliste de l'auto-duperie et fournit une conception unifiée des formes classique et inversée de l'auto-illusion. Or, ces arguments n'ont pas fait l'objet d'une étude détaillée. L'affectivisme remplit-il ses promesses quant au non-intentionnalisme et à l'unité explicative ? Cet article propose une nouvelle conception qui rend justice aux aspirations affectivistes. Selon notre théorie, la duperie de soi résulte de filtres affectifs, à savoir de l'évaluation de l'information à la lumière de nos buts ou préoccupations (la conception des filtres affectifs). Nous développons cette conception en portant une attention particulière aux mécanismes neurobiologiques sous-jacents à la duperie de soi, à savoir les marqueurs somatiques et la régulation dopaminergique. Décrire le phénomène à ce niveau permet de justifier la conception nonintentionnelle et d'offrir un modèle unifié de l'auto-duperie. Nous motivons cette approche en critiquant les principales théories affectivistes, à savoir l'idée que la duperie de soi aurait pour fonction de réduire l'anxiété ou serait motivée par l'anxiété. Les mécanismes affectifs éclairent la dynamique de la duperie de soi sans faire appel aux intentions, comme de nombreuses études empiriques sur les biais affectifs le démontrent. L'affectivisme tient donc ses promesses.

WHAT DOES EMOTION TEACH US ABOUT SELF-DECEPTION?

AFFECTIVE NEUROSCIENCE IN SUPPORT OF NON-INTENTIONALISM

FEDERICO LAURIA
CENTER FOR SCIENCE AND SOCIETY, COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK/SWISS CENTER FOR AFFECTIVE SCIENCES, UNIVERSITY OF GENEVA

DELPHINE PREISSMANN
CENTER FOR PSYCHIATRIC NEUROSCIENCE, DEPARTMENT OF PSYCHIATRY, LAUSANNE UNIVERSITY HOSPITAL, PRILLY, SWITZERLAND

ABSTRACT:

Intuitively, affect plays an indispensable role in self-deception's dynamic. Call this view "affectivism." Investigating affectivism matters, as affectivists argue that this conception favours the non-intentionalist approach to self-deception and offers a unified account of straight and twisted self-deception. However, this line of argument has not been scrutinized in detail, and there are reasons to doubt it. Does affectivism fulfill its promises of non-intentionalism and unity? We argue that it does, as long as affect's role in self-deception lies in affective filters—that is, in evaluation of information in light of one's concerns (the affective-filter view). We develop this conception by taking into consideration the underlying mechanisms governing self-deception, particularly the neurobiological mechanisms of somatic markers and dopamine regulation. Shifting the discussion to this level can fulfill the affectivist aspirations, as this approach clearly favours non-intentionalism and offers a unified account of self-deception. We support this claim by criticizing the main alternative affectivist account—namely, the views that self-deception functions to reduce anxiety or is motivated by anxiety. Describing self-deception's dynamic does not require intention; affect is sufficient if we use the insights of neuroscience and the psychology of affective bias to examine this issue. In this way, affectivism can fulfill its promises

RÉSUMÉ :

Intuitivement, l'affect joue un rôle indispensable dans la dynamique de l'autoduperie. Appelons cette conception « l'affectivisme ». Il importe d'examiner l'affectivisme, étant donné que les affectivistes soutiennent que cette conception favorise une approche non-intentionnaliste de l'auto-duperie et fournit une conception unifiée des formes classique et inversée de l'auto-illusion. Or, ces arguments n'ont pas fait l'objet d'une étude détaillée. L'affectivisme remplit-il ses promesses quant au non-intentionnalisme et à l'unité explicative ? Cet article propose une nouvelle conception qui rend justice aux aspirations affectivistes. Selon notre théorie, la duperie de soi résulte de filtres affectifs, à savoir de l'évaluation de l'information à la lumière de nos buts ou préoccupations (la conception des filtres affectifs). Nous développons cette conception en portant une attention particulière aux mécanismes neurobiologiques sous-jacents à la duperie de soi, à savoir les marqueurs somatiques et la régulation dopaminergique. Décrire le phénomène à ce niveau permet de justifier la conception nonintentionnelle et d'offrir un modèle unifié de l'auto-duperie. Nous motivons cette approche en critiquant les principales théories affectivistes, à savoir l'idée que la duperie de soi aurait pour fonction de réduire l'anxiété ou serait motivée par l'anxiété. Les mécanismes affectifs éclairent la dynamique de la duperie de soi sans faire appel aux intentions, comme de nombreuses études empiriques sur les biais affectifs le démontrent. L'affectivisme tient donc ses promesses.

Stevens has dedicated his life to rendering loyal service to Darlington Hall. He is obsessed with dignity. He believes that a perfect butler must be exclusively devoted to his profession, and he has lived his life accordingly. Confronted with rumours of Lord Darlington's Nazi sympathies, Stevens dismissed them as nonsense. He was utterly convinced of Lord Darlington's honesty. Years earlier, Stevens started to develop romantic feelings for the housekeeper Miss Kenton, and the feelings were mutual. Still, Stevens believed that their relationship was strictly professional, as it should be for a perfect butler. Subsequent to Miss Kenton's marriage to another man, Stevens ventures on a trip, with ample time to reflect. One day, he realizes that he has always loved Miss Kenton. He then fathoms that Lord Darlington is corrupt. This fills Stevens with regret; his whole life's purpose has been based on an illusion. Time has come to focus on what is left of his life.

So runs the plot of Ishiguro's novel *The Remains of the Day,* a story that dramatizes self-deception. For decades, Stevens's beliefs have been biased by his desire to be a perfect butler and have not been formed in light of actual evidence. Acknowledging his true feelings for Miss Kenton or his master's dishonesty would have devastated Stevens, as this would have been in stark conflict with his desire to live as a perfect butler. Stevens thus formed beliefs that appeased him and that aligned with his desire to be a perfect butler. The irony of the story and its dramatic character lie in the pernicious effects of self-deception and of its consolations: Stevens has wasted his life.

Intuitively, Stevens's tragedy can be understood, at least partly, in affective terms; he deceived himself to avoid distress. The prospect of pleasure is the crux of self-deception (Johnston, 1988; Barnes, 1997). At least, it is intuitive to think that Stevens's anxiety eased him into deceiving himself (Galeotti, 2016). Call "affectivism" the view that emotion or affect plays an indispensable role in self-deception's dynamic.

Affectivism offers a new conception of self-deception's dynamic, alongside the two main accounts: intentionalism and deflationism. A brief summary of each account will allow us to understand affectivism's relevance. Intentionalists claim that self-deceived subjects *intend*, albeit unconsciously, to form the deceptive beliefs (Davidson, 1982, 1985; Bermúdez, 1997, 2000). After all, self-deception seems to be analogous to interpersonal deception, which is intentional. By contrast, non-intentionalists deny that self-deception necessarily involves an intention to form the deceptive belief (Bach, 1981; Mele, 1997). Proponents of deflationism claim that deceptive beliefs are biased by desire *tout court* (Mele, 1997, 2001) like other biases, self-deception need not be intentional. Affectivism diverges from these accounts. Against intentionalists, affectivists argue that self-deception need not be intentional; in contrast with deflationists, they claim that *emotion* or *affect* also features in self-deception's dynamic and plays a role that is irreducible to that of desire.[1]

Let us assume, for the sake of argument, that emotions play an indispensable role in self-deception's dynamic. What would this teach us about self-deception's dynamic? This article tackles this question by examining affectivism through the lenses of two heated debates on self-deception. First, and this touches on the most vivid controversy concerning self-deception, affectivists claim that their view justifies non-intentionalism (Barnes, 1997; Lazar, 1999; Galeotti, 2016; Echano, 2017). This is the first promise of affectivism. Second, affectivists claim that their view illuminates the more recent puzzle of self-deception's unity. While straight self-deception results in a belief that squares with what one wants to be true (as in Stevens's case), twisted self-deception yields the belief in what one does *not* want to be true (Mele, 2003; Nelkin, 2002). For example, despite ample evidence to the contrary, Othello's anxiety leads him to believe that Desdemona is unfaithful, because he desperately wants her to be faithful. Straight and twisted self-deception result in irrational beliefs that are motivated by desire rather than founded on evidence. Thus, carving self-deception at the joints calls for an account that covers both straight and twisted cases, and affectivists claim that their view offers such an account (Lazar, 1999; Galeotti, 2016; Echano, 2017). Affectivism thus promises non-intentionalism and unity. Does it keep these promises? Scrutinizing affectivism's relevance to these two issues is important, as they are at the very core of self-deception's dynamic and invite us to capture the very route(s) of self-deception.[2]

There has been a recent surge of interest in the affective dimension of self-deception (Johnston, 1988; de Sousa, 1988; Barnes, 1997; Lazar, 1999; Sahdra and Thagard, 2003; Bayne and Fernandez, 2009; Correia, 2014; Galeotti, 2016; Echano, 2017). However, philosophers have paid little attention to the empirical literature on the subject. Now, these studies offer insights into self-deception's dynamic and the affectivist promises mentioned. To fill this lacuna, we propose a new affectivist approach—the "affective-filter view"—that illuminates affect's role in self-deception by describing the underlying mechanisms governing self-deception. We claim that affect's role in self-deception lies in affective filters of information—that is, in evaluation of information in light of our concerns. We develop this conception by integrating findings drawn from affective neuroscience, particularly on the mechanisms of somatic markers and dopamine regulation. We argue that describing the phenomenon at this neurobiological level fulfills the affectivist aspirations; this conception clearly favours non-intentionalism and offers an elegant, unified account of self-deception. It is time to leave the armchair and substantiate the thought that self-deception is "belief under influence."

The article is divided in seven sections. As a preliminary, section 1 clarifies the affectivist agenda. We then examine the main affectivist accounts, starting with the promise of unity: section 2 scrutinizes the claim that self-deception functions to reduce anxiety, while section 3 criticizes the claim that self-deception is motivated by anxiety. In section 4, we examine these accounts in light of the promise of non-intentionalism. As this discussion suggests refining the mechanisms involved in self-deception, we then present our affective-filter view, which

hinges on such mechanisms (§ 5), before showing how it fulfills the promises of non-intentionalism (§ 6) and unity (§ 7).

1. THE AFFECTIVIST AGENDA

Let us first consider the affectivist argument for non-intentionalism, as this sets the stage for a careful defence of the affectivist research program. The standard argument appeals to the influence of affect on belief (Kunda, 1999). We tend to form optimistic beliefs when we are happy and pessimistic beliefs when we are gloomy. Likewise, emotion biases belief. Beset by a burst of anger, Mary believes that Sam is unworthy of her affection; after her rage has vanished, she recognizes that her judgment was biased by emotion. Now—and this is the crux of the argument—affect typically biases belief in an *unintentional* manner. Given that affect biases deceptive beliefs, it follows that self-deception need not be intentional (Lazar, 1999; Correia, 2014).

Although this is a compelling argument, intentionalists will hardly be impressed by it. The argument rests on the assumption that self-deception operates analogously to unintentional affective biases. However, intentionalists dispute this assumption. They may grant that affect (e.g., moods) can bias belief in an unintentional manner. They even concede that motivated cognition can be unintentional, since wishful thinking is unintentional in their view (Bermúdez, 2000). That said, they think that self-deception differs from unintentional affective biases and operates analogously to *intentional* affective biases. For an example of the latter, consider the positivity effect: with age, people tend to focus on rewarding activities and to feel more positive emotions, which results in biased beliefs. This bias can be explained by top-down effect and intentional reappraisals (Reed and Carstensen, 2012). Consequently, a question arises: Why should we regard self-deception as analogous to unintentional bias, rather than to intentional bias? In the absence of an answer to this question, the affectivist argument begs the question. After all, intentionalists have never disputed emotion's role in self-deception, as emotions motivate the intention to form the deceived belief. Thus, the affective dynamic of self-deception does not undermine their claim.

To substantiate this line of skepticism, intentionalists may reiterate one of their main objections to non-intentionalism, the so-called selectivity problem. Consider Talbott's (1995, p. 60-61) seminal scenario:

> *Anxious Driving* – While driving his car, Bill notices that the brake pedal is not as firm as usual. He suspects that his car is not functioning properly. He feels anxious and stops to determine whether the car is functioning properly.

Bill desires his car to function properly. He is presented with sufficient evidence to the contrary. Still, he does not deceive himself. He feels anxious, and this motivates him to act. Why does Bill not deceive himself? Only in certain circum-

stances does desire lead to the formation of deceptive beliefs. The selectivity challenge consists in contrasting cases where desire results in self-deception with cases where it does not (the subject forms the rational belief). Now, intentionalists argue that deflationism cannot offer a satisfactory solution to this problem. The claim that desire biases belief is insufficient to distinguish between cases where desire results in deceptive beliefs and cases where it does not (see, however, Mele 2001). By contrast, intentionalists claim to have a ready answer: self-deception occurs only when the subject *intends* to form the deceptive belief (Bermúdez, 2017, 2000). In our example, Bill does not deceive himself, because he lacks the intention to form the deceptive belief.

Importantly, the objection is not simply that deflationism cannot adequately predict self-deception. Such a challenge would be intractable and largely an empirical issue (Mele, personal communication). To demonstrate why the selectivity problem differs from the issue of predicting self-deception, consider interpersonal deception. Prima facie, interpersonal deception involves the intention to deceive. This offers one way of drawing the line between cases where deception occurs and cases where it does not: deception occurs only when the subject intends to deceive. This, however, does not predict deception, as it does not specify when a subject will form the relevant intention. The selectivity problem thereby differs from concerns about prediction.

Let us assume that the selectivity problem is a legitimate objection to deflationism. A promising non-intentionalist account should be able to rebut it. Whether affectivism supports non-intentionalism thus depends on whether it can solve the selectivity problem. For argument's sake, we do not examine the intentionalist solution, nor do we consider alternative solutions to the problem (Pedrini, 2010; Jurjako, 2013); our only purpose is to refine the affectivist agenda. Our first desideratum is the following:

> *Selectivity*: Affectivism distinguishes the cases in which desires lead to deceptive beliefs from the cases in which it does not.

If we turn to the affectivist promise of unity, it appears that the spectre of intentionalism arises again. Intentionalists claim that the *intention* to form the deceptive belief unifies straight and twisted self-deception. Emotions, such as anxiety, could motivate such intention. Therefore, the influence of emotion does not undermine the intentionalist proposal; affectivists must provide further justification for their argument. For argument's sake, let us bracket any qualms about the soundness of this issue and set aside the intentionalist solution (see Lazar, 1999). We also ignore other potential solutions (Scott-Kakures, 2000; Nelkin, 2002), as discussing them is beyond the scope of this paper. Our second desideratum focuses on affectivism's merits on its own terms.

> *Unity*: Affectivism offers a unified account of straight/twisted self-deception.

The agenda for affectivism is thus set.

To guide our investigation, let us assume that self-deception is a process that results in deceptive beliefs. The role of affect may come into play at different phases of the process. Affect may feature in the output of the process, as in the claim that self-deception aims at pleasure (§ 2). Alternatively, affect could initiate the process, as in the idea that anxiety motivates self-deception (§ 3). Finally, affect could mediate desire's influence on belief and thereby play a role at the level of evaluating evidence (§ 5). These possibilities are distinct yet compatible with one another. Let us start by examining the main account that situates affect's role in the output.

2. THE HEDONIC DYNAMIC OF SELF-DECEPTION: UNITY

Intuitively, we deceive ourselves to avoid distress; the dynamics of self-deception are inherently hedonic. According to the main variant of this idea, self-deception's function is to reduce anxiety. For example, Stevens's belief's in his master's innocence alleviates his anxiety. To wit, the deceptive belief that p reduces anxiety about the nonsatisfaction of the desire that p (Johnston, 1988). Prima facie, this proposal fares well with straight self-deception.[3] However, it is hardly generalizable to twisted cases. For instance, Othello's belief in Desdemona's infidelity fails to reduce his anxiety about the matter; rather, it increases or, at least, sustains it.

In response to this difficulty, Barnes (1997) argues that self-deception functions to reduce some anxiety, where the anxiety may or may not correspond to the matter of the deceptive belief. Consider her example (Barnes, 1997, p. 41):

> *George's Regard* – John desires Mary's faithfulness. Out of anxiety, he believes that Mary is having an affair with George. Now, John badly desires that George have high regard for him, and he is very anxious about this. George has declined John's requests many times, but has always agreed to help Mary. John would be devastated if George had a higher regard for Mary; it would be a source of acute anxiety. By contrast, the belief that George and Mary are having an affair reduces John's anxiety about George's regard, because it is compatible with believing that George has equal regard for John. Hence, John deceives himself into believing that Mary is unfaithful.

This suggests that there is a perceived hedonic gain in twisted self-deception as well. The deceptive belief that p (Mary is having an affair with George) reduces anxiety about some other matter q (George has a higher regard for Mary) because the subject believes that, if p, then not q (Barnes 1997, p. 36). This is how Barnes captures self-deception's unity.

Let us raise two difficulties regarding the claim that, in twisted self-deception, the belief that p reduces anxiety about some other matter q.

First, we do not dispute that twisted self-deception may reduce anxiety, as in "George's Regard." However, it is doubtful that this proposal is generalizable (Echano, 2017), as this example suggests. Sally is anxious that Penelope has cancer. A sense of panic prompts Sally to believe that Penelope has cancer. Intuitively, Sally's belief is motivated by her anxiety about *this* matter. What other anxiety might the deceptive belief alleviate? This intuition is corroborated by empirical studies on the biases involved in anxiety (henceforth called "anxiety biases"), which correspond to, or partly overlap with, twisted self-deception. Anxious people detect threats more efficiently than controls do. The bias operates at the levels of (pre)attention and the interpretation of evidence (Cisler and Koster, 2010; Mogg and Bradley, 2016). Far from reducing anxiety, such a bias often leads to a state of generalized anxiety. It is therefore questionable to conceive of twisted self-deception as reducing anxiety.

Second, even if twisted self-deception results in anxiety reduction as proposed, this proposal fails to do justice to the specificity of twisted self-deception. On this proposal, twisted self-deception is modeled on, and somehow reduced to, straight self-deception. The deceptive belief reduces anxiety because subjects end up believing what they most desire to obtain. John believes that Mary is unfaithful to retain his belief about what he desires most—namely, George's regard. The anxiety reduction that occurs in twisted self-deception ultimately results from straight self-deception. Twisted self-deception is straight self-deception in disguise. However, it is unlikely or, at least, questionable that twisted self-deception is reducible to straight self-deception. One may capture the unity of self-deception at a more general level without reducing twisted self-deception to straight self-deception. One way to do so is to outline that both forms of self-deception involve similar mechanisms, which, however, operate in opposite manners. Consider optimism and pessimism as an analogy. It is intuitive to understand both phenomena through similar components, albeit ones that operate in opposite ways. By contrast, it would be counterintuitive to capture the unity of both phenomena by reducing pessimism to optimism. Given the partial overlap between optimism and straight self-deception as well as the close connection between pessimism and twisted self-deception, a nonreductive approach to twisted self-deception is an intuitive option. An account that captures the specificity of twisted self-deception in its own terms would thus have the upper hand.

Let us consider another variant of the proposal that does not suffer from the difficulties just raised, by elaborating on Sally's example.

> *Hypervigilant Sally* – Out of anxiety, Sally deceives herself into believing that Penelope has cancer. This motivates her to act to avoid the undesired state (she consults doctors, asks for a second opinion, etc.). It turns out that Penelope has appendicitis. What a relief!

On this variant, the deceptive belief alleviates anxiety by motivating the subject to reduce anxiety by acting (Barnes, 1997, p. 45). Whereas straight self-decep-

tion reduces anxiety at the time of the belief, twisted self-deception reduces anxiety in the future. On this proposal, twisted self-deception involves high anxiety concerning the matter of the deceptive belief, which squares with empirical studies. That being said, as a kind of hypervigilance and "bitter medicine" (Pears, 1986, p. 42-43), twisted self-deception reduces anxiety through its impact on action—that is, in a twisted manner.

However, does this proposal justify the claim that twisted self-deception functions to reduce anxiety? In fact, this proposal is consistent with a conception of self-deception as functioning to *sustain* or *increase* anxiety so as to ensure protection from threats.[4] On this interpretation, anxiety reduction would be a byproduct of twisted self-deception, but not its function. After all, the specificity of twisted self-deception consists in its mode of reducing anxiety: if anything, it reduces anxiety by sustaining it, as opposed to other ways of reducing anxiety, such as by forming the rational belief. It is thereby plausible to regard twisted self-deception as functioning to sustain anxiety. After all, the function of anxiety is arguably not to reduce anxiety, but rather to recognize threats and protect oneself through action. If twisted self-deception recruits anxiety's function, it is natural to think that it aims at vigilance and protection, rather than at anxiety reduction. Of course, there might be no way of determining whether anxiety reduction is the function or a mere byproduct of twisted self-deception. However, given that this reading of Barnes's proposal is compatible with a conception of twisted self-deception as functioning to sustain anxiety or protect oneself, it does not imply that twisted self-deception functions to reduce anxiety. Therefore, it is controversial whether anxiety reduction captures self-deception's unity. Strictly speaking, the dynamics of twisted self-deception may be anxious rather than hedonic, which suggests that we consider the second main affectivist account.

3. THE ANXIOUS DYNAMIC OF SELF-DECEPTION: UNITY

One natural suggestion is simply that anxiety motivates self-deception. This claim is neutral regarding self-deception's function and output. It situates anxiety's role at the input (Barnes, 1997) or in the mediation of the process. That anxiety drives self-deception is straightforward in twisted cases. As for straight self-deception, anxiety's role appears more clearly at the level of the treatment of evidence. Straight self-deception involves being presented with sufficient evidence that one's desire is doomed to frustration; one is presented with a threat to the satisfaction of a desire. Now, anxiety and, more generally, fear are dedicated to recognizing threats. When Melania is afraid of a bird flying in her direction, she experiences the situation as threatening (Tappolet, 2000); the same applies to anxiety, despite some differences. As straight self-deception is formed in the face of a threat, it thereby involves anxiety. This idea is thus compatible with the possibility that anxiety coincides with the initiation of the process, without anxiety being present beforehand. Stevens becomes anxious only when presented with threatening evidence. Consequently, that people may deceive themselves about matters that they were not anxious about beforehand does not undermine anxiety's role of motivating self-deception.

Still, does desire not bias deceptive beliefs so subtly that the threatening evidence is immediately reinterpreted in a reassuring way and anxiety does not arise (Mele, 2003)? This may prevent conscious anxiety from arising, but it is compatible with straight self-deception involving *unconscious* anxiety. Reinterpreting threatening evidence requires having identified it; this is precisely anxiety's role, and anxiety may play this role even if it is unconscious. This bears on the controversial issue of unconscious emotions. For argument's sake, let us grant that unconscious anxiety may play a role in self-deception, as we assume that affectivism is true. For our purposes, let us explain how appealing to anxiety's role of motivating self-deception seems to have the resources to capture its unity.

Galeotti (2016) argues that the unity of self-deception revolves around anxiety's role. In straight self-deception, the subject desires that *p*, and negatively appraises the evidence threatening *p*. This appraisal generates anxiety. In twisted cases, the subject desires that *p*, and irrationally appraises evidence as favouring not-*p* (in Galeotti's terms, the subject "misappraises" evidence). This also generates anxiety. In both cases, anxiety's role is situated at the level of the treatment of evidence. The next condition for self-deception consists in the subject's assessment of the costs of error (Friedrich, 1993; Klayman and Ha, 1987). In self-deception, subjects assess the costs of forming the deceptive belief as low, which explains why they form the belief. For instance, Stevens believes that his master is innocent, because he assesses that this belief affords immediate relief, while the opposite belief would cause him significant distress and thereby prove costly. Similarly, in twisted self-deception, Othello assesses the belief in Desdemona's fidelity as costly (for instance, it would result in his failure to take steps to remedy the situation, for instance by ensuring that Desdemona will be faithful in the future). Hence, he deceives himself and believes in Desdemona's infidelity. Self-deception's unity can be captured by the presence of anxiety, followed by the assessment of the costs of error (Galeotti, 2016, p. 96).

This account does justice to anxiety's role in self-deception without suffering from the pitfalls of the output approach. However, it leaves one matter unexplained. When does anxiety lead to straight, as opposed to twisted, self-deception? A promising account should capture the unity of self-deception, as well as the distinctive dynamics of straight and twisted self-deception. Now, the extent to which this proposal captures such a distinction is unclear, as anxiety can bias belief in each direction. Although the difference between straight and twisted self-deception could be captured by the influence of anxiety on the assessment of the costs of error, the question remains: When does anxiety influence the costs of error in one way as opposed to the other? Far from a fatal objection, this observation invites us to probe the mechanism by which anxiety leads to straight self-deception or twisted self-deception.

To be fair, Galeotti (2016, p. 96-97) does address this concern. She claims that straight and twisted self-deception involve different mechanisms; straight self-deception relies on confirmation bias, whereas probability neglect (considering the worst-case scenario) is responsible for twisted self-deception. However, this

does not offer a clear-cut contrast. One may equally conceive of twisted self-deception as involving confirmation bias due to anxiety. Alternatively, in straight self-deception, subjects might be described as displaying probability neglect, as they overlook the evidence supporting the most dreaded scenario. Can the affective dynamic of self-deception offer a unified account that captures the distinctive routes of self-deception?[5]

Let us take stock. The two main affectivist accounts fail to adequately capture the unity of self-deception. Two avenues suggest themselves. Intentionalists secure the unity (and diversity) of self-deception by invoking intentions to form the deceptive belief. Alternatively, we propose to refine the affective dynamic of self-deception and secure the affectivist aspirations by shifting the discussion to the neurobiological level. This same moral emerges from examining how the main affectivist accounts fare with regard to selectivity. Let us now turn to this issue.

4. ANXIETY AND SELECTIVITY

How do the hedonic or anxious dynamics of self-deception solve the selectivity problem? Stevens's anxiety explains why he deceived himself. Yet, it could also have led him to believe the exact opposite (that his master is dishonest), as it does at the end of the story. Likewise, Sally's anxiety explains her deceptive belief. Still, a rational person would not deceive herself in similar circumstances. So, when does anxiety lead to self-deception?

The anxiety-reduction account offers a principled answer to the problem. If the function of self-deception is anxiety reduction, it follows that self-deception would occur only when the deceptive belief is likely (or expected) to result in anxiety reduction. Without the prospect of hedonic gain, self-deception does not occur. In the "Anxious Driving" example, this idea provides a clear explanation of Bill's failure to self-deceive. Believing that his car is functioning well would not have reduced anxiety; it would, instead, have increased anxiety, as Bill would not have taken the necessary precautions to avoid an accident. A similar solution is at the heart of Galeotti's (2016) appeal to the costs of error. As observed, people do not deceive themselves when they assess the costs of error as high. Hence, Bill does not deceive himself, because he assesses the costs of error as high, notably because he thinks that he can act to remedy the situation.[6] Self-deception occurs only when people assess the situation as beyond their control (Galeotti, 2016; more on this in § 4).

This solution, in terms of (hedonic) costs of belief, is intuitive. However, it does not apply to what we call the "hard cases" for selectivity. In such cases, subjects assess the (hedonic) costs of error as low (notably for lack of control over the situation), but do not deceive themselves. Here is such a case, which is inspired by Bermúdez's (2000) observations, with some differences that are irrelevant for our purposes.

> *Guilty Son* – Don has been accused of treason; the evidence is ambiguous, but suggests that he is guilty. Don's parents, Mark and Juliet, desire

their son's innocence and are anxious about their son being guilty. Juliet believes that Don is innocent, and this thereby reduces her anxiety. By contrast, Mark does not deceive himself; he believes that his son is guilty, and this sustains his anxiety. He would prefer to believe the contrary, as this belief would appease him. However, the evidence speaks for itself.

This case reveals that the hedonic dynamic of self-deception fails to solve the selectivity problem. The belief in Don's innocence would alleviate Juliet and Mark's anxiety equally, whereas the belief in Don's guilt would devastate them. Given that the prospect of hedonic gain is the same for Juliet and Mark, there should be no difference with regard to self-deception. However, they differ in this respect. Why does Mark not believe that Don is innocent, when this would clearly alleviate his anxiety?[7] Intentionalists have a ready answer: Mark does not intend to form the deceptive belief and thereby does not deceive himself. The objection also applies to the solution in terms of costs of error. Mark assesses the costs of believing that Don is innocent as low. Whether Don is innocent is beyond Mark's control, so self-deception would not come with the high costs associated with the failure to take precautionary measures. Nonetheless, Mark does not deceive himself. Why?

It is important to distinguish this case from variations of it that are compatible with the solution at hand. Consider that Mark believes that forming the deceptive belief would be dangerous (e.g., Don might fool him in the future) or imagine that Mark thinks that he can act to improve the situation. These scenarios would elevate the costs of error and explain his failure to self-deceive. The problematic case is different. Mark and Juliet desire Don's innocence equally, and there are no further desires involved. Both are convinced that Don will not fool them and that they cannot remedy the situation. They concur that the deceptive belief would reduce their anxiety and that they have nothing to lose in deceiving themselves. However, Mark does not deceive himself. Why do people sometimes face an unwelcome reality? The main affectivist proposals cannot adequately solve the selectivity challenge. Rather than taking the intentionalist route, we can make progress by describing the underlying neural mechanisms governing the affective dynamics of self-deception.[8]

5. THE AFFECTIVE-FILTER VIEW

This section presents our conception of straight self-deception, which we then use to approach the issues of selectivity (§ 6) and of unity (§ 7). We claim that self-deception involves affective "filters" of information (Lauria, Preissmann and Clément, 2016). Let us start with a few clarifications.

The metaphor of filters of information points to the fact that people evaluate information. For instance, they assess the reliability of sources of information (Sperber at al. 2010). *Affective* filters consist in the evaluation of information in light of one's goals, such as pleasure or any other concern. In psychology, affec-

tive filters are the crux of the appraisal theory of emotion. On this view, emotions are elicited via a sequence of cognitive appraisals of the situation in light of one's goals (Lazarus, 1991; Scherer et al., 2001; Ellsworth, 2013). For instance, in fear, people typically appraise a situation as goal-obstructive (i.e., dangerous), as being in their control (i.e., escapable), etc. Our conception of self-deception relies on appraisals of this type.

Furthermore, we make significant use of neuroscientific findings on affective mechanisms involved in decision making and selective information processing. This mechanistic level of description is well suited to describing the very dynamic of self-deception, as it will appear.

As a consequence, our picture is a hybrid, integrating the psychological and the neurobiological levels of description into a philosophical view. Some components of our account spring from the armchair, while others refer to mechanisms studied in the empirical sciences. Our conception should thus be partly read as a conceptual truth (conditions [i]-[iv]) and partly read as an empirical claim (conditions [v]-[vii]). Let us now delve into the proposal.

Given that affective filters are assessments of information, our conception situates affect's role at the phase of the evaluation of evidence. More precisely, self-deception involves affective filters that take the form of four appraisals and two neurobiological mechanisms (the order is an expository one). In straight self-deception, a subject S desires that p, is presented with sufficient evidence favouring not p (henceforth "distressing evidence"), and forms the belief that p only if

(i)   assesses the distressing evidence as ambiguous (weight of evidence);

(ii) appraises the distressing evidence as having a significant negative impact on his or her well-being (affective coping);

(iii) appraises his or her control on the situation as low (coping potential); and

(iv) appraises the welcome situation p and the evidence for p as positive (affective coping).

Let us justify each condition. The first condition is the idea that self-deception precludes certainty about desire's frustration. Stevens would not deceive himself if he appraised the evidence as speaking unambiguously in favour of Lord Darlington's dishonesty. This would be more akin to delusion than self-deception. Of course, subjects might assess the evidence as ambiguous, even when the evidence clearly isn't ambiguous. This appraisal is epistemic rather than affective, yet it is importantly biased by affect (Lauria, Preissmann and Clément, 2016).

The first affective filter is spelled out in the second condition. As self-deceived subjects are presented with threatening evidence, self-deception involves a negative appraisal. Appraising a given situation as negative (e.g., as goal obtrusive, as unbearable) can arouse anxiety, sadness, or other negative emotions. In the

appraisal theory of emotion, a variety of specific appraisals are dedicated to this task (e.g., goal-conduciveness appraisal, affective-coping appraisal). They can operate unconsciously and may lead to conscious or unconscious instances of the emotions mentioned. We shall return to this momentarily.

The third condition concerns the idea that people appraise events in light of their own ability to act (coping-potential appraisal). For instance, sadness typically involves the appraisal that there is nothing one can do to remedy the situation. In self-deception, we appraise our coping potential as low; we appraise that we have little or no control over the distressing situation. Self-deceived subjects might appraise the situation as being in their control, yet reckon that acting on the situation would come at a critical cost. This explains why people do not deceive themselves when they think that they can act to neutralize the threat, as in the example "Anxious Driving." In such circumstances, it is natural to protect oneself by acting. After all, the matters about which people deceive themselves (personal relationships, health, intelligence, etc.) are typically matters that most would not appraise as being under their full control. Likewise, the populations especially prone to self-deception (e.g., addicts, terminal patients) concern conditions over which control is critically missing or believed to be absent (Martínez-González et al., 2016; Echarte et al., 2016). Finally, empirical studies suggest that people are less inclined to gather more information about a given disease when they consider the disease untreatable (Dawson et al., 2006); the best predictor of information gathering is the treatability (and not the severity) of the disease, as predicted by the third condition.

The fourth and final condition is the inverse of the second; it concerns the situation in which the desire is satisfied. Self-deceived subjects positively appraise this situation and the evidence that supports desire satisfaction. This takes the form of conscious or unconscious positive anticipation.

These conditions are necessary. They are justified conceptually and empirically (see Lauria, Preissmann and Clément 2016). However, they are insufficient, or, more to the point, this level of description does not adequately capture self-deception's dynamic. Consider the example "Guilty Son." Mark appraises the evidence in favour of Don's guilt as both ambiguous and devastating. He assesses his ability to remedy the situation as low. He positively appraises the situation in which Don is innocent. Nevertheless, he does not deceive himself. Our picture, so far, fails to explain how the positive appraisal takes precedence over the negative one; it fails to capture the dynamic relation between the appraisals. We therefore need an additional component or, at the least, some way of refining our account. This is where the neurobiological mechanisms enter the picture.

At the neurobiological level, straight self-deception involves the following conditions:

(v)   the appraisal of the distressing evidence is accompanied by negative somatic markers;

(vi)   the appraisal of the positive situation is accompanied by dopaminergic activity; and

(vii)   dopaminergic activity takes precedence over frontal activation and negative somatic markers in the processing of information.

The fifth condition correlates with the negative appraisal presented earlier (condition [iii]; for more on the relation, see below). Initially, somatic markers were intended to describe how people implicitly rely on affect when making decisions (Damasio, 1994). Negative affect automatically leads us to discard certain courses of action, by simulating the impact of options on well-being and by eliciting somatic states (e.g., hunches). This has been called "gut feeling unconscious intelligence" (Bechara, 1997; Gigerenzer, 2007, 2008). Broadly speaking, somatic markers refer to this mechanism and correspond to specific neural structures, particularly the ventromedial prefrontal cortex and the amygdala. For instance, patients with lesions in these regions suffer from emotional deficits that explain their inability to make optimal decisions. Likewise, addicts tend to ignore the negative signals of somatic markers in their decision making, which explains the persistence of the irrational behaviour. Similarly, experiments suggest that self-deceived people disregard the negative signals of somatic markers, unlike rational subjects (Peterson et al., 2002, 2003). This is corroborated by studies revealing that the neural structures that correspond to somatic markers are involved in self-deception (Westen et al., 2006). Somatic markers can account for the inhibition of the treatment of the distressing evidence in straight self-deception because their role is to discard further processing of negative information, as studies on decision making show.

Conversely, the mechanism of dopamine regulation accounts for the preferred treatment of positive information. Dopamine is the neurotransmitter of desire. It encodes reward anticipation and prediction errors, especially in the proximal future (Schultz, 1997; Schultz et al., 1998). It is heavily released in uncertainty and it modulates attention to cues that are relevant to desire's satisfaction. Dopaminergic deficits correlate with apathy, depression, and anxiety, as revealed in Parkinson's disease. Importantly, self-control relies on the balance between dopaminergic transmission and prefrontal-cortex activation. For instance, addiction is characterized by the predominance of dopaminergic activity over frontal activation (Heatherton and Wagner, 2011; Crews and Boettiger, 2009). The same holds for irrational behaviours or cognitions, such as hypersexuality, gambling behaviour, stereotypic behaviour, and delusions. Similarly, there is compelling evidence that self-deception involves a significant increase in dopaminergic transmission (Sharot et al., 2012; Delgado et al., 2005; Westen et al., 2006) and a decrease in frontal activation (McKay et al., 2013). Just as the precedence of dopamine partly explains addiction, it also illuminates the selective treatment of positive information in straight self-deception. The dominance of dopaminergic activity is central to understanding phenomena that revolve around the preference for immediate reward, such as addiction and straight self-deception,

even if they have long-term negative consequences. When people are uncertain and appraise a significant inevitable threat, somatic markers and dopamine protect them from forming the distressing belief.

Our proposal is neutral with regard to the exact relation between the appraisals and the neurobiological mechanisms described. It is compatible with the possibility that the appraisals are identical to the relevant neurobiological mechanisms, with the appraisals causing them, supervening on them, or being grounded in them. What matters for our purposes is that these neurobiological mechanisms capture how the positive information takes precedence over negative information in straight self-deception. By definition, these mechanisms describe how the affective part of our brain competes with the rational one (roughly, the prefrontal cortex) in the treatment of information, which can lead to a state of imbalance in addiction and in self-deception. To put it metaphorically, they describe the "hydraulics" of information processing and obey the principle of communicating vessels. In this sense, they are inherently dynamic.

As it appears, our conception differs in type from the other accounts examined. Strictly speaking, it is compatible with the hedonic dynamic of self-deception, although it does not imply this view. It refines the idea that self-deception is driven by anxiety, as it describes the underlying mechanisms governing its dynamic. Shifting to this level of description allows us to fulfill the affectivist aspirations.

6. THE AFFECTIVE-FILTER VIEW: SELECTIVITY

Not every desire results in self-deception, and our view explains why this is so. At the psychological level, three appraisals delineate the conditions in which desiring subjects do not deceive themselves. A desiring subject does not deceive herself in the presence of distressing evidence if

- (i) does not appraise the evidence as ambiguous;
- (ii) does not appraise the evidence as having a significant negative impact on S's well-being; and
- (iii) does not appraise S's coping potential as low.

The first condition correctly predicts that people cease to deceive themselves when distressing evidence accumulates, such that the evidence is no longer appraised as ambiguous. The second condition relies on the fact that the affective-coping appraisal is not an all-or-nothing matter. Subjects who estimate that they can bear with a distressing fact will not self-deceive. Regarding the third condition, we have already observed that self-deception does not occur when people appraise that they can act on situations. Consequently, the verdicts of various filters generate several routes out of self-deception.

However, as emphasized, the psychological appraisals are compatible with forming the rational belief. Therefore, staying at this level of description does not

solve the selectivity challenge, which is why our solution relies on neurobiological mechanisms as well.

Our solution can be summarized as follows: subjects do not deceive themselves if dopaminergic activity fails to take precedence over other neural structures, such as frontal activation and negative somatic markers. This accounts for the hard case of "*Guilty Son*." Mark appraises the situation as negative and as falling beyond his control, but does not deceive himself, because dopaminergic transmission fails to dominate other structures. This can happen for several reasons. For instance, subjects may suffer from dopaminergic deficits that are compatible with the retention of desire; they just render such desire inert, so to speak. This might explain why some subjects do not self-deceive. Alternatively, dopamine can fail to take precedence if people are hypersensitive to threats. Such people would not ignore the negative signals of somatic markers; somatic markers would triumph over dopamine. For instance, depression and anxiety involve acute sensitivity to threats via somatic markers, at the expense of dopaminergic activity (Surbey, 2011). Our view hereby offers a clear-cut contrast between cases where desire leads to self-deception and cases where it does not— in neurobiological terms and, particularly, in dopaminergic terms.

This solution captures the grain of truth of the alternative proposals examined, but does not fall prey to the same pitfalls. It does not imply that self-deception occurs only when it would reduce anxiety, which is a virtue (§ 2). In the absence of a predominance of dopaminergic activity, people do not self-deceive even when self-deception would reduce anxiety. Our solution also goes beyond the idea that self-deception occurs when the subject assesses the costs of error as low. On our view, the subject may assess the costs of error as low, yet not self-deceive if dopamine fails to dominate other neural structures. The neurobiological mechanisms explain when the assessment of the costs of error as low leads to self-deception. Although our proposal is compatible with the other affectivist solutions, shifting the discussion to the level of these neurobiological mechanisms has the advantage of capturing the process in inherently dynamic terms, given the imbalance between the rational/frontal and the affective brain regions described.

One might be skeptical. Our solution hinges on the dominance of dopaminergic activity in information processing. This raises the following question: Why does dopaminergic transmission take precedence in some cases only? In other words, the selectivity problem might arise again. Although dopamine and somatic markers are important predictors of self-deception, we concede that we have not explained when dopamine will triumph. However, as observed, the selectivity problem would be intractable if it required predicting self-deception. Our solution is satisfactory because appealing to dopaminergic transmission provides a contrast between cases in which desire results in deceptive beliefs and cases in which it does not.

However, the intentionalist spectre might arise once more. Why should our solution justify non-intentionalism? After all, the neurobiological mechanisms proposed are compatible with the intention of forming the deceived belief. Affective filters cut no ice. In response to this objection, let us observe that the affective filters described, such as the neurobiological mechanisms, operate automatically—that is, unconsciously and unintentionally. Somatic markers function to signal and simulate threats, whereas dopamine's function is partly to direct subjects' attention to cues that are relevant to desire's satisfaction. For these functions to be fulfilled, these mechanisms are better understood as operating unintentionally; they would lose their economical character if they involved the intention of forming beliefs. This is compatible with affective filters eliciting the intention to attend to relevant stimuli; this is where these biases are partly subject to control. However, intentionalists claim that self-deception involves the intention to form the deceptive belief—not merely the intention to attend to some information (Lynch, 2014). Moreover, given the balance between dopaminergic transmission and frontal activation, it is empirically implausible to regard self-deception as intentional. Its neural signature would involve significantly more frontal activation than it actually does, given that intentions to deceive should come with strong frontal activation, such as in interpersonal deception (Christ et al., 2008). Self-deception thus differs from other affective biases, like the positivity effect, that involve significant frontal activation. It aligns itself with unintentional affective influences on belief. The affective-filter view thereby offers empirical justification for non-intentionalism.

7. THE AFFECTIVE-FILTER VIEW: UNITY

How does our proposal apply to twisted self-deception? Recall that a promising account should not reduce twisted self-deception to straight self-deception (§ 2). Instead, it is preferable to conceive of twisted and straight self-deception as involving similar components that operate in opposing ways. This opens a path for an amendment of our proposal on straight self-deception, which will allow us to capture twisted cases. In straight self-deception, the evaluation of positive information takes precedence over that of distressing evidence via dopaminergic activity triumphing over somatic markers and other neural structures. Conversely, in twisted self-deception, the evaluation of distressing evidence takes precedence over that of positive evidence via negative somatic markers triumphing over dopamine and other neural structures. Straight and twisted self-deception involve the same components, but they differ in terms of the dominance of one over the other. More precisely, a subject S, who desires that p and is presented with sufficient evidence in favour of p, forms the belief that not-p, if and only if

(i)   appraises the evidence in favour of p as ambiguous;
(ii)   appraises the distressing evidence as negative;
(iii) appraises his or her coping potential as low;
(iv) appraises p and the evidence for p positively;

(v)   the appraisal of the distressing evidence is accompanied by negative somatic markers;

(vi) the appraisal of the positive evidence is accompanied by dopaminergic activity; and

(vii) negative somatic markers take precedence over frontal activation and dopaminergic activity in the processing of information.

The first, second, and fourth conditions were justified earlier. The third condition is more controversial. Isn't twisted self-deception compatible with appraising the situation as being within one's control, as it functions to protect oneself through action? Consider an example. Sarah deceives herself into believing that she has left the stove on, which ensures that she will check whether the stove is on. Doesn't she appraise her coping potential as high? Let us recall that the coping potential appraisal allows for degrees. In some cases, one appraises one's coping potential as low, even if one regards the situation, strictly speaking, as under one's control; acting may be costly or one may have only indirect control of the situation. Imagine that Sarah suspects that she left the stove on while she is at home. It is unlikely that she will deceive herself; rather, she will make sure that the stove is off, because she appraises her coping potential as high. This third condition is compatible with twisted self-deception functioning to protect oneself via action because the relevant actions ensure only indirect satisfaction of a desire.

The core of our proposal lies in the last components pertaining to the relation between the neurobiological mechanisms, especially the precedence of somatic markers over frontal activation. Common accounts of the anxiety bias square with the somatic-markers hypothesis. Anxious people regard their anxious hunches as evidence for certain beliefs (Mogg and Bradley, 2016). This corresponds to negative somatic markers, as hunches come with negative anticipation, as revealed by studies on decision making (Miu et al., 2008). Whereas the signals of negative somatic markers are discarded and block further processing of negative information in straight self-deception, subjects do not neglect the signals of negative somatic markers in twisted self-deception. On the contrary, the anticipation and simulation of threats take precedence over frontal activation (Cisler and Koster, 2010). This is compatible with the presence of dopaminergic transmission, notably because dopamine is released especially in cases of uncertainty and it increases attention to cues relevant to desire's satisfaction, even when these point toward desire's frustration. Still, in twisted self-deception, negative somatic markers trump dopaminergic transmission and frontal activation in the processing of information.

It appears that the only crucial difference between the dynamics of straight and twisted self-deception involves the last condition. Twisted self-deception is the inverted analogue of straight self-deception.

For these reasons, our proposal has advantages over competing accounts, while retaining their intuitive character. As observed, it does not imply that self-decep-

tion functions to reduce anxiety, so it does not suffer from the difficulties associated with this claim (§ 2). For instance, it is compatible with the idea that twisted self-deception aims at protection, because somatic markers and the neural structures of anxiety have this function. Moreover, the proposal substantiates the idea that self-deception is motivated by anxiety and explains the different routes that anxiety might take in self-deception (§ 3). It offers a clear-cut contrast between straight and twisted self-deception by describing the difference between them at the subpersonal level. Finally, for the reasons mentioned above, our account of twisted self-deception is clearly non-intentionalist. Somatic markers, along with the influence of anxiety on belief, operate at the early stages of processing. The neural structures responsible for the anxiety bias are far from corresponding to the frontal activation involved in intentional behaviour. It is therefore unlikely that twisted self-deception is intentional.

One might doubt it. As the proposal reduces twisted self-deception to beliefs formed under the influence of anxiety, does it truly capture the specificity of self-deception? How does it avoid generalizing to all types of affective bias? In our picture, straight and twisted self-deception both result in beliefs motivated by desire and formed through similar mechanisms, but operating in inverted fashion. This secures the unity of the phenomenon. By contrast, other affective biases need not involve these components. For instance, the influence of sadness on belief is not explained by dopamine, as revealed by studies on depressive realism (Surbey, 2011), and the negative biases of sadness do not rely on anticipation, as somatic markers do (Koster et al., 2010). Likewise, we have already mentioned how the positivity effect depends on other mechanisms. Of course, our components may partly feature in other emotional biases, given that they are central to protective mechanisms in general (Ansermet and Magistretti, 2017). Yet, as far as self-deception is concerned, they are the paramount ones.

Let us step back and consider a final objection concerning the role of emotion in our picture. The focus on the underlying mechanisms of self-deception might come at the price of eluding affect's role in self-deception. What, exactly, is emotion's role in self-deception, according to our picture? Does the picture truly do justice to *emotion*'s role in self-deception? The answer to this question depends on the vexed question of the relation between emotion and affective filters. Consider the relation between emotion and cognitive appraisals. One possibility is that emotions *are* cognitive appraisals, as in the idea that emotions are experiences of values (Tappolet, 2000). In that case, self-deception would involve emotions, such as anxiety and positive anticipation, as these correspond to the appraisals described. Unconscious instances of those emotions may play a role, as appraisals can be unconscious. Alternatively, appraisals might be conceived as a cause or a component of emotions, in which case emotion's role in self-deception would be less straightforward in our picture. Nonetheless, on this interpretation, affect would still play a role, through "proto-affective" phenomena. These phenomena are components of emotions and lead to full-fledged emotions only under some conditions (e.g., when a sufficient degree of integration is attained or when the subject is conscious of them [Ortony et al.,

2012]). For some authors, cognitive appraisals and the neurobiological mechanisms mentioned above are among proto-affective phenomena. The affective nature of these phenomena hinges on the fact that they constitute appraisals of situations in light of one's goals. Our conception is neutral with regard to the relation between emotion and affective filters. Whatever one's interpretation of the relation, affect's role consists in the assessment of information in light of personal concerns, whether this takes the form of discrete emotions or proto-affective phenomena.

CONCLUSION

Affectivism touches on key issues, such as the dynamic of self-deception, its unity, and its contribution to happiness. Surprisingly, it has been seldom scrutinized with the help of empirical findings, despite the insights that studies on affective biases provide into this issue. In this article, we have aimed to redress this imbalance. The examination of the main affectivist accounts has invited us to leave the armchair and to offer an empirically minded approach to the affective dynamic of self-deception.

We have argued that affect's role in self-deception is better understood at the phase of the evaluation of evidence. Understanding its role as the mere input or as the function of the process is less promising. We do not deny that affect may and often does play a role at these other levels. However, this role does not lead us very far with regard to the promises of affectivism. By contrast, the idea that self-deception involves evaluating information in light of one's concerns (the affective-filter view) fulfills the two promises of affectivism. First, our conception disentangles the latest challenge to non-intentionalism—namely, the selectivity problem—as the affective filters capture the selective treatment of information in non-intentionalist terms. Second, our approach offers an original account of twisted self-deception. Twisted self-deception involves the same affective filters as straight self-deception does, with the single difference being the predominance of one mechanism over the other. In our proposal, self-deception's dynamic may involve discrete emotions, such as anxiety and anticipated pleasure, or proto-affective phenomena. Be that as it may, the affective-filter view supports the idea that self-deception need not be intentional. The battle among dopamine, somatic markers, and frontal activation vindicates the thought that self-deception is "belief under influence." This conception could be developed further to tackle other types of motivated biases, such as wishful thinking, motivated information gathering, and repression, but this will wait for another occasion. Affective filters are central to self-deception's dynamic. Ultimately, the aspirations of affectivism are realized.

NOTES

[1] It is assumed that in this debate desires differ from emotions.

[2] For this reason, we do not consider the idea that emotions can be self-deceptive (De Sousa, 1978) or the claim that self-deception involves conflicting beliefs because it results in anxiety. These claims do not address the dynamics issue.

[3] As we assume that the product of self-deception is the welcome belief, we ignore doubts about whether straight self-deception results in anxiety reduction because it involves conflicting beliefs.

[4] See Scott-Kakures (2000, 2001) for the idea that self-deception functions to promote one's interests broadly speaking, with anxiety reduction being one of many goals.

[5] Echano (2017) claims that anxiety's role in twisted self-deception lies in triggering unwelcome hypotheses, just like desire triggers welcome hypotheses in straight self-deception. We do not consider this proposal, as it restricts the role of emotion to twisted self-deception only. Given anxiety's role in straight self-deception, we think that there is more room for emotion's role in self-deception.

[6] Ironically, Talbott (1995) offered a similar solution to the selectivity problem within his intentionalist framework. We shall not consider his argument in detail here, as the solution examined does not appeal to intention. See Scott-Kakures (2000, 2001) for a discussion.

[7] Barnes (1997, p. 80) acknowledges that the tendency to self-deceive can be trumped by other dispositions, such as the disposition to protect oneself from danger. However, this proposal does not apply to cases in which no action is available to protect oneself, as in "Guilty Son." Although other dispositions might trump the tendency to self-deceive, the problem consists precisely in specifying the conditions in which people self-deceive.

[8] We shall not discuss the computational model of the role of emotion in self-deception (Sahdra and Thagard, 2003), because the authors do not argue that their picture favours non-intentionalism (Sahdra and Thagard, 2003, p. 227-228), nor that it covers twisted self-deception (see, however, Thagard and Nussbaum [2014] for a computational model of twisted self-deception). That being said, our conception can be seen as a way of developing the computational model with the help of empirical findings.

REFERENCES

Ansermet, François, and Pierre Magistretti, *Biology of Freedom: Neural Plasticity, Experience, and the Unconscious*, New York, Other Press, 2017.

Bach, Kent, "An Analysis of Self-Deception," *Philosophy and Phenomenological Research*, vol. 41, no. 198, p. 1351-1370.

Barnes, Annette, *Seeing through Self-Deception*, Cambridge, Cambridge University Press, 1997.

Bayne, Tim, and Jordi Fernández (eds.), *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, New York, Psychology Press, 2009.

Bechara, Antoine, "Deciding Advantageously Before Knowing the Advantageous Strategy," *Science*, vol. 275.5304, 1997, p. 1293-1295.

Bermúdez, José Luis, "Defending Intentionalist Accounts of Self-Deception," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 107-108.

———, "Self Deception, Intentions, and Contradictory Beliefs," *Analysis*, vol. 60, no. 268, 2000, p. 309-319.

———, "Self-deception and Selectivity. Reply to Jurjako," *Croatian Journal of Philosophy*, vol. 17, no. 1, 2017, p. 91-95.

Christ, Shawn E., et al., "The Contributions of Prefrontal Cortex and Executive Control to Deception: Evidence from Activation Likelihood Estimate Meta-analyses," *Cerebral Cortex*, vol. 19, no. 7, 2008, p. 1557-1566.

Cisler, Josh M., and Ernst H, Koster, "Mechanisms of Attentional Biases towards Threat in Anxiety Disorders: An Integrative Review," *Clinical Psychology Review*, vol. 30, no. 2, 2010, p. 203-216.

Correia, Vasco, "From Self-Deception to Self-Control," *Croatian Journal of Philosophy*, vol. 14, no. 3, 2014, p. 309-323.

Crews, Fulton Timm, and Charlotte Ann Boettiger, "Impulsivity, Frontal Lobes and Risk for Addiction," *Pharmacology, Biochemistry and Behavior*, vol. 93, no. 3, 2009, p. 237-247.

Damasio, Antonio R., *Descartes' Error: Emotion, Reason and the Human Brain*, New York, Putnam, 1999.

Davidson, Donald, "Paradoxes of Irrationality," in Wollheim, Richard and James Hopkins (eds.), *Philosophical Essays on Freud*, Cambridge, Cambridge University Press, 1982,p. 79-92.

———, "Deception and Division," in LePore, Ernst and Brian McLaughlin (eds.), *Actions and Events*, New York, Basil Blackwell, 1985.

Dawson, Erica, Kenneth Savitsky, and David Dunning, "'Don't Tell Me, I Don't Want To Know': Understanding People's Reluctance To Obtain Medical Diagnostic Information," *Journal of Applied Social Psychology*, vol. 36, no. 3, 2006, p. 751-768.

Delgado, Mauricio R., et al., "An fMRI Study of Reward-Related Probability Learning," *Neuroimage*, vol. 24, no. 3, 2005, p. 862-873.

De Sousa, Ronald, "Self-Deceptive Emotions," *Journal of Philosophy*, vol. 75, no. 11, 1978, p. 684-697.

———, "Emotion and Self-Deception," in McLaughlin, Brian and Amelie O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley, University of California Press, 1988, p. 63-91.

Echano, Mario R., "The Motivating Influence of Emotion on Twisted Self-Deception," *Kritike*, vol. 11, no. 2, 2017, p. 104-120.

Echarte, Luis E., Javier Bernacer, Denis Larrivee, J. V. Oron, and Miguel Grijalba-Uche, "Self-Deception in Terminal Patients: Belief System at Stake," *Frontiers in Psychology*, vol. 7, 2016, p. 117.

Ellsworth, Phoebe C., "Appraisal Theory: Old and New Questions," *Emotion Review,* vol. 2, 2013, p. 125-131.

Friedrich, James, "Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena," *Psychological Review*, vol. 100, no. 2, 1993, p. 298-319.

Galeotti, Anna Elisabetta, "Straight and Twisted Self-Deception," *Phenomenology and Mind*, vol. 11, 2016, p. 90-99.

Gigerenzer, Gerd, *Gut Feelings: The Intelligence of the Unconscious*, New York, Viking, 2007.

———, "Why Heuristics Work," *Perspectives on Psychological Science*, vol. 3, no. 1, 2008, p. 20-29.

Heatherton, Todd F., and Dylan D. Wagner, "Cognitive Neuroscience of Self-Regulation Failure," *Trends in Cognitive Sciences*, vol. 15, no. 3, 2011, p. 132-139.

Ishiguro, Kazuo, *The Remains of the Day*, London, Faber and Faber, 1989.

Johnston, Mark, "Self-Deception and the Nature of Mind," in Brian McLaughlin and Amelie O. Rorty (eds.), *Perspectives on Self-Deception,* Berkeley, University of California Press, 1988, p. 63-91.

Jurjako, Marko, "Self-Deception and the Selectivity Problem," *Balkan Journal of Philosophy*, vol. 5, no. 2, 2013, p. 151-162.

Klayman, Joshua, and Young-Won Ha, "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Psychological Review,* vol. 94, 1987, p. 211-228.

Koster, Ernst H., Rudi De Raedt, Lemke Leyman, and Evi De Lissnyder, "Mood-Congruent Attention and Memory Bias in Dysphoria: Exploring the Coherence among Information-Processing Biases," *Behaviour Research and Therapy*, vol. 48, no. 3, 2010, p. 219-225.

Kunda, Ziva, "The Case for Motivated Reasoning," *Psychological Bulletin*, vol. 108, no. 3, 1990, p. 480-498.

Lauria, Federico, Delphine Preissmann, and Fabrice Clément, "Self-Deception as Affective Coping. An Empirical Perspective on Philosophical Issues," *Consciousness and Cognition*, vol. 41, 2016, p. 119-134.

Lazar, Ariela, "Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'," *Mind*, vol. 108, no. 430, 1999, p. 265-290.

Lazarus, Richard S., "Progress on a Cognitive-Motivational-Relational Theory of Emotion," *American Psychologist,* vol. 46, no. 8, 1991, p. 819-834.

Lynch, Kevin, "Self-Deception and Shifts of Attention," *Philosophical Explorations*, vol. 17, no. 1, 2014, p. 63-75.

Martínez-González, J. M., R. V. López, E. B. Iglesias, and A. Verdejo-García, "Self-Deception as a Mechanism for the Maintenance of Drug Addiction," *Psicothema*, vol. 28, no. 1, 2016, p. 13-19.

McKay, Ryan, et al., "Vestibular Stimulation Attenuates Unrealistic Optimism," *Cortex*, vol. 49, no. 8, 2013, p. 2272-2275.

Mele, Alfred R., "Real Self-Deception," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 91-102.

———, *Self-Deception Unmasked*, Princeton, Princeton University Press, 2000.

———, "Emotion and Desire in Self-Deception," in Anthony Hatzimoysis (ed.), Royal Institute of Philosophy Supplement, Cambridge, Cambridge University Press, 2003.

Miu, Andrei C., Renata M. Heilman, and Daniel Houser, "Anxiety Impairs Decision-Making: Psychophysiological Evidence from an Iowa Gambling Task," *Biological Psychology*, vol. 77, no. 3, 2008, p. 353-358.

Mogg, Karin, and Brendan P. Bradley, "Anxiety and Attention to Threat: Cognitive Mechanisms and Treatment with Attention Bias Modification," *Behaviour Research and Therapy*, vol. 87, 2016, p. 76-108.

Nelkin, Dana K., "Self–Deception, Motivation, and the Desire To Believe," *Pacific Philosophical Quarterly*, vol. 83, no. 4, 2002, p. 384-406.

Ortony, Andrew, Donald A. Norman, and William Revelle, "Affect and Proto-Affect in Effective Functioning," in J.-M. Fellous and M. A. Arbib (eds.), *Who Needs Emotions? The Brain Meets the Robot*, New York, Oxford University Press, 2012.

Pears, David, *Motivated Irrationality*, Oxford, Clarendon Press, 1986.

Pedrini, Patrizia, "Cognition and Desires: How To Solve the 'Selectivity Problem' for Self-Deception," in *Pratiche della cognizione*, Proceedings of the AISC2010, Italian Society for Cognitive Sciences, 2010.

Peterson, Jordan B., Erin Driver-Linn, and Colin G. DeYoung, "Self-Deception and Impaired Categorization of Anomaly," *Personality and Individual Differences*, vol. 33, no. 2, 2002, p. 327-340.

Peterson, Jordan B., et al., "Self-Deception and Failure To Modulate Responses Despite Accruing Evidence of Error," *Journal of Research in Personality*, vol. 37, no. 3, 2003, p. 205-223.

Reed, Andrew E., and Laura L. Carstensen, "The Theory behind the Age-Related Positivity Effect," *Frontiers in Psychology*, vol. 3, 2012, p. 339.

Sahdra, Baljinder, and Paul Thagard, "Self-Deception and Emotional Coherence," *Minds and Machines*, vol. 13, no. 2, 2003, p. 213-231.

Scherer, Klaus R., Angela Schorr, and Tom Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research*, Oxford, Oxford University Press, 2001.

Schultz, Wolfram, Peter Dayan, and P. Read Montague, "A Neural Substrate of Prediction and Reward," *Science*, vol. 275.5306, 1997, p. 1593-1599.

Schultz, Wolfram, Léon Tremblay, and Jeffery R. Hollerman, "Reward Prediction in Primate Basal Ganglia and Frontal Cortex," *Neuropharmacology*, vol. 37, no. 4-5, 1998, p. 421-429.

Scott Kakures, Dion, "Motivated Believing: Wishful and Unwelcome," *Noûs*, vol. 34, no. 3, 2000, p. 348-375.

Scott-Kakures, Dion, "High Anxiety: Barnes on What Moves the Unwelcome Believer," *Philosophical Psychology*, vol. 14, no. 3, 2001, p. 313-326.

Sharot, Tali, et al., "How Dopamine Enhances an Optimism Bias in Humans," *Current Biology*, vol. 22, no. 16, 2012, p. 1477-1481.

Sperber, Dan, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson, "Epistemic Vigilance," *Mind & Language*, vol. 25, no. 4, 2010, p. 359-393.

Surbey, Michele K., "Adaptive Significance of Low Levels of Self-Deception and Cooperation in Depression," *Evolution and Human Behavior*, vol. 32, no. 1, 2011, p. 29-40.

Talbott, William J., "Intentional Self-Deception in a Single Coherent Self," *Philosophy and Phenomenological Research*, vol. 55, no. 1, 1995, p. 27-74.

Tappolet, Christine, *Émotions et valeurs*, Paris, Presses Universitaires de France, 2000.

Thagard, Paul, and A. David Nussbaum, "Fear-Driven Inference: Mechanisms of Gut Overreaction," in Lorenzo Magnani (ed.), *Model-Based Reasoning in Science and Technology*, Studies in Applied Philosophy, Epistemology and Rational Ethics 8, Heidelberg, Springer, 2014, p. 43-53.

Westen, Drew, et al., "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*, vol. 18, no. 11, 2006, p. 1947-1958.