

UNE REVUE MULTIDISCIPLINAIRE SUR LES ENJEUX NORMATIFS  
DES POLITIQUES PUBLIQUES ET DES PRATIQUES SOCIALES.

# Les ateliers de l'éthique The Ethics Forum

A MULTIDISCIPLINARY JOURNAL ON THE NORMATIVE  
CHALLENGES OF PUBLIC POLICIES AND SOCIAL PRACTICES.

## DOSSIER :

### On Self-Deception

- 4-10 Introduction, **Anne Meylan**
- 11-24 Liberalizing Self-Deception: Replacing Paradigmatic-State Accounts of Self-Deception With a Dynamic View of the Self-Deceptive Process , **Patrizia Pedrini**
- 25-47 Self-Deceptive Resistance to Self-Knowledge **Graham Hubbs**
- 48-69 How to Tragically Deceive Yourself **Jakob Ohlhorst**
- 70-94 What Does Emotion Teach Us About Self-Deception?  
Affective Neuroscience in Support of Non-Intentionalism, **Federico Lauria and Delphine Preissmann**
- 95-118 Costly False Beliefs: What Self-Deception and Pragmatic Encroachment Can Tell Us About the Rationality of Beliefs, **Melanie Sarzano**
- 119-134 Responsibility for Self-Deception, **Marie van Loon**

UNE REVUE MULTIDISCIPLINAIRE SUR LES ENJEUX NORMATIFS  
DES POLITIQUES PUBLIQUES ET DES PRATIQUES SOCIALES.

# Les ateliers de l'éthique The Ethics Forum

A MULTIDISCIPLINARY JOURNAL ON THE NORMATIVE CHALLENGES  
OF PUBLIC POLICIES AND SOCIAL PRACTICES.

La revue est financée par le Conseil de recherches en sciences humaines (CRSH)  
et administrée par le Centre de recherche en éthique (CRE)

The journal is funded by the Social Sciences and Humanities Research Council (SSHRC)  
and administered by the Center for Research on Ethics (CRE)



## Attribution 4.0 International

Vous êtes libres de reproduire, distribuer et communiquer les textes de cette revue au public, transformer et créer à partir du matériel pour toute utilisation y compris commerciale, selon les conditions suivantes :

- Vous devez créditer l'Œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Œuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Œuvre.

Pour tous les détails, veuillez vous référer à l'adresse suivante :  
<https://creativecommons.org/licenses/by/4.0/legalcode>

You are free to copy and distribute all texts of this journal, transform, and build upon the material for any purpose, even commercially, under the following conditions:

- You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

For all details please refer to the following address:  
<https://creativecommons.org/licenses/by/4.0/legalcode>

**RÉDACTRICE EN CHEF/EDITOR**

Christine Tappolet, Université de Montréal

**COORDONNATEUR DE RÉDACTION/ADMINISTRATIVE EDITOR**

Jean-Philippe Royer, Université de Montréal

**COMITÉ EXÉCUTIF DE RÉDACTEURS/EXECUTIVE EDITORS**

Éthique fondamentale : Natalie Stoljar, Université McGill

Éthique et politique : Ryoa Chung, Université de Montréal

Éthique et santé : Kristin Voigt, Université McGill

Éthique et économie : Peter Dietsch, Université de Montréal

Éthique et environnement : Gregory Mikkelsen, Université McGill

**COMITÉ D'EXPERTS/BOARD OF REFEREES:**

Arash Abizadeh, Université McGill

Charles Blattberg, Université de Montréal

Marc-Antoine Dilhac, Université de Montréal

Francis Dupuis-Déri, Université du Québec à Montréal

Geneviève Fuji Johnson, Université Simon Fraser

Axel Gosseries, Université catholique de Louvain

Joseph Heath, Université de Toronto

Samia A. Hurst, Université de Genève

Julie Lavigne, Université du Québec à Montréal

Robert Leckey, Université McGill

Catriona Mackenzie, Macquarie University

Katie McShane, Colorado State University

Bruce Maxwell, UQTR

Pierre-Yves Néron, Université catholique de Lille

Wayne Norman, Duke University

Hervé Pourtois, Université catholique de Louvain

Vardit Ravitsky, Université de Montréal

Mauro Rossi, Université du Québec à Montréal

Christine Straehle, Université d'Ottawa

Daniel M. Weinstock, Université McGill

Jennifer Welchman, University of Alberta

Danielle Zwarthoed, Université catholique de Louvain

**NOTE AUX AUTEURS**

Merci de soumettre votre article à l'adresse  
[jean-philippe.royer@umontreal.ca](mailto:jean-philippe.royer@umontreal.ca)

Les articles seront évalués de manière anonyme par deux pairs. Les consignes aux auteurs se retrouvent sur le site de la revue ([www.lecre.umontreal.ca/categorie/ateliers\\_ethique\\_ethics\\_forum/](http://www.lecre.umontreal.ca/categorie/ateliers_ethique_ethics_forum/)). Tout article ne s'y conformant pas sera automatiquement refusé.

**GUIDELINES FOR AUTHORS**

Please submit your paper to: [jean-philippe.royer@umontreal.ca](mailto:jean-philippe.royer@umontreal.ca)

Articles are anonymously peer-reviewed. Instructions to authors are available on the journal website ([www.lecre.umontreal.ca/categorie/ateliers\\_ethique\\_ethics\\_forum/](http://www.lecre.umontreal.ca/categorie/ateliers_ethique_ethics_forum/)). Papers not following these will be automatically rejected.

# DOSSIER

## SELF-DECEPTION: NEW ANGLES

ANNE MEYLAN  
UNIVERSITY OF ZURICH

### INTRODUCTION

It is no wonder that self-deception has always sparked philosophers' interest. Self-deception is a very intriguing phenomenon from both the descriptive and the normative points of view. First, self-deception raises a number of descriptive problems. Should we model self-deception on others' deception and hold that the self-deceived subject *intends* to deceive herself? Or should we rather—granted that some motivational state must be part of the self-deceptive process—identify self-deception's motivational cause with a *desire*? Famously enough, the classical opposition between intentionalist (e.g., Davidson, 1986) and deflationist accounts (e.g., Mele, 1997, 2001; Nelkin, 2002) of self-deception revolves around these questions. Intentionalist and deflationist accounts face different sorts of worries and each seems to succeed where the other fails. For instance, deflationism has been accused of not faring as well as intentionalism with regards to the selectivity problem (Talbot, 1995). For its part, intentionalism—mostly because it does not allow the self-deceptive process to take place unknowingly—has been charged of raising paradoxes (the well-known “static” and “dynamic” paradoxes; see, e.g., Mele, 2001). Furthermore, self-deception typically involves a certain epistemic discomfort or tension (Funkhouser, 2005; Noordhof, 2009; Van Leeuwen, 2007). Subjects do not hold self-deceptive beliefs in the simple, wholehearted way in which they hold their other, non-self-deceptive beliefs. Most often, doubts nag at the back of their minds and prevent them from being completely at ease with their beliefs. Another descriptive problem is to explain this tension.

Self-deception is also *normatively* fascinating. Self-deception is often considered to be an irrational cognitive phenomenon. But what makes it irrational? Is it not, at least occasionally, acceptable to deceive oneself? According to Joseph Butler (1726/2006) and Adam Smith (1759/2002), self-deception is always morally reprehensible, mainly because we get used to casting a too-favourable light on the morality of our own actions by self-deceiving ourselves about them. This gradually corrupts our moral judgment and prevents us from correcting our moral mistakes. Much more recently, Van Leeuwen (2009) has claimed that self-deception is not even “egoistically good” since it does not make us happy.<sup>1</sup> In contrast, Barnes (1997) has argued that in some sufficiently difficult or costly circum-

stances, “the avoidance of a painful truth” (chapter 9, p. 165) is not always *prima facie* morally bad (even though, according to Barnes, the epistemic cowardice that goes along with self-deception is *prima facie* objectionable). Additionally, several psychologists and neuroscientists (see, e.g., Sharot, 2011; Taylor, 1989) have emphasized the positive effects that (positive) self-deceptive evaluations of ourselves have on our mental health. These evaluations even seem able to promote our ability “to care about others” and “to engage in productive and creative work” (Taylor and Brown, 1988, p. 193).

While the first four papers included in this special issue address some descriptive issues raised by self-deception, the last two articles deal rather with normative questions surrounding it.

Pedrini’s paper “Liberalizing Self-Deception: Replacing ‘Paradigmatic State Accounts’ of Self-Deception with a Dynamic View of the Self-Deceptive Process” suggests a change of paradigm. Philosophers should give up the “snapshot” conception of self-deception—that is, the currently prevailing assumption that self-deception is a stative phenomenon. Pedrini suggests that we should replace this classical conception with a dynamic and processual account. According to her proposal, self-deception is a process that is susceptible to including what she describes as “a multitude of highly tensive and unstable mental states” that are not only cognitive but also conative and affective.

The purpose of Hubbs’s contribution is also to illuminate the very nature of self-deception. In line with Barnes’s anxiety-avoidance account, he argues that self-deception results from the “tendency of the mind to avoid thinking unpleasant thoughts.” An additional, and to my knowledge, innovative element of Hubbs’s view is, however, as follows. When this tendency is satisfied—that is, when the subject holds the self-deceptive belief—she also takes the positive feeling (or pleasure) that comes with the self-deceptive belief to signify its warrant or its truth while it is, in fact, not the mark of truth but rather the mark of the believer believing what she wants to believe. In other words, the self-deceived subject confuses “the epistemic satisfaction of believing what is warranted” with “the thumotic satisfaction of believing what he wants to be true.” According to Hubbs’s paper, self-deception results from a phenomenologically describable confusion between two distinct feelings of satisfaction.

Ohlhorst’s article focuses on a specific form of self-deception, which he calls *tragic self-deception*. In tragic self-deception, Ohlhorst claims, the self-deceived subject holds a belief that is immune to all pieces of evidence, even compelling ones. But is it possible to dismiss compelling evidence? Does this not amount to madness or deep irrationality? Ohlhorst defends the possibility of tragic self-deception by bringing in the Wittgensteinian notion of hinges or certainties. Briefly, hinges are acceptances (for instance, that the earth exists) that constitute the necessary rock-bottom of our knowledge and, most importantly, that lie beyond evidential justification. Hinges have, according to Ohlhorst, their analogue in the affective domain. These are iHinges, acceptances that are so

crucial for our affective balance (for instance, the acceptance that your daughter is not a persistent murderer) that they also lie beyond evidential justification. Finally, iHinges are what makes genuine cases of tragic self-deception possible.

For some time already, discussions surrounding self-deception have included findings from empirical science. For instance, significant inspiration has been found in the psychological studies addressing our “bounded rationality” (see, e.g., Mele, 1999 and Scott-Kakures, 2001). Lauria’s affectivist filter view of self-deception extends this trend but also innovates on it by introducing specific results from neuroscience into the picture of self-deception. More precisely, Lauria’s paper defends an account on which affective mental states (e.g., emotions like fears, shame, etc.) play a crucial role in the self-deceptive process at the stage of the appraisal of the evidence. His view is supported by, among other things, the fact that such an appraisal is accompanied by a certain neurobiological mechanism—mainly, dopaminergic activity that takes precedence over neural structures, such as frontal activation and negative somatic markers. This neuroscientifically informed account of self-deception has the advantage furthermore—Lauria argues—of solving the problem of selectivity (Talbot, 1995) and of unifying “straight” and “twisted” self-deception (Lazar, 1997; Mele, 1999), while other affectivist accounts do not score as well on these two problems.

Deceiving yourself is generally considered to be irrational. But what makes self-deception irrational? Sarzano’s paper focuses on this question. She gets part of her inspiration from the epistemological literature on pragmatic encroachment. “Pragmatic encroachment” refers to the cluster of views according to which the possession of epistemic states like knowledge and justified beliefs does not exclusively depend on truth-related factors. Pragmatic considerations—most famously the costs and benefits of knowing/believing something (de Rose, 1992)—also seem to have an influence on whether one really knows or holds a rational belief. For instance, in some cases, the *costs* of being wrong about *p* are such that the right attitude to hold is, for instance, to suspend judgment about *p*. Now, as Sarzano writes, not only are the costs/benefits of holding a belief occasionally considered (Mele, 2001) to part of the mechanism that initiates self-deception, but the influence of such practical considerations on the production of the self-deceptive belief also seems to make the latter irrational. But, how could practical considerations be responsible for the *rationality* of our doxastic attitudes (at least, according to the pragmatic Encroachers) while explaining their irrationality in the case of self-deception? In the last part of her article, Sarzano suggests various answers to this question. She differentiates right and wrong ways in which practical considerations might influence our doxastic attitudes.

Another important normative question is whether we are responsible for deceiving ourselves; van Loon’s paper addresses this issue. According to McHugh’s account of doxastic responsibility (see, e.g., McHugh, 2017—which is currently one of the most influential accounts), our being responsible for our beliefs is a matter of these beliefs being responsive to our reasons. In her paper, van Loon

shows that one implication of Mele’s account of self-deception is that self-deceptive beliefs are always reasons responsive. According to van Loon, self-deceptive beliefs “à la Mele” always fulfil the crucial necessary condition for doxastic responsibility.

As it should now be clear, the six articles included in this special issue approach self-deception from different angles, bringing in notions, tools, and results from distinct research areas. The outcome, hopefully, is a collection of essays that renews the traditional debate surrounding self-deception and that opens several original lines of research.

## ACKNOWLEDGEMENTS

I would like to thank Jean-Philippe Royer for his remarkable work in the preparation of this special issue and Sophie Keeling for the proofreading of the introduction. Funding was provided by Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (Grant No. PP00P1\_157436).

## NOTES

<sup>1</sup> See also Baron (1988) for some other charges against self-deception.



## REFERENCES

- Baron, Marcia, "What is Wrong with Self-Deception," in B. McLaughlin and A. O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley, University of California Press, 1988, p. 431-449.
- Barnes, Annette, *Seeing through Self-Deception*, Cambridge, Cambridge University Press, 1997.
- Butler, *The Works of Bishop Butler*, D. E. White (ed.), Rochester, University of Rochester Press, 2006 [1726].
- Davidson Donald, "Deception and Division," in J. Elster (ed.), *The Multiple Self*, Cambridge, Cambridge University Press, 1986.
- DeRose, Keith, "Contextualism and Knowledge Attributions," *Philosophy and Phenomenological Research*, vol. 52, no. 4, 1992, p. 913-929.
- Fischer, John Martin and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge, Cambridge University Press, 1998.
- Funkhouser, Eric, "Do the Self-Deceived Get What They Want?," *Pacific Philosophical Quarterly*, vol. 86, no. 3, 2005, p. 295-312.
- Lazar, Ariela, "Self-Deception and the Desire to Believe," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 119-120.
- McHugh, Conor, "Attitudinal Control," *Synthese*, vol. 194, no. 8, 2017, p. 2745-2762.
- Mele, Alfred, "Real Self-Deception," *Behavioural and Brain Sciences*, vol. 20, no. 1, 1997, p. 91-136.
- , "Twisted Self-Deception," *Philosophical Psychology*, vol. 12, no. 2, 1999, p. 117-137.
- , *Self-Deception Unmasked*, Princeton, Princeton University Press, 2001.
- Nelkin, Dana, "Self-Deception, Motivation, and the Desire to Believe," *Pacific Philosophical Quarterly*, vol. 83, no. 4, 2002, p. 384-406.
- Noordhof, Paul, "The Essential Instability of Self-Deception," *Social Theory and Practice*, vol. 35, no. 1, 2009, p. 45-71.
- Scott-Kakures, Dion, "High Anxiety: Barnes on What Moves the Unwelcome Believer," *Philosophical Psychology*, vol. 14, no. 3, 2001, p. 313-326.
- Sharot, Tali, *The Optimism Bias: A Tour of the Irrationally Positive Brain*, New York, NY, Random House, 2011.
- Smith, Adam, *The Theory of Moral Sentiments*, Knud Haakonssen (ed.), Cambridge, Cambridge University Press, 2002 [1759].
- Talbott, William J., "Intentional Self-Deception in a Single Coherent Self," *Philosophy and Phenomenological Research*, vol. 55, no. 1, 1995, p. 27-74.
- Taylor, Shelley and Jonathon Brown, "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin*, vol. 103, no. 2, 1988, p. 193-210.

Taylor, Shelley, *Positive Illusions: Creative Self-Deception and the Healthy Mind*, New York, Basic Books, 1989.

Van Leeuwen, Neil, "The Product of Self-Deception," *Erkenntnis*, vol. 67, no. 3, 2007, p. 419-437.

———, "Self-Deception Won't Make You Happy," *Social Theory and Practice*, vol. 35, no. 1, 2009, p. 107-132.

# LIBERALIZING SELF-DECEPTION: REPLACING PARADIGMATIC-STATE ACCOUNTS OF SELF-DECEPTION WITH A DYNAMIC VIEW OF THE SELF-DECEPTIVE PROCESS

PATRIZIA PEDRINI  
UNIVERSITY OF FLORENCE, ITALY

## ABSTRACT:

In this paper, I argue that paradigmatic-state accounts of self-deception suffer from a problem of restrictedness that does not do justice to the complexities of the phenomenon. In particular, I argue that the very search for a paradigmatic state of self-deception greatly overlooks the dynamic dimension of the self-deceptive process, which allows the inclusion of more mental states than paradigmatic-state accounts consider. I will discuss the inadequacy of any such accounts, and I will argue that we should replace them with a dynamic view of self-deception that is more liberal regarding the mental states in which self-deceivers may find themselves.

## RÉSUMÉ :

Dans cet article, je soutiens que les explications de l'auto-illusion en termes d'un état « paradigmatique » souffrent d'un problème de limitation qui ne rend pas justice à la complexité du phénomène. Plus précisément, j'avance que la recherche même d'un état paradigmatique néglige tout à fait la dimension dynamique du processus d'auto-illusion, de sorte à inclure davantage d'états mentaux que ne le font les explications en termes d'un état paradigmatique. Après avoir démontré l'insuffisance de ces dernières, je proposerai que nous devrions les remplacer par une conception dynamique de l'auto-illusion qui serait plus flexible quant aux états mentaux potentiellement vécus par des personnes sous l'emprise de l'auto-illusion.

## 1. INTRODUCTION

The problem of the mental state in which self-deceivers find themselves has a long tradition. It can be traced back at least to Mele's early objections (1997, 2001) to Davidson's intentionalism (1985). Famously, Davidson's idea that self-deception must be intentional was criticized by Mele for leading to a couple of paradoxes, one of which is the "static paradox," as it is known.<sup>1</sup> It regards the mental state a self-deceiver is in when the self-deceptive process is successfully accomplished. Mele's argument is well known: if self-deception is intentional, and self-deceivers thus intend their self-deception, at the end of the self-deceptive process they must retain the belief that not- $p$ —that is, the belief about how things really stand—while also getting to believe the self-deceptive, desired falsity that  $p$ . If this is correct, then it seems that the final, resulting mental state in which self-deceivers find themselves is somehow paradoxical, amounting to both believing that  $p$  and also believing that not- $p$ .<sup>2</sup> Thus, Mele suggested that we should get rid of intentionalism altogether and resort to focusing on the motivational set of the subjects engaging in self-deception: since the subjects desire that not- $p$ , the desire in question biases their evaluation and selective search for evidence. This opens the door to a motivationally distorted treatment of the relevant data, which leads the subjects directly into believing that not- $p$ , without also retaining any belief that  $p$ .

However, although the solution offered by Mele may be a satisfactory way out of the static paradox, it has the drawback of describing the state of mind of the self-deceiver as quite peaceful: if a full-blown belief that  $p$  is successfully reached, then there is no trace of the psychological tension that seems, instead, to be highly typical of self-deception. For this tension is obviously due to the fact that the motivationally distorted self-deceptive process runs counter to evidence that is at hand or easily available that not- $p$ .

Of course, scholars are sensitive to the interplay of motivation and evidence, and to their opposing thrusts. And once the problem of the psychological tension created by this contrasting interaction has been posed as crucial, scholars have continued to investigate the nature of this resulting, final state of mind of self-deception, and have tried to keep psychological tension in the picture. To this end, they have advanced interesting and refined analyses of what this final state must be, virtually all requiring that a satisfactory account of self-deception must preserve and account for the psychological tension that is characteristic of the self-deceiver's final state of mind.

In this paper, I wish to raise the issue, again and afresh. I will argue that, while most accounts currently on offer are on the right track in their search for states of mind that can account for tension, I will also object that most of them are on a misleading track insofar as they look, or tend to look, for a paradigmatic, final mental state for self-deception. For virtually all accounts working with paradigmatic states for self-deception suffer from a potential flaw that should be carefully considered—namely, a certain restrictedness, at least in spirit. I say "at

least in spirit” because, if a state is presented as paradigmatic, that need not be necessary. Thus, there may be room for predicting other mental states in a descriptive account of the mental state a self-deceiver is in—a less paradigmatic one, say, and yet capable of satisfying the constraints set by the motivated self-deceiver’s struggle with opposing evidence. This seems to be especially true for all those accounts that offer only sufficient (and not also necessary) conditions for self-deception, such as Mele’s.<sup>3</sup> However, the rhetoric of most accounts addressing the problem of the paradigmatic mental state of self-deception may lead us to assume that, when self-deceiving, more often than not we are in a certain highly typical state, and to focus on that state of mind at the expense of other possibilities. My plan in this paper is to show that these other possibilities are not only live, but also quite common and highly typical as well. They are so common and typical that, as a matter of fact, once we have and keep this empirical truth in view, any claims regarding states that can be taken as paradigmatic become fatally weakened.

One might ask why there has been such a great focus on the problem of a paradigmatic state. Most likely, it is created by a bias in favour of a static state, which questionably guides the analysis toward what I will dub a “snapshot theory” of self-deception. Unpacking the metaphor, one discovers that a snapshot theory of self-deception seems to be designed to take a descriptive, static picture of the mental state a self-deceiver is in. But this focus on static states may be highly misleading and lead us to exclude other candidates for typical self-deceptive mental states.

Starting from these premises, I will argue that self-deception is a psychological process before it is an end state—a process set in motion by the force field created by motivation and evidence. Accordingly, I will offer an argument in favour of replacing what I call “paradigmatic-state accounts” of self-deception with a dynamic view of the self-deceptive process. Once we have seen the reasons in favour of this move, and once the move is made, we will be in a position to see that mental states that are taken to be paradigmatic of self-deception should be liberalized, so as to include all the variations allowed by the evolving combinations of the two factors of motivation and evidence. I will focus fully on motivation and evidence as the two fundamental constraints singling out the phenomenon of self-deception. I will explain how the pull of motivation and the thrust of reality create a force field that is dynamic, often in motion, strained by the variations in these two components that may be triggered by various contingent or noncontingent factors. If we bear it in mind that self-deception amounts to a dynamic psychological space determined by both motivation and evidence, then a dynamic view becomes a promising option. And this significantly changes our approach to the attitudes that are the products of the self-deceptive process. Such a move turns out to be liberalizing regarding the mental states that can be instantiated in self-deception and whose possibilities should be brought fully into the picture.

Here is the plan of the paper. In section 2 I will discuss the inadequacy of the paradigmatic-state accounts of self-deception, thus creating the premises for moving forward toward formulating an alternative view. In section 3 I will explore the most promising alternative—namely, a dynamic, more liberal theory of the self-deceptive process, which amounts to my proposed positive view. I will also address one obvious objection to my positive view, and I will conclude by briefly indicating what my approach suggests in terms of refining and applying the conceptual mental categories that prove to be useful for capturing self-deception.

## 2. THE INADEQUACY OF PARADIGMATIC-STATE ACCOUNTS OF SELF-DECEPTION

As we have seen, the problem of including and explaining psychological tension in a satisfactory account of self-deception has led several scholars to advance proposals suggesting that we should adopt an “attitude adjustment approach” (cf. Deweese-Boyd, 2017) regarding the mental state that is the product of self-deception. Since full-blown self-deceptive belief as a final product of self-deception seems unable to explain why the self-deceiver experiences psychological tension, we can choose to adopt alternatives. One possible alternative is to posit some quasi-doxastic or other nondoxastic attitudes towards the self-deceptive proposition. Candidates are hopes, suspicions, doubts, anxieties (Edwards, 2013), “besires” (Egan, 2009), pretense (Gendler, 2007), and imagination (Lazar, 1999). On this view, the subject is not required to end up believing a desired proposition. Rather, the subject can either entertain the hope that the desired proposition is true or suspect that the desired proposition is not true, or the subject can have doubts about its truth value or maybe also anxiously fear the possibility of it being false. All these attitudes seem to be able to explain why the self-deceiver is in a highly tensive state of mind: since the evidence points to a certain truth value of the desired proposition—that is, it suggests that it is at least likely that it is false—the subject struggles with such evidence and tries to see if the desired proposition can be true.

Another alternative is to alter the content of the proposition believed. We can do so without incurring any static paradox of the kind associated with the traditional intentionalist models of self-deception. For instance, Funkhouser (2005) suggests that self-deceivers have a second-order belief about their believing that  $p$ , while they do not believe that  $p$  at all. Reality shows them that  $p$  cannot be true or that  $p$  is unlikely; however, they process the evidence in a motivationally biased way, so that they can at least believe that they believe that  $p$ .<sup>4</sup> That creates a tension between the false second-order belief that they believe that  $p$ , along with the dispositions associated with having it, and the first-order belief that not- $p$ , along with the dispositions associated with it. Bilgrami (2006) also suggests that the tension is due to a conflict, this time between a fully authoritative, true, second-order belief about a completely transparent, first-order belief that  $p$ , and another first-order belief that not- $p$  that is, however, not transparent, and that does not generate any second-order belief that one also has the first-order belief

that not- $p$ . Here the tension is created either by the conflict between the transparent, first-order belief that not- $p$  and the opaque, first-order belief that  $p$ , or by the clash between the true, second-order belief that one has a first-order belief that  $p$  and the first-order belief that not- $p$ , or both, along with the complications created by the further dispositions associated with all of these beliefs.

More recently, Lynch (2012) has claimed that “wholeheartedly believing what one wants to be true may be rare in self-deception (p. 440), and that we should look for a “more fine-grained way of capturing the attitudes of subjects towards propositions than can be accomplished with the coarser apparatus of belief” (p. 438). Thus, he claims that unwarranted degrees of confidence in  $p$  are enough to explain tension.

He also argues that we have self-deception as long as the subject puts some different degree of confidence in  $p$  and in not- $p$ , whereas any phenomena in which the subject avoids altogether the question whether  $p$  are best captured as “escapism” (Lynch, 2012, p. 446). He draws on Longeway (1990) and describes escapism as a defence against reality. According to Lynch, “deep conflict cases” best represent escapism: while in self-deception there is a cognitive tension due to the different degrees of confidence placed by the subject in  $p$  and not- $p$ , in deep conflict cases we have a more profound tension that is behavioural. The subjects here are taking a greater risk by acting upon their attitudes than they take by just speaking and thinking (p. 435). Often the subjects engage in avoidance behavior. Such actions give us reasons to think that the subjects are not simply struggling cognitively with  $p$  and not- $p$ , but that they are investing in  $p$  in a way that shows that they must have found a way to avoid any vacillation as to whether  $p$ . According to Lynch, this may not be self-deception any longer—or at least not a paradigmatic case of it. Interestingly, however, he admits that it may not be a necessary truth that self-deception contains tension, and allows for the possibility that, at times, the subject may become fully convinced of the desired proposition that  $p$  (p. 442). I will come back to this later in the paper.

In the context of distinguishing willful ignorance from self-deception, Lynch (2016) is even more explicitly interested in singling out paradigmatic cases of self-deception. He is aware that “philosophical analysis of a phenomenon is challenging enough at the best of times, but it becomes all the more difficult when there is disagreement over what the paradigmatic cases are. Such is the situation, unfortunately, with regards to self-deception” (p. 513). However, he thinks that “there are some features that are generally recognized to be present in paradigmatic self-deception,” such as

- (1) the subject’s encounter with evidence indicating that some true proposition, not- $p$ , is true; and
- (2) the strong desire of the subject that  $p$  be true.

So, according to (1) and (2), in paradigmatic self-deception the subject encounters *unwelcome* evidence, indicating that not- $p$ . Lynch adds that “beyond that, disagreement persists, particularly with regard to the subject’s epistemic/doxastic relation with the truth” (p. 513).

Lynch then claims that approaches regarding paradigmatic cases of self-deception can be sorted into three categories (p. 513-514):

- (a) unwarranted-belief accounts, where the subject ends up self-deceptively believing that  $p$  (Mele, 1997, is an example of such an account);
- (b) implicit-knowledge accounts (e.g., Bach, 1981), where the subject does not believe that  $p$ , but recognizes the truth of not- $p$ , while such knowledge is “shunned, ignored, or kept out of mind, and the subject acts in various ways” as if the subject believed that  $p$ , though other behaviour may betray the knowledge that not- $p$  (p. 514); and
- (c) intermediate accounts, where the subject both believes that  $p$  and believes that not- $p$  (Davidson, 1985), or where what the subject believes remains indeterminate (e.g., Funkhouser, 2009).

All these approaches agree on the discrepancy between the attitude toward  $p$  held by the self-deceiver and the attitude the self-deceiver should have, given the available evidence. Lynch adds that it seems to be characteristic of self-deception that the subject encounters the countervailing evidence, while this is not typical of wishful ignorance. For in willful ignorance the subject successfully manages to avoid evidence altogether and does this voluntarily and intentionally. This guarantees that willful ignorance lacks the encounter with evidence that is typical of self-deception.

However, Lynch thinks that this is no conclusive evidence that willful ignorance is not a kind of self-deception after all. For the former could, for example, be a nonparadigmatic case of self-deception. Lynch contemplates this possibility because he is persuaded that, if we had an analysis that did not rely on any of the three views listed above, we could get different results. If such an analysis were available, maybe we could be in a position to see that willful ignorance and self-deception are two of a kind, notwithstanding their obvious differences. I have reasons to think that the theory I will develop could be a promising candidate for being such a view. But before I move on to it, it is important to see why all the paradigmatic-state accounts are inadequate.

All the approaches seen thus far lead us to think that there must be a characteristic, paradigmatic state the self-deceiver is in. They do so by looking statically at the final product of the self-deceptive process. Even if Lynch seems to be more liberal in admitting that more states than are predicted by a paradigmatic-state account could be instantiated by the self-deceiver, he ends up adopting a rhetoric suggestive of the importance of singling out the paradigmatic state. Inso-



far as all these views look statically at the allegedly final product of self-deception, they can be described as snapshot theories: that is, it is as if they take a static, instantaneous picture of the mental state that a self-deceiver is in at a certain time  $t$ , presumably taken to be representative of the central phase of self-deception, and try to unpack its features so as to meet the constraint of tension.

By doing so, however, these views lay themselves open to a quite obvious objection: how should we deal with the empirical discovery that, with regard to the self-deceptive process, self-deceivers are in mental states other than the paradigmatic one? What if these other states are capable of meeting the constraint of tension, too? Obviously, the answer will be that all those accounts suffer from a restrictedness that puts us on the wrong track in our attempts to give a description of the phenomenon, which is rich enough to include other possible, often empirically instantiated, mental products. This is exactly what I think happens if we go beyond the biasing search for a paradigmatic state and head toward a more accurate focus on the self-deceptive process as a whole.

The analysis I will propose in the next section is given over precisely to showing how we should frame our view of self-deception, by considering the process, its moving forces, its dynamics, and all its possible, evolving, mental products. We will see that there is a multitude of highly tensive and unstable mental states that can be instantiated by a self-deceiver, and which have been unjustifiably excluded by paradigmatic-state accounts. However, as long as we persist in trying to freeze self-deception in a certain state instantiated at a certain time  $t$ , we lose the chance to give citizenship to all those other mental states. Insofar as a dynamic view is interested in focusing on the process instead, it leads us to liberalize the varieties of mental states in which a self-deceiver may be. Let me then move on to an outline of such a dynamic, liberal view.

### 3. A DYNAMIC, LIBERAL VIEW OF SELF-DECEPTION

As we have seen, virtually all scholars who study self-deception agree that there are two main forces that set the process in motion—namely, motivation that  $p$  be true, and the thrust of evidence that points to not- $p$ . Both factors are active together, and most likely they create a force field that can be, and in fact often is, highly dynamic. For obviously these factors can vary, and co-vary, depending on contingencies that can occur over time. For instance, there may be times when self-deceived subjects feel more strongly the pull of their motivation that  $p$  be true. Accordingly, they may engage more intensely in their biased treatment of the evidence or of the hypotheses associated with what the evidence suggests. At other times, instead, the encounter with evidence that not- $p$  may be more pressing, either because further evidence is provided or because the evidence already possessed is now seen in a less prejudiced light or else because the motivation that  $p$  be true as such is weakened by other intervening factors that have nothing to do with evidence (e.g., a diminished interest in  $p$  being true on the part of the subject).

There are also presumably noncontingent factors that can intervene in the dynamic of the process. For instance, these may be certain quite stable features of the subject's psychology. For example, if a subject is well trained to treat evidence impartially, even if motivation may have more or less momentarily suspended, shunned, or weakened that epistemic virtue, the latter may at some point just spontaneously strike back and correct, at least for a while, the motivationally distorted treatment of evidence. This may not guarantee the subject's complete exit from self-deception, as it may be the case that the motivation that  $p$  is true strikes back in turn once again. But it can certainly change the specific mental state self-deceiver is in at least temporarily. Perhaps, before such epistemic virtue made its claims felt, the subject placed more confidence in  $p$  than he or she does now. But if the drives of motivation rise forcefully again, the subject may revert once again to a higher degree of confidence in  $p$ .

Let me now expand on the possible mental states a self-deceiver can experience with regard to the self-deceptive process. It may not simply be the case that degrees of confidence vary, as Lynch (2012) correctly diagnoses. It may also be the case that a subject can at times temporarily reach even a state of full-blown belief that  $p$ , while, owing to the variations in the factors described, the same subject can revert to less than that, even to the antipode of believing not- $p$ . Of course, given the dynamics at work, none of these attitudes seems set to last. Or else there may be times when the subject reaches a false second-order belief that he or she has a first-order believe that  $p$ , while truly believing that not- $p$ , as Funkhouser (2005) requires. And there may also be moments when the subject is in an intermediate, indeterminate state of mind, possibly even recognizing it as such.

As we see, the dynamic force field that self-deception amounts to can easily instantiate a vacillation between  $p$  and not- $p$ , one that can (and often does) include a variety of attitudes toward  $p$  and not- $p$ . No attitudes can be excluded in principle, and what temporal extension and qualitative intensity such vacillation may have is a totally empirical question. It all depends on the individual subject, that subject's specificities, and the contingent evolution of his or her practical and epistemic interaction with evidence and reality.

This vacillation over time is best captured, I think, as an "attitude seesaw." There is no reason to exclude from it a priori any of the states that have been indicated as paradigmatic by different scholars, including those that they judge nonparadigmatic, such as escapism. There seems to be no a priori reason why a subject could not at times engage in escapism as well, perhaps emerging from it after a while. Or a subject might just start with escapism and then move on to other less extreme attempts to deal with reality. Nothing in escapism suggests that it cannot be irreversible, nor is there anything in self-deception to suggest that escapism cannot be one of its more-or-less temporary outcomes. The same interpretative line may be applied to willful ignorance, I think. Willful ignorance may well be a phase in the self-deceptive process, earlier or later on—one that can later be

reversed by new variations within the force field, or that can be entered from another state.

Lynch gets close to this when he says that a self-deceiver may at times go as far as self-delusion. He thinks that self-delusion is not a paradigmatic case of self-deception, but he shows an awareness of the variations to which I am drawing attention. My proposal, however, is more radical, and certainly more liberal: there is no need to persist in looking for a paradigmatic state when it is clear that, by its very nature, the process of self-deception can be, and often definitely is, highly dynamic, evolving and varying according to the opposing forces at work in it.

It is apparent that, in a liberal view such as the one I am putting forward, tension is fully preserved. First, such a liberal view preserves the tension that has been associated with a certain single state, or set of states, by paradigmatic-state-account theorists: if and when the state occurs, it has the tension that its proponents correctly attribute to it. In addition, however, there is a further tension that my account guarantees, and it is not clear that paradigmatic-state accounts take it into account: namely, the tension that a more or less prolonged vacillation produces over time.<sup>5</sup> Note that more self-reflective subjects could also experience a sort of metacognitive tension—that is, they find themselves on an attitude seesaw that they can intuit as such. To be sure, however, even if subjects do not reach this metacognitive level of self-reflection, the psychological effect of going on such a seesaw may make them experience tension, presumably of a more opaque kind.<sup>6</sup>

I said that it is a totally empirical question whether a subject enters into any of the possible states that the force field of self-deception allows. Equally, it is a totally empirical question how long such an attitude seesaw can last. Only a case-by-case empirical analysis of specific self-deceptive processes in individual subjects can give an answer to this question.

Let me then briefly recapitulate the main tenets of my proposal.

When one formulates a theory of self-deception, there seems to be no renouncing a couple of necessary constraints—namely, that

- (i) the process is triggered by a motivational state that leads the subject to wish that things stand in a certain, desired way (*p*);  
and
- (ii) the subject driven by such motivation struggles with evidence, or easily accessible evidence, that suggests how things really stand.

The clash between motivation and evidence makes the process highly tense from a psychological point of view. Psychological tension is a descriptive feature of self-deception that virtually all scholars refuse to renounce; rather, they all

visibly want it to be preserved, predicted, and accounted for. When this *desideratum* is combined with the bias in favour of the search for a paradigmatic state, scholars may feel the pressure to look for one single, characteristic state for self-deception, which is also highly tensive and unstable in itself. This combination of drives has led to several competing theories of the characteristic state of mind of self-deception, where what is at stake is the kind of state that can best account for instability, while also satisfying our search for a paradigmatic state of self-deception.

However, we will see that if we subscribe to (i) and (ii), we must accept the consequence that whatever mental state turns out to be compatible with the combined action of (i) and (ii) must be considered characteristic of self-deception. And clearly, there is no reason why we should not extend this consequence to mental states that may turn out to be even quite distant from the paradigmatic one.

Since a wide variety of states are compatible with (i) and (ii), my proposal is that

(A) we should liberalize the types of mental states that are representative of self-deception; and

(B) to avoid remaining on a misleading track, we should replace any paradigmatic-state accounts of self-deception with a dynamic view of the self-deceptive process.

Lynch (2016) provided an argument to show that willful ignorance is not a case of self-deception, whether paradigmatic or nonparadigmatic; therefore, it is a different kind of phenomenon. I think, however, that if we liberalize self-deception along the lines suggested, we are in a position to see that it may be the case that willful ignorance is sometimes a phase along the process of self-deception. Willful ignorance may well be the kind of phenomenon that Lynch diagnoses it as, and may fully retain its characteristic. The two phenomena need not be conflated. Yet willful ignorance could surface along the liberalized self-deceptive process as a possible stage of it. It is also important to note that my liberal view of the self-deceptive process does not rely on any of the three views of self-deception that, according to Lynch, could be used to distinguish willful ignorance from self-deception. Since I am not relying on any of the three, I am in a position to include willful ignorance as a possible stage of self-deception, and to do so by Lynch's own standards. For I am not proposing an unwarranted-belief account, where the subject ends up believing that *p* self-deceptively (Mele, 1997, is an example of such an account). I am not proposing an implicit-knowledge account either (e.g., Bach, 1981), where the subject does not believe that *p*, but recognizes the truth of not-*p*, while that knowledge is "shunned, ignored, or kept out of mind," and the subject's behaviour suggests that he or she believes that *p*, though other behaviour may betray the knowledge that not-*p* (p. 514). Nor do I propose an intermediate account, where the subject both believes that *p* and believes that not-*p* (Davidson, 1985), or where what the subject believes remains indeterminate (e.g., Funkhouser, 2009). Rather, I am proposing a view

whose focus is on a process within which more than is dreamed of by our paradigmatic-state-account philosophy of self-deception may happen.

My sense is that there might be room to apply a liberal solution to twisted self-deception as well. Twist-self-deception (Mele, 1999) is typically described as a case of self-deception where a subject does not end up believing what he or she desires or wants to be true. Rather, the subject ends up believing what he or she fears and, in any case, does *not* want to be true. Even if investigating this would lead me too far from my present purposes, it seems credible that a twisted self-deceiver may experience an attitude seesaw fully compatible with (i) and (ii). If this is correct, then twisted self-deception should cease to be seen as a nonparadigmatic case of self-deception, as, Lynch notes, it is still generally considered (2016, p. 513).

One may wonder whether any liberal view is too liberal after all. That is to say, does a liberal view risk being too broad, thus erring on the side of overinclusiveness? Were a liberal view to include more states than really fall within the self-deceptive kind, we would lose the unity of the phenomenon, as well as its specificity.

I do not think that overinclusiveness is a genuine risk for a liberal view, except when liberality is completely unrestricted. But I set (i) and (ii) as constraints for self-deception. Thus, I have reason to think that as long as (i) and (ii) are active, none of the states that can possibly be entered by subjects during their self-deceptive attitude seesaw is outside the phenomenon of self-deception. Of course, should either (i) or (ii) stop being active, then the subjects might well enter another kind of phenomenon altogether, such as permanent self-delusion or, in the event of their exiting self-deception completely, full adherence to reality.

I conclude with a final remark on what my analysis seems to suggest in terms of the adequacy of the conceptual apparatus we deploy for capturing self-deception. Even if I agree with Lynch (2012) that we should ameliorate our conceptual categories and look for more fine-grained concepts to capture psychological reality (p. 438), I also think that we should be ready to apply the categories we already have more liberally whenever our psychology shows a complexity that no single traditional mental category can capture in isolation. Sometimes, psychological complexity just requires us not to force and freeze it into snapshots. Rather it clearly invites us to look more closely at the richness embedded both in its static dimension and in its temporally evolving dynamic.

## ACKNOWLEDGMENTS

I am grateful to the audience of the Latin Conference of Analytic Philosophy held in Paris at the Jean Nicod Institute in July, 2013, for helpful questions on an early outline of this argument.

I am also indebted to two anonymous reviewers of this journal for their valuable comments on an early version of this paper.

## NOTES

- <sup>1</sup> Famously, the other paradox is the “dynamic paradox.” Since it is not immediately useful for my argument, although it is closely connected to it, I will not go into it in this paper.
- <sup>2</sup> I will not address the partitioning solutions to the static paradox here. See Deweese-Boyd (2017).
- <sup>3</sup> I thank one anonymous reviewer of this journal for pointing this extremely crucial aspect out to me.
- <sup>4</sup> Funkhouser (2005) relies on Nelkin (2002) here. They both think that self-deceiver reaches a false, second-order belief that he or she believes that  $p$  because that self-deceiver is primarily moved by a mind-directed desire to believe that  $p$ , not by a world-directed desire that  $p$  be true. For views that go in the same direction, see also Holton (2001) and Fernández (2013). For comments on the mind-directed desire, see also Pedrini (2012; 2013).
- <sup>5</sup> I am grateful to one anonymous reviewer of this journal for pointing out the importance of emphasizing this variety of tension.
- <sup>6</sup> Cf. Pedrini (2018).

## REFERENCES

- Bach, Kent, "An Analysis of Self-Deception," *Philosophy and Phenomenological Research*, vol. 41, 1981, p. 351–370.
- Bilgrami, Akeel, *Self-Knowledge and Resentment*, Cambridge, MA, Harvard University Press, 2006.
- Davidson, Davidson, "Deception and Division," in *Actions and Events*, E. LePore and B. McLaughlin (eds.), New York, Basil Blackwell, 1985.
- Deweese-Boyd, Ian, "Self-Deception," *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL: <https://plato.stanford.edu/archives/fall2017/entries/self-deception/>
- Egan, Andy, "Imagination, Delusion, and Self-Deception," in Bayne, T. and J. Fernández (eds.), *Delusion, Self-Deception, and Affective Influences on Belief-Formation*, Hove, Sussex, Psychology Press, 2009.
- Edwards, Sophie, "Nondoxasticism about Self-Deception," *Dialectica*, vol. 67, no. 3, 2013, p. 265–282.
- Fernández, Jordi, "Self-Deception and Self-Knowledge," *Philosophical Studies*, vol. 162, no. 2, 2013, p. 379–400.
- Funkhouser, Eric, "Do the Self-Deceived Get What They Want?" *Pacific Philosophical Quarterly*, vol. 86, no. 3, 2005, p. 295–312.
- , "Self-Deception and the Limits of Folk Psychology," *Social Theory and Practice*, vol. 35, no. 1, 2009, p. 1–13.
- Gendler, Tamara S., "Self-Deception as Pretense," *Philosophical Perspectives*, vol. 21, no. 1, 2007, p. 231–258.
- Holton, Richard, "What is the Role of the Self in Self-Deception?" *Proceedings of the Aristotelian Society*, vol. 101, no. 1, 2001, p. 53–69.
- Lazar, Ariela, "Deceiving Oneself or Self-Deceived?" *Mind*, vol. 108, no. 430, 1999, p. 263–290.
- Longeway, John, "The Rationality of Self-Deception and Escapism," *Behavior and Philosophy*, vol. 18, no. 2, 1990, p. 1–19.
- Lynch, Kevin, "On the 'Tension' Inherent in Self-Deception," *Philosophical Psychology*, vol. 25, no. 3, 2012, p. 433–450.
- , "Willful Ignorance and Self-Deception," *Philosophical Studies*, vol. 173, no. 2, 2016, p. 505–523.
- Mele, Alfred, "Real Self-Deception," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 91–102.
- , "Twisted Self-Deception," *Philosophical Psychology*, vol. 12, no. 2, 1999, p. 117–137.
- , *Self-Deception Unmasked*, Princeton, Princeton University Press, 2001.

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

Pedrini, Patrizia, "What Does the Self-Deceiver Want?" *Humana.Mente: Journal of Philosophical Studies*, Pedrini, Patrizia (ed.), vol. 5, no. 20, 2012, p. 141-157.

———, *L'autoinganno. Che cos'è e come funziona*, Rome, Laterza, 2013.

———, "The 'Crux' of Internal Promptings," in Patrizia Pedrini and Julie Kirsch (eds.), *Third-Person Self-Knowledge, Self-Interpretation, and Narrative*, Springer, 2018.



# SELF-DECEPTIVE RESISTANCE TO SELF-KNOWLEDGE

GRAHAM HUBBS

ASSOCIATE PROFESSOR, UNIVERSITY OF IDAHO

## ABSTRACT:

Philosophical accounts of self-deception have tended to focus on what is necessary for one to be in a state of self-deception or how one might arrive at such a state. Less attention has been paid to explaining why, so often, self-deceived individuals resist the proper explanation of their condition. This resistance may not be necessary for self-deception, but it is common enough to be a proper explanandum of any adequate account of the phenomenon. The goals of this essay are to analyze this resistance, to argue for its importance to theories of self-deception, and to offer a view of self-deception that adequately accounts for it. The view's key idea is that, in at least some familiar cases, self-deceived individuals maintain their condition by confusing a nonepistemic satisfaction they take in their self-deceived beliefs for the epistemic satisfaction that is characteristic of warranted beliefs. Appealing to this confusion can explain both why these self-deceived individuals maintain their unwarranted belief and why they resist the proper explanation of their condition. If successful, the essay will illuminate the nature of belief by examining the limits of the believable.

## RÉSUMÉ :

Les explications philosophiques de l'auto-illusion ont eu tendance à mettre l'accent sur ce qui est nécessaire pour que quelqu'un soit considéré comme étant sous l'emprise de l'auto-illusion ou encore sur la façon dont quelqu'un parvient à un tel état. Moins d'efforts ont été dirigés vers les raisons pour lesquelles, si souvent, les individus sous l'emprise de l'auto-illusion opposent une résistance à l'explication véritable de leur condition. Cette résistance n'est peut-être pas essentielle à l'auto-illusion, mais elle est suffisamment courante pour constituer un explicandum approprié pour tout traitement adéquat du phénomène. Cet essai a pour buts d'analyser cette résistance, de défendre son importance pour les théories de l'auto-illusion, et de proposer une conception de l'auto-illusion qui en rend compte de manière adéquate. Cette conception repose sur l'idée suivante : au moins dans certains cas connus, les individus sous l'emprise de l'auto-illusion maintiennent leur condition en prenant la satisfaction non-épistémique qu'ils retirent de leurs croyances illusoire pour la satisfaction épistémique qui caractérise les croyances justifiées. C'est en faisant appel à cette confusion que l'on peut expliquer à la fois pourquoi ces individus conservent leur croyance infondée et pourquoi ils opposent une résistance à l'explication adéquate de leur condition. Si tant est qu'il y parvienne, cet essai éclairera la nature de la croyance en examinant les limites du croyable.

*What a fool believes he sees  
No wise man has the power to reason away  
– Michael McDonald and Kenny Loggins,  
“What a Fool Believes” (1978)*

A fool who cannot be swayed by reason may not be self-deceived, but often he is. Consider the fool of McDonald and Loggins’s song. He meets an old acquaintance who, he thinks, once longed for him romantically and might feel the same way again. She never had such feelings, and she never will; perhaps out of pity, she apologizes to the fool for the fact. Loggins and McDonald tell us that “as he rises to her apology, anybody else would surely know/ he’s watching her go.” Anybody else would surely know that she is going, and for good, because this is what the evidence overwhelmingly suggests. In spite of this evidence, the fool believes that she will return to him someday. On accounts of self-deception that Dion Scott-Kakures (2002) calls *deflationary*, little more needs to be said about the fool to depict him as self-deceived.<sup>1</sup> On Alfred Mele’s version of deflationism, we need only to add that the fool’s false belief results from him treating the relevant data in a motivationally biased way (Mele, 1997, p. 95). On Mark Johnston’s version, we need only to add that the belief is the result of a distinct sort of mental tropism (Johnston, 1988, p. 86). On Annette Barnes’s version, we must only add that what Johnston calls a tropism is specifically a form of anxiety avoidance and that the fool underestimates the causal impact that this has on the maintenance of his belief (Barnes, 1997, p. 117). These views are deflationary because they characterize the apparent goal-directedness of self-deception non-intentionally; they thus contrast with intentionalist views, such as Donald Davidson’s (Davidson, 2004a; 2004b; 2004c). Scott-Kakures’s own account takes a middle way between deflationism and intentionalism: he claims that in order to be self-deceived, the fool would have to engage in reflective reasoning that maintains his unwarranted belief (Scott-Kakures, 2002; 2009). On this view, the self-deceived fool would not intentionally bring himself into his condition, but nor would he fall into it as the result of some blind motivation.

Although these views disagree on the roles of intentions and reasoning in producing self-deception, they would all agree that the self-deceived fool believes that his acquaintance once had romantic feelings towards him and may feel this way again. Some recent accounts have called into question whether the self-deceived fool has any such belief. Eric Funkhouser and David Barrett would claim that, if the fool holds such a belief, then he is self-deluded but not self-deceived (Funkhouser, 2005; Funkhouser and Barrett, 2016; 2017). To be self-deceived, on their view, the fool would have to behave in a way that suggests he does not believe he has a chance romantically with his acquaintance. This behaviour is at odds with what the fool asserts—namely, that she might one day long romantically for him. To account for his nonlinguistic behavior, Funkhouser and Barrett would recommend that we ascribe to the fool the belief that she does not and will not care romantically for him; to account for his assertion to the contrary, we are to characterize him as falsely believing that he believes she

might want him one day. The difference here between self-delusion and self-deception is not merely terminological. Funkhouser explicitly asserts that self-deception is philosophically interesting in a way that self-delusion is not—in due course I will explain why. Jordi Fernández does not go as far as Funkhouser and Barrett, for Fernández allows cases they would describe as mere self-delusion to count as instances of self-deception (Fernández, 2013). Nevertheless, Fernández agrees with Funkhouser and Barrett that these cases are not of much philosophical interest; to be philosophically interesting, the fool must be characterized roughly in line with Funkhouser and Barrett's definition of self-deception. This requires that the fool hold the first-order belief that he has no romantic future with his acquaintance while simultaneously holding the false metabelief that he believes they will one day be together.<sup>2</sup>

Now it is one thing to explain what it takes for the fool to be self-deceived, which is a task pursued by all of the accounts just mentioned; it is another to explain why, so often, no one, no matter how wise, has the power to reason the self-deceived fool out of his condition. One might produce an intuitively plausible characterization of being self-deceived and perhaps even a causal account of how one arrives at the condition without explaining why, so often, self-deception resists rational revision. The resistance I have in mind may not be necessary for self-deception, but I believe it is common enough to be a proper explanandum of any adequate account of self-deception. The central goals of this essay are to analyze this resistance, to argue for its importance to theories of self-deception, and to offer a view of self-deception that adequately accounts for it.

To bring this resistance squarely into view, consider the following specific but straightforward way in which the wise man in the song might fail in his attempts to correct the self-deceived fool. Suppose the wise man points all of the relevant evidence out to the fool and asserts that, on its basis, it is reasonable to conclude that his acquaintance has never wanted him and will never want him. If the fool accepts this, then, at least for that moment, on any account of self-deception, he is no longer self-deceived. Assume this does not happen: the fool insists that, in spite of the evidence, his acquaintance wanted him once and may want him again. The wise man shakes his head and tells the fool that he says this only because admitting the opposite would be too painful. The wise man is wise, so he makes no assumptions about the fool's unconscious beliefs or higher-order beliefs; he simply notes and explains what the fool will not admit.

The fool can go one of two ways at this point. On the one hand, he can give in and accept the wise man's explanation. This would seem to force the fool to acknowledge that, as a first-order matter of fact, the acquaintance will never want him, but perhaps he can resist this; perhaps he can acknowledge that the evidence does not warrant what he claims about her, that he says what he does only because admitting otherwise would be too painful, yet still maintain that she has wanted him and might want him again. This case would be an instance of what some have called *epistemic akrasia*.<sup>3</sup> Like those who perform akratic actions, epistemically akratic individuals know they are being unreasonable; in

spite of this, they go on believing their unwarranted beliefs. We will consider this epistemic condition in several places over the course of the argument. At present, let us set it aside and suppose that the fool goes the other way and rejects the wise man's explanation. Without straightforwardly lying, the fool denies that there is anything he is failing to admit. The reason he says she has wanted him and might want him again has nothing to do with his desires, he asserts, though he grants that he does desire her affection. The reason he says these things, he claims, is that they are true. The fool, of course, is wrong, both about his acquaintance's feelings and about the reason he insists he knows what she feels. Call the former error *self-deceptive resistance to evidence*; this topic has been well examined in the literature on self-deception. The latter error, which I will call *self-deceptive resistance to self-knowledge*, has received less attention.<sup>4</sup> This resistance to self-knowledge constitutes a failure of self-explanation: by committing the error, the fool resists the proper explanation of why he claims (or, depending on one's view of self-deception, believes) that his acquaintance may one day want him again. This resistance is no accident, no simple mistake—as the wise man knows, it is part of an overall epistemic condition that enables the fool to go on believing what he does about his acquaintance.

The third section of this essay will provide an account of this self-deceptive resistance to self-knowledge. To work up to it, I will begin by discussing the views and arguments in Fernández (2009) and Funkhouser (2005).<sup>5</sup> Fernández might claim that my topic is not an interesting sort of self-deception, and Funkhouser might deny that it even counts as self-deception. My goal in responding to these complaints is to defend the claim that self-deceptive resistance to self-knowledge warrants “self-deceptive” as part of its label and to argue that this resistance is a proper explanandum for any account of self-deception. I will close this first section by showing that Funkhouser's and Fernández's accounts cannot explain the phenomenon. I will then turn to a discussion of deflationary and intentionalist views of self-deception. I will borrow from Scott-Kakures's discussion of deflationism and intentionalism, which I find insightful. Scott-Kakures's account improves upon these views, but in the end, I argue, it too lacks the resources to explain self-deceptive resistance to self-knowledge. The next section presents my account of the phenomenon. Its key idea—which, to the best of my knowledge, is novel to the literature on self-deception—is that self-deception involves a distinct sort of *confusion*. The sort of confusion I have in mind occurs when a person takes one thing to be something that it is not.<sup>6</sup> The fool, I will argue, confuses what I shall call the *epistemic satisfaction* of believing what is warranted with what I shall call the *thumotic satisfaction* that results from believing what he wants to be true. I will explain this distinction as well as my introduction of the term “thumotic” in section 3. At present, may it suffice to say the following: the fool thinks his belief is satisfying because he thinks it is warranted—i.e., he takes his satisfaction in the belief to be epistemic satisfaction—but he is confused, for the belief involves a different sort of satisfaction, the satisfaction one paradigmatically finds when one is esteemed or valued or admired in a way one wants, which I am calling thumotic satisfaction. My account here will draw upon recent work on the neurobiology of emotion by

Lisa Feldman Barrett (2017). I close the essay with some brief remarks on how the confusion account can explain so-called twisted cases of self-deception; this, I hope, will further elaborate the view.

The success of the confusion account turns ultimately on the accuracy of the characterization of the nature of belief in section 3. A key element of this characterization is the claim that the pleasure a belief might bring can partially determine whether or not one holds that belief, sometimes over and even against the belief's warrant. I think this is simply a fact, one that is readily demonstrated by the hopeful beliefs of some football supporters. Supporters may believe that this is the year for their team. The proper explanation for their holding this belief may in part be a matter of the evidence they consider—the season is young, and the team has not yet shown that it is bound to fail. This may, however, not be the whole story. To tell that story, one might also need to add that the hope brought by the belief is far more pleasant to the supporter than the doxastic alternatives, and that this too is part of the full explanation of the football supporters holding their belief. Because my account turns on a substantive view of some nonevidential but nonaccidental causes of belief maintenance, my discussion here may prove relevant to the reader who is unconcerned with self-deception but who is interested in doxastic voluntarism.<sup>7</sup> This essay will not investigate the possible connections between the confusion account and debates about believing at will. I mention the latter simply to flag a focus that it shares with the present discussion: the limits of the believable, as determined by the nature of belief.

## 1. AVOWAL AND CONFLICTING BEHAVIOUR IN SELF-DECEPTION

Let us start by considering the fool as we did at the outset—namely, as believing his acquaintance may one day want him again. As already noted, on Funkhouser's view this fool is self-deluded, not self-deceived; on Fernández's view this fool demonstrates only one of the two "remarkable features" of self-deception (Fernández, 2013, p. 381). The feature the fool demonstrates is what Fernández calls the *normativity of self-deception*, which is manifested by the fact that we are supposed to find his epistemic condition objectionable. I expect Funkhouser would agree that self-delusion involves some sort of normative failure. To be self-deceived, however, on Funkhouser's view, the fool would have to demonstrate the second of Fernández's features of self-deception—namely, the conflict feature.<sup>8</sup> There does not appear to be any conflict in the fool's overall psychological condition: he sincerely believes that his acquaintance may one day want him again. Both Funkhouser and Fernández think this case is importantly different from cases in which there is a conflict between a person's behaviour and what that person claims to believe. Consider, for example, Funkhouser's case of the self-deceived wife (Funkhouser, 2005, p. 302). She has ample evidence that her husband is having an affair with a friend. She claims to believe that he is faithful, but she behaves to the contrary; for example, she avoids driving by the friend's house at times when it seems likely that her husband might be there. Funkhouser thinks that the only way to make sense of the wife's driving behaviour is to attribute to her the belief that her husband is having an affair.

To make sense of her avowal to the contrary, we describe the wife as falsely believing that she believes her husband is faithful.

Adopting Funkhouser's labels for just this paragraph, we see that self-delusion and self-deception present different problems to explain. The problem with self-delusion is that evidence is not considered impartially, which leads to irrational belief formation. The problem with self-deception, by contrast, is that of squaring one's behaviour with one's conflicting avowals. There are at least two reasons one might find the latter problem more interesting than the former. Although Funkhouser and Fernández do not think of the problems this way, one might think the latter problem entails the former. If we treat the wife's avowal as an expression of what she believes, then not only does her behaviour fail to square with her avowal, but her belief also fails to square with the evidence, which includes her own behaviour. On this way of conceiving of the issue, self-deception is more problematic than self-delusion—it involves two problems, not just one—which might license finding the former more problematic than the latter. Funkhouser and Fernández do not characterize self-deception this way, however; again, they deny that the wife believes what she avows. I suspect this is due to their tacit acceptance of the following two principles of belief ascription. The first is that actions speak louder than words, so the tension between what the wife says and what she does is to be resolved by a belief ascription that explains her deeds, not her words. The second is a version of the principle of charity that prevents, as far as is possible, ascribing two flatly contradictory beliefs to the same person. They, it seems, take characterizing the wife's avowal without violating these principles to be more challenging than explaining irrational belief formation; this is the second possible reason one might find self-deception more interesting than self-delusion.<sup>9</sup>

I have no interest here in adjudicating what one should find more or less philosophically interesting, but I do want to argue, *contra* what is at least implied by Funkhouser's and Fernández's accounts, that there is something philosophically interesting about self-deceptive resistance to self-knowledge. First of all, even if resistance is not a necessary feature of self-deception, it is not uncommon for self-deceived individuals to be disposed to resist the proper explanation of their self-deceived condition. Funkhouser's cases suggest as much. For example, he describes the self-deceived wife as follows: "She laughs off the concerns of her girlfriends, and thinks to herself that Tony is certainly a faithful husband" (Funkhouser, 2005, p. 302). This demonstrates her self-deceptive resistance to the evidence, and her girlfriends lack the power to reason this resistance away. Now imagine one of these girlfriends saying, "Look, the only reason you insist Tony is faithful is that you can't bear to think otherwise, even though the evidence overwhelmingly suggests that he is cheating on you." It is surely possible, if not likely, that the self-deceived wife will dismiss this just as she laughed off their other concerns. This resistance to self-knowledge, then, is something that even a restricted view of self-deception such as Funkhouser's or Fernández's should seek to explain.

It is not clear, however, that they can explain the resistance without attributing a pair of contradictory beliefs to the wife. As we have already seen, they explicitly characterize her as believing that she believes that her husband is faithful. If she is not epistemically akratic—set this possibility aside—then she also must at least tacitly believe that her belief about what she believes is warranted. To believe the higher-order belief is unwarranted is simply to believe that she lacks the first-order belief; again, it is a central component of their view that she believes she has the favourable first-order belief about her husband. They are committed, however, to explaining her resistance to self-knowledge by characterizing her as also and simultaneously believing that her higher-order belief is unwarranted. This follows from the way they explain avoidance behaviour. They explain the wife's avoidance of driving by her friend's house by characterizing her as believing that her husband is cheating. Her resistance to self-knowledge also involves an object of avoidance: she is avoiding the fact that her belief that she believes her husband is faithful is unwarranted. According to their explanatory approach, she can avoid this fact only by believing that it is, indeed, a fact, so they are committed to characterizing her as believing that her higher-order belief about what she believes about her husband is unwarranted. If we are to avoid such contradictory ascriptions—and, again, this seems to be a principle that guides Funkhouser and Fernández—then this false-higher-order-belief approach to self-deception cannot characterize self-deceptive resistance to self-knowledge.<sup>10</sup>

## 2. INTENTIONALISM, DEFLATIONISM, AND REFLECTIVE REASONING

Funkhouser and Fernández are not alone in thinking that it is the internal psychological conflict of self-deception that makes the condition philosophically interesting. For example, in summarizing self-deception, Davidson says the following: “Finally, and it is this that makes self-deception a problem, the state that motivates self-deception and the state it produces coexist; in the strongest case, the belief that  $p$  not only causes a belief in the negation of  $p$ , but also sustains it” (Davidson, 2004c, p. 208). Unlike Funkhouser and Fernández, Davidson thinks that the conflict of self-deception exists between a pair of inconsistent first-order beliefs, one of which sustains the other. Nevertheless, all three agree that the internal psychological conflict of self-deception, whatever exactly it is, is the condition's most philosophically interesting element. Davidson's attempt to explain this conflict proceeds as follows. The self-deceived individual brings about an unwarranted belief through intentional activity, such as selectively attending to evidence that supports that belief. To explain how the unwarranted belief can be held even as the self-deceived individual goes on also holding its warranted contradictory, Davidson claims that the individual's mind is divided. Davidson insists that this division is functional;<sup>11</sup> he does this so as to allay worries of what we might call “homuncularism”—i.e., the idea that partitioning the mind requires partitioning the person into separate agents, each with its own agenda concerning what the person should believe.<sup>12</sup> The challenge of explaining the conflict of self-deception, then, is met by dividing the mind and attributing the contradictory beliefs to the separate parts.

Davidson's view is intentionalist because it claims that the self-deceived individual does something intentionally to bring about the belief that, though unwarranted, that individual prefers to hold. One need not divide the mind as Davidson does to hold an intentionalist view; indeed, Kent Bach holds an intentionalist view that agrees with Funkhouser, Barrett, and Fernández—namely, that the self-deceived individual does not believe the unwarranted but preferred belief (Bach, 1981).<sup>13</sup> Intentionalists provide a straightforward account of the way in which self-deceived individuals' motives can influence their beliefs: the motives produce intentional activities that allow the agents to ignore (if not eliminate) their unfavourable beliefs. A challenge for all intentionalists, then, is to explain how self-deceived individuals are able to do this while simultaneously acknowledging that the unfavourable beliefs are warranted by the evidence; as Scott-Kakures puts the point, the intentionalist is “under great pressure to claim that the contrary evidence is not *really* believed or that such evidence is somehow forgotten or otherwise pushed into inaccessibility” (Scott-Kakures, 2002, p. 577, italics in original). Scott-Kakures does not present the point as counting decisively against the intentionalist, but it provides a good reason for wondering whether one can account for self-deceived beliefs or avowals in less cognitively robust terms.

This is the project of the deflationist. There are, as noted at the outset, all sorts of deflationist views, but they all agree that self-deception is a sort of biased belief formation. Most views characterize this bias as motivated, but some deny that this is necessary (e.g., Patten, 2003). The challenge for these accounts, Scott-Kakures argues, is to distinguish self-deception from other sorts of biased belief formation, which do not seem to involve the characteristic tension of self-deception. The worry here is similar to the one that motivates Funkhouser and Barrett to distinguish self-delusion from self-deception.<sup>14</sup> For Scott-Kakures, however, the point is not that deflationist accounts pick out an uninteresting human condition; rather, it is that they cannot distinguish self-deception from the representational state a brute may enter under the effects of motivational bias. Scott-Kakures's example of such a brute is Bonnie the Cat, who is usually good at distinguishing cat-food sounds from non-cat-food sounds but who overreacts to non-cat-food sounds when she is very hungry. It seems natural to say that Bonnie's representations are affected in these situations by a motivational bias towards finding cat food; the challenge the deflationist faces is explaining how self-deception differs from Bonnie's biased representations.

The different limitations of intentionalism and deflationism are also shown by the difficulty each has in explaining self-deceptive resistance to self-knowledge. If the intentionalist is correct, then this resistance is intentional. For example, when the fool claims that he believes his acquaintance will return to him because this is what the evidence warrants, he is, on the intentionalist view, intentionally avoiding the correct explanation of his belief. Now the goal of avoiding a given sort of explanation is quite cognitively sophisticated; it is the sort of thing lawyers might do on behalf of their clients, and it is not something that Bonnie the Cat can so much as attempt. It is not clear how a person can intentionally



pursue such a sophisticated goal while sincerely denying doing so, but the intentionalist is committed to characterizing the self-deceived fool in this way, for such fools will sincerely deny the proper explanation of their resistance. The only hope for the intentionalist here, it seems, is the homuncularism that even Davidson wants to avoid. It is not clear that the deflationist fares any better, however; given the cognitive sophistication involved in avoiding the proper explanation of one's self-deceived condition, it is not clear how it could be pursued in anything less than an intentional manner. Deflationism may be an attractive alternative to intentionalism when it comes to explaining self-deceptive resistance to evidence, but its resources are strained if it is called on to explain the cognitively sophisticated resistance to self-knowledge.

Scott-Kakures's own view takes a middle path between intentionalism and deflationism.<sup>15</sup> According to Scott-Kakures, Bonnie the Cat is not self-deceived, because she plays no active role in generating or maintaining her biased representations. Self-deceived individuals, by contrast, play an active role in at least maintaining their unwarranted beliefs through reflective reasoning. Scott-Kakures grants that the unwarranted self-deceptive belief may be generated by non-self-deceptive means; what matters—and on this point, he is surely right—is the way in which the belief is maintained, not the way in which it is initially formed. On his view, for the fool in the song to be self-deceived, he must at least occasionally reflect on his beliefs about his acquaintance's feelings. When the fool does this, he must fail to reason clearly: his reasoning must be swayed towards believing what he prefers instead of what is warranted by the evidence.

Scott-Kakures explains this capacity for wrongful reasoning in terms of the psychological account of “pragmatic hypothesis testing.”<sup>16</sup> According to this account, we are to make sense of the fool's reasoning by identifying the different costs associated with false-positive and false-negative hypotheses concerning his acquaintance. Consider the proposition, “My acquaintance has had and may again have feelings for me,” conceived of as a hypothesis. If the fool settles in favour of this proposition and it is false, it is a false positive; if the fool settles against this proposition and it is true, then it is a false negative. From the fool's perspective, a false negative of this hypothesis would be worse than a false positive. If he does not believe it, but it turns out to be true, he will miss his chance with his acquaintance; if he believes it, but it turns out to be false, he will have given himself every chance with her, and he will have enjoyed temporarily believing in her affection.<sup>17</sup> This asymmetry of costs produces an asymmetry of acceptance thresholds for the positive and negative of the hypothesis: effectively, the fool tests the positive for sufficiency and the negative for necessity. He thus reasons his way to the false conclusion that he still has a chance with his acquaintance. Scott-Kakures is not the only writer to claim that reasoning plays a critical role in self-deception; he cites David Sanford (1988) as agreeing with him on this point. Sanford, however, depicts the reasoning involved as a false rationalization for one's belief, which serves to mask the genuine causes of one's self-deceived belief. Scott-Kakures argues that the reasoning conducted by the self-deceived individual functions not to disguise the genuine causes of the rele-

vant belief, but actually to maintain the belief; he says of the self-deceived individual that “her putative reasons are her reasons and they do, as a causal matter, explain why she has come to believe as she does” (Scott-Kakures, 2002, p. 597).<sup>18</sup>

It is surely correct that the reasoning performed by self-deceived individuals can sustain their unwarranted beliefs and thus play a causal role in the maintenance of these beliefs. It is not clear, however, that this can explain all of the beliefs that may be involved in self-deceptive resistance to self-knowledge. The putative reasons that constitute the fool’s false self-explanation may indeed have the function of sustaining his self-deception. The existence of these self-explanatory beliefs, *contra* Scott-Kakures’s account, is not explained by acts of reasoning that might have them as their conclusions. Their existence is explained by the fact that it would be too painful for the fool to admit the truth, either about what the evidence warrants believing about his acquaintance or about why he believes she will return to him one day. Scott-Kakures might think he can capture this fact by reiterating his point about acceptance thresholds, but appeal to these thresholds makes sense only if we think of the fool’s false self-explanatory beliefs as something he might treat as hypotheses. This may not be impossible, but there are certainly cases (which, I suspect, are typical) in which the causal role is reversed. We can think of the fool’s false self-explanation as comprising beliefs he has not adopted due to biased reasoning and does not sustain by such reasoning. For this fool, his self-beliefs do not have the status of hypotheses; they are not held on the basis of an examination of their epistemic credentials. The beliefs are prior to any reasoning about his acquaintance and they are prior to any reasoning about the cause of his beliefs about her. His false self-explanation expresses a presumption about the legitimacy of his reasoning, both about his acquaintance and about the cause of his beliefs about her. Commitment to such a false self-explanation is a condition of his reasoning counting as self-deceptive; it is not a product of any such reasoning.

If this is possible, then not even Scott-Kakures’s account can fully explain self-deceptive resistance to self-knowledge. For Scott-Kakures, this resistance would have to manifest itself as an openness to the possibility of the truth of the proper self-explanation, which is then resisted. If the fool is as I have described him, the object of resistance is the possibility that the proper self-explanation is true—again, his preferred self-explanation is taken as fact, not treated as a hypothesis. By focusing on the role that reasoning can play in self-deception, however, Scott-Kakures’s account points us in the right direction. The key is to see how we can become confused about what governs our mental activity as we rationalize our beliefs and actions, for it is in terms of this confusion, I think, that we are able to explain self-deceptive resistance to self-knowledge. I turn now to this.

### 3. CONFUSION AND RESISTANCE

My account starts with what I take to be an uncontroversial point: reasoning is a sort of agential mental activity.<sup>19</sup> As such, reasoning shares some metaphysical characteristics with other sorts of agential activity, including intentional

bodily action. Our agential activity, both mental and bodily, is a product of our nature as rational living beings. As specifically *living* beings, we have a general tendency to do what is satisfying and not to do what is dissatisfying. This tendency can affect the body by leading us to indulge in sexual, gastronomic, or drug-induced pleasure or to avoid arduous or fearful situations. This tendency can also affect the mind, and in similar ways—for example, in the same way that we have a tendency against putting ourselves in fearful or sad situations, we have a tendency against thinking fearful or sad thoughts. I am not denying that we sometimes tend towards or even fixate on unpleasant thoughts—more on this in section 4. At present, I am simply noting something that I think Johnston’s tropistic account and Barnes’s anxiety-avoidance account get right: there is a tendency of the mind to avoid thinking unpleasant thoughts.<sup>20</sup>

According to the constructivist account of emotion advanced by Lisa Feldman Barrett (2017, ch. 4), the basic affective components of all of these tendencies are surprisingly simple. They are the product of interoception and are two-dimensional: they all are valenced and so are more or less pleasant, and they all involve some level of arousal, the lowest being exemplified by the lethargy characteristic of deep depression. On her view, the affective components of anger and fear are the same—both are more unpleasant than pleasant and involve arousal rather than calmness or lethargy. The difference in subjective experience between these two emotional categories is primarily a matter of nonaffective interpretation, which is heavily influenced by a person’s understanding of emotional concepts. Affect can thus be common across emotional categories; it can also be common across perceptions of fact and mere contemplation of thoughts. On her account, the affect underlying severe ophidiophobes’ aversion upon seeing actual snakes has the same components—displeasure and arousal—as their aversion to merely thinking about snakes. Combining this line of thought with that of the previous paragraph yields the following: there is a tendency, in many if not most of us, to avoid unpleasant and arousing affect, whether that affect is part of one’s body’s engagement with the external world (including, e.g., actual snakes) or is merely part of one’s mind (including, e.g., mere thoughts of snakes). This holds for other agential tendencies as well (e.g., towards pleasure).

I take the following also to be uncontroversial: many if not most of us have a tendency to want the approval of at least some humans, whose opinion of us we value in some way. I call the satisfaction at which this tendency aims *thumotic*. The root of this term is the Ancient Greek “*thumos*,” which is commonly translated into English as “spirit” and is sometimes understood narrowly as a psychological faculty of anger or, still more narrowly, revenge. Not hamstrung by any English version of the term, I intend “thumotic” more broadly, in line with the variety of uses of “*thumos*” in Ancient Greek—I use it to pick out the sorts of self-satisfaction that follow from one’s perception of the positive judgments of others and the sorts of dissatisfaction that follow from corresponding negative judgments.<sup>21</sup> The satisfaction of many sorts of pride is, in my sense, thumotic. When one feels good in victory, the satisfaction is thumotic; when one basks in the praise of a superior, the satisfaction is thumotic. These are cases in which one

feels pleasant affect at having achieved a particular social standing—the good feeling follows from how one perceives oneself to be judged by others. Various sorts of admiration can also give rise to thumotic satisfaction: teachers may take thumotic satisfaction in the admiration of their students; someone fit and smartly dressed may take thumotic satisfaction in the desirous glances of others. The affect in all of these cases is pleasant; it seems typically to be arousing as well. Shame, guilt, and other emotions of social inadequacy or social failure involve thumotic dissatisfaction. The affect of these is unpleasant; when they manifest as a sort of sadness or depression, they are also low on arousal, not stimulating.

These remarks on thumotic satisfaction and dissatisfaction are, admittedly, quite rough—one might worry that these are ad hoc categories, and one might wonder what else belongs to them outside of the examples I have just listed. These are worthy concerns, which I will not address here. It will presently suffice if I can adequately distinguish the satisfaction one can take from the esteem of others from the satisfaction one can take in learning or discovering or understanding something, which I call *epistemic* satisfaction. The latter may seem like a strange sort of satisfaction, but it is found, I believe, in a variety of mundane mental activities. Consider the philosopher who wonders how to explain why we cannot form beliefs at will. The perplexity of the person troubled by this topic is accompanied by a distinct sort of epistemic tension, which that person seeks to resolve with a satisfying solution. Or consider the person who loves detective stories; this person takes pleasure in the explanatory tension posed by a good mystery and seeks to resolve this tension with a satisfying resolution to the story. In both of these cases, an explanatory tension persists until an answer is found that satisfactorily resolves the tension. The satisfaction in each case is epistemic, for it is satisfaction at having arrived at what one takes to be an understanding of the matter at hand. In these cases, the affect is pleasant and calming, for it resolves the arousal of the epistemic tension. To be clear, I am not claiming that all, or even most, epistemically satisfying states involve a perceived *feeling* of satisfaction. The satisfaction in question is often experienced as a sort of epistemic *ataraxia* or tranquility—many, perhaps most, epistemically satisfying states are those from which a feeling of epistemic dissatisfaction is absent.<sup>22</sup> The dissatisfaction in question is the genus to which the negative affect of cognitive dissonance belongs—satisfaction can simply be the absence of this feeling of negative affect.<sup>23</sup> Although it is easiest to exemplify epistemic satisfaction by focusing on cases in which a felt perplexity is resolved, no such feeling is necessary for a belief or explanation to be epistemically satisfying.<sup>24</sup>

If the difference between thumotic and epistemic satisfaction is not already clear, note that people sometimes seek to satisfy explanatory tensions even though they believe the resolution they seek might be otherwise unpleasant. Consider the person who is betrayed by a friend and who wants, among other things, an explanation that makes sense of the friend's betrayal. It seems wrong to think that this person aims at an overall condition of pleasure in wanting to understand the cause of the friend's betrayal, yet the betrayed person still wants an explanation that makes what the friend has done intelligible. The betrayed person wants an

explanation that is epistemically satisfying, even if it is one that also involves a feeling of disappointment, a sense of being insufficiently valued by one's friend. This latter feeling is one of thumotic dissatisfaction. Indeed, it may involve anger, perhaps at the friend for the betrayal, or perhaps at oneself for trusting someone who turns out to be untrustworthy. Even if the feeling is not the unpleasant arousal of anger, the dissatisfaction is thumotic: it is the displeasure, whatever the level of arousal, that follows from the negative evaluation implied by the friend's act of betrayal. The betrayed person may thus pursue an explanation that is epistemically satisfying even while expecting it will be thumotically dissatisfying.

Although we are capable of distinguishing epistemic satisfaction from thumotic satisfaction, we do not always exercise the capability. On occasion, the failure to exercise the capability can lead to a distinct sort of confusion, which in turn, I argue, can produce self-deception. I should be clear here about what I mean by "confusion" and its cognates. As stated in the introduction, as I understand the phenomenon, confusion occurs when a person takes one thing to be something that it is not. One may confuse instances of a common kind (as happens when, e.g., I confuse a person with that person's twin), or one may confuse instances of different kinds (as happens when, e.g., I confuse molybdenum with aluminum). One may be confused because one lacks the ability to discriminate between two discriminable items, but one may also be confused because one has a discriminative ability yet fails to utilize it properly on some occasion. One may be confused without knowing that one is confused, so confusion should be distinguished from the affective condition of feeling perplexed. If, for example, I do not know there is anything called "molybdenum," then I will not know when I confuse molybdenum with aluminum that I am confusing the two. Even if I do know that they are two distinct sorts of metals, I may on occasion confuse the one with the other without being aware in the least that I am doing so.

Many cases of self-deception, I think, result from an individual confusing thumotic satisfaction with epistemic satisfaction. These sorts of satisfaction can be confused when their underlying affective components are sufficiently similar. Consider the fool again. The thought that his acquaintance may return to him one day satisfies him. The valence of this thought is pleasant, and the valence of the alternative is unpleasant. The satisfaction he finds in this thought cannot be exclusively epistemic, for an impartial review of the evidence would lead anyone, himself included, to conclude that she never has wanted him and never will want him. To make the point clear, assume that, if the fool were given similar evidence about a different pair of people, he would immediately conclude that the admired individual has no reciprocal affection for her admirer. At least part of the satisfaction he finds in the thought is thumotic—it is the pleasure he takes in believing that she wants him. He takes the relevant pleasure, however, as an indication that the thought *is* true, not that it is what he *wants* to be true. The thought feels right to him: he confuses this feeling, which is the pleasure of thumotic satisfaction, with the affect of epistemic satisfaction. So confused, he believes the thought. (This is not to say that the thought must be completely devoid of epistemic satisfaction: more on this below.)<sup>25</sup>

Someone sympathetic to Scott-Kakures might complain here that the confusion account, at least as it has just been presented, does not capture the active role the self-deceived agent plays in maintaining his or her condition. In order for the confusion account to be adequate, the complaint here goes, it must depict self-deceived individuals as playing some agential role in maintaining their confusion, as that, according to the account, is the source of their self-deception. This complaint can be met, I think, if the view is augmented by adding the following conditional claim: if self-deceived individuals are asked why they hold their unwarranted beliefs, they will, without lying, typically provide epistemic reasons in support of them, and they will never acknowledge that the unwarranted beliefs are unwarranted. Were such individuals epistemically akratic, they would not offer any epistemic defence but instead would admit that their beliefs are unwarranted; as ever, let us set this possibility aside. Consider again the fool from Loggins and McDonald's song: because he is confused about the satisfaction he takes in his belief, he will not say that he holds it because it would be too painful to do otherwise; instead, he will offer reasons that he takes to warrant holding the belief. His disposition to defend and to support his self-belief is clearly agential—this is the sort of thing agents and agents alone do—so his condition is maintained, at least counterfactually, by his agency.

Appealing to confusion and counterfactuals as I have gives rise to another, perhaps more basic, worry. Most, if not all, cases of self-deception at least appear to be motivated. Confusion, however, is often the result of an accident; how then can my account capture the at least apparently motivational aspect of self-deception? To answer this worry, consider the order of explanation I have been presenting: the self-deceived individual finds some thought satisfying, he confusedly takes the satisfaction to be epistemic satisfaction when it is in fact thumotic, so he believes the thought, and is thereby disposed to explain the belief as he would other beliefs of his. It should be clear that the belief here is motivated; it is held not on the basis of epistemic credentials but instead for the thumotic satisfaction it brings. The question, then, is whether the confusion itself is motivated, or whether it is a mere accident, or whether there is some third way it might come about.

To understand self-deception as self-*deception*, I think we must understand the confusion as motivated. This, however, does not require that we see it as some strategy executed by the self-deceived individual. For a given individual's thought to be a self-deceived belief, the thought, obviously, must be a possible belief. As such, it is the sort of thought that can be epistemically satisfying. It also, however, must be the sort of thought that can be thumotically satisfying to the individual; were it not, the person's epistemic stance towards the thought would be simply and exclusively determined by the extent to which that person finds it epistemically satisfying. (Perhaps most of our beliefs are like this, devoid of any thumotic component, and perhaps this is why most of our beliefs are not candidates for self-deception.) If the thumotic satisfaction of such a thought can be maintained only by believing it, and if the affective consequence of giving up the thought would be intolerably dissatisfying, then nothing more is needed to

generate self-deceived confusion. The presence of these elements does not guarantee self-deception; it is a mark of epistemic courage to believe what is warranted even when the belief is thumotically or otherwise dissatisfying. To exercise this courage, however, one must clearly distinguish the different sorts of satisfaction that are present in a given belief. If one is not skilled at distinguishing these elements, then it is easy for one to settle on what is thumotically satisfying. Should this happen, then the resulting condition—including the confusion that sustains it—is motivated, for the result is that one believes what one wants.<sup>26</sup>

Describing the self-deceived individual as believing what is thumotically satisfying might give rise to yet another basic worry. One might complain that, although some thoughts are thumotically satisfying and others are not, only someone pathological would take the thumotic pleasure of some thought as an indication of its being true. Most take self-deception, however, to be a nonpathological sort of irrationality. If the confusion account succeeds only by making all self-deception pathological, the complaint concludes, then it is not plausible. The way to resist this complaint, I think, is to note that confusion-based cases of self-deception often (if not always) involve a genuine element of epistemic satisfaction in the overall condition. Unlike confusing, for example, aluminum with molybdenum, which involves the complete confusion of the one with the other, there is likely to be an element of epistemic satisfaction sustaining the fool's belief about his acquaintance. If he concocts a rationalization for the belief, that rationalization will supply reasons for it that, were they true and decisive, would warrant his holding it. The rationalization, like any chain of reasoning a person finds compelling, is epistemically satisfying. Indeed, this explains why the fool seizes on these reasons as a rationalization for his belief; because he takes the belief to be epistemically satisfying, and because the relevant reasons provide grounds for this epistemic satisfaction, he settles on them as his explanation for his belief. It is not pathological to hold a belief that one takes to be backed by epistemically satisfying reasons, so the fool's condition, although irrational, is not pathological.

These remarks concerning self-deceived rationalization bring the confusion account partially in line with Scott-Kakures's view. The views differ, however, on the way in which they explain self-deceptive resistance to self-knowledge. Specifically, they differ on how to account for the fool's false self-explanation of why he believes his acquaintance will return to him one day. As noted in the previous section, Scott-Kakures seems committed to claiming that biased hypothesis testing is the cause of the fool's belief that his belief about his acquaintance is evidentially grounded.<sup>27</sup> The fool's self-belief here, however, is not the result of hypothesis testing; rather, it expresses the presumption that his belief that his acquaintance will return to him one day is warranted. This self-belief, which concerns the proper explanation for his maintaining his belief about his acquaintance, may be only tacit. At least as we have been conceiving of the case, however, it must be there; it explains why he resists the wise man's (correct) explanation of his condition. This self-belief is an immediate result of

the epistemic satisfaction he mistakenly takes in his belief about his acquaintance: the fool believes the self-belief to be true because its truth is a condition of the truth of his belief about his acquaintance. As long as he takes his satisfaction in the latter to be, at base, epistemic, he will also take the former to be true. Scott-Kakures is right, I think, to insist that an adequate account of self-deception must explain the agent's role in maintaining the condition in non-intentional terms; I also think he is right to focus on reasoning as the primary means of this maintenance. His hypothesis-testing model, however, cannot correctly characterize the self-deceived individual's resistance to being told either that or why he is maintaining his self-deception. If a disposition to resist the proper explanation of one's self-deceived belief is a common feature of the condition—and I hope to have shown here that it is—then the confusion account should be preferred for its ability to explain it.

#### 4. CONCLUSION: TWISTED SELF-DECEPTION

I have argued that the confusion account can explain both what the fool believes and why no wise man has the power to reason him out of his condition. The fool's anticipation of the pain he would feel upon giving up his preferred belief drives both his maintenance of the belief and his insistence that it is warranted. Not all cases of self-deception are like the fool's, however; there are cases of self-deception in which the individual maintains an unwarranted belief whose valence, at least apparently, is unpleasant. Mele (1997, 1999, 2001) has dubbed this "twisted" self-deception. For an example, imagine a jealous husband who has no evidence that his wife is cheating on him and overwhelming evidence that she is faithful, yet who persists in believing that she is having an affair. This sort of case poses at least a *prima facie* challenge to views such as Mele's and Barnes's, for it is not immediately obvious what could motivate this husband to maintain his jealous belief, nor is it immediately obvious how this husband's condition reduces anxiety (indeed, it might seem to provoke it). Both Mele (1997, 2001, ch. 5) and Barnes (1997, p. 44-46) have sought to defend their views from this challenge. Funkhouser (2005, p. 307-309), Scott-Kakures (2009, p. 101-105), and Fernández (2013, p. 386) also consider twisted self-deception; for them, it is a phenomenon that a full account of self-deception should be able to explain.<sup>28</sup> The confusion account can, I think, explain the phenomenon, and showing how it does will help to elaborate the view, in part by saying a bit more about thumotic satisfaction. Let me close, then, with some remarks on how it may be used to approach twisted cases.

In all cases, twisted and nontwisted alike, the explanatory value of the confusion account comes out when considering resistance to self-knowledge. Imagine, then, that the jealous husband is correctly told that he does not think his wife is unfaithful because that is what his best estimate of the evidence suggests; rather, he believes what he does because, even though the belief is unpleasant, it is somehow satisfying to him. Suppose he resists, asking how such an unpleasant belief could be satisfying in any way. Here, we can answer him by first taking a clue from etymology. The term "satisfaction" derives from Latin terms that



signify the idea of doing (“*facere*”) enough (“*satis*”). There are all sorts of things that we can satisfy by doing enough: we can satisfy demands, expectations, contractual obligations, etc. These sorts of satisfaction need not involve pleasure. The satisfaction the jealous husband takes in his unpleasant belief, then, could be the thumotic satisfaction of angrily upholding a code of honour, which he confuses with willingly accepting an unpleasant truth. These are confusable, for both anger and considering unpleasant truths are negatively valenced. To be sure, there may be other ways of characterizing the jealous husband; all I want here is to sketch the kind of explanation the confusion account is positioned to give. It does not have to appeal to pleasure to make sense of satisfaction, so nor must it appeal to pleasure to make sense of confusion, self-deception, or resistance to self-knowledge. Reflecting on twisted cases, then, may help not only clarify the confusion account but also bring out its explanatory power.

## ACKNOWLEDGMENTS

The ideas in this paper have been developing for well over a decade—gratitude for helping to shape them goes to more people than I can now remember. I would like to explicitly thank Joe Camp, Matthew Chrisman, Peter Machamer, John McDowell, Sebastian Rödl, and Kieran Setiya for help at the beginning of this project. For help finishing it, I thank Todd Nagel, as well as the editors and referees of this journal.

## NOTES

- <sup>1</sup> Scott-Kakures adopts this term from Alfred Mele’s self-characterization (Scott-Kakures, p. 577, n. 3) but applies it more broadly than Mele does.
- <sup>2</sup> Funkhouser, Barrett, and Fernández are not the first to develop views along these lines. They all acknowledge the influence of Robert Audi (1985, 1988, 1997) and Kent Bach (1981, 1997) on their work.
- <sup>3</sup> See, e.g., Hookway (2001), Owens (2002), and Greco (2014).
- <sup>4</sup> Although little has been written about this resistance to self-knowledge, plenty has been written on self-deception as involving a failure of self-knowledge. Scott-Kakures, Funkhouser, and Fernández all present self-deception in this way, as do Sanford (1988), Cohen (1992), Holton (2001), and Bilgrami (2006).
- <sup>5</sup> I focus on Funkhouser (2005) rather than Funkhouser and Barrett (2016) because the former relates more immediately to the present discussion.
- <sup>6</sup> Joseph Camp (2002) has suggested that the sort of confusion I am discussing might be more perspicuously labeled “ontological confusion” (Camp, 2002, p. 3). I, like Camp, will stick to using the simpler “confusion.” My thinking about confusion owes much to Camp’s insightful work on the topic.
- <sup>7</sup> The literature on this topic is vast. See, *inter alia*, Alston (1988), Audi (2001), Bennett (1990), Chrisman (2008), Hieronymi (2006), Setiya (2008), Scott-Kakures (2000), Shah and Velleman (2005), and Williams (1970).
- <sup>8</sup> See Lynch (2012) for an extensive discussion of this feature.
- <sup>9</sup> If this last point is what leads them to find self-deception interesting, then they are not alone. Tamar Gendler (2010b, 2010c) has introduced the concept of “alief” to account for cases in which individuals act contrary to what they believe; one of her main explanatory goals is to account for these actions without characterizing the individuals as holding inconsistent pairs of beliefs. For critiques of this notion of alief, see Hubbs (2013) and Mandelbaum (2013). Gendler (2010a) does not explain self-deception in terms of aliefs; instead, the phenomenon is characterized as a sort of pretense. Whatever virtues this account might have, it will not help explain the fool’s self-deceptive resistance to self-knowledge—sustaining this resistance is not a matter of pretending.
- <sup>10</sup> For more problems with Funkhouser’s position as it is elaborated in Funkhouser and Barrett (2016), see Doody (2017); for a response, see Funkhouser and Barrett (2017).
- <sup>11</sup> On this, see Davidson (2004a, p. 185).
- <sup>12</sup> I take the term “homuncularism” from Johnston (1988).
- <sup>13</sup> Other, somewhat more recent intentionalist views can be found in Talbott (1995) and in Bermúdez (1997, 2000).
- <sup>14</sup> Similar, but not identical. Funkhouser and Barrett claim that “philosophers and psychologists have a hard time keeping the deception in self-deception” (Funkhouser and Barrett, 2017, p. 682). The deception mentioned here is distinguishable from the tension discussed above.
- <sup>15</sup> Scott-Kakures (2002), elaborated and developed in Scott-Kakures (2009).
- <sup>16</sup> Scott-Kakures is not alone here; see also Mele (2001), as well as Scott-Kakures (1996), which draws on and critiques Mele (1987). As Scott-Kakures (2009, p. 76, n. 12) notes, sources for this approach include Friedrich (1993) and Trope and Liberman (1996).

- <sup>17</sup>With something close to this last point in mind, McDonald and Loggins tell us, “What seems to be is always better than nothing.”
- <sup>18</sup>Scott-Kakures (2009) elaborates this view using the resources of cognitive dissonance theory—these elaborations are irrelevant to the criticism I pursue in this section (but cf. n. 23 n. 27).
- <sup>19</sup>For more on agential mental activity, see Soteriou (2005) and Soteriou and O’Brien (2009).
- <sup>20</sup>This is not a recent discovery: the fact that our minds are susceptible to forces that are standardly thought of as operating on the body is a major theme of Freud’s work. For a particularly illuminating discussion of the matter, see Freud (1911).
- <sup>21</sup>See Padel (1992, p. 27-30) on “*thumos*.” Drawing very loosely on the view Socrates presents in Plato’s *Republic*, I take thumotic satisfaction to be distinguishable from the appetitive “pleasures of food, drink, sex, and others that are closely akin to them” (Plato, 1992, p. 111) and from the epistemic satisfaction I discuss later in this section. Mine can only be a loose interpretation, however, as Socrates characterizes *thumos* as the part of the soul we “get angry with” (Plato, 1992, p. 111). I take the space between food, drink, and sex, on the one hand, and learning and knowledge, on the other, to leave room for social emotions other than anger.
- <sup>22</sup>Barrett agrees that an affective condition can be satisfying without involving a detectable feeling; “even a completely neutral feeling is affect” (Barrett, 2017, p. 72).
- <sup>23</sup>Scott-Kakures (2009) draws explicitly on the literature on cognitive dissonance to develop an account of self-deception. I return to this below; cf. n. 27.
- <sup>24</sup>Jonathan Lear (1988) discusses this tendency towards epistemic satisfaction as a desire for understanding; adapting a term from Melanie Klein, he calls this desire *epistemophilia* (see Lear, 1988, p. 3-10). I might have characterized this tendency as aiming at truth or knowledge or understanding, but I wish to avoid the debates that surround these topics.
- <sup>25</sup>I have focused here just on valence, but for the two sorts of satisfaction to be confusable, the levels of arousal will also need to be sufficiently similar. I suspect there are some ways of depicting the fool where the arousal is elevated and others where it is low. I set this aside. My point here is made, I hope, by understanding how the common valence—which in the fool’s case is pleasure—could be confused.
- <sup>26</sup>By using the virtue-theoretic language of “epistemic courage,” we can accurately locate, I think, the sort of responsibility that self-deceived individuals have for their condition. In saying this, I thus disagree with Neil Levy’s view that self-deception is a simple mistake that lacks any necessary connection to culpability (Levy, 2004). A full discussion of the normative questions surrounding self-deception is beyond the scope of the present essay. For more on the topic, see Mary van Loon’s contribution to this issue of *Les ateliers de l’éthique/The Ethics Forum*.
- <sup>27</sup>Also, as noted in n. 23, Scott-Kakures has elaborated this view by drawing on cognitive dissonance theory. This is an account of what might drive one to test a (biased) hypothesis, but it keeps the hypothesis-test model intact. Moreover, cognitive dissonance is necessarily a dissonance *between* two separate epistemic states. The confusion view on offer here concerns two sorts of satisfaction one might find in a *single* epistemic state.
- <sup>28</sup>For other discussions on twisted self-deception, see Nelkin (2002) and Michel and Newen (2010).

## REFERENCES

- Alston, William, "The Deontological Conception of Epistemic Justification," *Philosophical Perspectives*, vol. 2, p. 257–299, 1988.
- Audi, Robert., "Self-Deception and Rationality," in Martin, Mike (ed.), *Self-deception and Self-understanding*, Lawrence, University of Kansas Press, 1985, p. 169-94.
- , "Self-Deception, Rationalization, and Reasons for Acting," in McLaughlin, Brian and Rorty, Amélie Oksenberg (eds.), *Perspectives on Self-Deception*, Berkeley, University of California Press, 1988, p. 92-120.
- , "Self-Deception vs. Self-Caused Deception: A Comment on Professor Mele," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 104.
- , "Doxastic Voluntarism and the Ethics of Belief," in Steup, Matthias (ed.), *Knowledge, Truth, and Duty*, New York, Oxford University Press, 2001, p. 93-114.
- Bach, Kent, "An Analysis of Self-Deception," *Philosophy and Phenomenological Research*, vol. 41, no. 3, 1981, p. 351–70.
- , "Thinking and Believing in Self-Deception," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 105.
- Barnes, Annette, *Seeing through Self-Deception*, Cambridge, Cambridge University Press, 1997.
- Barrett, Lisa Feldman, *How Emotions Are Made*, New York, Mariner, 2017.
- Bennett, Jonathan, "Why Is Belief Involuntary?" *Analysis*, vol. 50, no. 2, 1990, p. 87–107.
- Bermúdez, José, "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. 4, 2000, p. 309-319.
- Bilgrami, Akeel, *Self-Knowledge and Resentment*, Cambridge, Harvard University Press, 2006.
- Camp, Joseph, *Confusion: A Study in the Theory of Knowledge*, Cambridge, Harvard University Press, 2002.
- Chrisman, Matthew, "Ought to Believe," *Journal of Philosophy*, vol. 105, no. 7, 2008, p. 346-370.
- Cohen, L. Jonathan, *An Essay on Belief and Acceptance*, Oxford, Clarendon Press, 1992.
- Davidson, Donald, "Paradoxes of Irrationality," in Donald Davidson, *Problems of Rationality*, Oxford, Clarendon Press, 2004, p. 169-188.
- , "Incoherence and Irrationality," in Donald Davidson, *Problems of Rationality*, Oxford, Clarendon Press, 2004, p. 189-198
- , "Deception and Division", in Donald Davidson, *Problems of Rationality*, Oxford, Clarendon Press, 2004, p. 199-212.

Doody, Paul, “Is There Evidence of Robust, Unconscious Self-Deception? A Reply to Funkhouser and Barrett,” *Philosophical Psychology*, vol. 30, no. 5, 2017, p. 657-676.

Freud, Sigmund, “Formulations on the Two Principles of Mental Functioning,” in Stratchey, James and Anna Freud, (collaborating trans.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, vol. 12, London, Hogarth Press, 1911/1966, p. 218-226.

Fernández, Jordi, “Self-Deception and Self-Knowledge,” *Philosophical Studies*, vol. 162, no. 2, 2013, p. 379–400.

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.

Funkhouser, Eric, “Do the Self-Deceived Get What They Want?,” *Pacific Philosophical Quarterly*, vol. 86, no. 3, 2005, p. 295–312.

Funkhouser, Eric and David Barrett, “Robust, Unconscious Self-Deception: Strategic and Flexible,” *Philosophical Psychology*, vol. 29, no. 5, 2016, p. 682-96.

———, “Reply to Doody,” *Philosophical Psychology*, vol. 30, no. 5, 2017, p. 677-681.

Gendler, Tamar, “Self-Deception as Pretense,” in Gendler, Tamar, *Intuition, Imagination, and Philosophic Methodology*, Oxford, Oxford University Press, 2010a, p. 155-178.

———, “Alief and Belief,” in Gendler, Tamar, *Intuition, Imagination, and Philosophic Methodology*, Oxford, Oxford University Press, 2010b, p. 255-281.

———, “Alief in Action and Reaction,” in Gendler, Tamar, *Intuition, Imagination, and Philosophic Methodology*, Oxford, Oxford University Press, p. 282-310, 2010c.

Greco, Daniel, “A Puzzle about Epistemic Akrasia,” *Philosophical Studies*, vol. 167, no. 2, 2014, p. 201-219.

Hieronymi, Pamela, “Controlling Attitudes,” *Pacific Philosophical Quarterly*, vol. 87, no. 1, 2006, p. 45–74.

Holton, Richard, “What Is the Role of the Self in Self-Deception?” *Proceedings of the Aristotelian Society*, vol. 101, no. 1, 2001, p. 53-69.

Hookway, Christopher, “Epistemic *Akrasia* and Epistemic Virtue,” in Fairweather, Abrol and Linda Zagzebski (eds.), *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*, Oxford, Oxford University Press, 2001, p. 178-99.

Hubbs, Graham, “Alief and Explanation,” *Metaphilosophy*, vol. 44, no. 5, 2013, p. 604-620.

Johnston, Mark, “Self-Deception and the Nature of Mind,” in McLaughlin, Brian and Amélie Oksenberg Rorty (eds.), *Perspectives on Self-Deception*, Berkeley, University of California Press, 1988, p. 63-191.

Lear, Jonathan, *Aristotle and the Desire to Understand*, Cambridge, Cambridge University Press, 1988.

- Levy, Neil, "Self-Deception and Moral Responsibility," *Ratio*, vol. 17, no. 3, 2004, p. 294-311.
- Lynch, Kevin, "On the 'Tension' Inherent in Self-Deception," *Philosophical Psychology*, vol. 25, no. 3, 2012, p. 433-450.
- Mandelbaum, Eric, "Against Alief," *Philosophical Studies*, vol. 165, no. 1, 2013, p. 197-211.
- Mele, Alfred, *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*, Oxford, Oxford University Press, 1987.
- , "Real Self-Deception," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 91-102.
- , "Twisted Self-Deception," *Philosophical Psychology*, vol. 12, no. 2, 1999, p. 117-137.
- , *Self-Deception Unmasked*, Princeton, Princeton University Press, 2001.
- Michel, Christoph, and Albert Newen, "Self-Deception as Pseudo-Rational Regulation of Belief," *Consciousness and Cognition*, vol. 19, no. 3, 2010, p. 731-44.
- Nelkin, Dana, "Self-Deception, Motivation, and the Desire to Believe," *Pacific Philosophical Quarterly*, vol. 83, no. 4, 2002, p. 384-406.
- Owens, David, "Epistemic Akrasia," *The Monist*, vol. 85, no. 3, 2002, p. 381-397.
- Padel, Ruth, *In and Out of the Mind: Greek Images of the Tragic Self*, Princeton, Princeton University Press, 1992.
- Patten, David, "How Do We Deceive Ourselves?" *Philosophical Psychology*, vol. 16, no. 2, 2003, p. 229-246.
- Plato, *The Republic*, G.M.A. Grube (trans.), Indianapolis, Hackett, 1992.
- Sanford, David, "Self-Deception as Rationalization," in McLaughlin, Brian and Amélie Oksenberg Rorty (eds.), *Perspectives on Self-Deception*, Berkeley, University of California Press, 1988, p. 157-170.
- Scott-Kakures, Dion, "Self-Deception and Internal Irrationality," *Philosophy and Phenomenological Research*, vol. 56, no. 1, 1996, p. 31-56.
- , "Motivated Believing: Wishful and Unwelcome," *Nous*, vol. 34, no. 3, 2000, 348-375.
- , "At 'Permanent Risk': Reasoning and Self-Knowledge in Self-Deception," *Philosophy and Phenomenological Research*, vol. 65, no. 3, 2002, p. 576-603.
- , "Unsettling Questions: Cognitive Dissonance in Self-Deception," *Social Theory and Practice*, vol. 35, no. 1, 2009, p. 73-106.
- Setiya, Kieran, "Believing at Will," *Midwest Studies in Philosophy*, vol. 32, 2008, p. 36-52.
- Shah, Nishi, and J. David Velleman, "Doxastic Deliberation," *Philosophical Review*, vol. 114, no. 4, 2005, p. 497-534.
- Soteriou, Matthew, "Mental Action and the Epistemology of Mind," *Nous*, vol. 39, no. 1, 2005, p. 83-105.

Soteriou, Matthew, and Lucy O'Brien (eds.), *Mental Action*, Oxford, Oxford University Press, 2009.

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

Trope, Yaacov, and Nira Liberman, "Social Hypothesis Testing: Cognitive and Motivational Mechanisms," in Kruglanski, Arie. W. and E. Tory Higgins (eds.), *Social Psychology: Handbook of Basic Principles*, New York, Guilford, 1996, p. 72-101.

Williams, Bernard, "Deciding To Believe," in Williams, Bernard, *Problems of the Self*, New York, Cambridge University Press, 1981, p. 136-151.

# HOW TO TRAGICALLY DECEIVE YOURSELF

JAKOB OHLHORST

GRADUATE STUDENT, UNIVERSITY OF COLOGNE

## ABSTRACT:

This paper introduces the concept of tragic self-deception. Taking the basic notion that self-deception is motivated belief against better evidence, I argue that there are extreme cases of self-deception even when the contrary evidence is compelling. These I call cases of tragic self-deception. Such strong evidence could be argued to exclude the possibility of self-deception; it would be a delusion instead. To sidestep this conclusion, I introduce the Wittgensteinian concept of certainties or hinges: acceptances that are beyond evidential justification. One particular type of certainties—iHinges, which are adopted for motivational reasons—explain the phenomenon of tragic self-deception: they warrant the subject's dismissal of the evidence without loss of rationality from the subject's point of view. Subsequently, I deal with some objections that can be raised against this account of self-deception.

## RÉSUMÉ :

Cet article présente le concept d'auto-illusion tragique. En prenant la notion de base selon laquelle l'auto-illusion consiste en une croyance motivée à l'encontre de meilleures preuves, je soutiens qu'il existe des cas extrêmes d'auto-illusion qui persistent même face à des preuves contradictoires incontestables. J'appelle ces cas : auto-illusion tragique. On pourrait soutenir que des preuves d'une telle force excluent la possibilité d'auto-illusion et qu'il s'agirait plutôt de délire. Afin d'éviter cette conclusion, j'introduis le concept wittgensteinien de certitudes ou de propositions charnières (hinges) : des admissions qui se situent au-delà d'une justification de nature probante. Un type de certitudes en particulier – iHinges, qui sont adoptées pour des raisons d'ordre motivationnel – rend compte du phénomène d'auto-illusion tragique. Ces certitudes justifient que le sujet rejette la preuve sans que cela implique, de son propre point de vue, une perte de rationalité. Subséquemment, je traite de certaines objections qui peuvent être soulevées contre cette explication de l'auto-illusion.



## INTRODUCTION

Self-deception is emotionally motivated belief against better evidence. This is a common way of glossing the phenomenon. It is a very expansive notion: for example, wishful thinking will also fall under this umbrella. There is a debate on whether self-deception needs to be intentional. I ignore this question, because intentional self-deception excludes certain phenomena that I take to be clear cases of self-deception (cf. Mele, 2001, p. 9).

I think that there is a special form of self-deception, which is usually overlooked. I call it *tragic self-deception*: in such cases the subject is able to dismiss *any*, even compelling, evidence. This is the sort of self-deception from which, for example, religious fanatics suffer: the tragically self-deceived are in that state because, to them, certain things *cannot* be true.

Tragic self-deception is overlooked because it does not fit with our usual epistemological concepts. Either the contrary evidence seems to be too weak, so that it leads to ordinary self-deception, or the evidence is considered to be too strong. In the latter case, the subject would have to be mad or delusional, rather than merely self-deceived, in holding such a belief.

Note that “delusional” is ambiguous. There is the medical sense—such a *pathological* delusion means that the subject suffers from some mental illness. I will come back to how this sense of “delusional” appears to exclude tragic self-deception in the case of compelling evidence. The other sense is loose talk: “he is delusional” just means there is something very wrong with his beliefs. This looser sense of “delusional” does not seem to demarcate a class of its own and is often synonymous with “self-deceived.” I shall therefore ignore it (cf. Bortolotti and Mameli, 2012).

Our usual concepts of self-deception provide only a limited arsenal of ways to avoid an abhorred conclusion: selective sampling, selective treatment, biased weighing, and misinterpretation of evidence (Mele, 1997). These self-deceptive tricks cannot dispel compelling evidence against the false belief. Therefore, we overlook the possibility of tragic self-deception. A regularly self-deceived subject confronted with sufficiently strong evidence would therefore snap out of it and find truth. Such a subject who failed to do so would be so deeply irrational that that subject would be pathologically delusional—his or her mind would have to be defective.

Nevertheless, I think we should take the possibility of tragic self-deception seriously. Indeed, if we extend our conceptual framework to include *hinges* or *certainties*, then we can account for tragic self-deception. Hinges were introduced by Ludwig Wittgenstein in *On Certainty* (1969). We can use the notion to account for how and why tragic self-deception arises. Additionally, we can explain away the impression that tragic self-deception is impossible and that it collapses into pathological delusion.

I will begin by presenting an example of what I take to be tragic self-deception. Indeed, I believe that such cases are quite common. I then explain how our understanding of rationality threatens the concept of tragic self-deception. I believe that this idea explains why tragic self-deception seems to collapse into delusion. I will then introduce certainties as the feature of our epistemological apparatus that is able to explain tragic self-deception.

My account of certainties loosely follows Wittgenstein's (1969). A peculiar form of such certainties can cause self-deception. I call them *iHinges*, and they are adopted for subjective motivational reasons. This paper shows how they explain the possibility of tragic self-deception, where the subject is absolutely resistant to contrary evidence.

While this account deals with the mentioned objection of irrationality, other objections can be raised. I conclude by explaining and responding to those other objections: that my definition of tragic self-deception seems to make it indistinguishable from pathological delusion as it is usually defined, that it does not account for stubbornness or the nagging doubts usually accompanying self-deception, and that it is too heavy-handed an approach for something so straightforward as self-deception.

## TRAGICALLY DECEIVING ONESELF

Tragic self-deception arises when a subject has a belief that is *immune to any, even compelling, evidence*. This is analogous to regular self-deception being mere resistance against evidence. What is meant by "immunity," and what do I mean by "compelling evidence"?

A belief is immune when there is no possible evidence that could rebut or undermine it. I name this *immunity to evidence* because the belief could be lost due to other psychological factors like being shocked or overwhelmed, but no *evidential* reasons could dislodge it.

Evidence is compelling when it is impossible to ignore, when it entails what it is evidence for, and when the entailment is also impossible to ignore. When neither the evidence nor its evidential relations can be ignored, this makes the evidence compelling: the evidence forces us to believe.

Now this seems to make tragic self-deception an inconsistent concept: how can we be immune to believing something that we're forced to believe? I will work out this inconsistency more precisely in the section on tragic self-deception and madness. Ultimately, I shall argue that the appearance of inconsistency is resolved by epistemic certainties. But first I will give an example of tragic self-deception and develop the notion to show that it is nevertheless rather natural.

MISBEGOTTEN: A young woman is accused of a heinous crime: the evidence laid out by the prosecutor against her is crushing and she has

confessed to it repeatedly. Notwithstanding this, her parents remain convinced of their baby's innocence. If they accepted that their daughter was a criminal, they would be devastated: in their eyes, they did everything for their daughter, they gave her their values, and their identity is built on their idea of their daughter's moral character.<sup>1</sup>

This is an example of what I named tragic self-deception—it is the first example that comes to my mind when I am thinking about self-deception in general. It fits the idea that self-deception occurs when people believe something against better evidence because they desire or fear it to be true.

I have to note, however, that I think that self-deception is a multifarious phenomenon—I am not certain that *all* self-deception is a matter of believing against better evidence.<sup>2</sup> Let us then take a closer look at the example.

The parents of “Misbegotten” are self-deceived. Most people would be called so if they believed something because of some motivation without appropriately accounting for their evidence. They believe that their daughter is innocent because they desire her to be so and because they fear the contrary. Meanwhile, such a desire has no epistemic force. Nevertheless, they dismiss the compelling contrary evidence they have: all that has been laid out in the judiciary proceedings.

Still, “Misbegotten” differs from run-of-the-mill examples used for self-deception as characterized by motivated resistance against evidence. Take the story of Sid and Roz as an example.<sup>3</sup> Sid is in love with Roz; Roz, however, does not share this sentiment. Because she fears hurting his feelings and because she nevertheless appreciates him as a friend, Roz does not simply rebuff him. Rather, she implicitly communicates in a way that any sensible person would understand that she is not interested. Unluckily, Sid is not sensible; he is in love. Given his emotional state, he misreads Roz's refusals as encouragements.

Sid has evidence against *p* but because of his motivational state—desiring Roz to love him—he misinterprets the situation and comes to believe that she also loves him. The difference from “Misbegotten” is that, if Roz had explicitly told Sid that she does not want to be with him, then Sid would snap out of it. He is not deluded either, so runs the thought, except in a hyperbolic sense. And he would not resist *any* evidence against the desired outcome: if Roz wrote him a letter, he would maybe want to double-check with her; but he would not out of hand dismiss it as a forgery sent to him by his malevolent competitors. That is, Sid is insensitive but not immune to evidence.

The parents in “Misbegotten”, however, are so deeply entrenched in their belief that no evidence could convince them otherwise. I therefore call “Misbegotten” a case of *tragic self-deception*—tragic, because it is characterized by total immunity to evidence.<sup>4</sup> Tragic self-deception is then a special kind of self-deception, which is characterized by its being emotionally motivated and immune to evidence—something that Mele's classical model (2001, p. 50-51) arguably cannot account for.

Given the above considerations, I shall follow Alfred Mele's (1983, p. 370; 2001, p. 50-51; 2009, p. 267) deflationist approach to self-deception in order to characterize tragic self-deception. The following characterization is a modification of Mele's:

- S enters *tragic* self-deception in acquiring a belief that p *iff*
- (1) the belief that p that S acquires is false;<sup>5</sup>
  - (2) S treats *the evidence* relevant, or at least seemingly relevant, to the truth value of p in a motivationally biased way;
  - (3) this biased treatment is a nondeviant cause of S's acquiring the belief that p; and
  - (4) the body of *evidence* possessed by S at the time *entails not-p*, S is *aware of the entailment*, and *the evidence is such that S cannot ignore it* (cf. Mele, 2009, p. 267).<sup>6</sup>

By *entailment* I do not mean mere material implication but rather strict entailment. It is the fourth condition that makes self-deception tragic, as it strengthens the evidential relation to entailment and makes the evidence impossible to ignore. Consequently, the subject is aware of the evidence and of what it entails.

For the general debate on self-deception, I shall be relying more or less on Mele's position and work (2001). I do not think that tragic self-deception and deflationist self-deception are the only types of the phenomenon. Nor do I think that the intention to deceive oneself is a prerequisite for being self-deceived. For, if we survey how the nonphilosophical literature describes self-deception, we find instances of motivated belief against evidence.<sup>7</sup> These are philosophically interesting phenomena, whether they are called self-deception or not. But if we consider the sort of examples I have mentioned, these seem (at least to me) well captured by the term "self-deception" or "tragic self-deception."

## REMAINING SUFFICIENTLY RATIONAL

### Egocentric and Objective Rationality

Tragic self-deception raises a whole nest of issues. Immediately, there is the question of rationality. Holistically speaking, if we take into account all the reasons a subject might have, tragic self-deception may be rational. It allows the subject to pursue his or her life as before. However, I am concerned with the epistemology of self-deception and focus on the self-deceived's epistemic rationality. There are two senses of epistemic rationality: objective and egocentric rationality.

Objectively, the self-deceived are clearly irrational. We do criticize people for having objectively irrational beliefs—for example, when an agent does not account for all the evidence that agent has available or when an agent believes mutually exclusive things. There is then a standard of rationality that everybody can be measured against. Richard Foley calls this "objective rationality" (Foley,

1991, p. 369), and it is circumscribed by how a knowledgeable outside observer would evaluate the situation.

But this is not everything. There also is an inside perspective on rationality, which evaluates the subject's own point of view:

We are sometimes interested in evaluating your decisions from your own egocentric perspective. Our aim is to assess whether or not you have lived up to your own standards ... or, perhaps, the perspective you would have had were you to have been carefully reflective. (Foley, 1991, p. 367–368)

Following Foley, I call this weaker notion *egocentric epistemic rationality*. We all have standards of coherence for ourselves: our beliefs should be compatible with each other and with the evidence we have available. If we become *aware* of incoherence, either between our beliefs and our evidence or amongst our beliefs, then we are required to revise either a belief or our interpretation of the evidence. If we failed to account for all the evidence we are aware of or to deal with our incoherence, we would become egocentrically irrational, believing things we believe to be incompatible.

There is the common idea that we cannot be egocentrically irrational if we want to preserve our sanity. That is, we as mentally healthy subjects are incapable of believing contradictions while being aware of their contradictoriness.<sup>8</sup> In the same way, clear and immediate evidence against our beliefs cannot be ignored just like that—if we did, something would be wrong with our minds. Becoming consciously egocentrically irrational appears to be a form of madness; it implies a mind in disarray.

Even if a consciously egocentrically irrational person were not crazy, it still would be extraordinary. Could we make sense of what that person was doing and thinking?<sup>9</sup> If somebody seems to be consciously egocentrically irrational—that is, holding inconsistent beliefs while aware of the inconsistency, not accounting for evidence while recognizing it, but otherwise behaving normally (if that is possible)—then we need some explanation for what is going on. I do not think that a sane subject can be consciously egocentrically irrational.

### Tragic Self-Deception or Madness?

So, we are also subject to egocentric rationality. The household of our beliefs is subject to diverse constraints. Anything we are or would, upon reflection, become aware of falling under these constraints constitutes our egocentric rationality. As mentioned, we have to appropriately account for the evidence that we are aware of, and we cannot believe a contradiction while being aware of it; otherwise, we are consciously egocentrically irrational.

Those who are tragically self-deceived seem to be consciously egocentrically irrational—they dismiss evidence that they recognize to be compelling. Conscious egocentric irrationality implies that the subject is mad. If all instances of tragic self-deception are instances of madness, then what good is the concept? It's just madness. This reasoning creates the appearance that cases of tragic self-deception with sane subjects are impossible. Consequently, tragic self-deception does threaten a subject's egocentric epistemic rationality and thereby itself as a concept.<sup>10</sup>

In tragic self-deception, the available evidence is compelling. The subject cannot ignore that evidence without becoming consciously egocentrically irrational. But if that subject accounted for the evidence by making the appropriate inferences, this would lead to a contradiction with the strongly confessed self-deceptive beliefs. The subject would again be consciously egocentrically irrational. How could a subject be tragically self-deceived without being egocentrically irrational, given that as long as such a subject holds on to a self-deceptive belief he or she will end up in an apparent incoherence?

Prima facie, there seem to be two options on how to interpret what is going on here: either the subject ends up egocentrically irrational due to the compelling evidence and his or her immunity to it, or the evidence was not actually as compelling as it was made out to be.<sup>11, 12</sup>

The first option means that the subject is out of his or her right mind. Conscious egocentric irrationality is not possible for a sane subject—consequently, the consciously egocentrically irrational person must suffer from some mental illness. Most probably, that person is delusional. Delusion is defined as “a false belief ... that is firmly held despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary” (DSM-5, 2013, p. 819). Now, if tragic self-deception were nothing but a symptom of mental illness, calling it a kind of “self-deception” would be unwarranted.

The other option is that actually the evidence is not compelling. If that is the case, then this is not a case of *tragic* self-deception. Rather it is regular evidence and the subject is just ordinarily self-deceived.

Notwithstanding this appearance, I want to defend the notion of tragic self-deception as something that occurs with fanatics or in cases like “Misbegotten”. I want to maintain that a sane subject can stay both egocentrically rational and immune to any evidence, even if that evidence is compelling. This is possible because the evidence may be compelling, but it cannot be *certain*. In a way, I concede that compelling evidence is not as compelling as it might seem at first sight. Evidence cannot compel us to just any belief, even when we have accurately assessed it. That would be an unrealistic requirement. Evidence may always be defeated. The evidential relation is one of entailment, and every modus ponens has its modus tollens. Finally, there is a class of doxastic states that defeat any evidence. These states are called *certainties* or *hinges*.

## HINGING

### Regularly

I shall argue that tragic self-deception is grounded in a peculiar feature of our epistemologies: *hinges* or *certainties*—two terms that I use interchangeably. Historically speaking, hinges go back to Ludwig Wittgenstein’s *On Certainty* (1969). But they have recently regained popularity in internalist epistemology (cf. Coliva and Moyale-Sharroch, 2016).

Certainties or hinges are acceptances<sup>13</sup> that are beyond evidential justification. This immunity against evidence stems from their peculiar role: they are so general or fundamental that one cannot possibly find any noncircular evidence for or against them. They are the acceptances that are attacked and rendered visible by sceptical arguments. Wittgenstein gives the following example:

91. If Moore says he knows the earth existed etc., most of us will grant him that it has existed all that time, and also believe him when he says he is convinced of it. Has he also got the right ground for his conviction? (Wittgenstein, 1969, p. 14e)

As Wittgenstein points out, there is noncircular evidence neither *for* nor *against* such propositions. They are beyond the game of giving reasons. Consequently, they are also beyond doubt. For Wittgenstein, genuine doubt is controlled by evidence—we cannot really doubt something that is beyond evidence; it would just be empty gesturing (Wittgenstein, 1969, p. 18e).

These evidentially ungrounded certainties form the “rock-bottom” on which the whole building of our knowing and believing is erected (Wittgenstein, 1969, p. 33e). Wittgenstein uses the bedrock metaphor because of the hinges’ solidity—as mentioned, we cannot really doubt hinges, and we do not make sense if we do because of their foundational role. Accepting certainties is necessary for evidential relations, communicative interaction, and practical projects to work. Without hinges, we would not be able to investigate, tell, or do anything intentionally.

So how do hinges work? Mostly they are hidden and implicit. They form the fixed background in front of which we operate. They are the things we take for granted and on which our epistemic activity turns. But in peculiar situations, they can come to the fore. Consider the following example.

ROSE: As you read this paper, you see how a rose appears out of thin air, falls to the ground, and shatters. The shards dissolve into thin air again. You do not trust your eyes and refuse to believe that a rose really appeared and disappeared just now.

I would like to point out that the evidence in “Rose” is compelling, you *see* the rose appearing—if the evidence is not strong enough for your tastes, strengthen

it as you like. The certainty that is active in this example can be formulated as follows: “midsized massive objects do not just appear and disappear.” This is certain. You would have a hard time furnishing evidence for it. And evidence against it, such as “Rose”, is dismissed even if it is compelling. Indeed, the certainty *warrants* dismissing any evidence against it.

For, if you were to accept the evidence from “Rose”, you would lose the ability to know many things about objects—especially if the rose’s appearance remained a unique or random event. Your memory would become useless; you could not tell others about things elsewhere. You would not even be able to intentionally get milk from the fridge the way you usually do, as it could have disappeared in the meantime. You would only ever be able to intentionally go check whether there is still milk. Depending on your milk management, that might be what you actually do, and, in that case, imagine yourself to be more reliable.

The constancy of objects is not the only certainty. There are large swaths of interconnected propositions that can be taken to be certain—for example, the belief that the world is more or less as we perceive it to be, the idea that causal relations hold between certain events, or the conviction that other living beings also have consciousness. These certainties inform our world picture (Wittgenstein, 1969, p. 24e). We need such certainties, lest we be unable to act intentionally or to believe genuinely. We would not be able to interact with the world in any way. Hinges are a necessary condition for epistemic and practical agency.

To use a slogan: without certainties, mind-to-world, world-to-mind, and mind-to-mind adaptation would break down. That is, neither could the mind adopt a picture of the world nor could it act upon the world or exchange thoughts with other minds. It is therefore both epistemically and practically *rational*<sup>14</sup> to treat certain fundamental beliefs as certainties or hinges. From this particular role flows the epistemic rationality of dismissing any evidence that is incompatible with our hinges.

The attentive Wittgensteinian will already have realized that I do not follow an orthodox interpretation of what hinges are, because I treat them as logically related to other beliefs and maybe even truth apt instead of animal (a rule of our form of life) and beyond evaluation (cf. Moyal-Sharrock, 2016). Indeed, I am not at all certain that this epistemic account of certainties (cf. Kusch, 2016) accounts for Wittgenstein’s own view. Nonetheless, I believe that such an epistemic account of hinges is the way to go; it is a powerful tool. In other words, I am not wedded to the noble enterprise of reconstructing Wittgenstein’s own philosophy.

### Variant Hinges

Things are however not as simple as they may have seemed until now: there is no natural class of certainties. Being a hinge is not an essential feature of propositions. Rather, this depends on the environment in which we live and the kinds of organisms that we are: a member of the San people living in the thirteenth



century has and needs a different world picture than an Indian Brahman from the third century, and their world pictures are profoundly different from a dolphin's.

A salient example of how hinges can be under tension or shift is the theory of relativity. Wittgenstein was aware that we changed hinges in that case. Notably, he would refuse to say that people's former certainty that time and space are independent measures was *mistaken*. Rather, they were "confused" about what time and space are—namely, something that, for lack of better words, bends and stretches (Wittgenstein, 1969, p. 39e). Indeed, if you look at how we think, talk, and behave today, we still largely treat space and time as independent. Another example that Wittgenstein raises is that of faith:

Catholics believe as well that in certain circumstances a wafer completely changes its nature, and at the same time that all evidence proves the contrary. And so if Moore said 'I know that this is wine and not blood', Catholics would contradict him. (Wittgenstein, 1969, p. 32e)

These examples show that there is no salient criterion of which propositions should be hinges—it is contingent for any proposition *p* whether it is certain or not. Rather, hinges arise out of how we treat certain propositions. We take them to be certain and indubitable—we take them for granted. Hinges fulfil a certain functional role in the households of our believing and they are hinges because they play this functional role.

### Life Is Easier on iHinge

This opens up space for propositions to be treated like certainties that have little to do with grounding our epistemic, communicative, and practical activities. People do not treat only the existence of the world, or the acceptance that other members of their community will understand what they say, as certain. There may be also more subjective emotional or existential hinges.

Tragic self-deception results from this category of subjective hinges. If we compare the examples of the "Rose" and the misbegotten daughter, then the parallels become clear. In both cases, there is apparent and compelling evidence for some proposition *p*. But the agent dismisses this evidence instead of adapting her beliefs to it. She does this because, from her point of view, things just *could not* be as the evidence would indicate—that is, *p*.

Things could not be such because the agent is *certain* of propositions that imply that not-*p*. Therefore, the evidence cannot be accurate. Thus, the agent's hinges are the grounds on which she dismisses the apparent evidence. Interestingly enough, neither with regular hinges nor in tragic self-deception does the certainty need to be held explicitly. The certainty may become apparent only if we asked the subjects why they dismissed their evidence. The mechanism works automatically.

Further, the certainties preserve the agents' egocentric epistemic rationality when they dismiss even compelling evidence that goes against their hinges. Their functional role as certainties make the dismissal of any contrary evidence egocentrically rational. The same also holds for tragic self-deception: the tragically self-deceived may dismiss any evidence going against their fundamental certainties, without becoming egocentrically irrational. Their self-deceiving hinges parasitize the egocentric rationality-preserving function of regular hinges. I shall call the hinges leading to tragic self-deception *iHinges*,<sup>15</sup> given their central role for our self-image.

What makes certainties into *iHinges*? It's the basis on which they are adopted. As mentioned, regular hinges serve to make action, communication, and inquiry as such possible. They are a structural requirement. Without them, we would end up with an infinite regress or a circularity of evidence justifying beliefs about evidence. All our beliefs and actions would lose their rational grounds. Wittgenstein therefore likes to call them "logical" requirements (Wittgenstein, 1969, p. 7e, 20e, 58e).

*iHinges* do not play the same fundamental role. They are necessary for our lives in another way. Rather than being necessary for our world picture, they are hinges necessary for our self-image. *iHinges* are the response to questions like "What sort of person am I?" and not to questions that are dealt with by other hinges such as "What is a person?"

Take, for example, the conviction that the people close to you genuinely care about you and vice versa. Evidence for this is extraordinarily hard to come by. All of their behaviour is compatible with the fact that, ultimately, it is only their own enjoyment they seek in interaction with you. Indeed, there are people who argue for a crude *Homo oeconomicus* anthropology, claiming that all actions simply result from an egocentric utility calculus.<sup>16</sup>

Nevertheless, we trust ourselves and our close friends to spend time with us and support us not because they expect some corresponding return value from this. In Kantian terms, we have a hinge that to our close friends we are an end in ourselves and not a means to something else. Without this certainty, the very notion of friendship would fall apart. Imagine how empty the idea would be. The only way to avoid this hinge is by believing that you have no friends at all.<sup>17</sup>

In other words, *iHinges* may help give sense to our lives. They tie together a self-image, helping us to deal with our desires, doubts, and fears. *iHinges* play an existential role. If such a certainty becomes unhinged and is accepted to be false or doubtful, it is not our *very capability* to interact with the world that is at stake. We still have access to all the categories required to act and believe rationally.

Instead, what happens if we lose an *iHinge* is that we fall prey to an existential crisis: many maxims on which we have put our stakes until now, many evaluations that we took to be certain, and many ideas about who we are would fall

apart. While it would not hinder us from doing or investigating things by robbing us of the prerequisite world picture, it would undermine our motivations, values, and ideas, thus throwing us into lethargy.

We accept iHinges to avoid such existential devastation. They are a psychological or motivational rather than an epistemic necessity for living our lives. In sum, a certainty is an iHinge if and only if it has all the structural trappings of a certainty – immunity to evidence, implications for a wide class of other beliefs, and so on – but it is adopted because it is necessary for preserving our motivations and self-image rather than for preserving our epistemic and practical agency as such. One consequence from this is that, from the point of view of objective rationality, our iHinges ought to be susceptible to evidence. The fact that they are not, because of their existential role, generates problems like tragic self-deception.

### PICK YOUR IHINGE FOR YOUR TRAGEDY

iHinges lead to the motivated manipulation of evidence in tragic self-deception. Thus, in “Misbegotten”, the parents have invested everything in their daughter: they instilled her with their morals and values, they cared for her, and scolded her when she did something wrong. In sum, they built their lives around their idea of their daughter.<sup>18</sup> This means that, ultimately, their daughter’s moral integrity has become more important to them than their own epistemic access to the world. If they accepted that she did bad things, their world would break down. Much of what they had done and believed since she was born would lose its meaning.

At first sight, as in any case of tragic self-deception, this appears to be egocentrically epistemically irrational. In her parents’ eyes, the young woman could not do any wrong. Notwithstanding this, the evidence is clear and compelling: she *did* do something wrong. But for the parents, their daughter’s innocence is certain. This allows them to dismiss the unpleasant evidence as fabricated or misleading without a loss of egocentric rationality. Through their iHinge, they tragically deceive themselves in the face of compelling evidence.

They do this as follows: They are aware of the evidence laid out by the prosecutor, and they are aware that this evidence entails that their daughter is guilty. It is compelling, and they cannot simply ignore it. But the parents are also certain that their daughter is innocent. Nothing can epistemically trump a certainty; therefore, their iHinge that the daughter is innocent defeats the evidence from their egocentric point of view. Either the iHinge rebuts the evidence by *modus tollens* or the evidential relation is somehow undercut. How the evidence is defeated will depend on the circumstances.

Consequently, from their egocentric point of view they deal with the evidence in an appropriate manner. They dismiss evidence that in their eyes cannot be accurate in favour of a certainty. The certainty even warrants taking some

account of why the evidence is undercut or rebutted to be the most plausible explanation. This allows them to preserve their egocentric epistemic rationality in tragic self-deception and, consequently, to avoid delusion. In short, we are tragically self-deceived when an iHinge is incompatible with the world and there is compelling evidence for this incompatibility.

This, then, is the basic idea of tragic self-deception: certainties that are adopted for motivational reasons lead to dismissal of accurate evidence. I will now respond to four objections that can be raised against the iHinge view of tragic self-deception. The first three have been raised against Mele's deflationist stance but apply to my view, too. The last one specifically attacks the approach of using hinges to account for self-deception.

### Nagging Doubts

Some authors argue that self-deception is essentially characterized by a certain inner tension (Audi, 1997, p. 104; Bach, 1997, p. 105; Losonsky, 1997, p. 121). In other words, when we are self-deceived, we will be aware that something is amiss. We will frequently revisit our deceptive belief and have "nagging doubts" (Losonsky, 1997, p. 121) about it. This point has been raised in response to a paper by Alfred Mele, where he argues that self-deception as investigated by empirical psychology arises out of our biases (Mele, 1997, p. 93-95), a view that does not account for these supposedly essential nagging doubts. The idea behind the objection is that the self-deceived subject is somehow implicitly aware of his or her deceit, hence the nagging doubts.

In tragic self-deception, there can be no such nagging doubts at all. Doubt is *excluded* by the very nature of certainty—iHinges need to be beyond doubt to fulfil their role. Nevertheless, a form of tension remains in tragic self-deception, but it is an outer rather than an inner tension. As Bach observes, the self-deceived subject is at stark odds with reality—"truth is dangerously close at hand" (Bach, 1997, p. 105). First, the subject's belief is false; second, contrary evidence is easy to come by. The self-deceived subject will therefore quite frequently come up against such contrary evidence—having to dismiss it *each time*. This is an outer tension.

Consider how someone would react to not only a single, but repeated occurrences of "Rose": suddenly roses start popping up and disappearing all over the place. The first few times, she would probably simply dismiss it; but with time she would start treating the events differently. She would start to ask others whether they had experienced the same thing as she has, she might seek medical help, and she might even try to examine the phenomenon more closely. In other words, she would start to display the same behaviour as someone doubting—mostly her own sanity. She would be subject to outer tension.

Analogously, those who are tragically self-deceived—always coming up against contrary evidence, always having to do the epistemic work of dismissing it—

may at a certain point start to display such doubting behaviour. Maybe they would not doubt their own sanity but rather the sanity of people in their environment. While this phenomenon is not the inner tension that Mele's critics demand, it may still suffice as an explanation for why there are "doubts" accompanying self-deception.<sup>19</sup>

## Delusion

Like deflationist self-deception, the notion of tragic self-deception raises a further spectre: that tragic self-deception is mere delusion. More specifically, the question is, why would cases of tragic self-deception not simply be cases of *pathological* delusion? The worry also arises for other accounts that take resistance against evidence as a starting point to account for self-deception. This is so because diagnoses for delusion are also characterized by immunity against evidence:

Delusion. A false belief based on incorrect inference about external reality that is firmly held despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary. (DSM-5, 2013, p. 819)

I emphasize the pathological side of delusion because in ordinary language someone who is self-deceived may be called deluded without being a case requiring psychiatric treatment. Tragic self-deception skirts the abyss of pathological delusion. It is epistemically so irrational that it has itself the air of being pathological; the tragically self-deceived looks like a case for a therapist—for example, a psychoanalyst. Nevertheless, I think the tragically self-deceived are quite sane—there is no work for the medically trained psychiatrist.

A nice example that demonstrates the fine line between delusion and tragic self-deception is the romantic partner who, out of jealousy, is self-deceived about being cheated on. How exactly does this self-deceived person differ from someone who needs psychiatric help for morbid jealousy? Morbid jealousy is a syndrome characterized by "a range of irrational thoughts and emotions, together with associated unacceptable or extreme behaviour, in which the dominant theme is a preoccupation with a partner's sexual unfaithfulness based on unfounded evidence" (Kingham and Gordon, 2004, p. 207).

Apart from the "unacceptable or extreme behaviour," someone who is tragically self-deceived fits the morbidly jealous's profile. Epistemically speaking, the self-deceived person is indiscernible from the delusional patient in this case. So, is tragic self-deception nevertheless just a form of delusion?

While this is worrying at first sight, we should keep in mind that psychiatry is anything but an exact science.<sup>20</sup> Furthermore, delusions do not form an aetiologically unified class of phenomena. This is in contrast to illnesses like schizophrenia or depression, each of which is unified by a range of symptoms and

neurological characteristics. Already morbid jealousy itself is only the surface description of several other illnesses—for example, it occurs frequently in cases of schizophrenia (Kingham and Gordon, 2004, p. 207).

I surmise that, rather than being a phenomenon of the same kind and at the same level of generality as delusion, self-deception plays the role of a symptom *of* delusion. That is, delusions will always be accompanied by behaviour that can be described as deflationary self-deception, but there can also be isolated self-deception that is not a symptom.<sup>21</sup>

This would explain why the two are hard to separate but nevertheless distinct. An additional argument may be, that for someone to be diagnosed as pathologically deluded, their agency and ability to live a normal life would have to be seriously impaired. Arguably, cases of tragic self-deception exactly succeed at preserving agency and a *partially* normal life. We can push this line of thought further: one of the strategies in clinical psychiatry for distinguishing pathological behaviour from merely unusual behaviour is whether it poses a threat to either the subject or the subject's environment.<sup>22</sup> The fine line separating delusion from tragic self-deception would then be how much danger the particular case poses.

This differentiation is rather tentative. But given the main goal of this paper—introducing the phenomenon of tragic self-deception—it may be sufficient to show how delusion may tie in with tragic self-deception. Note, however, that, given the multifariousness of delusion, each syndrome must be treated separately, and that the moves we made for morbid jealousy will not necessarily work universally (cf. Mele 2006, p. 116-123).

### Stubbornness

A different, though similar, difficulty is how to differentiate tragic self-deception from mere stubbornness about being right. Kevin Lynch (2013) raised the problem and proposed a viable solution in the same go. The problem, as with delusion, is that stubbornness is intuitively distinct from self-deception, even though it is also characterized by motivated immunity against evidence. A stubborn person does not care about the evidence and will stick to his or her guns.

Lynch brings to attention the idea that acceptances can be adopted *because of* emotional attitudes towards their content or they can be adopted *independently* of content for some other reason. According to Lynch, stubbornness is content neutral (Lynch, 2013, p. 1342–1344). In tragic self-deception, the subject cares about what he or she adopts as a certainty—the subject has an emotional attitude towards what he or she accepts, be it fear, desire, or something else. Meanwhile, in stubbornness, as introduced by Lynch, the content of the beliefs adopted is irrelevant. It is about being right in general, not about any particular belief content. In contrast, in tragic self-deception, we are certain about certain propositions.

That is, if Lynchean stubbornness about being right also operates on certainties, then the relevant hinges will not be about some specific content. Rather, they would take the form “Those who do not share my ideas are intellectually careless” or something in this vein. This hinge would then warrant the dismissal of any counterargument.

Tragic self-deception turns on certainties that are adopted because the hinges’ content fulfils a certain motivational, psychological, or existential role for us. This distinguishes it from stubbornness as proposed by Lynch, for which the content of beliefs does not play any role or only a secondary one. Meanwhile, nothing seems to preclude the possibility that stubbornness as introduced by Lynch is a form of self-deception.

### Cannons at Sparrows

Another issue may be the hinge approach itself. Using hinges to deal with self-deception may look like shooting cannons at sparrows. Bringing out the heavy apparatus of certainties to deal with something as straightforward as self-deception when we have so many other options may seem overblown. Are we then exaggerating?

I have two things to say about such general doubts: First, if we grant the notion of hinges a role in our epistemology anyway, then the apparatus is already there, and it will play a pervasive role rather than stay confined to some peripheral phenomena. Second, certainty is not as heavy as it might seem: while there are these very fundamental hinges that ground inquiry and agency, there are more everyday, lighter, ways to acquire something similar to certainties. For example, we take things for granted all the time: you do not always check whether your bicycle brakes still work and whether there is enough pressure in your tires—you just hop on (cf. Wright, 2004, p. 190).

Another line of this objection is to argue that we do not need hinges. We have everything we need with the well-researched menagerie of biases that psychology has to offer. So why use another category, certainties? First, biases are only distorting effects. That is, they strengthen or weaken our credences. Arguably, such distortion effects cannot account for the complete dismissal of strong evidence as occurs in tragic cases of self-deception.

Second, even if biases were able to account for the discarding of compelling unavoidable evidence, they do not give the same kind of explanation as hinges do. Biases are psychological mechanisms; they leave open how a subject treats his or her beliefs, epistemically speaking. We are always subject to epistemic pressures and demands. The dismissal of compelling evidence demands some epistemic explanation if anything does. Biases do not go far enough in that job because they underdetermine what happens in tragic self-deception—so we need something else to lift that explanatory weight: certainties.

Another take on this objection is to argue that tragically self-deceived subjects *take their evidence in a way that* will allow them to infer the self-deceptive conclusion. The flippant response to this would be that, if they are able to reinterpret their evidence, then arguably this is not sufficiently compelling evidence for tragic cases. But I do not think that this would hold up. Rather, I would argue that in cases of tragic self-deception, people do not refuse the evidence by pointing to other independent evidence. This simply is not what happens in tragic self-deception.

The parents in “Misbegotten” cannot point to any strong enough positive evidence for why their child would *never* do such a thing, however privileged it may be; no past evidence would be able to trump or rebut the current *compelling* evidence. At least I do not know what such past evidence would have to look like to warrant dismissing the compelling current evidence. Instead, it just could not be otherwise to them—it is inconceivable.

Additionally, self-deceived subjects would read their past evidence differently in the light of an iHinge. As mentioned, mere biased treatment of past evidence without certainties would not suffice to generate tragic cases of self-deception because the counterevidence could not be compelling in such cases. Something needs to explain the strength of the skewed interpretation of the past: iHinges do this job.

A different issue is borderline cases. There are hinges that are adopted as much out of epistemic curiosity as out of personal pride or desire, which makes it hard to determine whether the subject is self-deceived or not. Take for example Albert Einstein’s refusal to accept the indeterminacy of quantum phenomena. That “God does not play dice” is a classic example of a certainty. Now take some fictitious counterpart of his, Twinstein, who stakes a lot of his ego on his supposedly correct insights into the nature of reality. Is Twinstein tragically self-deceived? It is hard to say. However, I do not think we need to worry. As vagueness appears in many mental phenomena, especially defects, it is not as troubling as it may seem at first sight: we just have to live with it.

## LOOKING BEYOND

What lessons are there to be drawn from the hinge-account of tragic self-deception? First, it shows that and how one can be tragically self-deceived without being pathologically delusional or egocentrically irrational. Cases where a subject proves to be absolutely resistant to evidence constitute an important subclass of self-deception. Tragic self-deception shows how extreme the phenomenon can get. Additionally, hinges explain how the subject can be in such a state without becoming incoherent or egocentrically irrational—thereby avoiding the open and apparent contradictions that certain patients of delusion run into. Even somebody tragically self-deceived will appear to be more objectively rational than someone with, for example, anosognosia. Anosognosia is one symptom of hemiplegia among others, where patients do not recognize that



half their body is paralyzed. They confabulate in an often-inconsistent manner to explain away their handicap and their incoherent behaviour.

Furthermore, iHinges give us a tool to diagnose what exactly is going on in tragic self-deception. Notably, they give us a structure of beliefs that tragically self-deceived subjects have. This contributes to understanding what is happening in such cases: how subjects start treating certain beliefs as fundamental and thus are led astray. Indeed, iHinges may be at work in an even broader range of cases of self-deception than only the most extreme cases. In the long run, we may also be able to differentiate different types of self-deception according to the epistemic mechanism at play.

Finally, the account of tragic self-deception as a manifestation of iHinges is able to tie in self-deception with internalist epistemology. Given its structure, internalism always is subject to a certain pressure to account for what is going wrong when something goes wrong with our rationality. My account can be used to explain how tragic self-deception is possible for an egocentrically rational subject. From an internalist point of view, the difficulties are especially pressing because tragic self-deception poses a threat to internalist accounts, raising the spectre of egocentric irrationality.

## NOTES

- <sup>1</sup> A similar example is considered in Bortolotti and Mameli (2012) and Murphy (2012).
- <sup>2</sup> For example, Pascal's-Wager-style undertakings of bringing oneself to believe something can be argued to be a typical though different instance of self-deception (Jones, 1998, p. 167).
- <sup>3</sup> One of Mele's favourite examples, as the following not necessarily complete list shows (Mele, 1987, p. 125; 2001, p. 26; 2006, p. 110; 2009, p. 262; 2010, p. 746).
- <sup>4</sup> It is their fate to believe this. Tragedy is always based on fate.
- <sup>5</sup> This is part of Mele's characterization, but I would not strictly exclude the possibility of being self-deceived even while believing the truth.
- <sup>6</sup> This characterization deviates from Mele's on several counts (i.e., the *italics*): first, he simply gave sufficient conditions for cases of self-deception, while I give sufficient and necessary conditions for *tragic* self-deception. Second, Mele used "data" instead of "evidence." I am carefully optimistic that experiential data can be adequately captured by some corresponding fine-grained proposition. I will therefore talk about evidence as propositional. Third, the fourth condition is modified so as to make the evidence compelling.
- <sup>7</sup> Consider for example, how supporters of US president Donald Trump are occasionally described by the press: because of their disdain for Hillary Clinton and their attachment to their hero, they manage to not recognize his moral faults (e.g., Allen, 2016; Shapiro, 2016; Douthat, 2017).
- <sup>8</sup> The impossibility of believing contradictions is not uncontested (see, for example, Priest, 1986, p. 102).
- <sup>9</sup> By this *egocentric irrationality*, I do not mean a dialetheist toying around with liar sentences or metaphysical postulates about the nature of god. These are metalinguistic phenomena.
- <sup>10</sup> Noordhof (2003, p. 83-88) argues for something similar, though on different grounds. If the subject is aware of the contrary evidence, but still maintains his or her belief, then that subject cannot be self-deceived because self-deception essentially involves instability in the face of contrary evidence. His example (Noordhof, 2003, p. 76) is structurally similar to my MISBEGOTTEN, so I would call it another case of tragic self-deception. Meanwhile, Noordhof appeals to the intuition that we cannot be self-deceived if we are aware of the contrary evidence, because there would apparently be no deception. I will not defend my position against this claim, but simply point out that he takes a fine-grained view of what self-deception is, while I take the broader deflationary approach (Mele, 2001) that bases itself on the circumstance that instances of motivated resistance against evidence can be and are described as self-deception.
- <sup>11</sup> A third possibility is that the subject is not immune and loses his or her belief, but then that subject would not be self-deceived anymore.
- <sup>12</sup> I thank my anonymous reviewers for pressing me on this point.
- <sup>13</sup> It is disputed whether hinges are beliefs, as many take beliefs to be essentially guided by evidence. I therefore use the more neutral notion of acceptance.
- <sup>14</sup> The epistemic rationality in that case is however atypical: it is not evidential but rather consequentialist or, maybe, transcendental.
- <sup>15</sup> Thanks to Nikolaj Pedersen for this apt nomenclature.
- <sup>16</sup> This is a hinge in its own right that often cannot to be dislodged by any amount of argument.
- <sup>17</sup> This hinge is compatible with the idea that friends betray each other. Betrayal presupposes a friendship to be betrayed.
- <sup>18</sup> Without making their daughter the centre of their life, they would hardly get so far off-track in their self-deception.
- <sup>19</sup> What I describe here is only a disposition to display doubting behaviour; however, I doubt that one effectively must be under permanent tension in order to count as self-deceived.
- <sup>20</sup> This can, for example, be seen with the constantly changing catalogue of syndromes and with the structure of their definitions: "S suffers from I if she/he fulfils at least x of the following y criteria."

- <sup>21</sup> Mele takes a similar strategy, pointing out the multifariousness of morbid jealousy. Either the phenomena of self-deception and delusion belong to entirely different classes or self-deception is a symptom of the delusion (Mele, 2006, p. 121). See also Bortolotti and Mameli (2012).
- <sup>22</sup> Take as an example the phenomenon of senile dementia: there are many patients where no psychiatric treatment is necessary. They live their lives as they've done the past ten, twenty years—with habit protecting them from the dangers of their condition. But at a certain point they may start forgetting more crucial things—e.g., to turn off the stove—or begin to wander aimlessly and get lost, thereby beginning to pose a threat.

## REFERENCES

Allen, Cynthia M., "Self-Deception is the Key to Justifying a Vote for Trump," *Fort Worth Star-Telegram*, August 4, 2016.

Alston, William P., "The Deontological Conception of Epistemic Justification," *Philosophical Perspectives*, vol. 2, 1988, p. 257-299.

American Psychiatric Association (ed.), *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, 5th edition, Washington, D.C., American Psychiatric Publishing, 2013, p. 819.

Audi, Robert, "Self-Deception vs. Self-Caused Deception: A Comment on Professor Mele," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 104.

Bach, Kent, "Thinking and Believing in Self-Deception," *Behavioral and Brain Sciences*, vol. 20, 1997, p. 105.

Bortolotti, Lisa, "Delusion," in N. Zalta, Edward (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2016 edition, 2013, URL: <https://plato.stanford.edu/archives/spr2016/entries/delusion/>.

Bortolotti, Lisa, and Matteo Mameli, "Self-Deception, Delusion and the Boundaries of Folk Psychology," *Humana.mente*, vol. 5, no. 20, 2012, p. 203-221.

Coliva, Annalisa, and Danièle Moyal-Sharrock (eds.), *Hinge Epistemology*, Leiden, Brill, 2016.

Douthat, Ross, "The Bannon Revolution," *The New York Times*, October 17, 2017, URL: <https://www.nytimes.com/2017/10/11/opinion/steve-bannon-revolution.html>.

Foley, Robert, "Rationality , Belief and Commitment," *Synthese*, vol. 89, no. 3, 1991, p. 365-392.

Jones, Ward E., "Religious Conversion, Self-Deception, and Pascal's Wager," *Journal of the History of Philosophy*, vol. 36, no. 2, 1998, p. 167-188.

Kingham, Michael, and Harvey Gordon, "Aspects of Morbid Jealousy," *Advances in Psychiatric Treatment*, vol. 10, no. 3, 2004, p. 207-215.

Kusch, Martin, "Wittgenstein on Mathematics and Certainties," in Annalisa Coliva and Danièle Moyal-Sharrock (eds.), *Hinge Epistemology*, Leiden, Brill, 2016, p. 48-71.

Losonsky, Michael, "Self-Deceivers' Intentions and Possessions," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 121-122.

Lynch, Kevin, "Self-Deception and Stubborn Belief," *Erkenntnis*, vol. 78, no. 6, 2013, p. 1337-1345.

Mele, Alfred, "Self-Deception," *Philosophical Quarterly*, vol. 33, no. 133, 1983, p. 365-377.

———, *Irrationality: An Essay on Akrasia, Self-Deception, Self-Control*, Oxford, Oxford University Press, 1987.

———, "Real Self-Deception," *Behavioral and Brain Sciences*, vol. 20, 1997, p. 91-136.

———, *Self-Deception Unmasked*, Princeton, Princeton University Press, 2001.

———, “Self-Deception and Delusions,” *European Journal of Analytic Philosophy*, vol. 2, no. 1, 2006, p. 109-124.

———, “Have I Unmasked Self-Deception or Am I Self-Deceived?” in Martin, Clancy (ed.), *The Philosophy of Deception*, Oxford, Oxford University Press, 2009, p. 260-276.

———, “Approaching Self-Deception: How Robert Audi and I Part Company,” *Consciousness and Cognition*, vol. 19, no. 3, 2010, p. 745-750.

Moyal-Sharrock, Danièle, “The Animal in Epistemology: Wittgenstein’s Enactivist Solution to the Problem of Regress,” in Coliva, Annalisa and Danièle Moyal-Sharrock (eds.), *Hinge Epistemology*, Leiden, Brill, 2016, p. 24-48.

Murphy, Dominic, “The Folk Epistemology of Delusions,” *Neuroethics*, vol. 5, no. 1, 2012, p. 19-22.

Noordhof, Paul, “Self-Deception, Interpretation and Consciousness,” *Philosophy and Phenomenological Research*, vol. 67, no. 1, 2003, p. 75-100.

Priest, Graham, “Contradiction, Belief and Rationality,” *Proceedings of the Aristotelian Society*, vol. 86, 1986, p. 99-116.

Shapiro, Ben, “You Can’t Pretend Trump’s Flaws Away,” *National Review*, October, 2016, URL: <http://www.nationalreview.com/article/440742/donald-trump-supporters-self-delusion>.

Wittgenstein, Ludwig, *Über Gewissheit = On Certainty*, Oxford, Blackwell, 1969.

Wright, Crispin, “Warrant for Nothing (and Foundations for Free)?” *Aristotelian Society Supplementary Volume*, vol. 78, no. 1, 2004, p. 167-212.

# 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 appealed 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.<sup>1</sup>

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.<sup>2</sup>

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.<sup>3</sup> 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.<sup>4</sup> 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?<sup>5</sup>

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.<sup>6</sup> 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?<sup>7</sup> 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.<sup>8</sup>

## 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 et 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) *S* assesses the distressing evidence as ambiguous (weight of evidence);
- (ii) *S* appraises the distressing evidence as having a significant negative impact on his or her well-being (affective coping);
- (iii) *S* appraises his or her control on the situation as low (coping potential); and
- (iv) *S* 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 obstructive, 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) S does not appraise the evidence as ambiguous;
- (ii) S does not appraise the evidence as having a significant negative impact on S's well-being; and
- (iii) S 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) *S* appraises the evidence in favour of *p* as ambiguous;
- (ii) *S* appraises the distressing evidence as negative;
- (iii) *S* appraises his or her coping potential as low;
- (iv) *S* 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.

## ACKNOWLEDGMENTS

This article has been partly presented at the conference *Self-Deception: What It Is and What It Is Worth* (University of Basel, October 25-27, 2017, The Cognitive Irrationality Project) and at the Conference of the Italian Society for Analytic Philosophy (Novara, University of Oriental Piedmont, September 4-7 2018). We would like to thank the organizers and participants of these conferences. In particular, we wish to express our gratitude to Alfred Mele, Anne Meylan, Elisabetta Galeotti, Christine Tappolet, Neil Van Leeuwen, Patrizia Pedrini, Elisabeth Pacherie, Dana Nelkin, Martina Orlandi, Marie van Loon, Melanie Sarzano, and two anonymous reviewers for their fruitful feedback.

## NOTES

- <sup>1</sup> It is assumed that in this debate desires differ from emotions.
- <sup>2</sup> 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.
- <sup>3</sup> 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.
- <sup>4</sup> 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.
- <sup>5</sup> 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.
- <sup>6</sup> 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.
- <sup>7</sup> 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.
- <sup>8</sup> 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.

Heatherston, 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 Mascaró, 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.

# COSTLY FALSE BELIEFS: WHAT SELF-DECEPTION AND PRAGMATIC ENCROACHMENT CAN TELL US ABOUT THE RATIONALITY OF BELIEFS

MELANIE SARZANO

PHD STUDENT, UNIVERSITY OF ZURICH

## ABSTRACT:

In this paper, I compare cases of self-deception and cases of pragmatic encroachment and argue that confronting these cases generates a dilemma about rationality. This dilemma turns on the idea that subjects are motivated to avoid costly false beliefs, and that both cases of self-deception and cases of pragmatic encroachment are caused by an interest to avoid forming costly false beliefs. Even though both types of cases can be explained by the same belief-formation mechanism, only self-deceptive beliefs are irrational: the subjects depicted in high-stakes cases typically used in debates on pragmatic encroachment are, on the contrary, rational. If we find ourselves drawn to this dilemma, we are forced either to accept—against most views presented in the literature—that self-deception is rational or to accept that pragmatic encroachment is irrational. Assuming that both conclusions are undesirable, I argue that this dilemma can be solved. In order to solve this dilemma, I suggest and review several hypotheses aimed at explaining the difference in rationality between the two types of cases, the result of which being that the irrationality of self-deceptive beliefs does not entirely depend on their being formed via a motivationally biased process.

## RÉSUMÉ :

Dans cet article, je compare les cas classiques de duperie de soi aux cas que l'on trouve dans les débats sur la question de l'empiètement pragmatique et défends l'idée selon laquelle ces deux types de cas peuvent être compris comme étant produits par un même mécanisme visant à éviter la formation de croyances fausses coûteuses. Cette comparaison nous mène naturellement à former un dilemme à propos de la rationalité des croyances. Le dilemme repose sur l'idée que bien que ce mécanisme mène à la formation de croyances irrationnelles dans les cas de duperie de soi, il ne semble pas affecter la rationalité du sujet dans les cas d'empiètement pragmatique : alors que les sujets autodupés sont irrationnels, les sujets décrits dans les cas d'empiètement pragmatique ne le sont pas. Pour résoudre ce dilemme sans rejeter les présupposés selon lesquels les croyances issues de la duperie de soi sont irrationnelles et que les cas sur lesquels repose l'empiètement pragmatique sont rationnels, je propose plusieurs hypothèses visant à expliquer cette différence, prouvant ainsi que ce dilemme n'est qu'apparent et que l'irrationalité de la duperie de soi ne peut uniquement dépendre de ce mécanisme sous l'influence de considérations pratiques.

## 0. INTRODUCTION

Sometimes, subjects hold irrational, motivated beliefs in the face of evidence to the contrary, mistreat the evidence at hand, and seem impervious to any evidence contradicting their desire. Such subjects are likely to be self-deceived.

In reaction to more traditional views, *motivationists* about self-deception have recently argued that self-deceptive beliefs can be understood as *motivationally biased beliefs*: they are the result of the subjects' motivation, such as a desire or an emotion, that biases their treatment of the evidence, leading them to form this irrational belief (Johnston, 1988; Barnes, 1997; Funkhouser, 2005; Scott-Kakures, 2002; 2012; Lazar, 1999; Mele, 1997; 1999; 2001; Nelkin, 2002).

In this paper, I argue that, understood this way, self-deception shares interesting similarities with high-stakes cases borrowed from the literature on pragmatic encroachment. Despite these striking similarities, these two types of cases have rarely, if ever, been compared in either part of the literature.<sup>1</sup> Although these two types of cases differ in many ways, I argue that they can both be explained by reference to the same underlying belief-formation mechanism: a mechanism characterized by the influence of practical factors—in particular, the influence of costly errors. While Mele (1997; 1999; 2001) refers to this mechanism to explain the distorted ways in which subjects come to form irrational, self-deceptive beliefs, the same mechanism can be invoked to explain the subject's epistemic behaviour in rational high-stakes cases.

My argument rests upon the assumption that self-deceptive beliefs generated by this type of belief-formation mechanism are irrational, whereas high-stakes cases typically aren't. If this mechanism really does participate in making the former, but not the latter, irrational, we then face a dilemma about the rationality of beliefs. According to this dilemma—or, as we will see, what I take to be an *apparent* dilemma—we are forced either to accept that self-deception is rational or to accept that high-stakes cases are irrational.<sup>2</sup> In order to solve this dilemma, I suggest and review several hypotheses to explain this difference in rationality, and I argue that the best way of articulating the difference is by drawing a line between two types of motivation, or two types of costs influencing the belief-formation process. If this line of thought is correct and if there really is a way of dissolving the dilemma, then the irrationality of self-deceptive beliefs, and of irrational beliefs in general, cannot be explained merely by the fact that the beliefs in question result from a motivationally biased process.

In section I, I introduce self-deception as presented in the motivationist framework. In section II, I turn to classical high-stakes cases borrowed from the literature on pragmatic encroachment and put forward the idea that, contrary to cases of self-deception, these cases are cases in which the subject's epistemic state is rational. In section III, I begin by showing how both sets of cases can in fact be explained by referring to the same type of belief-formation mechanism: Friedrich-Trope-Liberman (FTL) model of lay hypothesis testing. On this basis,



I argue that, if we assume that this belief-formation process is what causes these beliefs to be irrational, then we are led to a dilemma about the rationality of beliefs. Finally, in section IV, I suggest several hypotheses for explaining this difference in rationality and evaluate them in turn, finally putting forward a solution to the dilemma.

## I. SELF-DECEPTION

Some beliefs are undoubtedly irrational. Self-deceptive beliefs are of this type, they are beliefs held in the face of evidence to the contrary. Such cases typically involve subjects refusing to believe that their unfaithful partner is cheating on them, parents resisting the fact that their children aren't as perfect as they take them to be, and lovers mistakenly believing, despite continuous rejection, that their love is requited. In all of these cases, subjects aren't merely hoping or wishing reality were different: they seem to strongly believe that reality is such, that it matches their desires and respond to evidence in a very biased way, despite the facts being clear to everyone but themselves.

There are two main families of theories about self-deception. On the classical, *intentionalist* view (Davidson, 1985; 2004; Pears, 1984; Rorty, 1988; Talbot, 1995), self-deception is understood literally, as a case of deception in which the deceiver and the deceived are one and the same person. This model is closely built on our common understanding of interpersonal deception, of cases in which a first subject intentionally tricks another into believing something false. On the intentionalist view, then, subjects are self-deceived only if they intentionally deceive themselves and, at least on most intentionalist accounts, hold contradictory beliefs (an initial belief that not- $p$  and a self-deceived belief that  $p$ ). This approach has famously led to intricate puzzles, for pretty obvious reasons; what can easily be applied to two distinct minds in cases of interpersonal deception becomes overly complicated and tricky when applied to a single mind. How does a single mind intentionally deceive itself, since it is probably aware of its own intention to do so? And how is it possible for a single subject to come to believe  $p$  when that subject already believes, and knows, that this belief is false?

In response to these thorny issues, *motivationists* (sometimes also called *non-intentionalists*) have recently argued that self-deception need not involve an intention or any contradictory beliefs. On the motivationist view, all that is required for a self-deceptive belief to be formed is a motivational factor influencing the subject's belief-formation process (Johnston, 1988; Barnes, 1997; Funkhouser, 2005; Scott-Kakures, 2002; 2012; Lazar, 1999; Mele, 2001; Nelkin, 2002). Here, it is the subjects' desire or emotion that biases their belief formation and influences their treatment of the evidence, thereby resulting in them holding a self-deceptive, irrational, belief. Because the intentionalist view is problematic in many aspects, I will here assume that the motivationist framework is correct and work with the following—fairly uncontroversial—characterization of self-deception primarily inspired by Alfred Mele's deflationary

account (1997; 1999; 2001). On Mele's account, a subject *S* is self-deceived in believing a proposition *p* if the following conditions obtain:

- (a) *unwarranted belief*  
*S*'s belief that *p* is false (or at least, unwarranted).
- (b) *doxastic alternative*<sup>3</sup>  
*S* possesses, or has been in contact with, evidence supporting the belief that not-*p*.
- (c) *motivated belief*  
*S*'s belief that *p* is the result of a motivationally biased process.

These conditions capture what is essential to most motivationist accounts of self-deception: (a) that the self-deceptive belief is unwarranted (i.e., that the subject *should not* in fact believe *p* given the evidence at hand but believes *p* nevertheless);<sup>4</sup> (b) that the subject has been presented, or has been in contact with, evidence supporting the exact opposite of what he or she currently believes (i.e., that, given the evidence, he or she should in fact believe not-*p*);<sup>5</sup> and (c) that *S*'s belief that *p* is a direct consequence of the subject's motivation that *p*. To illustrate this view, consider the following example.

*Fernando*

Fernando is a soft-hearted postdoc who happens to be madly in love with Steve. His fondness for Steve has grown into an obsessive love over the years and it is clear that he is now craving to see his love returned. The two men accidentally run into each other every few months at conferences and workshops, make small talk, and act politely. Steve has, so far, never shown any sign of mutual attraction and his small talk and polite behaviour only ever warranted Fernando to believe that Steve simply appreciates him as a colleague. In fact, it should even be clear to Fernando, after several refusals to have coffee together, or spend time together, that Steve isn't romantically interested in Fernando. Nevertheless, Fernando fallaciously interprets Steve's every little gesture and word as conclusive proof of their impending love affair.

In this example, (a) Fernando unwarrantedly believes that his love is requited, (b) when he should in fact believe the exact opposite (i.e., that Steve isn't romantically interested), and, what is more, (c) Fernando's belief is formed and maintained through a motivationally biased process that influences his interpretation of the evidence at hand. In other words, Fernando is self-deceived, and sadly Steve isn't about to declare his love to him.

As I said, it is central to motivationist accounts that the subject mistreats the evidence at hand (cf. condition [c]) or is insensitive to evidence. According to Mele (1997; 1999; 2001), there are several ways in which a subject can mistreat evidence. In the aforementioned case, Fernando is mistreating the evidence,

mainly by ignoring the evidence against Steve being romantically interested, as well as by freely interpreting Steve's behaviour as conclusive proof that Steve loves him.

More generally, as Mele (2001) rightly describes, self-deceived subjects mistreat evidence in a number of ways. For example, self-deceived subjects may *positively misinterpret evidence* (interpret evidence that doesn't support *p* as evidence supporting *p*), they may *negatively misinterpret evidence* (interpret as not supporting *p* evidence that in fact supports *p*), they may *selectively focus on or attend to evidence* according to whether it supports *p* (selectively focus on evidence supporting their belief that *p* and/or fail to attend to evidence counting against *p*), or they may *selectively gather evidence* (overlook available evidence counting against *p* or actively search for less accessible evidence supporting *p*). According to Mele, these are all ways in which subjects' desire (or motivation) may lead them to misinterpret the evidence at hand, depending on the particular cases.

Motivationists, Mele included, generally assume that it is a *desire that p* that plays this biasing role. Nevertheless, there are variations and disagreements amongst motivationists regarding what kind of motivation can trigger this type of fallacious reasoning. Nelkin (2002), for example, argues that it is a *desire to believe p* rather than a desire that *p* that is at play in self-deception. Mele himself also admits, in his article on twisted self-deception (1999), that an emotion (e.g., jealousy) can play a similar biasing role that leads a subject to form a self-deceptive belief that that subject in fact wished were false.

How exactly these motivations, desires, and emotions might psychologically play a role in leading the subject to mistreat evidence is amply discussed in the literature (see Mele, 1997; 1999; 2001). In section III, we will focus on what Mele identifies as one of the underlying psychological mechanisms leading self-deceived subjects to mistreat evidence. But before turning to this aspect of self-deception, I will first present a different type of cases: high-stakes cases borrowed from the literature on pragmatic encroachment. As we will see, contrary to self-deceptive subjects, subjects presented in this type of case do not seem so irrational.

## II. PRAGMATIC ENCROACHMENT

Contrary to what traditional views in epistemology assume, defenders of pragmatic encroachment argue that knowledge doesn't purely depend on truth-related factors but also varies according to the subject's practical interests—i.e., what is at stake for the subject in a given situation<sup>6</sup> (Stanley, 2005; Hawthorne, 2004; Fantl and McGrath, 2002; 2007; Hookway, 1990). In other words, advocates of the pragmatic-encroachment thesis defend the idea that a difference in the subject's practical circumstances, that a subject's interests and related risks, can *encroach* on the subject's epistemic state. This, of course, drastically departs

from the default view in epistemology, *purism*, according to which only truth-related factors such as truth, justification, reliability, and so on determine alone whether a subject is in a position to know.<sup>7</sup>

One way of motivating this original position is by appealing to a set of cases, the most famous of which are the so-called *bank cases*, originally presented by DeRose (1992).<sup>8</sup>

*Low Stakes.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. It is not important that they do so, as they have no impending bills. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Hannah remembers the bank being open on Saturday morning a few weeks ago, so she says, ‘Fortunately, it will be open tomorrow, so we can just come back.’ In fact, Hannah is right—the bank will be open on Saturday.

*High Stakes.* Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since their mortgage payment is due on Sunday, they have very little in their account, and they are on the brink of foreclosure, it is very important that they deposit their paychecks by Saturday. But as they drive past the bank, they notice that the lines inside are very long, as they often are on Friday afternoons. Hannah remembers the bank being open on Saturday morning a few weeks ago, so she says, ‘Fortunately, it will be open tomorrow, so we can just come back.’ In fact, Hannah is right—the bank will be open on Saturday. (Schroeder, 2012)

The reason why these cases support pragmatic encroachment is because our intuitions about whether the subject knows whether the bank will be open seem to shift from one case to the other. While we all tend to agree that Hannah knows that the bank will be open in the low-stakes case, many tend to disagree that Hannah is in a position to know that the bank will be open when the stakes are raised. Only, to admit this shift in intuition, to say that what is at stake for Hannah and her wife affects whether Hannah is in a position to know or not, amounts to denying purism and conceding that practical factors actually play a role in determining whether a subject is in a position to know.<sup>9</sup> For, if purism were true, Hannah would know that the bank will be open on Saturday morning both in the low-stakes case and in the high-stakes case, since the only variation between the cases is what is at stake for Hannah and her wife—i.e., the practical costs of being mistaken. Evidence, on the contrary, is stable across the cases.

This shifting intuition can, of course, be explained in a variety of ways. One might reject the validity of these intuitions and argue in accordance with purism that, if the subject knows *p* in one case, the subject must also know *p* in the other.

Another might argue that what is actually going on in the high stakes case isn't that the subject isn't in a position to know, or that the subject doesn't believe the target proposition, but rather that the subject isn't in a position to act rationally on the basis of this belief. There are indeed many ways of dealing with these cases, some of which are sympathetic to purism, others plainly rejecting it. As Engel (2009) notices, there are even different varieties of pragmatic encroachment, varying along two dimensions: (i) the kind of epistemic notion upon which these factors impede (whether it is knowledge, justification, belief, or rationality); and (ii) the degree of this influence.

Since my focus here concerns the rationality of beliefs, I will follow Schroeder's (2012) contribution, where he articulates pragmatic encroachment as a thesis about the rationality of beliefs. In his paper, Schroeder explores how we can make sense of the idea that practical considerations could encroach on knowledge. His solution to this puzzling idea is to say that practical considerations don't directly affect knowledge; rather, there is pragmatic encroachment on knowledge only because there is pragmatic encroachment on epistemic rationality. Given that epistemic rationality is a condition of knowledge, if high stakes defeat epistemic rationality, they defeat knowledge by defeating epistemic rationality. On his view, then, in the high-stakes case Hannah isn't in a position to know because she isn't in a position to rationally hold the belief that the bank will be open on Saturday.

As I said, pragmatic encroachment is a controversial position, and I do not wish to defend it here, since this would go far beyond the purposes of this paper. All I need to assume for the purposes of this paper is that there are at least some high-stakes cases in which the subjects would be more rational to suspend their belief, given the high stakes, than to believe  $p$ , and that, were they to believe  $p$ , their belief would be irrational. If you are not convinced that the bank cases meet these criteria, consider the following case.

*Molly*

Molly is about to pick some mushrooms in the forest to prepare the evening dinner for her family. She has good evidence that most if not all mushrooms in this area are edible (let us say her grandmother is a knowledgeable mushroom hunter who told her so). All the same, Molly is aware that many mushrooms are highly poisonous and that mistakes in identifying them can happen. Because Molly is motivated to avoid making a lethal mistake (she might die or poison her family as a result of this mistake), Molly suspends her belief and decides to check her *Mushroom Book* before serving dinner to find out whether the mushrooms she picked are edible.

In this case, there is less controversy about what exactly is going on, since it is explicitly mentioned that Molly suspends her belief. All the same, I think this is a plausible and ordinary case: sometimes, when the stakes are high, we suspend belief for the very reason that mistakenly holding the belief in question could

bear disastrous consequences. Following Schroeder's (2012) argument, as well as our intuitions, I think we can easily assume that there is something significantly more rational, if not plainly rational, in Molly's suspension of belief contrasted with Fernando's self-deceptive belief described in section I.<sup>10</sup>

Bearing this in mind, we will now see how comparing cases of self-deception with high-stakes cases of this type can generate a(n apparent) dilemma about the rationality of beliefs. I will thus begin by presenting the belief-formation mechanism that Mele (1997; 1999; 2001) appeals to, to explain how self-deceptive beliefs are formed. I will then argue that this very mechanism can also explain high-stakes cases.

### III. THE (APPARENT) DILEMMA

As I suggested, both self-deception and high-stakes cases can be explained with reference to the same type of belief-formation mechanism, a type of belief-formation process under the influence of pragmatic considerations. This mechanism might not be sufficient for explaining all the puzzling aspects of self-deception, nor might it be capable of explaining all types of cases. All I claim is that it is sufficient to explain at least some cases, and that this is sufficient for generating a dilemma about the rationality of beliefs. What is more, it also seems to be this pragmatic influence that renders the resulting belief irrational. Only, if this is the case, then it becomes unclear why high-stakes cases aren't cases in which the subject is being irrational—a conclusion that I deem unlikely. Whereas self-deceptive beliefs are considered a paradigm of epistemic irrationality, high-stakes cases are (as we saw section II) at least somewhat more rational than cases of self-deception. Before articulating the dilemma, I will thus begin by explaining why I think both types of cases can be explained by the same type of mechanism.

In describing the psychological mechanisms underlying the various ways in which a subject may be led to misinterpret evidence in a motivationally biased way, Mele (1997; 1999; 2001) refers to what he labels the *Friedrich-Trope-Liberman (FTL) model* of lay hypothesis testing and ordinary reasoning, inspired by Friedrich's (1993) and Trope and Liberman's (1996) findings in psychology. The FTL model, alongside a variety of biases Mele mentions (1997; 1999; 2001), contributes to explaining the distorted ways in which self-deceived subjects mistreat evidence, just as Fernando does in this example.

According to the FTL model, subjects are pragmatic thinkers whose cognitions are first and foremost concerned with minimizing costly errors rather than driven by an interest for truth. This way of testing hypotheses and forming beliefs incorporates the subject's practical interests into the way in which the subject searches, gathers evidence, and forms or rejects beliefs. This mechanism for hypothesis testing and belief formation isn't purely truth oriented in the way scientific hypothesis testing would be. On the contrary, the type of hypothesis testing displayed in everyday reasoning is largely influenced by the practical costs thinkers associate with the possibility of forming a false belief or rejecting

a true one. The subjects' motivation for avoiding costly errors is implemented in their belief-formation process through a system of thresholds for belief formation and rejection that determine the amount of evidence required for forming or rejecting a belief. These two thresholds—which needn't be symmetrical—vary according to the costs of mistakenly believing (forming a false belief) or mistakenly rejecting a belief (rejecting a true belief), as well as according to the costs of information (that is, the costs of gathering further evidence, including resources, time, and effort) (Trope and Liberman, 1996; Mele, 2001). For example, a subject who associates high costs with falsely believing  $p$  will have a high threshold for forming this belief. If the costs of information are reasonable relative to the costs of falsely believing  $p$ , the subject will search for more evidence about this matter. The costs of falsely believing something aren't limited to material consequences. As Friedrich (1993) notes, self-serving motives, such as maintaining one's self-esteem, can also play a role in the determination of errors to be avoided. Thus, the practical costs of falsely believing include a wide range of costs, including psychological ones (those linked to the subject's self-image, for instance) as well as material ones (such as poisoning).<sup>11</sup>

To see how this model functions, consider the following example. Imagine you are walking in your back garden and notice a few rhubarb leaves growing there. You know that this plant is somehow edible, but you aren't quite sure which part of the plant you can actually eat, or whether some of it might be indigestible. You also remember eating delicious rhubarb pie but cannot quite remember which part of the plant had been used in the preparation. You form the following hypothesis.

(H) *Rhubarb leaves are edible.*

Knowing that the risks of forming a belief on the basis of this hypothesis might be fatal (you are aware that some plants are poisonous), you are cautious and do not jump to conclusions. Instead, you test the hypothesis by searching for disconfirming evidence—that is, evidence that rhubarb leaves are poisonous—for this will give you the best chances of avoiding forming a false belief. And because the costs of falsely believing that rhubarb leaves are edible are high (i.e., the risk of suffering some sort of food poisoning), the amount of evidence you gather before forming the belief in question is important: you do not want to make a mistake about this particular matter and poison yourself or your guests. In other words, the higher the stakes or the costs of falsely believing  $p$ , the more evidence one needs in order to form a given belief.

The FTL model, as Mele points out, participates in the formation of self-deceptive beliefs. If we apply this model to a case of self-deception such as the one presented above, we can see how the model predicts and explains the distorted ways in which some self-deceived subject forms his or her belief. In this particular case, Fernando is self-deceived in believing that Steve is in love with him. Fernando not forming the appropriate belief on the basis of the evidence can be explained by the fact that he associates high costs with forming the belief that

Steve isn't romantically interested, and low costs with falsely believing that his love is requited. The truth (not- $p$ ) might break Fernando's heart and lead to severe emotional consequences for him, whereas the costs of falsely believing that his love is requited are low, as the belief serves only to maintain him in a happy bubble.<sup>12</sup> These costs of believing not- $p$  (i.e., that Steve isn't in love with him) thus set a high threshold for forming the appropriate belief, and the amount of evidence required for believing not- $p$  is correspondingly high. If this is correct, then at least some cases of self-deception, if not all, can be explained by referring to this type of belief mechanism as the one described by the FTL model.<sup>13</sup>

Self-deceptive beliefs, we have seen, are formed through a motivationally biased process. The FTL model presented above explains how a subject can be led to mistreat evidence in order to avoid forming costly false beliefs.<sup>14</sup> This mechanism for belief formation participates in making the belief irrational since it causes the subject to not form the appropriate belief. The motivation to avoid forming costly false beliefs—as well as, more generally speaking, the influence of practical considerations on beliefs—is typically deemed irrational. This suggests that the FTL model can at least explain some cases of self-deception: the costs of falsely believing function as a motivation to avoid certain undesired consequences that influences the subject's treatment of evidence and leads the subject to form a self-deceptive, irrational, belief.

Interestingly, these high-stakes cases can easily be explained by referring to the FTL model, with the very same mechanism that was responsible for the generation of irrational beliefs in cases of self-deception. In high-stakes cases such as the one described above, the costs associated with falsely believing  $p$  are high: Hannah and Sarah will be in a very uncomfortable situation if they fail to deposit their paycheques before Sunday. Since these costs are high, the threshold for belief formation is equally high and the amount of evidence required for believing that the bank will be open on Saturday is greater than that in the low-stakes case. This can explain why the amount of evidence can be sufficient for belief formation in the low-stakes case, without being sufficient when the stakes are high. Hannah's concern for avoiding costly false beliefs influences her threshold for belief formation and leads her to suspend her belief until further evidence is gathered. The same goes for Molly: her position is perfectly analogous to the example used to illustrate the FTL model. It is because the costs associated with falsely believing that the mushrooms are edible are extremely high that the threshold for belief formation is equally high. This results in Molly not forming the belief that the mushrooms are edible until she gathers further evidence from her *Mushroom Book*.

But, contrary to cases of self-deception such as the one described previously, the stakes here influence Hannah and Molly's beliefs in a rational way. Indeed, as I said, cases of this sort are used in the context of debates on pragmatic encroachment to bring about the intuition that there's a good reason for subjects to suspend their belief and search for further evidence. In what follows, I rely on



the intuition that in such cases the perceived costs of falsely believing actually have a positive effect on the subject's belief, in the sense that, even if we understand high-stakes cases by referring to the system of adaptive thresholds described by the FTL model, this does not affect the subject's rationality. Cases of self-deception thus provide us with an example in which pragmatic factors (in particular, the costs of forming a false belief) influence the subject's belief formation. This influence is what makes the belief irrational and insensitive to evidence. Take the example of Fernando again. Fernando believes (falsely) that Steve is in love with him. This belief results from the fact that Fernando associates high costs with falsely believing that Steve doesn't love him and associates low costs with believing that Steve does love him. According to the standard understanding of self-deception, what makes self-deceptive beliefs irrational lies precisely in the fact that the beliefs are formed through a motivationally biased process: a process under the influence of practical factors (such as desires, the motivation to avoid costly false beliefs, and so on), an influence that leads the subject to mistreat the evidence at hand. But, as we have seen with cases of pragmatic encroachment, the same mechanism can influence a subject's belief without affecting that subject's rationality. On the contrary, this influence seems warranted: when the stakes are high, it is more rational for subjects to suspend their belief until further evidence is gathered, rather than holding onto it. Indeed, were Hannah or Molly to believe  $p$  in the high-stakes situation—that is, were the stakes to not influence her belief—her belief would be irrational. In these cases, it might be more rational for Hannah or Molly to suspend her belief. Both sets of cases are cases in which the practical costs of forming false beliefs affect the subject's belief through what seems to be a system of thresholds for belief formation and rejection. But, while in the first case this influence seems to make the subject's belief irrational,<sup>15</sup> in the second this influence does not undermine the subject's rationality. The puzzling aspect of this comparison is precisely that this biased mechanism is what determines the rationality of the belief in the first case. If we accept these claims, we are then led to the following dilemma about rationality.

*Dilemma*

*Either* this mechanism produces irrational beliefs, and both self-deception and cases of pragmatic encroachment are irrational  
*or* this mechanism affects beliefs in a rational way, and both cases of pragmatic encroachment and cases of self-deception are rational.

There are of course three options for responding to this dilemma. We can decide either to embrace either the first horn or the second or to reject the dilemma by explaining why it doesn't hold. In what follows, I assume that neither horn is likely (that self-deception is indisputably irrational and that there is something rational—or at least significantly more rational—about suspending one's belief in high-stakes cases). If we want to find a way out of this dilemma, we need to deny that the rationality of one's belief has anything to do with being formed via the mechanism of adaptive thresholds described by the FTL model. If this line of thought is correct, the dilemma is merely apparent, and the irrationality of

self-deception lies elsewhere: it has to be found in a feature of self-deception that cases of pragmatic encroachment do not possess.

#### IV. HYPOTHESES AND SOLUTIONS

In this final section I spell out several differences between the two cases to find out which element, present in self-deception but not in cases of pragmatic encroachment, is responsible for the irrationality of self-deceptive beliefs. The hypotheses listed in the table below represent some, and hopefully all, of the relevant differences between the two cases. I do not intend to say that these hypotheses are mutually exclusive or that they cannot be combined to furnish a full explanation of this difference in rationality, nor do I mean to say that they bear no logical connection to one another. As we will see, I think some of the hypotheses are in fact related.

<i>Hypothesis</i>	<i>Self-deception</i>	<i>Pragmatic Encroachment</i>
H1	Falsity	Truth
H2	Formation	Suspension
H3	Quick reasoning	Slow reasoning
H4	Costs of believing $p$	Costs of $p$ being false

##### *Hypothesis 1: Truth versus falsity*

According to the first hypothesis, the fact that in cases of self-deception the subject's belief is false, whereas in cases of pragmatic encroachment the belief is true, can explain why cases of self-deception are irrational and high stakes cases are not. This difference might seem like a very trivial and simplistic one, for, as we know, falsity of the belief isn't a necessary condition for self-deception. It is indeed easy to see how, even though in most cases the self-deceptive belief will happen to be false, it could also accidentally be true. Irrationality and falsity are independent from one another, and a subject could be warranted to believe  $p$  even though  $p$  is false.

##### *Hypothesis 2: Suspension versus formation*

One can draw attention to the sort of impact these practical costs have on the subject's belief. In one case, self-deception, the subject's belief results from practical considerations, whereas in the case of pragmatic encroachment the subject merely suspends his or her belief. In his (2012) article, Schroeder argues that reasons to withhold belief are necessarily nonevidential. He writes:

Why is it that reasons to withhold cannot be evidence? It is because the evidence is exhausted by evidence which supports  $p$  and evidence which

supports  $\sim p$ . Consequently, the reasons to withhold must come from somewhere else. So they cannot be evidence. (Schroeder, 2012)

If Schroeder is correct, and if reasons to believe must be evidential, whereas reasons to withhold belief can only be practical, then this might explain the difference between the two sets of cases. This argument would allow us to explain why the influence of practical costs makes beliefs irrational in the case of self-deception, but not in typical cases of pragmatic encroachment, if, as I said, cases of self-deception are cases in which subjects *form* beliefs, and cases of pragmatic encroachment are cases in which subjects *suspend* beliefs. But is this true?

Although I think all cases of the second type—that is, cases of pragmatic encroachment—are cases in which subjects suspend their belief, the question of whether there can be cases of self-deception in which the self-deceptive subjects also suspend their belief is less straightforward. Indeed, despite self-deception being mostly defined as a process by which a belief is *formed*, applying only to cases in which a subject is self-deceived in believing  $p$ , this definition might be too restrictive (cf. the definition in section I). It does not seem to be true that one cannot be self-deceived in simply refusing to believe something, in suspending one's belief in the face of strong evidence instead of forming the warranted belief. There are cases in which the subject suspends belief for practical reasons, but these cases fail to be rational, and these cases are cases of self-deception. Here is an example of this sort.

*Suspension Jo*

Jo is meeting Laurie for a drink by the lakeside. She has very good reasons to believe that this is a date and that Laurie is romantically invested: he regularly sends her handwritten letters and red roses and bakes her delicious cookies. Their best friend also confirmed that Laurie does this only when he's madly in love. Nevertheless, Jo refuses to believe that Laurie is romantically interested, and suspends belief. Her reason for doing so is that the mere thought of him liking her in return is overwhelming and puts her in a general state of emotional distress.

Although the case doesn't fit with the definition of self-deception presented above (insofar as the definition mentions only cases in which the subject forms or holds a belief), I think some might still hold the intuition that Jo is self-deceiving in suspending belief. Imagine a discussion between Jo and her sister Beth: Beth is insisting that Laurie is in love with Jo, but Jo refuses to see the evidence supporting this belief. Would Beth be tempted to tell her sister that she is self-deceived? Maybe. Maybe not. Some might want to argue that self-deception necessarily involves forming a belief and that merely suspending one's belief, however irrational this may be, cannot constitute a genuine case of self-deception. However, it isn't crucial that we agree on calling this a case of self-deception; all we need is to agree that Jo is being irrational in refusing to believe that Laurie is romantically interested in her, which I think she is. If so, then the third

hypothesis does not account for the intuitive difference in rationality between the two sets of cases.

*Hypothesis 3: Quick versus slow*<sup>16</sup>

Another way of explaining the difference in rationality between the two cases is by referring to the way in which the beliefs are formed—i.e., the type of reasoning by which they are formed. One could argue that for a belief to be self-deceptive it should not be sufficient that the belief-formation process be influenced by the costs of falsely believing. In addition to this, the belief should be formed through a certain type of reasoning. Accordingly, the threshold variation isn't what causes beliefs to be irrational in cases of self-deception; rather, it is the biased treatment of the evidence that causes the belief to be irrational. Mele (1997; 1999; 2001) refers to several cognitive biases that can be triggered by desires or emotions and that might participate in the formation of self-deceptive beliefs. Such biases include the confirmation bias, the vividness-of-information bias (Mele, 1997; 1999; 2001), and even the availability-heuristics bias (Mele, 1997; 1999; 2001).

By contrast, subjects in cases of pragmatic encroachment do not go through this type of process, and this is why their attitude isn't irrational. A more precise formulation of this idea can be given by referring to dual-process theory (Evans, 2010; Evans and Frankish, 2009; Frankish, 2010). Mainly developed by Tversky and Kahneman (1974), dual-process theory provides a schematic vision of our cognition divided into two systems, two distinct ways of reasoning and treating information: system 1 and system 2. The first system is intuitive: it is quick, implicit, and effortless. The second process on the contrary is reflective; it is slower and relies on controlled, analytic thinking. Cognitive biases are understood as belonging to system 1: they allow us to treat information effectively while investing little cognitive effort, but can lead to errors. System 2 is more rational, but requires more cognitive effort and time.

This second hypothesis draws upon the dual-process theory to explain what might be irrational about self-deception and rational in cases of pragmatic encroachment. The relevant difference between these cases has nothing to do with motivation. In fact, the difference has to do only with the way in which this motivation affects our thinking. In cases of self-deception, what makes the belief irrational is that it results from a certain type of reasoning process, an intuitive, associative way of treating information, that results in a false belief. It is the motivation, the emotion, or the desire that leads the subject to think in system 1: the self-deceptive subjects don't reason analytically. And this is what differentiates them from the subjects in cases of pragmatic encroachment. Subjects in cases of pragmatic encroachment engage in slow, analytical thinking. Their motivation for reevaluating their belief might be a practical one, but their thinking isn't biased by their motivation. It isn't the belief in itself that is irrational, but only the process by which it is formed. According to this hypothesis, the influence of the practical costs of falsely believing in one's thresholds for belief

acceptance and rejection doesn't play a role in determining whether the resulting belief is rational or not: what matters is the type of reasoning or thinking that leads to this belief.

This hypothesis fails to explain the difference for the following reason: many rational beliefs are formed through system 1; therefore, relying on biases and heuristics does not necessarily make a subject's belief irrational. If this is correct, as I think it is, then why would this type of reasoning be what determines the belief's rationality only in this specific case? Since the answer to this question forces us to search for a further criterion, it cannot be a convincing solution to the dilemma.

*Hypothesis 4: Costs of believing  $p$  versus costs of  $p$  being false*

According to the fourth hypothesis, the relevant difference between the cases concerns the kind of practical costs at play. Indeed, cases of self-deception are cases in which the costs influencing the subject's threshold for belief are those related to holding the belief itself rather than the costs related to falsely believing  $p$ . In other words, what influences the subject's belief doesn't depend on the falsity of the belief; it merely depends on holding the belief regardless of whether it might be true or false. What matters to the self-deceived subject is to *believe*  $p$  no matter what the evidence suggests, regardless of whether  $p$  might be true or not. In the high-stakes cases, on the contrary, Hannah is concerned with the truth of  $p$ —what she wants to find out is whether  $p$  is true or not. Although the amount of effort invested in the inquiry is influenced by pragmatic considerations (the practical costs of falsely believing  $p$ ), her interest is to reduce the possibility of making a costly error—what matters to her is that her belief be true.

This distinction in fact neatly overlaps one offered by Jordan (1996), who presents two types of pragmatic arguments for belief formation: *truth-dependent* and *truth-independent* arguments. Truth-dependent arguments are pragmatic arguments for believing something because, if  $p$  happens to be true, then the practical benefits of believing  $p$  will be great. They are truth dependent precisely because these practical benefits depend on  $p$  being true. If  $p$  turned out to be false, then there would be no benefits to holding the belief in question. Truth-independent arguments, on the contrary, are pragmatic arguments for believing  $p$ , the benefits of which do not depend on  $p$  being true: the benefits gained by believing  $p$  hold whether or not  $p$  turns out to be true (Jordan, 1996). If you think of Fernando again, you will realize that what influences his threshold for belief formation doesn't depend on what will happen were he to believe falsely, but depends on the emotional costs of believing that Steve doesn't love him, something he is afraid to admit. The self-deceived subject isn't interested in finding out whether  $p$  is true or not—this is precisely not the point of their inquiry.

*A winning hypothesis?*

What conclusions can we draw from the evaluation of these hypotheses? Hypothesis 1 isn't convincing as it merely indicates an accidental feature of self-

deception. Hypothesis 2 does set up an interesting basis for explaining the difference between the cases, but assumes that there are no cases of self-deception in which the subject suspends belief, which is misleading. It is difficult to see how hypothesis 3 could ground a proper difference in rationality since many rational beliefs are formed through automatic reasoning. But as we will see, maybe hypotheses 2 and 3 haven't said their last word. Finally, hypothesis 4, I think, has more potential. First, it smoothly applies to both cases: In the case of self-deception, the subject's belief is influenced by the costs of believing  $p$  (regardless of whether  $p$  is true or not) rather than by the costs of falsely believing  $p$ . In the high-stakes case, on the contrary, Hannah's and Molly's thresholds for belief aren't influenced by the costs of believing but by the costs of falsely believing  $p$ . But this doesn't seem to be the full story.

Further support for this idea can be found in Kunda's (1990) and Kruglanski's (Kruglanski, 1980; Kruglanski and Ajzen, 1983; Kruglanski and Klar, 1987) works on motivated reasoning. In her famous paper, Kunda also distinguishes between two types of motivations: the *motivation to arrive at an accurate conclusion* (whatever the conclusion may be) *versus the motivation to arrive at a particular, directional, conclusion*.<sup>17</sup> While the first "enhances use of those beliefs and strategies that are considered most appropriate," the second "enhances use of those that are considered most likely to yield the desired conclusion" (Kunda, 1990). If Kunda is correct in this, then the type of motivation also influences the type of reasoning, not exactly in the sense described under hypothesis 2, but in the following way: directionally motivated subjects will tend to rely on ways of reasoning that allow them to "construct seemingly reasonable justifications for these conclusions" (Kunda, 1990). Subjects whose reasoning is directionally motivated do not only tend to rely on biased reasoning, but they also seem to "pick and choose" reasoning strategies likely to lead them to form the desired belief.

This points to another interesting difference between self-deceptive subjects and high-stakes subjects. In cases of self-deception, subjects do not recognize their motivations as being part of the reason why they come to believe  $p$ . On the contrary, they often seem unaware of this causal connection. Self-deceived people typically take the evidence to support their false belief, whereas it seems part of high-stakes subjects' position to be aware of the fact that they are suspending belief for practical reasons.

Finally, and given what I have just said, although I did argue that there were irrational cases of suspension of belief, it might well be the case that practical reasons bear a rational influence on beliefs only in cases of suspension.<sup>18</sup> In other words, this would mean that, although it could be warranted to suspend one's belief about whether  $p$  for practical reasons, one could never rationally believe  $p$  for practical reasons (no matter what type of reasoning one is using). In fact, this is compatible with what Schroeder (2012) suggests when he writes that reasons to withhold can only be nonevidential. The final question would thus be the following: Would we deem it rational for a subject to suspend belief for

truth-independent reasons? I suspect not. In the light of the points presented above, we could thus add the following: it might be rational to withhold belief for practical reasons, if these reasons are truth-dependent in the sense specified by Jordan (1996).<sup>19</sup>

These comments might, of course, merely constitute the sketch of an answer. They might even, as a matter of fact, raise more questions than they dare to answer. All the same, I hope to have shown that the suggested dilemma doesn't really hold, that, despite these similarities, self-deception and high-stakes cases differ in significant ways.

That being said, let us finally make note of a few implications for theories of self-deception and pragmatic encroachment. First, we need to recognize that, although Mele might be correct in recognizing the FTL model as a mechanism by which we can explain self-deception, it seems misleading to think that any type of motivation might play this role. Indeed, the motivation at play in cases of self-deception seems more likely to be a motivation to avoid forming beliefs *tout court*, rather than a motivation to avoid forming costly *false* beliefs. This might be a reason to prefer a motivationist account, such as Nelkin's (2002), that argues that self-deception should be defined as a desire to believe *p* rather than a desire that *p*, as most motivationists put it, or an account such as the one suggested by Lauria et al. (2016), according to which self-deception is best understood as a type of affective coping. These accounts, as well as the argument laid out above, seem to suggest that the motivation inherent to self-deception is a motivation linked to the costs of believing rather than related to the truth of *p*, something that isn't obvious in Mele's account.

Second, this should also have implications for our conception of what makes self-deceptive beliefs irrational. Indeed, as we have seen, the FTL model in itself isn't sufficient for explaining why we deem self-deceptive beliefs to be irrational since the same mechanism also leads to rational cases. As I suggested, it is important to distinguish between the costs of believing and the costs of believing falsely, and thereby acting as if *p*. Although I here suggested that this difference can explain why self-deception is irrational, whereas cases of pragmatic encroachment are not, there is more to say about how and why, precisely, these two types of influences ground rationality.

Third, on the pragmatic encroachment side of the discussion, this proves that the cases used to support the thesis lack detail. Although we do seem to have the intuition that these subjects aren't self-deceived, the lack of precision about what type of practical considerations and reasoning might be compatible with this rationality allows us to question the belief processes involved. In the absence of detail, one might just as well use this as an argument against pragmatic encroachment in order to show why these cases are perfectly analogous to self-deception.

Last but not least, I think this discussion has established that speaking in terms of the influence of practical factors alone isn't sufficient for explaining why a

belief might be irrational. Without heading towards the pragmatists' camp and saying that it is rational to believe whatever maximizes utility, for example (see Rinard, 2015; 2017 for a discussion and defence of pragmatism), I claim that it seems insufficient, even within a more evidentialist framework, to simply posit that only evidential considerations play a role in determining the rationality of one's beliefs.

## CONCLUSION

I have argued that cases of self-deception and cases of pragmatic encroachment can be explained by reference to the same mechanism—namely, the FTL model for belief formation and lay hypothesis testing. This mechanism shows that a subject's motivation for avoiding costly false beliefs not only explains the way in which the belief is formed, but also seems to explain why this belief is irrational—insofar as the motivation to avoid costly false beliefs thereby leads the subject to mistreat the evidence at hand. On this basis, I argued that by accepting this we are forced into a dilemma about the rationality of beliefs according to which either self-deception is rational or cases of pragmatic encroachment are irrational. Finally, I presented several hypotheses about how to solve this dilemma and argued that the type of motivation at play holds a central role in distinguishing the two types of cases.



## ACKNOWLEDGEMENTS

I am grateful to Natalie Ashton, Jie Gao, Marie van Loon, Alfred Mele, Anne Meylan and Jennifer Nagel, as well as the two anonymous referees at *Les ateliers de l'éthique/The Ethics Forum* for their insightful comments. Earlier versions of this article were presented at “Self-Deception: What It Is and What It’s Worth,” University of Basel; “Ateliers du GRE,” Collège de France; the “Epistemic and Practical Rationality” workshop, University of Fribourg; and the “European Epistemology Network” conference, Vrije Universiteit, Amsterdam. I thank the audiences at these venues for the stimulating discussions that followed these presentations. Research for this article was supported by the Swiss National Science Foundation (SNSF) Professorship grant “Irrationality” #PP00P1\_157436 and the Doc.Mobility grant “Believing Under the Influence” #P1BSP1\_181672.

## NOTES

- <sup>1</sup> To my knowledge, Gao (n.d.) is the only author who points out and discusses this similarity.
- <sup>2</sup> By “high-stakes cases” I mean to say cases described and used in the literature on pragmatic encroachment. The cases are often used as a motivation for rejecting purism, the idea that knowledge or other epistemic states are purely truth related (see section II for more detail).
- <sup>3</sup> This second condition helps mainly to distinguish self-deception from wishful thinking, often defined as cases in which the subject is only unwarranted in believing *p* (Szabados, 1973; Van Leeuwen, 2007).
- <sup>4</sup> In one of his footnotes (2006, p. 115), Mele writes “the requirement that *p* be false is purely semantic. By definition, one is *deceived in* believing that *p* only if *p* is false, the same is true of being *self-deceived in* believing that *p*. The requirement does not imply that *p*’s being false has special importance for the dynamics of self-deception. Biased treatment of data may sometimes result in someone’s believing an improbable proposition, *p*, that happens to be true” (see also Mele, 1987, p. 127-128).
- <sup>5</sup> Van Leeuwen (2007) describes “not-*p*” as the “doxastic alternative”. Attitudes towards the doxastic alternative vary from one theory of self-deception to another. On Mele’s view, it is not necessary that the subject *believes* the doxastic alternative to be self-deceived. However, condition (b) seems to describe the self-deceived subject as somehow possessing evidence supporting the doxastic alternative.
- <sup>6</sup> Most of versions of pragmatic encroachment primarily focus on “stakes.” Nonetheless, some philosophers (see Anderson, 2015; and Gerken, 2011) argue that other types of factors such as urgency, the availability of alternative evidence, social rules, and conventions can play a similar role.
- <sup>7</sup> Pragmatic encroachment on knowledge, for example, must be understood as a thesis about the metaphysics of knowledge rather than as a thesis about the pragmatics of the verb “to know.”
- <sup>8</sup> For similar cases, see Cohen, 1999; Fantl and McGrath, 2002; and McGrath, 2018.
- <sup>9</sup> The traditional position is usually referred to as purism. Purism can be spelled out as follows: “For any two possible subjects *S* and *S*’, if *S* and *S*’ are alike with respect to the strength of their epistemic position regarding a true proposition *p*, then *S* and *S*’ are alike with respect to being in a position to know that *p*” (Fantl and McGrath, 2007).
- <sup>10</sup> One could, of course, reject the assumption I am establishing: that is, that there are no such cases of subjects suspending belief in high-stakes situations because believing *p* in these circumstances would be irrational. If this were the case, there would be no dilemma concerning the rationality of beliefs, for the rationality of the high-stakes subjects wouldn’t concern their beliefs, but their action. Against this objection, one could invoke the tight link between

action and belief, or between action and knowledge. For example, functionalists about beliefs may argue that to believe  $p$  is to be disposed to act as if  $p$  (see Ganson, 2007) for a discussion of the double function of beliefs in relation to this issue. Defenders of pragmatic encroachment often invoke something they call the “knowledge-action principle.” Roughly put, this principle states that if  $S$  knows  $p$  then  $S$  is in a position to act as if  $p$  (see Fantl and McGrath, 2002; Williamson, 2005; Stanley and Hawthorne, 2008, for different formulations of this idea). This principle is used to reinforce the idea that if  $S$  isn’t in a position to act when the stakes are high, then  $S$  doesn’t know that  $p$ .

- <sup>11</sup> A similar idea can be found in James’s (1897) work, in which he mentions “two duties in inquiry”: (i) avoiding false beliefs and (ii) forming true beliefs. Depending on which of these two duties the subject takes to be his or her primary concern, the subject will alter his or her inquiry and treatment of the evidence relative to whether  $p$ . If the subject is primarily concerned with (i) avoiding falsehood, the amount of evidence required for believing  $p$  will be greater, whereas if the subject’s primary concern is (ii) forming true beliefs—and relieving himself or herself from a state of agnosticism—that subject will then form a belief on weaker grounds. More recently, Nagel introduced two psychological elements that influence belief formation and epistemic inquiry: epistemic anxiety (2010) and need for closure (2008). These two “forces” vary in function of the practical interests of the subject in a given situation (Nagel, 2008). Epistemic anxiety is the emotive response resulting from the perceived costs in being mistaken about a particular matter; the response consists in the subjects regulating their cognitive effort and adapting their cognitive strategy by relying on more deliberate and controlled cognition (Nagel, 2008; cf. Tversky and Kahneman, 1974). One’s level of “need for closure” on the other hand (cf. Kruglanski and Webster, 1996) refers to the threshold of belief (or desired “levels of confidence”; Nagel, 2010) at which a subject settles and forms a given belief.
- <sup>12</sup> There might be costs other than the ones mentioned here. These costs could range from further subjective, psychological costs to more objective costs. For example, one could consider the costs of causing distress to Steve by misinterpreting his behaviour. I think the issue of determining how to narrow down the relevant costs hasn’t yet been completely clarified: do these costs depend only on the subject’s interests and primary concerns? For purposes of simplicity, let’s here assume that the relevant costs are the ones described above.
- <sup>13</sup> This mechanism described as the FTL model in Mele’s words is in fact very close to the general functioning of adaptive cognition. Roughly speaking, adaptive cognition is the idea that our cognition, the way in which we treat information, test hypothesis, and form beliefs is cognitively adaptive in the following sense: “agents adapt their cognitive efforts” (and resources) “to how they represent the practical factors relevant to the task at hand” (Gerken, 2017). This means that the ways in which agents perceive their practical situation influence their cognition in a significant way. Although there is significant disagreement, as Gerken (2017) notices, “there is a wide agreement that our metacognitive procedures adapt the cognitive resources that we deploy for a given task to how they represent the practical factors associated with it.” The two central aspects of adaptive cognition are the following: (i) how much cognitive effort one is willing to allocate to a given cognitive task as well as (ii) how much evidence one needs in order to form or reject a given belief, both very according to one’s practical situation.
- <sup>14</sup> I thank the reviewers for noting that twisted self-deception may present a challenge to any account heavily relying on this type of mechanism. However, many accounts of self-deception face this challenge, and it might be the case that Mele takes FTL to sometimes, but not necessarily, play a role in the formation of self-deceptive beliefs. And it is sufficient for our argument that FTL sometimes produces irrational beliefs.
- <sup>15</sup> Although Mele does rely on the FTL model to explain how self-deceptive beliefs come about, it might be the case that neither Trope nor Friedrich nor Liberman would agree with the idea that, according to the FTL model (which, I recall, is a generalization provided by Mele, 1997; 1999; 2001), the motivation to avoid costly false beliefs may result in irrational beliefs. This

might be true, for example, for anyone working with a slightly different notion of rationality (i.e., an evolutionary notion, for example) than the one assumed throughout this discussion.

<sup>16</sup> I am grateful to Alfred Mele for suggesting this third hypothesis.

<sup>17</sup> For more work on this distinction see Kruglanski, 1980; Kruglanski and Ajzen, 1983; Kruglanski and Klar, 1987; see also Chaiken, Liberman, and Eagly, 1989; Pyszczynski and Greenberg, 1987.

<sup>18</sup> I here set aside the pragmatist (or nonevidentialist) idea that it can be rational to believe for nonevidential reasons because accepting the truth of pragmatism is incompatible with my beginning assumption that self-deception is irrational. Indeed, if we accept that it can be rational to believe  $p$  because believing  $p$  leads to positive consequences, for example (cf. Rinard, 2015; 2017), it then becomes unclear why we would still consider most cases of self-deception irrational.

<sup>19</sup> It could be interesting to think about Pascal's Wager here. Pascal's wagerer decides to believe in God because he or she thinks believing in God will lead to positive consequences whatever the truth is, whereas disbelieving (whether this means believing that God doesn't exist or suspending belief) will either lead to no consequences or lead to negative ones. Overall, this could be understood as an FTL belief. However, I do not think this would qualify as a self-deceptive belief for the following reasons: First, being the product of FTL isn't sufficient for qualifying as a self-deceptive belief (cf. conditions given in section 1). It isn't obvious—at least not to me—that Pascal's wagerer should in fact believe  $\text{not-}p$  (that God doesn't exist) in the sense that he or she has been presented with sufficient evidence for being warranted in believing  $\text{not-}p$ . Second, these final considerations about the difference between self-deception and high-stakes cases also shed light upon the kind of reasoning underlying self-deception, and it does not seem to me that Pascal's wagerer is similar in this respect. One might still want to argue that forming a belief for pragmatic reasons of this sort—regardless of whether one has evidence supporting this belief—is irrational, even though it might not be self-deceptive.

## REFERENCES

- Anderson, Charity, "On the Intimate Relationship of Knowledge and Action," *Episteme*, vol. 12, no. 3, 2015, p. 343-353.
- Barnes, Annette, *Seeing through Self-Deception*, New-York, Cambridge University Press, 1997.
- Chaiken, Shelly, Akiva Liberman, and Alice H. Eagly, "Heuristics and Systematic Information Processing within and beyond the Persuasion Context," in Uleman, J. S. and J. A. Bargh (eds.), *Unintended Thought: Limits of Awareness, Intention, and Control*, New York, Guilford Press, 1989, p. 212-252.
- Cohen, Stewart, "Contextualism, Skepticism, and The Structure of Reasons," *Philosophical Perspectives*, vol. 13, 1999, p. 57-89.
- Davidson, Donald, "Deception and Division," in LePore, E. and B. McLaughlin (eds.), *Actions and Events*, New York, Basil Blackwell, 1985.
- , "Paradoxes of Irrationality," Oxford, Clarendon Press, 2004.
- DeRose, Keith, "Contextualism and Knowledge Attributions," *Philosophy and Phenomenological Research*, vol. 52, no. 4, 1992, p. 913-929.
- Engel, Pascal, "Pragmatic Encroachment and Epistemic Value," in Haddock, A., A. Millar, and D. Pritchard (eds.) *Epistemic Value*, Oxford, Oxford University Press, 2009.
- Evans, Jonathan St. B. T., *Thinking Twice: Two Minds in One Brain*, Oxford, Oxford University Press, 2010.
- Evans, Jonathan St. B. T., and Keith Frankish, *In Two Minds: Dual Processes and Beyond*, Oxford University Press, Oxford, 2009.
- Fantl, Jeremy, and Matthew McGrath, "Evidence, Pragmatics, and Justification," *Philosophical Review*, vol. 11, no. 1, 2002, p. 67-94.
- , "On Pragmatic Encroachment in Epistemology," *Philosophy and Phenomenological Research*, vol. 75, no. 3, 2007, p. 558-589.
- Frankish, Keith, "Dual-Process and Dual Theories of Reasoning," *Philosophy Compass*, vol. 5, no. 10, 2010, p. 914-926.
- Friedrich, James, "Primary Error and Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena," *Psychological Review*, vol. 100, no. 2, 1993, p. 298-319.
- Funkhouser, Eric, "Do the Self-Deceived Get What They Want?" *Pacific Philosophical Quarterly*, vol. 86, no. 3, 2005, p. 295-312.
- Gao, Jie, "Self-Deception and Pragmatic Encroachment, a Dilemma for Epistemic Rationality," manuscript, n.d.
- Ganson, Dorit, "Evidentialism and Pragmatic Constraint on Outright Belief," *Philosophical Studies*, vol. 139, no. 3, 2008, p. 441-458.

- Gerken, Mikkel, "Warrant and Action," *Synthese*, vol. 178, no. 3, 2011, p. 529-547.
- Hawthorne, John, *Knowledge and Lotteries*, Oxford, Oxford University Press, 2004.
- Hookway, Christopher, *Scepticism*, London, Routledge, 1990.
- James, William, *The Will To Believe, and Other Essays in Popular Philosophy*, New York, Longmans, 1897.
- Johnston, Mark, "Self-Deception and the Nature of the Mind," McLaughlin, B. and A. Rorty (eds.), 1988, p. 63-91.
- Jordan, Jeff, "Pragmatic Arguments and Belief," *American Philosophical Quarterly*, vol. 33, no. 4, 1996, p. 409-420.
- Kahneman, Daniel, and Amos Tversky, "Judgement Under Uncertainty: Heuristics and Biases," *Science*, vol. 185, no. 4157, 1974, p. 1124-1131.
- Kunda, Ziva, "The Case for Motivated Reasoning," *Psychological Bulletin*, vol. 108, no. 3, 1990, p. 480-498.
- Kruglanski, Arie W., "Lay Epistemology Process and Contents," *Psychological Review*, vol. 87, p. 70-87, 1980.
- Kruglanski, Arie W., and Icek Ajzen, "Bias and Error in Human Judgment," *European Journal of Social Psychology*, vol. 13, no. 1, 1983, p. 1-44.
- Kruglanski, Arie W., and Yechiel Klar, "A View from the Bridge: Synthesizing the Consistency and Attribution Paradigms from a Lay Epistemic Perspective," *European Journal of Social Psychology*, vol. 17, no. 2, 1987, p. 211-241.
- Lazar, Ariela, "Deceiving Oneself or Self-Deceived? On the Formation of Beliefs Under the Influence," *Mind*, vol. 108, no. 430, 1999, p. 265-290.
- Lauria, Federico, Delphine Preissmann, and Fabrice Clément, "Self-Deception as Affective Coping: An Empirical Perspective to Philosophical Issues," *Consciousness and Cognition*, vol. 41, 2016, p. 119-134.
- McGrath, Matthew, "Defeating Pragmatic Encroachment," *Synthese*, vol. 195, no. 7, 2018, p. 3051-3064.
- Mele, Alfred, "Real Self-Deception," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 91-136.
- , "Twisted Self-Deception," *Philosophical Psychology*, vol. 12, no. 2, 1999, p. 117-137.
- , *Self-Deception Unmasked*, Princeton, Princeton University Press, 2001.
- , "Self-Deception and Delusions," *European Journal of Analytic Philosophy*, vol. 2, no. 1, 2006, p. 109-124.
- Nagel, Jennifer, "Knowledge Ascriptions and the Psychological Consequences of Changing Stakes," *Australasian Journal of Philosophy*, vol. 86, no. 2, 2008, p. 279-294.

———, “Epistemic Anxiety and Adaptive Invariantism,” *Philosophical Perspectives*, vol. 24, no. 1, 2010, p. 407-435.

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

Pears, David, *Motivated Irrationality*, Oxford, Clarendon Press, 1984.

Pyszczynski, Tom, and Jeff Greenberg, “Toward an Integration of Cognitive and Motivational Perspectives on Social Inference: A Biased Hypothesis-Testing Model,” in Berkowitz, L. (ed.), *Advances in Experimental Social Psychology*, New York, Academic Press, vol. 20, 1987, p. 297-340.

Rinard, Susanna, “Against the New Evidentialists,” *Philosophical Issues*, vol. 25, no. 1, 2015, p. 208-223.

Rinard, Susanna, “No Exception for Belief,” *Philosophy and Phenomenological Research*, vol. 94, no. 1, 2017, p. 121-143.

Rorty, Amelie, “The Deceptive-Self: Liars, Layers, and Liars2,” in McLaughlin, B., and A. Rorty (eds.), 1988, p. 11-28.

Schroeder, Mark, “Stakes, Withholding, and Pragmatic Encroachment on Knowledge,” *Philosophical Studies*, vol. 160, no. 2, 2012, p. 265–285.

Scott-Kakures, Dion, “At ‘Permanent Risk’: Reasoning and Self-Knowledge in Self-Deception,” *Philosophy and Phenomenological Research*, vol. 65, no. 3, 2002, p. 576-603.

———, “Can You Succeed in Intentionally Deceiving Yourself?” *Humana.Mente Journal of Philosophical Studies*, vol. 5, no. 20, 2012, p.17-39.

Stanley, Jason, *Knowledge and Practical Interests*, Oxford, Oxford University Press, 2005.

Stanley, Jason and John Hawthorne, “Knowledge and Action,” *Journal of Philosophy*, vol. 105, no. 10, 2008, p. 571-590.

Szabados, Béla, “Wishful Thinking and Self-Deception,” *Analysis*, vol. 33, no. 6, 1973, p. 201-205.

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

Trope, Yaacov, and Akiva Liberman, “Social Hypothesis Testing: Cognitive and Motivational Mechanisms,” in E. Higgins, and A. Kruglanski (eds.), *Social Psychology Handbook of Basic Principles*, New York, 1996.

Van Leeuwen, Neil, “The Product of Self-Deception,” *Erkenntnis*, vol. 67, no. 3, 2007, p. 419-437.

# RESPONSIBILITY FOR SELF-DECEPTION

MARIE VAN LOON

PHD STUDENT, UNIVERSITY OF ZURICH

## ABSTRACT:

In this paper, I argue that Alfred Mele's conception of self-deception is such that it always fulfils the reasons-responsiveness condition for doxastic responsibility. This is because self-deceptive mechanisms of belief formation are such that the kind of beliefs they bring about are the kind of beliefs that fulfil the criteria for doxastic responsibility from epistemic reasons responsiveness. I explain why in this paper. Mele describes the relation of the subject to the evidence as a biased relation. The subject does not simply believe on the basis of evidence, but on the basis of manipulated evidence. Mele puts forward four ways in which the subject does this. The subject could misinterpret positively or negatively, selectively focus, or gather evidence. Through these ways of manipulation, the evidence is framed such that the final product constitutes evidence on the basis of which the subject may believe a proposition that fits that subject's desire that P. Whichever form of manipulation the subject uses, the evidence against P must be neutralized in one way or another. Successful neutralization of the evidence requires the ability to recognize what the evidence supports and the ability to react to it. These abilities consist precisely in the two parts of the reasons-responsiveness condition, reasons receptivity and reasons reactivity. In that sense, self-deceptive beliefs always fulfil the reasons-responsiveness condition for doxastic responsibility. However, given that reasons responsiveness is only a necessary condition for doxastic responsibility, this does not mean that self-deceived subjects are always responsible for their belief.

## RÉSUMÉ :

Dans cet article, je soutiens que la conception d'auto-illusion chez Alfred Mele remplit toujours l'une des conditions de la responsabilité doxastique, à savoir la « sensibilité aux raisons » (reasons-responsiveness). Il en est ainsi car les mécanismes d'auto-illusion dans la formation de croyances produisent des types de croyances qui remplissent les critères pour la responsabilité doxastique quant à la sensibilité aux raisons épistémiques. J'explique pourquoi dans cet article. Mele décrit la relation du sujet à la preuve comme biaisée. Le sujet ne croit pas seulement sur la base de preuves, mais de preuves manipulées. Mele avance quatre façons qu'a le sujet de faire ceci. Le sujet peut mal interpréter positivement ou négativement, focaliser de façon sélective, ou accumuler des preuves. Par ces formes de manipulations, la preuve est formulée de sorte qu'elle produise un fondement pour la croyance en une proposition qui s'accorde avec le désir du sujet que P. Peu importe la forme de manipulation qu'emploie le sujet, la preuve contre P doit être neutralisée d'une façon ou d'une autre. Une neutralisation réussie de la preuve requiert la capacité de reconnaître ce que soutient la preuve et la capacité d'y réagir. Ces capacités consistent précisément en ces deux parties de la condition de la sensibilité aux raisons, soit la réceptivité et la réactivité aux raisons. En ce sens, les croyances d'auto-illusion remplissent toujours la condition de la sensibilité aux raisons pour la responsabilité doxastique. Toutefois, étant donné que la sensibilité aux raisons n'est une condition nécessaire que pour la responsabilité doxastique, cela ne veut pas dire que les sujets souffrant d'auto-illusion sont toujours responsables de leurs croyances.

## I. INTRODUCTION

In the same way that we reproach the ignorant subject, “You should have known,” we reproach the self-deceiver, “You shouldn’t be self-deceived.” Both the ignorant subject and the self-deceived subject make an *epistemic*<sup>1</sup> mistake, for which they are to blame at least epistemically and sometimes morally. Since blame requires responsibility, holding self-deceivers responsible for their self-deceptive belief must be appropriate. If these assumptions are correct, then there should be a way of explaining why self-deceptive beliefs are beliefs for which we are responsible.

According to Alfred Mele, self-deception need not be conceived on the model of interpersonal deception but may simply be understood as a motivationally biased belief (Mele, 1997; 1999; 2001; 2006). Because, according to such a conception, self-deception consists in, among other things, holding a belief, a theory of doxastic responsibility should be able to explain how we can be responsible for self-deceptive beliefs. Among the various competing theories of doxastic responsibility, *reasons responsiveness* (McHugh, 2013; 2014; 2015) offers a necessary condition for doxastic responsibility.

In this paper, I argue that Mele’s conception of self-deception is such that it always fulfils the reasons-responsiveness condition for doxastic responsibility. This is because self-deceptive mechanisms of belief formation are such that the kind of beliefs they bring about are the kind of beliefs that fulfil the criteria for doxastic responsibility and more particularly the criteria for doxastic responsibility from epistemic reasons responsiveness. I explain why in this paper. Mele describes the relation of the subject to the evidence as a biased relation. The subject does not simply believe on the basis of evidence, but on the basis of manipulated evidence. Mele puts forward four ways in which the subject does this. The subject could misinterpret positively or negatively, selectively focus, or gather evidence. What matters is that the evidence is framed such that the final product constitutes evidence on the basis of which the subject may believe a certain proposition. This strategy requires avoiding coming into contact with data that would disprove *P*. If the subject does not manage to avoid it, then that subject might instead misinterpret such information as counting in favour of *P* or at least as not counting against *P*. Whichever form of manipulation the subject uses, the evidence against *P* must be neutralized in one way or another. Successful neutralization of the evidence requires the ability to recognize what the evidence supports and the ability to react to it.

Reasons responsiveness requires that subjects would recognize evidence against *P* if they were presented with it and that they would react to this evidence if they were presented with it. These conditions that require the subject to recognize and react to evidence in counterfactual scenarios express an ability that is expected from the subject, and it is the possession of this ability that makes a belief reasons responsive.



I argue that, in order to successfully self-deceive, an epistemic subject must possess this ability. Possessing this ability makes a belief reasons responsive. In that sense, self-deceptive beliefs always fulfil the reasons-responsiveness condition for doxastic responsibility. However, given that reasons responsiveness is only a necessary condition for doxastic responsibility, this does not mean that self-deceived subjects are always responsible for their belief.

In order to show this, I proceed in the following way. In section II, I summarize the reasons-responsiveness account of doxastic responsibility. In section III, I offer a reminder of Mele's theory of self-deception. In sections IV, V, and VI, I apply reasons responsiveness to self-deception.

## II. EPISTEMIC REASONS RESPONSIVENESS

According to the reasons-responsiveness account of doxastic responsibility defended by Conor McHugh (2013; 2014; 2017), subjects are responsible for their beliefs and other attitudes in virtue of a feature they must possess: reasons responsiveness.

I speak of doxastic responsibility here because I take it that the object under consideration is a belief and that we cannot assess the responsibility for self-deception without using an adequate theory that pertains to its particular object. Most authors take on assessing responsibility for self-deception by asking whether subjects are morally responsible for it. I differ in that I first want to ground moral responsibility in doxastic responsibility.

One dominant account of doxastic responsibility is McHugh's application of Fischer and Ravizza's reasons-responsiveness account (Fischer and Ravizza, 1998) to the epistemic case, epistemic reasons responsiveness. Whereas Fischer and Ravizza's reasons responsiveness provides necessary conditions for responsibility for actions, McHugh provides us with necessary conditions for beliefs. The crucial difference is that, in the epistemic version of reasons responsiveness, we are not after any reasons that agents would recognize for holding their belief and that constitute the basis on which they form, maintain, or revise their belief, but we are specifically after epistemic reasons—that is, the evidence that agents would recognize for holding their belief and that constitute the basis on which they form, maintain, or revise their belief. The difference is important because practical reasons and epistemic reasons do not function in the same way with regard to belief. It is indeed frequently admitted that we cannot believe a proposition for practical reasons. Thus, the only reasons that are pertinent to the responsiveness of belief are epistemic reasons. Therefore, there is an important difference in applying epistemic reasons responsiveness to self-deception rather than simple reasons responsiveness to it—not only because these theories are different but also because of the reason why there is an *epistemic* reasons responsiveness in the first place: we cannot ask agents to be responsive to nonepistemic reasons in order for them to be responsible for their belief.

Levy (2004) and DeWeese-Boyd (2007) and Nelkin (2013), who discuss the application of reasons-responsiveness accounts of moral responsibility to self-deception, seem rather concerned with the actions and omissions (mental and nonmental) that lead to or maintain a state of self-deception.<sup>2</sup> In this vein, they are concerned with responsibility for self-deception in virtue of the indirect control we have over these beliefs. Such indirect control operates through actions (and omissions). Given that these authors are asking whether we can be held responsible for self-deception in virtue of these actions and omissions that surround self-deception, it makes perfect sense to turn to the Fischer and Ravizza version of reasons responsiveness, since they, too are interested in responsibility for actions. I am concerned not with questions of responsibility for the actions that lead to and follow from being in a state of self-deception, but with the state itself—that is, the self-deceptive belief. For this reason, McHugh’s reasons responsiveness, since it concerns doxastic responsibility directly, is better fitted for the task of asking whether and why subjects may be held responsible for their self-deceptive belief.

Theories of doxastic responsibility try to explain why there can be responsibility for beliefs despite doxastic involuntarism. Doxastic involuntarism is the thesis that beliefs are not under our *direct control* (Williams, 1973; Alston, 1988). We may work to abandon them or form new ones by means of other actions or attitudes, but not in the same way that we are able to raise an arm at will. Because we lack direct control over beliefs, it may seem as if we can never be held responsible for our beliefs. Several solutions to this problem have been offered in the literature. Most<sup>3</sup> theories of doxastic responsibility aim at grounding responsibility for beliefs on a basis that does not involve direct control. Some argue that what is needed is a form of indirect control (Meylan, 2013; 2017; Peels, 2013; 2017), others argue that we are responsible for beliefs if they are intentional (Steup, 2012), and others explain that we are responsible for beliefs in virtue of the fact that beliefs are the kind of attitude that reflect our take on the world (Hieronymi, 2008).

Reasons responsiveness remains the most agnostic of these theories with regard to the exact nature of control over beliefs; it endorses neither direct nor indirect control. In this sense, the type of control reasons responsiveness requires for one to be responsible for one’s beliefs is minimal. Mele’s conception of self-deception as motivationally biased belief intuitively doesn’t seem to require any kind of doxastic control that is not minimal, whether direct or indirect. Given that reasons responsiveness requires only a minimal form of control, this makes it perfectly well suited to accommodate Mele’s model of self-deception.

McHugh proposes a form of control that is distinct from direct control and indirect control and which applies specifically to beliefs and other attitudes: *attitudinal control* (McHugh, 2017). Nevertheless, this particular notion of control is a very minimal one, more akin to a kind of sensitivity to evidence. *S* has attitudinal control over *S*’s belief that *P* only if<sup>4</sup> *S*’s belief is responsive to reasons. *S*’s belief that *P* is responsive to reasons only if both of the following conditions hold.

#### Reasons Receptivity

*S* would recognize epistemic reasons<sup>5</sup> to believe *P* if presented with such reasons to believe *P*.

#### Reasons Reactivity

*S* would react to epistemic reasons *S* has to believe *P*.

Reasons receptivity and reasons reactivity together form epistemic reasons responsiveness (McHugh, 2013; 2014; 2015). Reasons responsiveness is a necessary condition for doxastic responsibility.

A belief fulfils the reasons-receptivity condition if in a sufficiently wide range of counterfactual scenarios relevant to the actual scenario the subject recognizes evidence counting in favour of or against *P*. Being reasons receptive amounts to being able to recognize that some evidence would count in favour of or against *P* if one were presented with such evidence. For example, suppose I believe that the teapot on my desk contains Earl Grey. My belief is reasons receptive if, were there Lapsang Souchong in the teapot, I would recognize the smoky scent emanating from the teapot as evidence counting against the belief that the teapot is full of Earl Grey.

A belief fulfils the reasons-reactivity condition if the subject, when faced with evidence to the contrary, reacts to that evidence by revising the belief, and would do so in a sufficiently wide range of counterfactual scenarios relevant to the actual scenario. For example, take again my belief that the teapot on my desk contains Earl Grey. Given that the smoky scent emanating from the teapot would be evidence supporting the belief that the teapot contains Lapsang Souchong and not Earl Grey, I am reactive to evidence if I abandon the belief that the teapot contains Earl Grey in reaction to this new piece of evidence.

Beliefs that are not reasons responsive are beliefs for which we are not responsible—beliefs involved in paranoia, for example (McHugh, 2014). In these cases, the belief fails to be reasons receptive. The subject would not recognize evidence supporting not-*P* if he or she were presented with such evidence in a wide range of counterfactual scenarios relevant to the actual scenario. For example, take the true belief that there is someone sitting behind me in the café. If in a wide range of relevant counterfactual scenarios I turn around, see no one at the table behind me, and yet do not recognize this as evidence that there is no one sitting behind me, I am not receptive to evidence. This is one way to fail to be reasons responsive. Another way to fail to be reasons responsive is to fail to be reactive to reasons. For McHugh, what he calls a “repressed belief” typically fails to react to recognized evidence. By “repressed beliefs,” McHugh means implicit prejudiced beliefs, for example (McHugh, 2017; p. 2752). In spite of having the ability to recognize reasons against their belief, subjects are not able to revise it. Here the subjects are not reasons reactive, and thus not reasons responsive. Thus, on the reasons-responsiveness account of doxastic responsibility, the paranoid subjects and subjects with repressed beliefs are not responsible for their beliefs.

In the next section, I turn my attention to self-deception itself and more precisely to Alfred Mele's account, of which I provide an overview.

### III. MELEAN SELF-DECEPTION

The following description of self-deception is generally agreed upon in the literature. A subject is self-deceived in believing a certain proposition when that subject seems to persist in believing this proposition in spite of the evidence he or she has against this proposition. In the majority of cases present in the literature, self-deceptive beliefs are emotionally significant for the subject: beliefs about the faithfulness of partners, beliefs about the morality of your loved ones, beliefs about our own value or emotional states, and the like. Mele proposes the following set of jointly sufficient conditions for self-deception (Mele, 1997; 1999; 2001; 2006):

- 1) The belief that  $p$  which  $S$  acquires is false.
- 2)  $S$  treats data relevant, or at least seemingly relevant, to the truth value of  $p$  in a motivationally biased way.
- 3) This biased treatment is a non-deviant cause of  $S$ 's acquiring the belief that  $p$ .
- 4) The body of data possessed by  $S$  at the time provides greater warrant for  $\neg p$  than for  $p$ .

As will become apparent, if reasons responsiveness is to be found in self-deception, its locus must be situated in the way the self-deceived subject treats data or evidence: indeed, in explaining doxastic responsibility, reasons responsiveness focuses on the way in which the subject recognizes and reacts to evidence. For this reason, I focus on Mele's second condition for self-deception and the way in which one treats the evidence.

Note that Mele mentions no desire in his conditions. The subject's desire that  $P$ , however, seems to be implicit in the second condition, which states that the subject treats the evidence in a *motivationally* biased way. For the present purpose, I will take the subject's motivation to be a desire. "Biased" picks out the manipulative strategies at work in the formation or sustaining of the subject's beliefs. Take the following example.

#### *Annie and Alvy*

Annie wishes she were still in love with Alvy. Because of her desire, she interprets the frequent knots in her stomach in his proximity as evidence that she is still in love—whilst in reality these knots are evidence of a growing anxiety in his presence—and forms the belief that she is still in love with Alvy. She also ignores the fact that she has lost interest in what Alvy does and frequently avoids physical contact. Annie is self-deceived in believing that she is still in love with Alvy.

Here Annie (1) forms the false belief that she is still in love with Alvy; (2) treats the evidence relevant to whether she is still being in love with Alvy in a motivationally biased way, by interpreting, because of her desire to still be in love with him, the knots in her stomach as evidence of her still being in love with him; (3) treats the evidence in such a way that it nondeviantly causes her to form the belief that she is still in love with Alvy; and (4) possesses a body of evidence that provides greater warrant for believing that she is not in love with Alvy anymore than for believing that she is still in love with Alvy.

Condition (2) requires that the self-deceived subject treat evidence in a motivationally biased way. This means that the subject acquires a belief that has been formed thanks to some biasing strategies. The biasing strategies in play may result in, but are not restricted to, the following effects: negative misinterpretation of the evidence, positive misinterpretation of the evidence, selective focusing on the evidence, and selective evidence gathering (Mele, 2001). In what follows, I provide the details of these ways of manipulating the evidence and see how these might play out in “Annie and Alvy.”

In cases of negative misinterpretation, the subject does not count as supporting not-*P* (where *P* stands for “Annie is still in love with Alvy”) evidence that would be recognized as supporting not-*P* in the absence of her desire that *P* (Mele, 2007). In “Annie and Alvy,” there must be evidence that Annie would usually count as supporting not-*P* but that she does not count, in this scenario, as supporting not-*P*—e.g., that her heart does not race when Alvy is nearby.

In cases of positive misinterpretation, the subject counts as supporting *P* evidence that would be recognized as supporting not-*P* in the absence of his or her desire that *P* (Mele, 2007). Again, in “Annie and Alvy,” there could also be evidence that Annie would usually count as supporting not-*P* but that she counts, in this scenario, as supporting *P*—e.g., the knots in her stomach.

In cases of selective focusing, the subject fails to focus on evidence that that subject would usually count as supporting not-*P* in the absence of his or her desire and focuses on evidence that seems to support *P* (Mele, 2007). Recurring panic attacks in the presence of Alvy would be an example of evidence that Annie would usually count as supporting not-*P* in normal circumstances but that she ignores in the actual scenario, where she is self-deceived. Instead she focuses on evidence that seems to support *P*—e.g., that she enjoyed Alvy’s last steamed lobster.

In cases of selective evidence gathering, the subject overlooks evidence supporting not-*P* that would have been easy to obtain in the absence of his or her desire that *P* and finds evidence seemingly supporting *P* that would have been hard to find in the absence of his or her desire that *P*. For example, Annie might avoid opening her journal and instead spend time looking at pictures of their last holiday together. The crucial difference with selective focusing on evidence is that, whereas the latter consists in cognitive operations (e.g., remembering, forgetting,

ignoring, etc.), the former seems to consist in concrete actions (e.g., searching, collecting, looking, reading or not reading one's journal, etc.).

Negative misinterpretation, positive misinterpretation, selective focusing, and selective evidence gathering all consist in successfully manipulating the evidence in such a way that it fits the subject's desire that *P*.

Mele's account captures perfectly well a feature of the phenomenon I have described at the beginning of this section. This is the feature of self-deception that consists in the subject's being in touch with the evidence in one way or another. The subject seems indeed to entertain a paradoxical relation with what that subject believes to be the facts. As I have shown in my presentation of reasons responsiveness, the subject's relation to the evidence is the locus of doxastic responsibility. Indeed, for a subject to be considered responsible for his or her belief, that belief must be reasons responsive. Another way of saying this is that subjects may be held responsible for their belief only if they follow the norms that govern belief formation, maintenance, and revision. According to these norms, subjects should be able to recognize (the right kind of) reasons for their belief if presented with them and should be able to react to these same reasons. This norm concerns the subject's relation to evidence. Thus, if we are to assess the reasons responsiveness of self-deception—to assess whether subjects fulfil a necessary condition for being held responsible for their self-deceptive beliefs—we will have to take a closer look at the feature of self-deception that pertains to the subjects' relation to the evidence. This is what I do in the next section. If it turns out that in cases of self-deception a subject would indeed recognize and react to epistemic reasons in a wide and relevant range of counterfactual scenarios, then the subject's belief is reasons responsive and moreover fulfils a necessary condition for doxastic responsibility.

#### IV. ACTUAL-SEQUENCE MECHANISM

Before I go on to ask whether self-deceptive beliefs fulfil both reasons receptivity and reasons reactivity, it is important to clarify the range of counterfactual scenarios that must be examined in order to determine whether self-deceptive beliefs meet the reasons-responsiveness condition. Delimiting the relevant range of counterfactual scenarios depends on the mechanism of belief formation in the actual sequence, also known as *the actual-sequence mechanism*. In this section, I identify this mechanism as the manipulation of the evidence that takes place in self-deception and that enables self-deceived subjects to form a belief that fits their desire.

What determines the relevance of the counterfactual scenario is the mechanism of belief formation used in the actual scenario, *the actual-sequence mechanism*. The mechanism must be the same across a wide range of counterfactual scenarios. It is however unclear what the conditions for the individuation of the actual-sequence mechanism are (Ginet, 2006; McKenna, 2013), and defining them remains to be done. This is a problem not only for McHugh's epistemic reasons

responsiveness, but for Fischer and Ravizza's original reasons-responsiveness account, too, from which he borrows.

Fischer and Ravizza themselves provide the following explanation:

We must confess that we do not have any general way of specifying when two kinds of mechanisms are the same. This is a potential problem for our approach; it will have to be considered carefully by the reader. But rather than attempting to say much by way of giving an account of mechanism individuation, we shall simply rely on the fact that people have intuitions about fairly clear cases of "same kind of mechanism" and "different kind of mechanism". For example, we rely on the intuitive judgement that the normal mechanism of practical reasoning is different from deliberations that are induced by significant direct electronic manipulation of the brain, hypnosis, subliminal advertising, and so forth. (Fischer and Ravizza, 1998, p. 40)

The individuation of the actual-sequence mechanism is a general difficulty for reasons responsiveness, and my concern is not to solve it here. Instead I will rely on McHugh's own indication regarding the epistemic case, that "the actual sequence mechanism must be owned by the agent" and "might be things like perception, memory and reasoning" (McHugh, 2013, p. 143).

One option is to narrow the mechanism down to these simple mechanisms—i.e., perception, memory, reasoning, etc.—by excluding the subject's desire and biases from the mechanism. The problem with this option is that, if the relevant counterfactual scenarios are those in which the subject is not motivationally biased, then the relevant counterfactual scenarios are scenarios where the subject is not self-deceived in the Melean sense. This is rather odd and gives the impression that we wouldn't be assessing the reasons responsiveness of self-deceptive beliefs only, but those of non-self-deceptive beliefs as well.

It might also be that the actual sequence mechanism consists in the interpretation of evidence in a motivationally biased way—in short, that it consists in the manipulation of the evidence (i.e., negative misinterpretation, positive misinterpretation, selective focusing, and selective gathering). This claim entails that the manipulation of the evidence is a mechanism in its own right, akin to mechanisms such as perception, memory, and reasoning. One might object that, even though manipulation might count as a type of mechanism, there seems to be an important difference between manipulation, on the one hand, and perception, memory, and reasoning on the other: contrarily to perception, memory, and reasoning, manipulation is a composite of simpler mechanisms. It might be, indeed, that manipulation of the evidence requires perceiving the evidence in the first place. However, the same holds for memory. Therefore, if McHugh identifies memory as a legitimate mechanism of belief formation, then conferring the same status onto manipulation should not be an issue. That manipulation constitutes a mechanism in its own right does not seem too unreasonable;

it is after all the way in which a subject acquires a self-deceptive belief. In the same way that a subject comes to believe by means of perceiving, the self-deceived subject comes to believe by means of manipulating the evidence. When we ask which mechanism of belief formation led the subject to believe deceitfully, it is more likely that we point at the manipulation of the evidence rather than at a noncomposite form of mechanism.

I have identified the actual-sequence mechanism in the case of self-deception with the manipulation of the evidence. What has not been determined is whether the manipulation of the evidence should remain of the very same kind (that is, fixed) throughout the counterfactual scenarios (e.g., positive misinterpretation)—this is manipulation narrowly conceived. Alternatively, the manipulation of the evidence could be allowed to adjust to different ways of becoming motivationally biased (e.g., positive misinterpretation at  $w_1$ , negative misinterpretation at  $w_2$ , selective focusing at  $w_3$ , ...,  $w_n$ , where  $w_1$ ,  $w_2$ , and  $w_3$ , ...,  $w_n$ , are relevant counterfactual scenarios)—this is manipulation broadly conceived. I presuppose that allowing the kind of manipulation to vary across the counterfactual scenarios is analogous to allowing the way one visually perceives to vary (e.g., wears glasses at  $w_1$ , squints at  $w_2$ , etc.), hence the general advice for the individuation of the actual-sequence mechanism is not violated. It would make little sense not to allow the kind of manipulation to vary. After all, if the kind of manipulation does not vary depending on the epistemic reasons or on the evidence with which the subject would be confronted, the relevant range of counterfactual scenarios would be overly restricted. That is, the range would be restricted to counterfactual scenarios in which, for example, the subjects positively misinterpret their evidence in order to believe what they desire. Because it only makes sense that the evidence counting against what the subject desires would be positively misinterpreted by the subject, the range of counterfactual scenarios would also be restricted to counterfactual scenarios in which the subject is confronted with evidence counting against the subject's desire (if we're looking at belief formation) or belief (if we're looking at belief maintenance). A range of counterfactual scenarios thus restricted is in fact not relevant, in the sense that the very point of thinking about reasons responsiveness in counterfactual terms is to look at what would happen if the subject were confronted with evidence that is different than in the actual scenario. Let us add here as well that what matters when it comes to individuation of the actual-sequence mechanism is that it would clearly be the agent's or the subject's own—that acting or believing due to a certain mechanism would be distinct from acting or believing due to a mechanism that is not the agent's or subject's own, having been implanted in his or her brain by an external person, for example. For these reasons, I will assume that the ways in which the subject manipulates the evidence should be allowed to vary.

## V. REASONS RECEPTIVITY

In this section, I assess the reasons receptivity of self-deceptive beliefs. For any  $S$ , where  $S$  is a self-deceived subject, to fulfil the reasons-receptivity require-



ment, the following proposition must be true: *S* is self-deceived and *S* would recognize evidence against *P* across a wide range of relevant counterfactual scenarios. In other words, it must be the case that “*S* is self-deceived” and “*S* would recognize evidence against *P* across a wide range of relevant counterfactual scenarios” can be true at the same time.

Let’s start with the actual scenario. In the actual scenario, *S* becomes self-deceived, let’s suppose, by means of positive misinterpretation. *S*’s desire that *P* leads *S* to count as supporting *P* evidence that supports not-*P*. For example, Annie’s desire to still be in love with Alvy leads her to interpret her stomach knots as support for the belief that she’s still in love, though the evidence in fact supports the belief that she’s not in love anymore (let’s say the knots are caused by her anxiety). For Annie to be receptive to reasons, there must a wide range of relevant counterfactual scenarios in which she would recognize evidence supporting not-*P* (that she’s not in love with Alvy anymore). The basic idea is that Annie must be sensitive to alternative evidence across these scenarios.

If the actual-sequence mechanism is the manipulation of the evidence, then the relevant range of counterfactual scenarios are the ones where Annie comes to believe a proposition by manipulating the evidence. In these counterfactual scenarios, evidence, whether it supports *P* in some or not-*P* in others, is manipulated and interpreted as supporting *P*. Because Annie comes to believe *P* in all the relevant counterfactual scenarios, it might look as though, at first sight, her belief is not receptive to reasons. That she does might result in Annie’s belief not being reactive to reasons, but for now let us simply look at the receptivity of her belief.

If we identify the mechanism of belief formation as the manipulation of the evidence itself, the relevant range of counterfactual scenarios are the ones where *S* comes to believe that *P* by manipulating the evidence in various ways. Thus, for each variation in the evidence, during which the desire that *P* remains constant, the way of manipulating the evidence is adjusted to the evidence at hand. In these counterfactual scenarios, *S* recognizes evidence supporting *P*, or evidence supporting not-*P* *in order to be further taken as supporting P*. In “Annie and Alvy,” the evidence across the counterfactual scenarios might vary: at *w*<sub>1</sub>, Annie has knots in her stomach, at *w*<sub>2</sub> she does not, at *w*<sub>3</sub> she has a panic attack, etc. In each counterfactual scenario, the way of interpreting the evidence changes so as to adjust to what is required in order to satisfy Annie’s desire to still be in love with Alvy. At *w*<sub>2</sub>, for example, Annie might resort to negative misinterpretation by not taking the absence of knots in her stomach as evidence supporting that she is not in love. Evidence at *w*<sub>3</sub> might call for selective focusing. The point is that such adaptation requires the ability to recognize evidence. Thus, according to a conception of the manipulation of the evidence broadly conceived, self-deceptive beliefs fulfil reasons receptivity. If they fulfil reasons reactivity as well, then self-deceptive beliefs are responsive to reasons.

## VI. REASONS REACTIVITY

I now turn to reasons reactivity. For  $S$ , where  $S$  is a self-deceived subject, to fulfil the reasons-reactivity requirement, the following proposition must be true:  $S$  would react to evidence against  $P$  across a wide range of relevant counterfactual scenarios. In other words, it must be the case that “ $S$  would react to evidence against  $P$  across a wide range of relevant counterfactual scenarios” must be true at the same time. I hinted at the fact, in the previous section, that, because the self-deceived subject forms the belief that  $P$  no matter the evidence he or she recognizes, it might look as if self-deceptive beliefs fail to meet the reasons-reactivity part of reasons responsiveness.

In the actual scenario,  $S$  becomes self-deceived, let us suppose again, by means of positive misinterpretation.  $S$ 's desire that  $P$  leads  $S$  to count as supporting  $P$  evidence that supports not- $P$  and further to form the belief that  $P$ . Annie's desire to still be in love with Alvy would lead her to count her stomach knots as supporting the belief that she is still in love, though the evidence in fact supports the belief that she is not in love anymore (suppose the knots are caused by her anxiety), and, on this basis, she would form the belief that she is still in love. For Annie to be reactive to evidence, there must a wide range of relevant counterfactual scenarios in which she reacts to evidence she recognizes as supporting not- $P$  by forming a belief accordingly. In fact, Annie would react to evidence she recognizes as supporting not- $P$  by forming a belief *contrary* to what she recognizes as evidence. This way of failing to meet reasons reactivity, if it is indeed the case, has little to do with the way in which subjects are typically said to fail to meet this condition. Indeed, as we've seen earlier, what McHugh has in mind are rather cases of what he calls “repressed beliefs,” beliefs whose content the subject rejects while maintaining his or her belief. But this is not what is happening in the case of self-deceptive beliefs. It is not that the subject would not react to evidence going against what that subject desires. Successful self-deception in fact requires that the subjects react to what they recognize to be evidence against  $P$ , except that they would react contrarily to this evidence. For McHugh, in order to fulfil reasons reactivity, the subject's belief should be such that in a wide range of counterfactual scenarios, the subject would react to evidence “by forming the doxastic attitude she takes them to call for” (McHugh, 2017, p. 2751). In the case of self-deception, evidence recognized as counting against  $P$  calls for forming or maintaining the belief that  $P$ . Although this way of reacting to evidence is epistemically vicious, it does not show that the subject lacks the capacity to react to evidence and, in this sense, it does not infringe on the subject's responsibility for his or her belief.

The proposition “ $S$  would react to evidence he or she recognizes against  $P$  across a wide range of relevant counterfactual scenarios” is true. Therefore, self-deception satisfies reasons reactivity.

## VII. CONCLUSION

Self-deceptive beliefs fulfil both reasons receptivity and reasons reactivity. Consequently, self-deceptive beliefs fulfil reasons responsiveness and thus meet a necessary condition for doxastic responsibility. If self-deceptive beliefs are reasons responsive, depending on whether or not one is willing to accept reasons responsiveness as a sufficient as well as necessary condition for doxastic responsibility, then we may be held responsible for our self-deceptive beliefs. However, reasons responsiveness is only a necessary condition for doxastic responsibility. This leaves room for sometimes not being held responsible for self-deceptive beliefs, since reasons responsiveness is not sufficient.

I will not take on the task of supplementing McHugh's account by showing that the reasons-responsiveness condition is also sufficient for responsibility. I do think the latter idea is plausible, however. Before concluding, I would like to at least provide some support for this idea. The following rough picture of responsibility is generally agreed upon: an epistemic condition and a control condition are required and, if fulfilled, add up to responsibility. Each theory then works out these conditions in further detail. We have seen that the case of belief poses a challenge to this picture of responsibility as it is not under our direct voluntary control. Instead McHugh offers a condition for responsibility that depends on an ability the subject should possess. This ability, as we've seen, consists in recognizing alternative evidence to one's belief and reacting accordingly if presented with it. This ability, reasons responsiveness, includes an epistemic component—in requiring a certain awareness on the part of the subject vis-à-vis their reasons for believing—and, for lack of a *control* component, a component that draws on the proper functioning of the subject's mechanism of belief formation. Once we concede that there can be responsibility without direct voluntary control, these two components of reasons responsiveness together seem to exhaust what can be required of a responsible belief.

To conclude, I have argued that, at least on Mele's account of self-deception, self-deceptive beliefs always fulfil the reasons-responsiveness condition for doxastic responsibility. This is because the self-deceived subject's relation to evidence is such that the subject would recognize and react to evidence against *P* if that subject were presented with it. I believe not only that this feature of self-deception is present in Mele's account of self-deception, but that we find it in any account of self-deception that posits a similar relation to the evidence on the part of the subject. In this sense, self-deceived subjects always fulfil a necessary condition for doxastic responsibility.

## ACKNOWLEDGEMENTS

I am grateful to Anne Meylan, Melanie Sarzano, Benoit Gaultier, and two anonymous referees at *Les Ateliers de l'Éthique/The Ethics Forum* for their constructive and insightful comments, which helped improve this article. I also thank the audiences of the “New Perspectives on Self-Deception” workshop at the University of Montreal, of the “Self-Deception: What It Is and What It Is Worth” conference at the University of Basel, of the “Epistemic and Practical Rationality” workshop at the University of Fribourg, and of the “Ateliers du Groupe de Recherche en Épistémologie” at the Collège de France, and especially Pascal Engel, Veli Mitova, Dana Nelkin, Sarah Stroud, and Christine Tappolet for their helpful questions and comments. Research for this article was supported by the Swiss National Science Foundation (SNSF) Professorship grant “Cognitive Irrationality” # PP00P1\_157436 and the Doc.Mobility grant “Irrational, Yet Responsible” # P1BSP1\_178606.

## NOTES

- <sup>1</sup> Epistemic at least. I remain agnostic on whether self-deception further consists in a *moral* mistake. In this sense, I follow Neil Levy, who argues that we shouldn't be presumed culpable for self-deception (Levy, 2004). However, I make a step towards disagreeing with him on whether we shouldn't necessarily hold subjects responsible for their self-deception, since I argue that self-deceptive beliefs always fulfil a necessary condition for responsibility. What is being reproached of the subjects is that they should have believed better at least according to epistemic norms. Plausibly, self-deception also constitutes a moral mistake: we might reproach the subjects for believing something they shouldn't according to moral norms as well.
- <sup>2</sup> Nelkin rightly points out that, in order to ultimately gain a proper understanding what responsibility for self-deception is, we should “get clear about exactly what the self-deceiver is supposed to be responsible for. We should distinguish between the process of self-deception, the immediate product of self-deception, and its more indirect consequences” (Nelkin, 2013, p. 129). She herself proposes several ways in which one might deal with the questions, one of them being to understand self-deception as a case of culpable negligence. Note also that, as she underlines, her approach to reasons responsiveness is agent based rather than mechanism based (Nelkin, 2013, p. 126).
- <sup>3</sup> Some argue that it is in fact possible to believe at will (Peels, 2015) or intentionally (Steup, 2017).
- <sup>4</sup> It is not clear whether McHugh takes reasons responsiveness to be both necessary and sufficient for responsibility. In McHugh, 2013, McHugh mentions that reasons responsiveness might also be sufficient.
- <sup>5</sup> In McHugh, 2014, McHugh specifies the kind of reasons in play here. They are what he calls “object-directed reasons”—that is, reasons “for being in some doxastic state that are such because they pertain to whether its content is true or false” (McHugh, 2014, p. 16) or reasons that consist in “considerations that pertain to the world's being as it is represented in the state's content” (ibid., p. 16-17). In other words, when *S* is epistemically reasons responsive, the reasons to which *S* responds are of the sort that make it epistemically warranted for *S* to believe that *P*.

## REFERENCES

- Alston, William, "The Deontological Conception of Epistemic Justification," *Philosophical Perspectives*, vol. 2, 1988, p. 257-299.
- DeWeese-Boyd, Ian, "Taking Care: Self-Deception, Culpability and Control," *Teorema*, vol. 23, no. 3, 2007, p. 161-176.
- Fischer, John M., and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge, Cambridge University Press, 1998.
- Ginet, Carl, "Working with Fischer and Ravizza's Account of Moral Responsibility," *Journal of Ethics*, vol. 10, no. 3, 2005, p. 229-253.
- Hieronymi, Pamela, "Responsibility for Believing," *Synthese*, vol. 161, no. 3, 2008, p. 357-373.
- Levy, Neil, "Self-Deception and Moral Responsibility," *Ratio*, vol. 17, no. 3, 2004, p. 294-311.
- McHugh, Conor, "Epistemic Responsibility and Doxastic Agency," *Philosophical Issues*, vol. 23, no. 1, 2013, p. 132-157.
- , "Exercising Doxastic Freedom," *Philosophy and Phenomenological Research*, vol. 88, no. 1, 2014, p. 1-37.
- , "Attitudinal Control," *Synthese*, vol. 194, no. 8, 2017, p. 2745-2762.
- McKenna, Michael, "Reasons-Responsiveness, Agents and Mechanisms," in *Oxford Studies in Agency and Responsibility*, vol. 1, 2013, p. 151-183.
- Mele, Alfred, "Real Self-Deception," *Behavioral and Brain Sciences*, vol. 20, no. 1, 1997, p. 91-102.
- , "Twisted Self-Deception," *Philosophical Psychology*, vol. 12, no. 2, 1999, p. 117-137.
- , *Self-Deception Unmasked*, Princeton, Princeton University Press, 2001.
- , "Self-Deception and Delusions," *European Journal of Analytic Philosophy*, vol. 2, no. 1, 2006, p. 109-124.
- Meylan, Anne, "The Legitimacy of Intellectual Praise and Blame," *Journal of Philosophical Research*, vol. 40, 2015, p. 189-203.
- , "The Consequential Control of Doxastic Responsibility," *Theoria*, vol. 83, no. 1, 2017, p. 4-28.
- Nelkin, Dana, "Responsibility and Self-Deception," *Humana.Mente: Journal of Philosophical Studies*, vol. 5, no. 20, 2012, p. 117-139.
- Peels, Rik, "Does Doxastic Responsibility Entail the Ability To Believe Otherwise?" *Synthese*, vol. 190, no. 17, 2013, p. 3651-3669.
- , "Believing at Will Is Possible," *Australasian Journal of Philosophy*, vol. 93, no. 3, 2015, p. 1018.

———, *Responsible Belief: A Theory in Ethics and Epistemology*, Oxford, Oxford University Press, 2017.

Steup, Matthias, “Belief Control and Intentionality,” *Synthese*, vol. 188, no. 2, 2012, p. 145-163.

———, “Believing Intentionally,” *Synthese*, vol. 194, no. 8, 2017, p. 2673-2694.

Williams, Bernard, “Deciding To Believe,” in *Problems of the Self*, Cambridge, Cambridge University Press, 1973, p. 136-151.