THE ANALYSIS OF SELF-DECEPTION: REHABILITATING THE TRADITIONALIST ACCOUNT

By

Vladimir Krstić

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Philosophy, the University of Auckland, 2018.

The University of Auckland,

2018

Abstract

Traditionalists affirm that in self-deception I intend to deceive myself; but, on the standard account of interpersonal deception, according to which deceiver intend to make their target believe a falsehood, traditionalism generates paradoxes, arising from the fact that I will surely know that I want to make myself believe a falsehood.

In this thesis, I argue that these well-known paradoxes need not arise under my *manipulativist* account of deception. In particular, I defend traditionalism about self-deception by showing that what causes paradoxes is not the idea that self-deception is an *intra*personal analogue of *inter*personal deception but rather our incorrect conceptions of deception, interpersonal deception, and lying. I show, by way of counterexamples, that the essence of deception is not ending up epistemically worse off but rather being intentionally manipulated into forming or retaining a certain truth-evaluable mental state. Vitally, contra the standard view, the deceiver need not hold that the mental state he intends to produce in the deceived involves a falsehood; in fact, the targeted effect may involve a proposition believed by the deceiver to be true or about whose truth the deceiver suspends belief.

Any phenomenon rightly called self-deception, I argue, involves an action in which the person intentionally manipulates (by way of a trick) her own way of forming or retaining some truth-evaluable mental state of her own (belief or a belief-like thought). Thus, on the manipulativist view, a self-deceiver may non-paradoxically intend to deceive himself: the relevant intention is to affect a particular truth-evaluable thought by way of non-deviantly manipulating his agential use of his own cognitive capacities.

While all cases of self-deception involve intentional manipulation of the self by the self, how this manipulative trick works out varies in different kinds of cases. In some cases, the self-deceptive manipulative trick is not intentional under that description, a description that captures the so-called *deflationary* view.

Although the manipulativist account allows for numerous different strategies of self-deception, and is consistent with both traditionalism and deflationism, it cannot, by itself, vindicate the idea of lying to oneself. On the standard conception, lying to myself is practically impossible (if the mind is not partitioned) and this is because the very forming the intention to lie to myself – as including the intention to make myself believe my own lie – seems very unlikely. Arguably, people will not φ if they think that φ-ing is impossible. I argue that some kinds of behaviour typically understood as self-deception are actually cases in which the person is lying to herself without the intention to deceive herself; rather, she avoids legitimising an unfavourable state of affairs by openly affirming its contrary. It may also be the case, however, that some lies to oneself are aimed at deceiving – this is because, as I argue, liars need not assert what they believe is false; hence, by lying to myself, I may intend to increase my confidence in the belief I already hold.

Finally, the motive for self-deception is always a matter of preserving the person's endorsed idealised picture of reality. This account allows me to explain all possible cases of self-deception. In the last chapter, I apply my theory to some paradigmatic cases first outlined in my introduction.

Table of Contents

I. INTRODUCTION

1. Four Main Cases

The following four examples – Hillary, Maria, Nicole, and Zelda – are thought by various people to be examples of self-deception.[1] I will use them as a basis for my discussion (other cases will also be discussed).

> **Hillary** possesses sufficient evidence that her husband Bill is having an affair with Monica. It's practically all over the news and yet she keeps consistently arguing that Bill is not having an affair and her other behaviour is in harmony with what she says. This does not seem to be ignorance, simple denial, or an innocent cognitive error, however. Hillary is a very smart person, she must have recognized the signs, and the very thought of Bill having an affair must be painful to her – they've been dating since college. Therefore, one might say that she actually realised that Bill is having an affair with Monica and that, wanting to break out from this distressing realisation, she deceived herself into believing what she wants to believe.

Some people, and we call them traditionalists about self-deception, tend to think that Hillary has deceived herself in the way just described because her behaviour and her epistemic situation – namely, her intelligence and available evidence – are best described as an *intra*personal analogue of *inter*personal deception. On the standard view, in interpersonal deception, the deceiver intentionally causes that the deceived ends up believing what the deceiver believes to be false and wants the deceived to believe. Likewise, Hillary must have noticed the evidence and formed the relevant yet distressing belief that Bill is having an affair. And, wanting to break out from believing what causes her distress, the rationale is, she decided to deceive herself into believing what she wants to be true (but believes to be false).

On this model of self-deception, a person believes or knows one thing and, due to motivation, intentionally gets herself to believe another, typically favourable, proposition. The deception is such that either the initial belief is replaced with the belief one ends up with

[1] The third example is by Funkhouser [2005: 302], the first two are modified versions of that example combined with two real characters. The reason for mixing reality with fiction is purely technical: I assume that the reader will find it easier to memorise names of famous people – I personally find this much easier. The names from the first example are known to everyone. The second example is of a famous bodybuilder, actor, and a former governor of California. That said, please notice that this thesis is not promoting any political options; I am just using things that are common knowledge in order to make reading easier. The fourth example is based on two separate cases of paranoid jealousy merged and presented as one. The characters of Scott and Zelda Fitzgerald, and their friend Ernest Hemingway are, of course, real. Finally, for the sake of simplicity, I describe these states as beliefs but we can imagine the agents from my examples entertaining a kind of belief-like thoughts, i.e., thoughts that manifest themselves in a way beliefs manifest themselves.

or is buried deep down so that it does not manifest itself in behaviour, visibly at least. Furthermore, everything looks as though the initial unfavourable, or even rather disturbing, belief is somehow responsible for the belief with which the person ended up – believing what I do not want to be the case or to believe gives me a reason to deceive myself. Finally, I am a self-deceiver only if this was done in the face of evidence that is by my own epistemic standards inadequate for the belief to be formed, which is what Hillary seems to have done.

Then, there is a case of Maria, who seems to be a self-deceiver but who does not appear to have violated her own standards of rationality and belief-formation, who did not intend to deceive herself, and who never realised the truth.

> **Maria** possesses much evidence that her husband, Arnold, is having an affair with one of his attractive female co-workers. Arnold has lost sexual interest in Maria, he has protected his phone with a password, and she sometimes picks up what appear to be subtle love signals her husband exchanges with this female co-worker. Yet, Maria intentionally sets these distressing pieces of information aside and, as a result, retains the false belief that Arnold is not having an affair. However, she does not intend to deceive herself (since she never suspected an affair, we think that she lacks the motive for deception). Rather, she retained her belief by what looks like explaining away pieces of information that cause her distress. In a sense, she seems to be acting just like scholars who only give weight to claims that support their own position and avoid those that may undermine it.

Unlike Hillary, Maria did not intend to deceive herself. Nevertheless, many people tend to see this case as involving self-deception, and they do this for two reasons. The first is that Maria ended up 'being deceived,' i.e., believing a falsehood, and the second that her state of deception was caused by her intentional action – she intended to avoid distressing information. The idea is that because she ended up being deceived and because deception was caused by her own intentional action, she has deceived herself unintentionally by acting intentionally.

The third case is the case of Nicole, an agent who exhibits a rather peculiar behaviour.

> **Nicole** possesses much evidence that her husband Tony is having an affair with her friend Rachel. Nicole's other friends have reported to her that Tony's car is often seen parked in Rachel's driveway, at times when he claims to be with his friends or at work. Tony has lost sexual interest in Nicole, and other suspicious behaviour provides sufficient evidence for Nicole to be more than sceptical. Yet she laughs off the concerns of her girlfriends and thinks to herself that Tony certainly is a faithful husband. ('After all, I am still an intelligent, charming, and attractive woman – certainly more so than Rachel!') Yet, in the evenings when Tony claims to be with his friends, Nicole avoids driving by Rachel's house – even when it requires her to drive out of her way.

There are a number of similarities between this and the Hillary case and these similarities are a strong incentive to regards this case as involving self-deception. For example, we may say that Nicole is deceiving herself because her behaviour and her doxastic situation fit the so-called 'lying model' of self-deception, she behaves just like Hillary (Bach 2009: 782). This model describes a person who believes or even knows one thing and hopes to get herself to believe another, typically favourable and antithetical proposition. Furthermore, just as in Hillary's case, everything looks as though the initial distressing belief is somehow responsible for the thought that corresponds to Nicole's sincere avowals while the acquiring of the false thought was done in the face of evidence inadequate by Nicole's own epistemic standards. However, and this is vital, unlike Hillary, Nicole's non-verbal behaviour contradicts her sincere avowals. Nicole does not change her mind in the sense of having one belief replaced by another. Rather, the favoured antithetical thought appears to be added to her stock of beliefs, which then results in inconsistent behaviour. Meanwhile, the original belief keeps on affecting some aspects of Nicole's behaviour without her consciously recognizing this influence or the belief's existence. The problem with this case is that we do not know what, if anything, Nicole believes.

Finally, Zelda, the fourth of our supposed self-deceivers, seems to be in the worst situation.

> **Zelda** mistakenly believes that she possesses plenty of evidence that her husband Scott is having an affair with his new close friend Ernest. Although Scott denies vigorously that they are having an affair, they are getting along great and that can't be right – she is quite sure of that. In fact, the more she thinks about it, the more evidence she finds. Thinking about that one time they were in Paris and met Ernest, she remembered how Scott went to the bathroom in the middle of the night and then she became convinced that he actually went to Ernest. Cheating fits Scott, she reasons, because he's such an unlikely person to do it and their marriage has been falling apart for months now. Zelda does not exhibit any signs of delusion or schizophrenia: she does not, for instance, think that Ernest is leaving Scott secret messages in newspapers, etc. (see Sass 1994: 5). Therefore, we cannot say that she is deluded but she does seem to be self-deceived. However, she certainly does not want Scott to be having an affair or that she believes that. Even worse, her belief makes her desperate.

According to the description of the case, Zelda has undergone what I will call unwelcome self-deception, in which a person who has deceived herself into believing that *p* has no desire that *p* be true.[2] Her case is important not because it involves a new mechanism of self-

[2] Mele [e.g., 1997: 132] calls this type of self-deception *twisted* but I find this term misleading: the word twisted implies that this action is a deviation from a rule. There is nothing odd with unwelcome self-deception, I will argue. It is only that the content of the product is not something the agent would like to be the case. Van Leeuwen [2007a: 425] calls this kind of self-deception *dreadful* self-deception. Nevertheless, although Zelda's

deception or an interesting doxastic situation of the self-deceiver (this is where the case is analogous to the Maria case), but rather because the product is something the person does not want. In fact, the whole description of the case is meant to highlight not the *how* aspect but rather the *why* aspect of her situation. Therefore, in analysing this case, I will focus on the 'why' question. Why did she deceive herself into believing something she does not want to be the case? This is the main problem here. Any successful theory of self-deception, or of any kind of self-deception, must be able to explain both the cases of welcome self-deception, when we get what we want, and the cases of unwelcome self-deception, when we do not.

Another extremely important feature of Zelda type cases is that we do not intuitively see the agents from these cases as initially knowing the truth – this would not make sense. That is to say, it does not make sense to say that Zelda deceives herself into thinking that Scott is having an affair, thereby causing herself significant distress, if she knows that he is not having an affair.

1.1. Understanding Self-Deception

Each of these examples may be described as involving self-deception, i.e., an intentional action by which the self deceives itself, and yet each case is perplexing in its own way. For example, we said that Hillary acted so that she abandoned her newly formed yet disturbing true belief in order to continue 'living in a lie.' This case seems like a textbook example of self-deception and it fits the traditional understanding of self-deception, the so-called *traditional* or *traditionalist view*, which takes self-deception to be modelled on interpersonal deception (Subchapter III-1). Specifically, if interpersonal deception involves person A intentionally getting person B to believe what A believes to be false, and Hillary is here both A and B, then, because Hillary successfully acted on her intention to cause herself to believe a favourable falsehood, she has deceived herself in a manner perfectly consistent with interpersonal deception.

However, if Hillary already believed that Bill is having an affair, and if this is why she initiated the deception, then how can she successfully deceive herself that he is *not* having an affair? A prerequisite of successful deception is that the victim does not know the truth. Furthermore, according to this traditionalist description, Hillary knows that she intends to deceive herself into believing something she takes to be false and you cannot deceive those

---

case is indeed dreadful, many other cases are not. These agents simply end up believing a proposition they would not welcome. Therefore, I adopt 'unwelcome,' the name proposed by Barnes [1997].

who know that you are trying to deceive them. Given these two problems, not many philosophers are traditionalists.

In the second example, Maria performed an intentional action as a result of which she ended up believing a falsehood, but it was not her intention to end up believing a falsehood. Biased scholars do not intend to adhere to bad arguments; they intend to adhere to *their* arguments. Therefore, the problem with this example, which fits with the account of the so-called *deflationary view* (Subchapter III-2), is not that we do not see how Maria successfully got herself believing a falsehood – people do this all the time. Rather, the problem is that Maria did not *intend* to cause herself to believe a falsehood, which is why her action does not appear to be intentional under the relevant description, i.e., 'deceiving myself.'

One reason for thinking self-deception is analogous to intentional interpersonal deception (hereafter just 'interpersonal deception') is that this helps us to distinguish self-deception from mere error – since the acquisition and maintenance of the false belief is intentional not accidental (Deweese-Boyd 2016). Many philosophers have argued that deception must be intentional (e.g., Carson [2010]; Saul [2012]; Faulkner [2013]; see Mahon [2008]) – they typically reserve the term 'mislead' to cover cases of causing false beliefs unintentionally – and some of these philosophers are also deflationists about self-deception (e.g., Mele [e.g., 1997]; Barnes [1997]; Galeotti [2012]). However, while it understands self-deception as caused by an intentional action, the deflationary account seems to be failing to capture the *right* description under which the action is intentional. Let me explain this concern further.

According to Davidson's [1963] theory of action, if I φ-ed intentionally, this must be because I intended to φ; φ-ing cannot be an unintended consequence of my action. If I flip the switch in order to illuminate the room and thereby unintentionally alert the prowler, my action is not intended as an action of alerting the prowler but as illuminating the room. Likewise, even though Maria acted intentionally and the action resulted in her believing a falsehood, because she did not act on the intention to deceive herself or to retain her false belief, Maria's action seems to be intentional as 'gathering evidence "my way"' rather than as 'deceiving myself' or even as 'making myself retain my belief.' Thus, Maria is guilty of forming her belief on insufficient evidence but (self-)deception here seems to be an unintended side-effect of the latter two. It follows, then, that Maria merely misled herself unintentionally into believing that Arnold is not having an affair; she did not deceive herself.

This worry, however, is not as strong as it initially appears. There is a sense in which we may rightly say that I intentionally φ-ed even though I did not know that what I am doing counts as φ-ing. Suppose that you ask me to play American football with you and your friends, a game I know nothing of. As the game progresses, it happens that I am in the open and so you throw me the ball, I catch it, and you shout 'Run behind the opponents' goal line and touch the ground!' At your command, I run towards the opponents' goal line and I touch the ground. In this case, I ran towards the opponents' goal line and touched the ground intentionally but I did not realise that I am doing this in order to score a touchdown, which my touching the ground effectively is. My action may rightly be called scoring a touchdown even though I did not know that what I am doing counts as scoring a touchdown. The *general* deflationist idea is that self-deceivers perform actions of this sort: they intentionally perform an action that results in their ending up deceived and, while they do not know what they have done, they have nonetheless done it.

However, while working on a plausible general idea, there is a sense in which deflationism falls short of giving us the right account of self-deception. Genuine deceivers must be insincere towards their victim and this vital element is missing from the way in which we have been interpreting Maria's behaviour and, as I argue in Subchapter III-2, from specific deflationist theoretical renderings of this general idea. Insincerity is vital for self-deception for two reasons. First, it keeps 'deceiving' in 'self-deceiving,' since there is no deception without insincerity. Second, it puts into perspective the intuition that self-deceivers somehow must violate their own norms of reasoning; that is, self-deceivers form or retain beliefs or belief-like thoughts they (as the deceived) would not have formed or retained had they (as the deceiver) not been insincere towards themselves (as the deceived). Not staying true to your own standard of reasoning, therefore, keeps the dichotomy between the deceiver and the deceived (which *is* a genuine hallmark of any kind of other-deception) present in self-deception.

The following, then, is the real problem of most deflationary views: they do not give us a theory according to which (i) self-deceivers are violating their own norms of reasoning caused by (ii) their insincerity towards themselves. Granted that faithfully modelling self-deception on interpersonal deception indeed produces an untenable account of self-deception, these two requirements should constitute a bare minimum self-deceivers must satisfy. Otherwise, we will effectively make a complete break with 'deceiving' when describing

'self-deceiving.' This, of course, is not a *sufficient* reason to abandon deflationism but it is a reason to prefer a view that, all things being equal, captures the analogy with interpersonal deception better, which would be a version of the traditional view that does not generate the paradoxes. That is to say, i.e., unless we find a way in which *Maria* satisfies (i) and (ii) (and, in V-3, I argue that it does), her behaviour should be understood either as not involving self-deception or as self-deception under the traditionalist description.

The third case describes a situation in which someone's avowals (Nicole laughs off her friends' claims) contradict her behaviour (Nicole also avoids driving by Rachel's house). Nicole proclaims – even to herself, which should alert us to the fact that she is sincere in her communication to others – that Tony is not having an affair with Rachel but she behaves in certain circumstances in a manner that strongly indicates that she believes that he is. Philosophers who take this to be a case of self-deception – and I will argue that they might be wrong, Nicole is lying to herself without the intention to deceive herself – also take it that internal conflict or, at least, tension is inherent in self-deception.

On the face of it, however, it may rather seem that Nicole is not deceiving herself but that she is merely going through an unsuccessful denial or repression – she does not appear to be deceived, she does not get anything wrong. The problem is that, even if we grant that Nicole wants and intends to deceive herself – she keeps telling herself that she is better than Rachel – she does not seem to have really taken the bait. She rather seems to be in a state that results from a failed attempt at deceiving oneself. This concern is the primary reason to consider revising the concept of self-deception; the resulting solutions can be classified as belonging to the *revisionist view* (Subchapter III-3), a family of solutions that reconceptualise self-deception in some important way.

The concern that Nicole does not seem to believe (consistently) what she asserts is sometimes resolved by saying that she merely believes that she believes what she asserts but that this second-order belief is false; plausibly, she merely wishes that Tony is not having an affair while actually believing that he is. Also, it could be that the mental state on which she reports is some hybrid mental state, something in between belief and imagination, or a kind of pretence, and so on. The main feature of these and similar explanations is that the self-deceiver has false beliefs about herself; in this case, Nicole is interpreted as having a false belief about what she believes. Accordingly, self-deception may also involve a type of

deception *about* the self, and some think, principally psychologists, that it *just is* deception about the self – no more, no less (III-3a).

An alternative revisionist strategy is claiming that the product of self-deception need not be a belief at all but some other metal state – plausibly, a hybrid one – that can be misidentified as a belief. Similarly, it could be that Nicole believes neither of the propositions. Rather, we should refrain from attributing or denying a belief when the dispositions the subject manifests do not warrant a determinate attribution. The conflict is not between beliefs but between dispositions, components of a belief. Considering these explanations, the correct description of this case would then be a *failed* attempt at deception by the self; since Nicole did not get what she wanted to get (III-3b). Incidentally, on this view, self-deception also involves a type of deception about the self.

Revisionist solutions are quite exotic and they sit uneasily with our folk-psychological practices but they seem to be a necessity; traditionalism simply asks too much of self-deceivers whereas deflationism asks too little. However, most revisionist strategies struggle with cases of full beliefs, in which the agents correctly take themselves as believing. Finally, Zelda's behaviour remains a puzzle to be solved. With so many substantially different solutions on the table, one may easily lose track of the phenomenon under scrutiny and start seeing self-deception in everything that may involve the self and some sort of cognitive error. And this does happen quite often, I argue in chapter III. So, what should we do now? I say we go back to the beginning, to the origin of the problem: namely, the conceptual analysis of interpersonal deception and lying.

1.2. Where Should We Go From Here

Often, identifying the person's action can be a significant aid in understanding her behaviour and relevant mental states. However, as we have seen, saying that I am deceiving myself may just create more problems. Accordingly, we may ask: Is 'self-deception' a concept coined with the purpose of explaining a class of actions, or did it just emerge from an unwitting combination of two terms that actually don't make much sense together? That is to say, the worry is that the concept is internally incoherent (e.g., Szabados [1974: 52]; Borge [2003: 1]; Porcher [2012: 68]).

In this thesis, I argue that 'self-deception' is not an internally incoherent concept but that it nevertheless is incoherent relative to our current best theories of lying and deception.

Specifically, I admit that the current traditional account of self-deception is paradoxical and untenable but then argue that this is only because its foundations, standard theories of lying and interpersonal deception, are incorrect. Therefore, I will argue that *traditionalism about self-deception* is correct in taking it that self-deception is modelled on interpersonal deception and that it typically involves lying to oneself, and I will defend this proposal by rejecting *'traditionalism' about lying and interpersonal deception* (notice, all other theories of self-deception are also considering the same traditional accounts of lying and interpersonal deception). I will deny, in particular, that deceivers must intend to make the target believe a falsehood, and that liars must assert what they believe to be false, or that they must intend to deceive.

On my reading, once purged of incorrect theories of lying and interpersonal deception, traditionalism will become our best theory of self-deception: it will generate no paradoxes or problems, it will correspond to our intuitions, and, unlike many other theories, it will avoid generating problematic implications. However, and this is also important, this is not to deny the merits of other theories or to argue that only a traditional account can explain self-deception. Rather, the main idea is that traditionalism, as offering the most successful account of self-deception, should come as our first choice in explaining a particular case that appears to involve it but that, if the traditionalist solution does not seem to be fitting, we should seek other solutions – since people may sometimes deceive themselves by way of intentionally acting but not acting with the intention to deceive themselves.

In this introduction, I will sketch some main problems traditionalist theories face (Section 1.2.1), and explain how my proposed account avoids these problems while keeping the advantages of traditionalism (Section 1.2.2).

1.2.1 Received Traditionalism about Self-Deception

In determining how the terms 'deceive' and 'deceived' should be understood, philosophers – and not just traditionalists – typically take the lexical approach. In a standard use, '(being) deceived' is taken to be equivalent to '(being) mistaken,' as for example in the sentence 'Unless I am deceived, I left my keys in my car' (Mele 1997b: 92). It supposedly follows that '(to) deceive' means 'to cause to believe what is (believed by the deceiver as) false' (Oxford English Dictionary). Accordingly, a deceiver, D_R, deceives the deceived, D_D, into believing that *p* only if D_R both knows (or believes, at least) that *p* is false and intentionally causes D_D

to believe that *p*. On this view, D_R would not have intended to deceive D_D into believing that *p* had D_R believed that *p* is true – some even think that this is conceptually impossible.

In turn, if self-deception can be modelled on interpersonal deception, as we traditionalists claim (and I see no reason for a traditionalist to completely deny deflationism), a successful self-deceiver – e.g., Hillary – should intentionally cause herself into believing a falsehood while actively knowing or believing that she is causing herself to believe a falsehood, which seems impossible if Hillary's mind is not compartmentalised in some relevant way. Even in interpersonal deception, D_D would not bite the hook if D_D knew D_R's intention to deceive her, and, of course, a unified-mind-Hillary must know that she intends to deceive herself. This is the so-called *Strategy (or Strategic) paradox* of self-deception (Mele 1987: 138). Furthermore, because in self-deception Hillary is both the deceiver, D_R, (who knows the truth) and the deceived, D_D, (who ends up believing a falsehood), she should at some point believe both that *p* and that ~*p*, which is also impossible if her mind is unified and reasonably coherent. Even more, the true belief is the reason why she ended up with a false belief. This is the so-called *Doxastic paradox*. These paradoxes are the main reasons you would not want to be a traditionalist (unless, perhaps, you are a paraconsistent logician or a dialetheist).

The typical next step of a traditionalist is to posit some kind of divisions and thereby resolve the paradoxes. There are two general groups of partitioning theories, one that posits psychological partitions and another that discriminates between person-stages, which is temporal partitioning of the person, and each group can be further divided. I will focus on psychological partitioning.

(1) *Psychological Partitioning theories*

a) *Hierarchical divisions*:
   i. *Hard*: the mind has 'levels' of consciousness, which operate relatively independently (e.g. Freud's conscious, unconscious self, and the censor), or involves a dissociation of mental processes (most notably, Pierre Janet, also Von Hippel and Trivers [2011]). In any case, some parts operate in a way that is inaccessible to consciousness. However, these states may still be operational in the sense of affecting processes in the consciousness in which case they hinder the mind's normal functioning.[3]
   ii. *Mild*: the mind has a limited working capacity, thus not all mental states are accessible to one's working memory. This division is not organisational and the

[3] For some Freudian accounts of self-deception, see Pataki [1997], Lockie [2003]; against, see Sartre [1978: 50–54], Gardner [1993: 40–58]. On Pierre Janet, see Van der Hart and Horst [1989], Braude [1995].

mind is not intentionally selective; this division is the result of the mind's inability to manage all data simultaneously, not of the mind's inability to cope emotionally with some data or stressful memories (e.g., Egan [2008]; Davies and Egan [2013]; similarly, Gardner [1993]).

iii. *Subtle*: the mind has different hierarchically ordered parts, which ultimately work as one and are co-conscious (e.g., Plato's tripartite soul).[4]

b) *Horizontal divisions*:

i. *Hard*: the mind is divided into subsystems or sub-personal goal-directed structures of some kind; they overlap considerably but can produce conflicting outputs, are not hierarchically ordered, and can house both conscious and unconscious mental states (Davidson [1986, 1997]; Pears [1982, 1986, 1991]; Mijović-Prelec and Prelec [2010]; Greve and Wentura [2010]; Kurzban [2011]).[5]

ii. *Mild*: the mind is divided into corpuses of beliefs, which act as storages for mental states and not separate centres of agency. That is, unlike subsystems, which can prompt actions, corpuses of beliefs are inert and are not agent-like. The agent stores beliefs in corpuses according to their association – similar beliefs or beliefs about similar topics are grouped in the same corpus (Egan [2008]; Davies and Egan [2013]; possibly, Gardner [1993]). A similar view is that memory is context-dependent (Cialdini et al. [1991]; Brown and Kendrick [1997]).

(2) *Temporal Partitioning theories* (positing logically distinct subjects): the mind at any given time is unified and divisions apply only diachronically; that is, there are different temporal stages of the same unified self (Sorensen [1984]; Bermúdez [2000]; see Braude [1995:157]).

Theories of psychological partitioning are not mutually exclusive. For example, the Mild hierarchical and the Mild horizontal divisions work nicely together and they can be combined with any Temporal Partitioning view. Nonetheless, combining moderate partitioning views cannot solve problems that traditionalists face. Those who decide to deceive themselves, according to what the standard theory of interpersonal deception implies, will at some point have to hold both beliefs in their working memory. Therefore, the mental division, it seems, must be as such that the relevant sub-system or compartment has some characteristics of an agent, which would allow it to act as a deceiver. However, that compartment would have to be capable of many things we would think only a real person is capable of; even worse, it would be able to affect the behaviour of rational agents in such a way that rational agents, as

[4] There are views that Plato defines parts as psychological subjects (Bobonich [2002], Lorenz [2006]) but this does not strike me as correct – the unity of the soul is often highlighted by Plato. Price [2009: 9], I think, explains Plato's view nicely: 'The contents of each part are described as being *automatically* open to view by the others. If appetite 'has no access' to reason's judgements about overall well-being, this is not because they are hidden from it, but because they are beyond its capacity.'

[5] Against the subsystems view, among many, see, Heil [1989, 1993], Gardner [1993: 59–76, 80].

main systems, are not in control of their behaviour. In addition, the suggested solution that one temporal stage of a person deceived a distant other stage, by hiding some information, for example, seems to fall short of self-deception according to many philosophers (e.g., McLaughlin [1988: 31–32]; Pataki [1997: 308–310]).

It seems as if only Hard Horizontal compartmentalization theories can explain self-deception on the traditional understanding. Relevant explanations say that, although we are incoherent as whole agents when we are deceiving ourselves, our semi-autonomous compartments are internally coherent and they are the ones that facilitate the deception – e.g., one subsystem houses one belief, while another its counterpart.

Hard psychological partitioning theories, horizontal or vertical, are mostly abandoned in contemporary psychology and philosophy but they are not without advocates. For example, Lockie [2003] proposes an argument in favour of positing a *dynamic unconscious*, which assumes active parts of some kind with their own motivational interests not necessarily known to other parts. These parts interact in various ways that may involve concealing, deceiving, etc. and this is apparently how we may explain self-deception. According to the argument in favour of positing a dynamic unconscious, we need to ask ourselves three questions. The first is 'Do I use some version of this notion to help make sense of myself and others?' and the answer is 'Yes.' The second question is 'Does academic psychology, as presently conceived, make space for the notion of a dynamic unconscious?' and the answer is 'No.' The third question, then, is actually a dilemma 'If (1) and (2) don't match up, who is right and who is wrong?' and, the argument is, this dilemma should make us think that mainstream psychology is wrong on this point (Lockie 2003: 129).

The problem with this argument is that it is based on a false dilemma. If (1) and (2) don't match up, that does not entail that one of them is false. (1) refers to our everyday practices, which by no means have to be correct or even precise in order to be successfully applied. In our conversation, we say that the sun came up to make sense of our experience of the sunrise, and astronomers would say the same thing as well, yet expressing ourselves regularly in this way to make sense of our experience has no relevance for our scientific theories. In a similar way, we may speak about our unconscious beliefs without theoretical commitments to the existence of 'the unconscious.'

Another reason to posit divisions in the mind is the traditional account of self-deception. In this thesis, however, I will argue that one can be a traditionalist about self-deception and understand it as an action in which a person intentionally deceives herself without positing *any* divisions of the mind. In fact, that claim will not depend on any theory of the mind and it will not have the need to appeal to factors such as second-order mental states, errors of self-knowledge, and similar. However, it will not exclude these factors as factors that play specific – yet minor – roles in irrationality and it will allow that some other cases of irrationality do demand some kind of divisions in the mind. That is, I will not defend holism of the mental and argue that the mind is perfectly coherent; rather, I will argue only that the successfulness of self-deception does not depend on any of the factors listed above and that thus we do not need them to explain self-deception.

My lying or deceiving myself is not much more problematic than my lying or deceiving you. Therefore, to resolve paradoxes of self-deception and to rehabilitate traditionalism is nothing other than to propose better theories of lying and interpersonal deception. In the next section, I will briefly summarise my solution. The complete novel theories of interpersonal deception and lying on which my solution depends will be proposed in chapters II-1 and IV respectively.

1.2.2 Traditionalism about Self-deception, My Version

When I started analysing self-deception, the phenomenon seemed so mysterious, interesting, and insufficiently analysed. However, when I got a bit deeper into this matter, I realised that philosophers have offered an abundance of interesting and innovative theories of self-deception; one might easily conclude that nothing new could be said about self-deception.

I partly agree with this: nothing substantially new could be said about self-deception from within the current theoretical framework; however, there is much to be said about the theoretical framework used to explain self-deception. The best – simplest and most effective – strategy of understanding self-deception rests on developing satisfactory theories of lying and interpersonal deception that will not generate problems when applied to self-deception. With these theories in place, understanding self-deception will become much easier.

In other words, I say that the real solution to the traditional paradoxes of self-deception should not be based on resolving the paradoxes themselves; rather, the real solution is in

making sure those paradoxes do not arise in the first place. In a sense, paradoxes of self-deception can be seen as a *reductio ad absurdum* of dominant theories of lying and interpersonal deception. What *causes* the paradoxes is our incorrect understanding of lying and interpersonal deception and not some feature intrinsic and exclusive to self-deception as a phenomenon. Preventing paradoxes of self-deception from emerging is what I consider to be the new contribution I can make against the background of established efforts, since received accounts take theories of lying and interpersonal deception as given and correct. Here is a short illustration of my point.

The Doxastic paradox arises because 'self-deception is like lying; there is intentional behaviour which aims to produce a belief the agent does not, when he institutes the behaviour, share' (Davidson 1986: 207). Therefore, a self-deceiver should, at least for a moment, believe both that $\sim p$, as a liar, and that p, as the one being lied to. The Strategy paradox follows from this idea combined with the view that a liar must intend to deceive: if I intend to deceive you into believing what I believe to be false, I must hide my intention from you. However, this precise form of deceit cannot be practised on oneself, since 'it would require doing something with the intention that that very intention should not be recognized by the intender' (Davidson 1986: 208). In turn, if intentions require reasons to act, people may not be able to form the intention to φ if they think that φ-ing is impossible, and thus even forming the intention to deceive myself turns out to be impossible.

But, would it not be nice if liars do not have to assert what they believe to be false? If I could lie to myself by asserting what I believe to be true (but which is, in fact, false), then the doxastic paradox would not arise. Moreover, I might even succeed in making myself more confident in what I believe to be true even though I know that I am lying to myself – and making myself thus epistemically worse off rightly counts as deception. Finally, if liars need not intend to deceive, then some people are mistakenly taken to be self-deceivers; they may be self-liars who do not intend to deceive themselves. This proposal nicely captures the Nicole case nicely: she does appear to be lying to herself and – for the reasons we have discussed – it makes perfect sense of her not to intend to make herself believe her own lie.

These are some of the ways in which I seek to resolve the paradoxes and fully rehabilitate traditionalism. However, I do not claim that all cases of self-deception need to be modelled on lying and interpersonal deception. My aim is just to show that traditionalism should be preferred.

2. Dialectics, Methodology, and Structure

In chapter II, I will propose and defend a novel account of deception; I call it the manipulativist view. On this view, which generates no paradoxes of self-deception, the essence of deception is manipulating the epistemic processes of the one who is deceived, not – necessarily – making the deceived person epistemically worse off (by 'epistemic processes' I mean 'processes of thought formation and revision'). In the literature review, chapter III, I will engage with other solutions in a way that might be seen as superficial: I will highlight some flaws that could be resolved by introducing some additional theoretical apparatus. However, I appeal to caution here: I will not be attacking straw man arguments in order to show that my position should be favoured. My argument does not depend on the thesis that other explanations of self-deception are incorrect. Allow me to explain.

My main negative argument is that other accounts of self-deception either (1) do not fit well with our intuitions about the phenomenon and our folk-psychological conceptual analysis, or (2) they need to posit some 'strange' forces in our minds, or (3) they fall into certain 'traps,' such relying on *ad hoc* hypotheses, being pseudo-scientific, etc., or they even exhibit more than one of these three kinds of flaws. I acknowledge that these problems need not amount to decisive objections to existing theories. Quite plausibly, some of our intuitions about self-deception should be set aside or revised (and I too argue that some of our intuitions, mainly those relevant to lying and deception or to cases such as Nicole, should be revised); or, again, it may be that we do need to understand the mind as somehow divided. Nevertheless, I think the fact that these flaws arise with other, non-traditionalist, accounts of self-deception does somewhat reduce the appeal of those theories.

More importantly, I will give reasons for thinking that the account I will propose raises no such concerns and that therefore it should be preferred over its rivals. For example, it is consistent with our basic intuition that people can lie to themselves or deceive themselves and with our relevant linguistic practices (we could say 'Stop lying to yourself, Nicole'); it is much simpler than any other theory of self-deception; it does not rest on any specific conception of the mind; it does not require additional *ad hoc* hypotheses; and, finally, it can non-problematically explain *all* cases of self-deception. In fact, the only problem may be that my account proposes too many options for explaining a self-deceiver's behaviour in a very simple manner. A theory this successful may prevail while allowing that other theories are not altogether bad.

In short, my aim is that, even if all other theories are defensible – and many of them seem convincing but are so complicated that only brilliant philosophers could manage to deceive themselves in those ways – the reader will judge that the account I propose should be preferred as the most successful one because it can explain *all* the data and it can do that in the *simplest* way, while being faithful to our intuitions. What is my proposed overall account? It is my task in this thesis to develop it, so let me now outline the structure of my discussion. My preferred explanations of each of the four cases with which I began will thus have to wait – I apologise for the suspense!

In subchapter II-1, I will defend my already mentioned claim that deception is basically a manipulation of the thought-forming processes of the one who is deceived. Then, I will present various possible strategies of self-deception (II-2). This will be beneficial for our discussion because we will see that they all share a common feature: they all involve a certain kind of epistemic self-manipulation, but not all of them involve the person ending up epistemically worse off. I conclude by proposing a general manipulativist theory of self-deception.

Chapter III involves a thorough literature review: III-1 discusses traditionalism, III-2 deflationism, and III-3 revisionism. I highlight problems each of these theories face and show that most of them can explain only a limited class of the types of phenomena we may refer to as self-deception. The manipulativist view makes the garden variety cases of self-deception non-paradoxical, but some problems still remain. As I argue, Nicole, in our third example, seems to be lying to herself; but this is impossible on the traditional account of lying, since it says that liars must intend to deceive and assert what they believe to be false and you cannot intend to make yourself believe as true what you already believe as false. Therefore, in chapter IV, I revise the traditional theory of lying by claiming that liars need not intend to deceive nor assert what they believe is false. On this view, Nicole can lie to herself as much as she wants. In chapter V, I provide simple and compelling explanations of each of the four cases introduced above and of other cases introduced in the course of the discussion.

II. SOLVING THE PARADOXES BEFORE THEY ARISE

1. The Main Idea

In this thesis, I defend traditionalism about self-deception (the view that self-deception is modelled on interpersonal deception) but argue against some standard views of lying and interpersonal deception. The reason for this approach is simple: Unless the mind is mentally partitioned, modelling self-deception on interpersonal deception as standardly understood generates paradoxes. One paradox arises from the idea that the deceiver (D_R henceforth) must intend to make the deceived (D_D) believe what D_R believes to be false. Because self-deceivers are both D_R and D_D, they should (at least briefly) simultaneously believe a clear contradiction – as D_R they should believe that *p* and as D_D that ~*p* – which is paradoxical. Another paradox arises from the intuition that people have a standing intention not to be deceived. It follows that self-deceivers should somehow be capable of hiding their intention from themselves, which is again paradoxical. Because of these paradoxes, philosophers have either abandoned modelling self-deception on interpersonal deception (among many, Mele [1997, 2010]; Audi [1982, 1989]; Barnes [1997]; Patten [2003]; Fernández [2013]; Lynch [2012]) or accepted some conception of the mind as mentally partitioned (e.g., Pears [1982]; Davidson [1986]; Pataki [1997]; Lockie [2003]; Mijović-Prelec and Prelec [2010]).

I believe that these paradoxes result from our incorrect theories of deception, all of which require that D_D ends up epistemically worse off or is prevented from becoming as epistemically well off as they could have been. In contrast, I will provide cases of interpersonal deception in which D_D ends up epistemically better off. In triple bluffs, for instance, D_R deceives D_D into believing a true proposition believed by D_R to be true. Received accounts of deception cannot explain these cases and thus should be abandoned.

II-1. Deception

1. Introduction

The main aim of this chapter is producing a theory of deception that successfully explains all cases of deception, including self-deception when modelled on interpersonal deception. I will argue that deception should be understood as involving deceiver's (D_R) manipulation of the deceived's (D_D) agential use of his or her own cognitive capacities. In interpersonal deception, this manipulation is for the purpose of causing the formation or retention of certain truth-evaluable mental states of D_D, whereas, in non-human deception, the purpose is generating certain behavioural responses in D_D. On this 'manipulativist' view, deception need not involve misinforming D_D and can even involve informing D_D, and be performed for D_D's own good and at the expense of D_R.

When applied to self-deception, manipulativism allows that self-deceivers may intentionally manipulate themselves in order to form or retain a certain truth-evaluable mental state while thinking that this manipulation is, in some way, for their own good. No paradoxes arise on this account: people do not have a standing intention not to be manipulated for their own good.

In what follows, I will first present received accounts of deception starting with the most restrictive and ending with the least restrictive. I will divide this discussion into two parts. In the first (2.1), I present the view of deception taken by the philosophers who analyse self-deception and in the second (2.2), I will present theories of deception proposed by scholars who specialise in deception – the reason is that the latter propose less restrictive accounts of deception. Then, I will present the manipulativist theory of deception and introduce cases which only this theory can explain correctly. These cases will show that even the least restrictive received theory of deception leaves some cases of deception unexplained. I thus make the case that you should be a manipulativist about deception.

2. Received Theories of Deception

2.1. Deception as Understood by Those Who Analyse Self-Deception

Many philosophers understand self-deception in terms of the self-deceived person ending up with an epistemically unjustified belief as a result of her motivated epistemic negligence

(e.g., Bayne and Fernández 2009). In short, you are deceived because you ended up epistemically worse off, and you are *self*-deceived because it is your fault. However, there are many ways in which one may understand self-deceivers' ending up epistemically worse off, i.e., 'deceived.' In this section, I present the understanding of 'deception' by scholars who, in analysing self-deception, focus on intentional deception and take the lexical approach in defining how the terms 'deceive' and 'deceived' should be understood. In the next section, I inspect more moderate approaches to deception and I discuss deception in general, not just intentional deception.

In the lexical approach, we start with a definition of a term, using the dictionary or common usage as a guide. In line with this, Mele [1997: 92] most notably writes that, in a standard use, '(being) deceived' is usually taken to be equivalent to '(being) *mistaken*.' That is, I am deceived only if what I believe is false (e.g., Deweese-Boyd [2016]; Raikka [2007: 522]; below), as, for example, in the sentence 'Unless I am deceived, I left my keys in my car.' Accordingly, 'to deceive' is 'to cause to believe what is false.'[6] However, this popular approach conflates the *verb* '(to) deceive' and the *adjective* 'deceived.' And it is this conflation, together with a specific interpretation of the adjective as 'believing falsely,' that generates a popular view according to which *a deceiver intentionally causes someone to believe something that is not true* (Oxford English Dictionary [2017]: 'deceive'; Carson [2010: 47–49]).

I call this approach to interpersonal deception the *hard-line approach* because those who take it hold that B is not deceived if the belief B ended up with is not false. Many prominent philosophers find this position not only appealing but undeniable. In fact, the hard-line thesis is a necessary ingredient of most theories of self-deception. Specifically, those according to which 'what makes a state self-deceptive … is not its etiology, but its object' (Patten [2003: 231]; similarly, Audi [1989: 249]) and those that take it that self-deception just is motivated forming and maintaining false beliefs about oneself, which is a common view in psychology (e.g., Gur and Sackeim [1979]; Von Hippel and Trivers [2011: 3–5]). If the relevant belief is not false, a proponent of these views will say that the agent is not self-deceived.

The hard-liners seem to be reading too much into deception. Fallis [2015: 388] argues that hard-liners have failed to tell us *why* there is deception only when the deceived ends up with

[6] In self-deception, this false belief can also be a (second-order) belief about my mental states (e.g., Holton [2001]; Patten [2003]; Funkhouser [2005]; Fernández [2013]).

a false belief (see Mahon 2008). Moreover, they commit a fallacy of equivocation; they conflate two different senses of the adjective 'deceived.' The first sense is 'having a false belief' and the second sense is the sense of being '*tricked into* believing something' – that *p* is false is at most an implicature, not an entailment, in the latter case.

We have good reasons for holding that the question whether *p* is true or false is tangential to the question whether you deceived me into thinking that it is true. For one, this is consistent with our linguistic practices (and I will present in Section 3.2 examples of interpersonal deception that do not involve ending up with false beliefs). Being deceived in the sense of having a false belief or in having your purpose (etc.) frustrated, as in 'Never was expectation more completely deceived,' is the *figurative* (3b) sense of 'deceived' (Oxford English Dictionary). And since it relies on the figurative meaning of being deceived, 'to cause to believe what is false' cannot be the primary meaning of 'deceive.' Philosophers who identify being deceived with having a false belief are incorrectly taking the metaphorical meaning to be the primary meaning of the adjective.[7]

It would be natural to soften the hard-line position by retreating to the claim that deception requires that the belief produced in the deceived is at least *believed* to be false by the deceiver. Most philosophers – Van Leeuwen [2013: 4753], for example – understand this conviction as a 'truth about deception.' This is a 'truth' because it is thought that I cannot form the intention to deceive you otherwise if this condition is not satisfied. However, this less restrictive view still gives rise to a paradoxical theory of self-deception: I made myself believe my own lie. The paradoxicality arises from the conviction that D_R must *believe* that the belief they intend to make D_D believe is false (Carson 2010: 51), not from *p* actually being false.

When applied to self-deception, this view entails that the same person (i) successfully intentionally deceives herself into believing as true (ii) something she already believes to be false, which gives a clearly paradoxical theory of self-deception. Without dividing the mind, we cannot explain how a single person can be taken in by her effort to deceive herself into believing something she already believes is false. This person should, by an act of apparent

[7] In fact, 'being deceived' used solely in the sense of having a false belief should be interpreted as a personification: by saying that I am deceived about the whereabouts of my keys, I am personifying the process that generates my false belief by describing it as if it were the action of another agent who has tricked me into believing what I believe.[7] Accordingly, 'Unless I am deceived, I left my keys in my car,' actually communicates something like 'Unless *my memory* deceives me, I left my keys in my car.' The implicature is that the possibly false belief is not a result of my epistemic negligence.

miracle, abandon the belief she takes to be true in favour of the one she takes to be false – this is the so-called Doxastic paradox. We need a miracle not only because one cannot believe inconsistent beliefs but also because I will know that I want to make myself believe what I already believe to be false and thus I should not bite the hook. Therefore, if self-deception is modelled on interpersonal deception, self-deceivers must hide their own intention from themselves – the so-called Strategy paradox of self-deception.

Delivering a very problematic conception of self-deception is not the main flaw of the hard-line view; rather, it is itself a very narrow understanding of deception, as we will see in the next section. It is a surprise that it is so popular among those who analyse self-deception.

2.2. Deception as Understood by Those Who Analyse Deception

According to what I will call the *epistemic conception* of deception, I am a victim of deception only if that action has made me epistemically worse off. The phrase 'ending up epistemically worse off' can be understood in various ways. One way is to understand it as equal to ending up believing a falsehood, which is what hard-liners hold.

A more plausible view, I will call it the *moderate epistemic conception*, understands being epistemically worse off so that it need only amount to being misinformed on purpose. The state of being misinformed on purpose can result from communicating half-truths or from transmitting misinformative signals (receiving these signals decreases the probability of the actual world state) but it excludes failures on the part of the deceived, such as a failure to remember where you left your keys that may cause a false belief. On this view, the intentional deceiver may but need not intend to make the deceived believe what the deceiver believes to be false; rather, they may just intend to misinform the deceived.

Along these lines, McWhirter [2016], for instance, argued that a message can be deceptive even when true (or half-true) if it is misused, i.e., if it is used in a different way than normal. McWhirter's position is subtle because, while misuse does not entail that the message itself carries false information, misused messages may leave D_D misinformed – ending up misinformed may be the result of D_D interpreting the message according to how it is normally used by the relevant population or how it was used by D_R in the past. For example, by saying that Ronaldo is a football player, I may misinform a person from the USA or Canada that Ronaldo plays what Europeans call 'American football.' What is deceptive here is not the message itself but the way it was used in the given context. Alternatively, if I have always used the term 'soccer player' in communication with my friends, my sudden undisclosed shift

to ‘football player’ when referring to Ronaldo may misinform them that I am talking about an ‘American football player.’

Nevertheless, developing this proposal into a general theory of deception as a matter of being made epistemically worse off is somewhat problematic. In order to account for animal deception (where we cannot appeal to intentions), the moderate epistemic conception must posit that deception should bring some detriment to D_D and/or that D_R benefits from deception (Searcy and Nowicki [2005: 5]; Skyrms [2010]; Wilson and Angilletta [2015: 206–207]; Fallis [2015]; McWhirter [2016: 717]; Shea, Godfrey-Smith, Cao [2017: 18]). These additions, however, result in the view being too narrow.

Cases in which D_D does not suffer any harm are common even in the animal world – many cases of symbiosis that involve deception do not bring harm to D_D, they just provide little or no reward (e.g., Bond and Robinson [1988: 298]; Waterman and Bidartondo [2008: 1086, 1091]) – and there are plenty of cases in which human deceivers knowingly suffer harm from their deception. In *Pretty Village, Pretty Flame* (Cobra Films, 1996) anti-war movie, a petty criminal and a womaniser, but nonetheless a great guy, named Velja pretends to be his brother Milos, a well-mannered archaeology student, in front of the military police who came to forcefully take Milos to the front. By intentionally making them believe that he is Milos, Velja got willingly conscripted as Milos and taken to the frontlines, where he was killed. In this case, the deceiver is the one who willingly suffers detriment while those who got deceived suffer no harm at all.[8] A satisfactory definition of deception should be able to account for these possibilities.

Fallis and Lewis [2017: 5] have recently replied to this objection by arguing that ‘the sender’s *genes* … benefit from misleading the receiver’ and that therefore the *sender benefit* condition should remain. This reply works well against Artiga and Paternotte’s [2017] hypothetical cases of animal deception but not against various plausible variations of the Velja case. We may simply imagine that the military police came to pick up the son of Velja’s neighbour and that Velja decides to save the boy from the horrors of war. We may also imagine that the military police came to pick up a guy Velja never saw, that Velja was drunk, and that he decided to pretend to be that guy thinking in his drunken trance that it would be awesome to go to war. Examples in which the sender’s genes do not benefit from

[8] See also white lies and benevolent lies in Erat and Gneezy [2012] or Artiga and Paternotte [2017: 10–11, n. 10].

deception could be found in the animal world as well: suppose that a dog, Bingo, (for no benefit) befriends a cat, Bongo, and that, one day, in order to save Bongo from another, much bigger, aggressive dog, Bingo barks aggressively for the purpose of scaring and saving Bongo. In this way, Bingo misinforms Bongo that he is the predator but he saves Bongo from the real predator and endangers himself – since the predator is a much bigger, aggressive dog.

Recently, Artiga and Paternotte [2017] have offered another general account of deception that intelligently avoids reference to harm/benefit. They call it the *functionalist definition of deception*, which says that M is a deceptive state iff [2017: 13]:

1) M has the (etiological) function of causing a misinformative state (or failing to acquire a particular piece of information)[9]
2) M leads to a misinformative state.

According to the functionalist definition, then, D_D is a victim of deception if and only if D_D was misinformed by M, which was M's function as specified by conditions 1 and 2, whereas D_R is a deceiver if and only if D_R generated M for this purpose. The functionalist approach has many virtues. For example, it neither requires any harm (apart from epistemic) to befall D_D nor that D_R benefits from the deception; also, M may be misinformative simply by deviating from normal practice rather than containing some false information. Because it requires neither that D_R benefits from the deception and D_D is harmed, nor that misinforming is systematic (*pace*, e.g., Skyrms 2010), this account can explain all possible cases of animal deception, a feature other general accounts of deception do not have. This feature reinforces the functionalist view.

However, both the moderate epistemic conception and the functionalist definition are problematic when applied to *intra*personal deception and so they do not help with one part of my project, namely, demystifying self-deception. They both entail that self-deceivers intend to make themselves epistemically worse off, by misusing a message in a way in which they themselves will misinterpret it or by misinforming themselves in some other way, and thus both paradoxes of self-deception re-emerge: you cannot misinform yourself.

[9] The bracketed condition entails that D_D was merely prevented from becoming as epistemically well off as they could have been. However, the authors themselves claim that (italics added) '*the central* idea of the functional account of deception is that deceptive states have the function of producing *misinformation*' (Artiga and Paternotte 2017: 13), which is making D_D epistemically worse off; hence, I classify their view as in-between the moderate and the weak epistemic conception (which I present below).

Finally, there is what I call the *weak epistemic conception* of deception, which only requires that D_D was prevented from becoming as epistemically well off as they could have been (e.g., Chisholm and Feehan [1977]; Lynch M. [2009: 190]; Lackey [2013]; Fallis [2015]). On this view, D_D was epistemically harmed not by being made to end up having epistemically less than what they had but rather by being prevented from becoming as epistemically well off as they could have been. As an example of this view, I will briefly present Chisholm and Feehan's (C&F henceforth) seminal theory of intentional deception.

According to C&F [1977], I can deceive you by acting intentionally, which is an act of *commission*, or by intentionally refraining from acting, an act of *omission*. For instance, I may give you a piece of false information or refrain from giving you a vital piece of true information. There are four acts of commission and four acts of omission [1977: 144–145], which can cause that you (1) end up with a brand new false belief, (2) continue to have an existing false belief, or (3) fail to acquire a true belief. When the deception contributes causally to acquiring or continuing to believe the false belief that *p*, they call it *positive*; when it contributes causally to D_D ceasing to believe the true belief that ~*p* or preventing D_D from acquiring the true belief that ~*p*, they call it *negative*. Finally, deception is *simpliciter* when it actually causes a change (positive or negative) in belief and it is *secundum quid* if it sustains a state of belief that ought to be changed.

Some philosophers – e.g., Lynch M. [2009: 191], Lackey [2013], and Fallis [2015: 389] – agree with C&F that allowing someone to continue without the true belief that ~*p* or preventing them from acquiring it (all cases of negative deception *secundum quid*), may count as deception. Nonetheless, the majority of philosophers – e.g., Mahon [2007: 187–188], Carson [2010: 56], and Lynch K. [2016: 519, n. 15, 520] – find this to be a mistaken idea. In fact, most philosophers think that preventing someone from becoming epistemically better off is not deception and they would not count C&F's situations 2 (left with a false belief) and 3 (failed to acquire a true belief) as involving deception.

Carson, for example, calls an action that prevents you from acquiring a true belief 'keeping in the dark,' which he distinguishes from deceiving. He [2010: 54] writes that, by keeping you in the dark, 'I have prevented you from gaining certain information. However, I have not caused you to acquire or retain any false beliefs [a victim of keeping in the dark may even be left without a belief], and, therefore, I have not deceived you about anything.' Even though this argument assumes the (arguably too strong) hard-line thesis that deception must

produce false belief, the general idea is good: victims of keeping in the dark need not be misinformed, they may just be insufficiently informed and this does not seem to be sufficient for deception.

Here I just note that, when applied to self-deception, the weak epistemic conception is also problematic. Since self-deceivers are both deceivers and those who are deceived it follows that, as deceivers, they already are epistemically better off. Accordingly, self-deception is again paradoxical if modelled on interpersonal deception: it requires that self-deceivers make themselves forget some specific piece of information.

This concludes the first part of my analysis. I presented various theories of deception starting with the most restrictive one, the hard-line epistemic view, and ending with the least restrictive one, the weak epistemic view. In Section 3.1, I will present the manipulativist account of deception – given my background goal of demystifying self-deception, the account will be focused on intentional (interpersonal) deception. Following that, I start my argument, which will unfold as follows.

In Section 3.2, I introduce what I argue are obvious instances of interpersonal intentional deception in which the victim ends up epistemically better off. These cases will show that all the theories discussed above are too narrow; they offer support for the view that the essence of deception is a certain kind of manipulation of the exercise of cognitive capacities. In Section 3.3, I will argue that the theory I propose can explain deception across the whole biological world. Therefore, I will conclude that the right account of deception is a manipulativist one.

Furthermore, because the view that sees deception principally as manipulation of D_D's exercise of cognitive capacities does not require that D_D ends epistemically worse off or suffers any harm – D_D can even benefit from deception – it may non-problematically be applied to self-deception: I may allow myself to deceive (i.e., manipulate) myself if I think that this is for my own benefit.

3. The Manipulativist View

3.1. Presenting the View

I propose *the manipulativist conception* of deception, *the manipulativist view*, which sees manipulation (by way of a trick) of the exercise of cognitive capacities as the distinguishing

feature of deception. I will offer two versions of the view, a definition of intentional deception and a definition of deception in general. Then, I proceed to clarify some important points. This is a substantially novel proposal so I ask for patience; I will discuss many concerns that may arise in sections 3.2 and 3.3. According to the manipulativist view, D_R intentionally deceived D_D by way of φ-ing, if and only if

1) D_R φ-ed intending to thereby generate a specific φ-relevant response on the part of D_D,
2) D_R's φ-ing causally contributes towards a non-deviant manipulation of D_D's agential (or agent-like) use of D_D's own cognitive capacities, and
3) Because of 2, D_D responded autonomously to φ-ing in a relevant way.

The first thing to notice is that this proposal does not require D_R to benefit from deception or D_D to suffer a detriment. Secondly, having the intended result defined as 'a φ-relevant response' allows that human agents may deceive non-human agent-like beings (beings capable of producing behaviour describable in terms of them exercising their own agential autonomy) as well as human agents. When D_D is a human, generating the φ-relevant response is limited to affecting D_D's truth-evaluable mental states, such as beliefs, acceptances, assessments, or judgements. In particular, D_R typically intends to causally contribute to D_D forming or retaining a certain thought or to affect the degree of confidence with which D_D accepts a proposition (see 4). For instance, when, by playing dead, I make an enemy soldier judge or form a belief that I am dead or takes this to be sufficiently probably not to bother to check whether I really am dead, I deceive that soldier on this view. When D_D is not a human, the φ-relevant response would involve some relevant complex behaviour that results from D_D's exercising its cognitive capacities. For example, when I deter a bear by simulating death, this is also deception.

When cashed out in functionalist terms, the manipulativist definition accommodates deception performed by a non-human agent or agent-like being, which turns it into a definition of deception in general. According to the general manipulativist account, D_R deceived D_D if and only if

1. D_R utilised an evolutionary adaptation φ (a trait, behaviour, a state, etc.),
2. The function of φ is generating a specific φ-relevant response on part of D_D by means of causally contributing towards a non-deviant manipulation of D_D's agential (or agent-like) use of its own cognitive capacities,
3. Because of 2, D_D responded autonomously to φ in a relevant way.

Under this general definition, a non-human agent or agent-like being may be D_R because deceivers need not act intentionally; rather, they utilise an evolutionary adaptation. And D_D

can be a non-human agent or agent-like being because the φ-relevant response that makes D_D a victim of deception may be purely behavioural. Accordingly, when the Western hog-nosed snake deters predators by simulating death (Beane et al. [2014: 171, 174]; also Platt [1969]), this is deception because deterring predators is the function of the snake's evolutionary adaptation while the deceived predators may but need not form the thought that the snake is dead; rather, they may just instinctively and automatically respond to the cognitive input. Hence, the snake may equally deceive me or a colony of fire ants – in both cases, the φ-relevant response has the effect of leaving the snake unharmed. Likewise, when an insect tricks predators into by-passing it by looking like a leaf (i.e., by utilising a trait), this is a deception for the same reasons. Incidentally, being capable of intentional deception may also be considered as an evolutionary adaptation; thus, intentionally deceiving by φ-ing should be covered too, making this a fully general definition.

I now proceed to make some important clarifications. I clarify conditions 2 and 3 in turn. I start by providing a rough account of manipulation.

3.2. Condition 2

I understand 'manipulation' in a semi-technical sense developed from definitions provided by Oxford English Dictionary and Merriam-Webster's collegiate dictionary.

MANIPULATION

> Managing or directing the victim in a skilful manner, i.e., by way of a (concealed) *trick* or (concealed) *interference*, tampering. The exercise of subtle, secret, oblique, or indirect influence or control over the victim by cunning or sneaky means.

Not every kind of manipulation is relevant for deception. The relevant manipulation needs to be *non-deviant* and it must target D_D's *agential (or agent-like) use of* D_D's own cognitive capacities, not those capacities themselves. There are four important characteristics of manipulation that I want to discuss here: (1) it must be non-deviant, (2) the victim must be an agent or at least sufficiently agent-like, (3) the manipulation targets D_D's agential (agent-like) use of its mechanisms rather than those mechanisms themselves, and (4) this use may be specific to the agent; that is, D_D may have its own specific way of responding to specific stimuli and the manipulation targets that.[10] I proceed to explain each of these characteristics.

[10] The fifth characteristic is that it must be *concealed* from the target. I discuss this in Section 4 of this subchapter.

The manipulation needs to be non-deviant in the sense that it does not involve a violation of D_D's agential or cognitive autonomy: D_R will not be a deceiver, for instance, if D_R drugs D_D in order to make him see pink elephants (Fuller [1976: 23]; Mahon [2007: 185; 2008]). Likewise, deterring a bear by spraying gas in its face or inserting paralyzing venom into your victim is not deception, although it manipulates the exercise of the victim's cognitive processes or mechanisms. In these cases, the victim's relevant cognitive processes are impaired or disabled.

In addition, the relevant kind of manipulation characteristic of deception is of D_D's *agential* (*agent-like*) *use of* D_D's own cognitive capacities, not of those capacities (i.e., cognitive processes or mechanisms) themselves. When, by playing dead, I deceive enemy soldiers into thinking that I am dead, this is in virtue of the fact that I manipulate the way in which the soldiers *interpret* what they perceive. I 'push them' towards making this specific judgement; I do not manipulate their perception and create an illusion. Therefore, creating an illusion, such as in a magic trick, will not count as involving deception on the manipulativist view, even though there is no 'deviant' impairment of the audience's cognitive capacities, in the sense described above. Perhaps any kind of manipulation can be broadly understood as deceptive, but being deceived cannot be understood as merely being manipulated into generating any relevant response. Rather, the response must be as such that, had there not been the trick, I could have responded differently by exercising my autonomy. Consider this example.

Some studies suggest that myco-heterotrophic orchids have the ability to 'to recruit and manipulate otherwise free-living fungi into a mycorrhizal relationship [a symbiotic relationship between fungi and plants]' (Waterman and Bidartondo 2008: 1091). There is a sense in which we may say that orchids are successfully engaged in deception, which is how biologists describe this adaptation, but it seems hard to say that fungi are deceived by this adaptation even though they do respond in a relevant way. On my manipulativist account of deception, the φ-relevant response must belong to that kind of responses that minimally involve agent-like control and, therefore, to be deceived, D_D must be capable, at least in principle, of autonomously responding otherwise in exercising the relevant cognitive capacities.

The deceived subject is described as an agent-like being rather than an agent because insects or fish can be deceived, and it is not clear whether they are agents. For instance, when

a firefly perceives a flash, its response seems to be automatic (nothing like considering reasons is going on). However, this is not to say that the response is completely causally determined by the perception of the flash in a way a mechanical process is determined by the press of a button. The firefly, at least in principle, could not descend once perceiving a mating signal and, in this sense, the firefly, even if not an agent, is sufficiently agent-like to count as deceived if he descends. Agent-like beings have dispositions to act differently and, if a being lacks dispositions to act differently to φ, it cannot be deceived in the main sense of the term; rather, it can be deceived only in a metaphorical sense, in which a skilful liar deceives the polygraph – here, 'deceives' stands for 'makes his lie go undetected.' For the sake of simplicity, I will refer to agent-like beings as agents.

The third characteristic, the manipulation is of the agent's *use* of its own cognitive capacities, may also seem problematic (for the sake of simplicity, I will refer to agent-like beings as 'agents'). There is a (weak epistemic) sense in which a magic trick is deceptive – the audience does not know how the trick was performed (they are kept in the dark with respect to this information). Even so, this kind of ignorance should not count as deception either: members of the audience know that the assistant was not sawn in half and they know that they do not know what the trick is. Now, contrary to my position, and in line with Lackey [2013: 241–242], Fallis [2014: 87; see 2015: 390] writes that

> A magician is being deceptive even if he simply conceals from the audience how the trick was done. He does not have to create a false belief in the audience that he has actually sawn his assistant in half. Moreover, it is not immediately clear why such 'negative deception' [leaving without a belief] should not count as deception. It involves the same sort of manipulation of someone's epistemic state as does 'positive deception' [adding a false belief]. Why should it matter so much that the suboptimal epistemic state that she ends up in is ignorance rather than false belief?

The idea here is that, while the audience members know that they are victims of a trick, they do not know what the trick is. And because they would have been made epistemically better off had the trick not been concealed, this is deception on the weak epistemic conception. I, however, argue that the magician was deceptive only in a broad, figurative, sense; namely, he has deceived 'their eyes,' not the audience members themselves. They were deceived in the same sense in which my eyes 'deceive' me in the Müller-Lyer illusion – I know that the lines are of the same length but I still see some of them as longer.

By concealing the vital part of the trick, the magician has directly manipulated the audience members' cognitive mechanisms thereby bypassing these people as agents.

Therefore, he did not manipulate the audience members themselves in the way characteristic of deception. A soldier can decide not to trust his eyes and check whether I really am dead, a firefly has a disposition not to descend after perceiving a flash, but the magician's audience does not have that option. Even after being told what the trick is, many will still see that the assistant was sawn in half. Again, we cannot say that the victim is deceived if she cannot react otherwise.

The exercise of agential control, then, is vital for being a victim of deception. Suppose that I am addicted to smoking and that I make a solid decision to quit. Aiming to undermine my effort, you keep finding ways to make me think about smoking. These thoughts now enter my reasoning and, due to my rapidly increasing desire to have a cigarette, I capitulate and surrender to my petty urges and, as a result, change my former decision. This should not count as deception because addiction or mental health problems obstruct agential control. My addiction significantly contributed to reaching my decision to go back to smoking.

3.2.1. 'Personal use' of One's Own Cognitive Capacities

The final, fourth characteristic of the relevant kind of manipulation is that agential control should be understood as *personal use*. Deceptive manipulations do not causally contribute to D_D violating some general norms; rather, they target D_D's *personal* use of his or her own cognitive capacities. Allow me to introduce this point with an oversimplified example of interpersonal deception.

> Suppose that I think that 2 plus 2 is 5 and that you use your knowledge of this to make me believe that my wife will meet me at 5 p.m. rather than 4 p.m. You say that my wife said that she is available two hours after her 2 p.m. meeting, which is, in fact, true. In this way, once I discover I came to the date one hour late and disappoint my wife, you can blame me for the error.

On the manipulativist view, you have deceived me into believing that we will meet at 5 by using my bad mathematical skills – i.e., you have directed me towards a false conclusion by letting me go on reasoning badly. It is true, furthermore, that by letting me reason badly from your true assertion, I have ended up epistemically worse off, but the manipulation (which *is* the deception) is not constituted by *that* fact, but rather by the fact that you used my poor reasoning to get the intended result. In the same way, that is, by manipulating my poor probabilistic reasoning, you could deceive me into thinking that I should invest money in your financial scheme (you would let me go on reasoning badly from true premises), or that I should place a bet, which is what schemers and bookmakers do all the time.

That deception targets and manipulates the deceived's 'way of thinking,' or, in non-human deception, her 'way of autonomously responding to stimuli,' is an important feature of the manipulativist view. Consider the following case.

> **Superstitious Philosopher False**: I am superstitious and I believe that bad luck comes in threes but I know that you are not superstitious. Even though you are not superstitious and you know that I know that, you may still use your knowledge of this to make me believe that I will fail if I take the upcoming driving test by naming two examples of bad luck I recently had. If you do this because, let's say, you want our dad to give his car to you rather than me and not because you really think that I will fail, this is a straightforward case of deception where the function of deception is producing a false belief.

> **Superstitious Philosopher True:** Suppose now that you *justifiably and truly believe* that I will fail a test if I take it, which I plan to do even though I have not properly prepared for it, and that, by manipulating me into having this true belief, you actually make me reschedule the test and so get a chance to prepare for it thus making me overall better off, both epistemically and otherwise. Furthermore, since I know that you are not superstitious (and you know that I know that, etc.), you will not make me believe falsely that you believe that I will fail the test because I just had two strokes of bad luck. Rather, you will just make me go from a state in which I have a false belief to a state in which I have a true belief, and you will do this by performing the same action as the one that made me acquire a false belief in place of a true belief in the *False* condition.

A manipulativist will gladly call this deception, even though you did not misinform me about anything. By contrast, because I ended up epistemically better off (I acquired a true belief in the place of a false one) and I did not acquire a false belief (I know that you are not superstitious), no epistemic conception of deception will say that *Superstitious Philosopher True* involves deception. I take this to be an incorrect verdict: you tricked me into believing something I would not have believed had it not been for your trick; therefore, this is deception. The only difference between the *True* and the *False* variation of *Superstitious Philosopher* is that in the former, I ended up with a true belief and in the latter with a false one, but this difference is irrelevant to the issue whether you deceived me into believing it and thus whether this is deception. I will come to this point again. In the next section, I explain condition 3.

3.3. 'Φ–relevant response'

I start with Skyrms's [2010: 75] famous example (see Bond and Robinson [1988: 298]; Fallis [2015: 377–378]), which I then develop into a case of human interpersonal deception.

> **Fireflies**: Fireflies use their light for sexual signalling. While flying over meadows, male fireflies flash a species-specific signal. For instance, the Photinus firefly produces a yellow-green flash whereas the Pyractomena firefly produces an amber flash. Accordingly, if a female

> Photinus on the ground gives the proper sort of answering flashes, the male Photinus descends and they mate. An exception to this practice is the behaviour of female fireflies of the genus Photuris. When one of these fireflies observes the flash of a male of the genus Photinus, she may mimic the female signals of the male's species and, if she does this, it is in order to lure him in and eat him.

In this example, there are three possible female fireflies (Photinus, Photuris, and Pyractomena) but only two possible signals (yellow-green and amber). The amber flash raises the probability of a Pyractomena firefly female to 1 and lowers the probability of Photinus or Photuris to 0. However, this is a state in which there are not enough messages to cover all possible situations (Shea, Godfrey-Smith, Cao 2017: 13) and therefore the yellow-green flash, by eliminating only females Pyractomena, raises the probability of the presence of *both* a female Photinus and a female Photuris (Skyrms 2010: 75). Thus, according to Skyrms, even when it is transmitted by the right female, the yellow-green flash produces misinformation.

Let us now consider an analogue in which the 'fireflies' are spies or undercover agents. The following example is inspired by the *Spy Hard* (Hollywood Pictures, 1996) movie. I will use it to argue that a person may be deceived into judging that *p* without actually forming the belief that *p*.[11]

> **Spy Hard**: Dick Steele is a spy who intends to make another spy, Miss Veronique Ukrinsky, reveal her position. By capturing an enemy spy, Dick Steele obtained the enemy's book of codes and signals. This is how he discovered Veronique. By mimicking a signal of an enemy spy, Dick decides to trick Veronique into revealing her position – presumably, she will signal him back assuming that the signal is sent by her fellow spy. Dick knows that he may succeed in doing this even if the signal does not make Veronique believe that she is engaging in a communication with a friendly spy. And, in fact, Veronique has been having an incredibly strong feeling all day that they have been compromised – she had a nightmare in which she was captured – and this makes her very much convinced that the signal was sent by an enemy spy. Yet, Veronique realises that her firm conviction is irrational and she knows that it is much more probable that the signal was sent by a friendly spy. Therefore, she judges that the signal comes from a friendly spy (or that *all the chances are* that the signal comes from a friendly spy) and that she should respond to it notwithstanding that she still firmly believes otherwise.

The manipulativist view non-problematically explains these two cases as deception. The female Photuris deceives the male Photinus into descending to mate because, utilising her evolutionary adaptation to transmit the relevant signal, she successfully and non-deviantly manipulates the male Photinus's agent-like use of its own cognitive capacities into generating the relevant response; she manipulates him into descending to mate. The male firefly, at least

[11] On the distinction between judging that *p* and believing that *p*, see Scanlon [1998: 25, 35] and Peacocke [1998: 90].

in principle, may exercise its agent-like capacity and not descend, but if it descends, this is because it was deceived by the female Photuris.

Likewise, by signalling, Dick Steele non-deviantly manipulated Veronique's agential use of her own cognitive capacities (he led her into making the relevant judgement in the face of the belief to the contrary) and because his manipulation causally contributed to Veronique's falsely judging that the signal comes from a friendly spy, Veronique was deceived by Dick. Importantly, Veronique exercised her agential control and decided to disobey her 'gut feeling' while forming this judgement. If Veronique had an akratic belief that the plane will crash but nevertheless remained seated during the landing, she would have exercised the same kind of agential control: she would judge against her belief and act on what she judges to be true.

The view I defend is this: *It is the successful concealed non-deviant manipulation of the deceived's agential control (i.e., use) of his, her, or its own cognitive capacities that makes deception what it is: this applies to human and non-human agents or agent-like beings equally, and it is what makes the manipulativist view generalizable across species*.

Having said that, I would like to make a point that is very much relevant for my project. If we go back to the *Fireflies* example, we may rightly say that the female Photuris is engaged in deception (and that the case is thus a case of *deception*) but I argue that we cannot rightly say that she is actually *deceiving* the male Photinus by flashing the signal the male registers as the mating signal. When this female produces a yellow-green flash, this flash for her means 'food.' For a male Photinus, however, this flash means 'mating.' The vital point is that the female does not produce this flash *because* it means 'mating' for the male Photinus but rather because it means 'food' for her. Her behaviour has a purpose and she produces the signal for that purpose but the purpose for which she produces it is 'having the food come down to me' and not 'sending the mating signal in order to get food.' Therefore, she does not seem to be a deceiver in the full sense of the word. The purpose of her behaviour and the purpose of the male's behaviour have become aligned in the course of evolution at the male Photinus's expense and to her benefit.

Applying this insight to self-deception, we get that, when some of my sub-personal mechanisms purposefully cause that I end up epistemically worse off, this is sufficient for deception and the self (i.e., myself) is the victim of this deception in the same sense in which

the male firefly is a victim of deception, but this behaviour does not involve an action of deceiving. That is, this seems to be a kind of self-*deception* that does not involve self-*deceiving*. Possibly, we could call this phenomenon 'deception *of* the self' rather than self-deception. In this way, we would discriminate cases in which self-deception *happens to us* from those in which we actually *deceive* ourselves. On this classification, then, deception of the self is an *intra*personal analogue of *other*-deception (namely, animal deception) and self-deception is an *intra*personal analogue of inter*personal* deception.

In this thesis, I will focus on cases in which self-deception is something we *do*, i.e., in which we, i.e., ourselves as main agents not our cognitive mechanisms, actively deceive ourselves, not because I think that only these cases count as self-deception but rather because a genuine traditionalist will model self-deception on interpersonal deception, not on animal-deception. The cases of self-deception modelled on interpersonal deception are the most problematic ones and if traditionalism can explain those cases, then it can also explain cases categorised here as deception *of* the self.

In the next section (4), I will introduce more cases of interpersonal deception in which D_R deceives D_D into believing a proposition that is true and believed by D_R to be true. In most of the cases I present below, D_R does not intend to harm D_D; in fact, D_R acts on the intention to help D_D and to make D_D epistemically better off. I will argue that only the manipulativist view can correctly explain these cases. What all cases of deception have in common is the trick by which the deceiver non-deviantly manipulates the deceived's agential use of their own cognitive capacities with the aim of getting the relevant response from the deceived. I maintain, therefore, you should be a manipulativist about deception in general. Then (Subsection 4.1), I will show how being a manipulativist about deception will make understanding self-deception much less problematic: we should, then, be manipulativists about self-deception too.

4. Here's Why You Should be a Manipulativist about Deception

My first case is a case of double bluffing, most famously analysed by St. Augustine [1952/385: 57] (see Fallis 2011: 362–363; forthcoming). I will modify this example so that the double bluffer is Paul the Apostle, the victim is a Christian running from prosecutors, and the time of the event is the morning after Paul's epiphany of Christ (when his name was

Saul), an event in which a fierce persecutor of Christians converted into a zealous advocate of Christianity. I will present the story from Paul's perspective, you be the victim.

> **Double Persecutors (DP)**: I know that the road you intend to take is besieged by persecutors of Christians and that you do not trust me. Even though we both pretend that we know nothing about each other, you know who I am – I am Saul, the famous persecutor of Christians – but you do not know that I am a changed man and you definitely would not believe my story of conversion if I tell it to you. Notwithstanding your more than justified distrust, I am now very much concerned about your welfare. Therefore, I tell you that there are *no* persecutors on that road, hoping that, because you do not trust me at all, you will, contrary to what I say, form the belief that persecutors *are* on the road and, accordingly, not go down that road and remain safe. I want to help you and, in fact, I do help you: convinced that I am lying, you decide to take another road.

In *DP*, because you expected me to lie, I 'pretended to lie' (Vincent and Castelfranchi [1981: 764–766]; Faulkner [2007: 537]) and, by 'pretending to lie,' I tricked you into coming to believe the truth.[12] Specifically, I used your expecting of me to lie to direct your reasoning towards the truth. Save for the fact that I believe that the belief that I have intentionally caused in you is true, everything else fits the C&F description of positive deception *simpliciter*: as an intended result of my action, you have gone from the state of not having a belief concerning whether *p* to having the belief that *p*. Therefore, the case suggests that the thesis that deceivers must intend to cause what they take to be the false belief that *p* in their victims and the idea that deceivers benefit at the cost of their victims are both incorrect. None of the theories that rely on the sender (D_R) benefit and receiver (D_D) detriment can explain this simple straightforward case of deception. In contrast, because I have successfully manipulated your own use of your cognitive capacities, the manipulativist view correctly counts *DP* as deceiving you into believing that *p*.

Nevertheless, I still asserted what I believe to be false and thus I made an assertion contrary to the norms of assertion; I misused it.[13] Misusing an assertion is insufficient for deception but it is sufficient for producing misinformation (I made you believe falsely that I want you to believe what I say) and, by producing misinformation, I have made you epistemically worse off. Thus, the functionalist account seems to be explaining the case

[12] Fallis [2010: 11, 14; 2010: 11] sees this as a case of an obvious lie rather than of pretending to lie. The issue whether this is a lie is irrelevant for my argument; hence, I take no stance on it.

[13] The same concern arises with respect to cases in which the deceiver invents evidence or a testimony (e.g., Barnes [1997: 8–9]; Lynch M. [2009: 109–191]; and Lynch K. [2010]). Suppose I know that *p* is true but that you only believe Jones. I lie to you that Jones told me that *p*, deceiving you thereby into believing that *p* even though *p* is true (Lynch K 2010: 1074). The problem is that, although I make you believe what I believe is true, i.e., that *p*, I do this by asserting what I believe is false, i.e., that Jones told me that *p*.

correctly. Be that as it may, I will argue that whether this feature of *DP* really makes it possible for the functionalist account to correctly count the case as involving deception depends exclusively on whether the *function* of my false assertion was (i) producing a *false* belief that I want you to believe that there are no persecutors on the road or (ii) producing a *true* belief that there are persecutors on the road. If the answer is the latter, then the manipulativist conception is in a better position to explain *BP* than the functionalist conception. In fact, the latter cannot explain it. Before I discuss this issue in detail, I offer a case of a triple bluff (see Faulkner [2007: 536–537]; Fallis [forthcoming]).

> **Triple Persecutors (TP)**: As in *DP*, I know that you justifiably think that I want to deceive you and I mean to do you good. However, in *TP*, you know that I know that you do not trust me (but not that I mean to do you good), and I know that you know this. I also know that you would not trust me even if I tell you that you should trust me explaining why you should trust me. Therefore, wanting to prevent you from taking that road, I tell you that the road *is* besieged by persecutors hoping that you will think not that I am lying but that I am trying to double bluff you and that you will, therefore, conclude that the road is besieged by persecutors, as I asserted.

In *TP*, I assert that *p* hoping that I will thereby trick you into concluding that *p* so that you come to believe the truth. This is also positive deception *simpliciter*: by asserting what I know to be true, I trick you into believing what you did not believe and what you would not have believed had I refrained from so acting. You will find yourself tricked into believing that *p* by means of deception, notwithstanding that the belief is true and believed by me to be true. And, because I do not assert anything false, I do not misuse my assertion, as in *DP*. However, although I do not directly produce misinformation by misusing my assertion, as in *DP*, I causally contribute to producing misinformation: you infer from my true assertion that I want you to believe that that the road is not besieged by persecutors, which is false. The difference, then, is that my assertion is false in *DP* and in *TP* it is true but misleading. Therefore, it appears as if the functionalist account correctly explains both *DP* and *TP* by appealing to your ending up epistemically less well off as you could have been, i.e., had I been completely honest with you.

I do not think that this is the correct interpretation of *DP* and especially *TP*. Only the manipulativist view captures what *constitutes* deception in these examples. The first thing to note is that I am not trying to develop counterexamples to the received conceptions of deception in which the deceived is fully informed, since it is impossible to be deceived if you are fully informed and this is not the claim that I defend. The manipulativist view merely

holds that what constitutes deception is a particular kind of manipulation – a trick, interference, etc. (see 3.2) – meant to affect your truth-evaluable thought and not keeping in the dark or producing misinformation.

It follows that not every manipulative trick or interference is deceptive. Non-deceptive manipulative tricks are those in which the target is fully informed about the relevant topic; they are common in jokes which require the audience to participate (e.g., 'knock-knock' or 'dad' jokes), non-literal speech (metaphors, phrases, figurative speech, and so forth), or warfare. Consider the following strategy of non-deceptive manipulation (Sorensen 2010: 612–613).

> During World War II, American soldiers would fight only if it was common knowledge that other soldiers would not defect. Knowing that the American soldiers were racially divided, Japanese propagandists assured African American (AA) defectors that they would be well treated. Although the Japanese realized that the AA soldiers knew they would not be well treated and that they would realise that this is just a trick meant to plant distrust within their ranks, the Japanese hoped that white soldiers would not know that the AA soldiers knew this. Even if none of the white soldiers were racists, the white soldiers would believe that some of their white comrades were racists. The mere spectre of racism was enough to undermine common knowledge that AA soldiers would not defect. As a result of this general distrust, the AA units were withdrawn and the Japanese propagandists succeeded in reducing American troop strength.

As Sorensen [2010: 613] correctly argues, this lie is non-deceptive (since everybody knew it is a lie and what its purpose was, no one was misinformed by it); however, it involved a manipulative trick whereby the Japanese propagandists reduced the American troop strength. The trick is non-deceptive because it is not hidden from its targets: each soldier knew that Japanese were lying and that AA soldiers would not be well treated and they knew that this was a trick meant to stir up distrust that already existed within their ranks. The trick was successful not because it was hidden but rather because soldiers did not know whether all other soldiers knew this. That is, the trick was successful because it had set in motion the racism that was already present and, for that, it need not have been deceptive.

Whether the trick is hidden from the target makes all the difference between deceptive and non-deceptive manipulation. Deceptive manipulative trick needs to be hidden, *at first* at least, in order to be successful. And this is the only sense in which D_D was not made epistemically well off as he or she could have been. Here is one such manipulative deceptive trick.

Stranger: When I was in high school, I read Camus's *Stranger* and I absolutely hated the book. This was exclusively because I found the main character, Meursault, disgusting. His indifference towards everything except for himself was intolerable. However, I later learned that this character and the whole atmosphere of the *Stranger* is not a product of Camus's destructive and opportunistic cynicism but rather of his constructive and creative irony. Camus used irony to make me reject the whole 'stranger' attitude towards life and others as sheer nonsense. He tricked me into forming a certain judgement and a relevant stance towards life. He tricked me into despising alienation.

Stranger involves a deceptive manipulation even though I ended up with a belief which (presumably) both is true and taken by Camus as true. The case highlights two features of deceptive manipulative tricks and other kinds of deceptive manipulative interferences. For the deceptive manipulation to work, I had to be misinformed about the author's intentions. However, once it performed its function, the trick can be revealed: I did not change my mind once I learned the truth but only because I chose not to. The same applies to *TP*: once you learn that the road is besieged by persecutors and that I knew that you expected a double bluff on my part, *and only then*, I will be able to disclose the full truth to you successfully.

That I am not in a position to be fully informative in *BP* and *TP* is a vitally important feature of these cases that typically gets left unnoticed. Highlighting this feature is my second point. If I try to tell you the whole truth right away, you will not believe me: since my story of conversion is unbelievable and you rightly do not trust me in *TP*, I am *not* in a position in which I could fully inform you and thus, in *TP*, *I did not* prevent you from becoming as well off as you could have; I never had the power to make you perfectly well off with respect to our question under discussion. It is *your* incorrect understanding of the situation, rather than my intention to misinform you, that actually caused that you end up misinformed about what I want you to believe. You prevented yourself from becoming as well off as you could have.

There is a great similarity between my situation in *TP* and that of the female Photinus: neither of us can avoid misinforming our receivers. This female is misinformative because there is no signal that could cover only the state in which she is present. Similarly, in the situation as described in *TP*, no signal could successfully cover the state in which I want to help you because I just got converted due to having an epiphany of Christ. In particular, if I tell you that I know that you expect a double bluff and that I want to help you because I found God last night, you may infer that I am doing this in order to make you expect a triple bluff and you may, therefore, expect a quadruple bluff on my part. Accordingly, I will not know

whether to *quadruple* or *quintuple* bluff you, and so on, and so on, *ad infinitum* (Mahon 2016: 26).

Full disclosure is impossible in the context in which you expect me to be dishonest. Just like in the Fireflies example, where the yellow-green eliminates only females Pyractomena and is not thus fully informative despite it being the correct signal, my telling you the truth (sending the correct signal) would also not be fully informative in *TP* – you will not believe what I say. I cannot avoid making you believe a falsehood. I can only choose which falsehood I will make you believe.

The difference between the deception involved in *Stranger* and in *TP* is that Camus, the deceiver in the former example, produces no misinformation: I drew no inferences as to what were the author's intentions or beliefs: I was seeing the events described in the book through Meursault's eyes and I was strictly analysing Meursault's actions and thoughts. Rather, Camus merely strategically withheld his trick from me and mere withholding information cannot constitute deception. We have a good reason, therefore, to consider *Stranger* as a counterexample to the functionalist definition of deception: the deceptive manipulative trick produces no misinformation.

The third thing to note concerning cases such as *TP* and *Stranger* is that other conceptions of deception simply cannot non-problematically allow true beliefs to be products of deception even when this involves producing misinformation; namely, they cannot explain *TP* and *Stranger* in a way that does not produce false positives. They cannot explain *Stranger* because Camus produces no misinformation. And they cannot explain *TP* without generating false positives because, even though you did end up misinformed in some sense, you ended up epistemically better off in the most important sense: the intended *product* is a true belief. Your false belief that I want you to believe that that the road is not besieged by persecutors is just a *means* of generating this product. That is, the *function of* this misinformation in *TP* is *generating information*; the function is *not* leading you into a misinformative state, as required by the functionalist definition (Artiga and Paternotte 2017: 13). Purposefully producing misinformation is not what constitutes deception; rather, it is what you *do* with that misinformation that constitutes deception. This brings me to my fourth and final point.

Taking it that the functionalist definition can explain *TP* is to confuse means with ends, which then generates further problems. In the Fireflies example, the male firefly descends

only to find the predator female Photuris there: this is why the signal leads to a misinformative state thus fulfilling its function. In contrast, you avoid the road that is really besieged by persecutors: my true assertion does not lead to a misinformative state; rather, the function of my assertion is leading to an informative state – you are deceived when you get what you want, a safe road to pass. If we simply take it that, in deception, the function of misinformation could be generating information (without the relevant manipulation), then the behaviour of the female Photinus will count as deception: she raises the probability of the predator female Photuris and the function of this misinformation is generating information, i.e., the function is disclosing her location to the male. But this is surely incorrect. Therefore, we should hold that what constitutes deception is manipulating the victim by way of using her personal way of responding to this misinformation – since this is how the predator female gets what she 'wants.' In contrast, the right female does not need a trick to get what she 'wants.'

Considering the four points that I have made, *TP* is a case of deception that received theories of deception cannot non-problematically explain. But, even if *TP* has left you unconvinced that you should be a manipulativist, there are cases of deception in which the deceiver does not produce any misinformation. *Stranger* is one such case and others can also be imagined and, in fact, I have already presented one. The *Superstitious Philosopher True* (Section 3.2.1), a case in which you make me believe that I will fail the test and the belief is true but I would not have believed this had you not tricked me, counts as deception according to manipulativism. In this case, I do not believe falsely that you are superstitious. The only thing that was hidden from me is your trick, your subtle manipulation of my superstition, which is why you are being deceptive.

Keeping the trick a secret is a part of the trick; it is not the function of the trick and, because it is not a function of the trick, it cannot be considered as the result of deception. Because the secrecy of the trick is not the result of deception but rather a means of deceiving, no received epistemic conception of deception, including the functionalist view, can correctly count *Stranger* and *Superstitious Philosopher True* as involving deception. In contrast, the manipulativist account explains these cases correctly: because you have intentionally and non-deviantly manipulated my agential use of my cognitive capacities, and because you got the result you wanted, you are a deceiver (D_R), I am the deceived (D_D), and these cases involve interpersonal deception.

These cases can be modified to show that C&F [1977] is not only not too wide (as argued by many philosophers) but that it is too narrow. In particular, C&F (which gives three possible products of deception) leaves out the following types of interpersonal deception: (I) actions by which D_R contributes causally towards D_D's (Ia) acquiring or (Ib) retaining a *true* belief (or truth-evaluable thought); and (II) actions by which D_R contributes causally either to (IIa) preventing D_D from acquiring a *false* belief or to (IIb) D_D's ceasing to have a *false* belief. Also missing are cases in which D_R causally contributes to D_D's (III) ending up with an inadequate degree of belief (confidence in a proposition). The confidence level will become either (IIIa) *higher* due to manipulation (D_D is more confident in *p* where *p* may be true or false), or (IIIb) *lower* due to manipulation (less confident), or even (IIIc) *unchanged* due to manipulation (it remained the same).

DP, *TP*, *Stranger*, and *Superstitious Philosopher True* involve actions of type (Ia), while other variations of deception can be generated from them – I focus on *TP*. Suppose that you already believed that the road is besieged by persecutors (when there are persecutors) but you needed a confirmation, which you aimed to obtain from my response to your question. I know that you only seek confirmation and, knowing that you expect a double bluff, I decide to triple bluff you: I assert that the road is besieged by persecutors hoping that this will make you retain your belief. Accordingly, my deception would cause that you retain a true belief, which is deception type (Ib). We can also imagine that you made me more confident in believing my true belief, which was your intention. Thus, we have a case of intentional deception of the type (IIIa), the variation in which *p* is true. Suppose now that you suspect that your belief that there are persecutors on the road should be abandoned and you decide to seek confirmation from my answer. My triple bluff will causally contribute to your keeping the same confidence in your true belief, which is an action of the type (IIIc), the variation in which *p* is true.

Conversely, suppose that you believe that the road is *not* besieged by persecutors (when there are persecutors) and that you want to check whether your belief is true by analysing my reply. My triple bluff will contribute causally to your ceasing to have a false belief, which is deception type (IIb). We may also imagine that my action did not make you abandon your belief immediately, you had some second thoughts, but that it did make you less confident in believing falsely that the road is not besieged by persecutors. This gives us an example of deception of type (IIIb), the variation in which *p* is false. Finally, you could be without a

belief on the topic but going in the direction of forming the false belief that the road is not besieged by persecutors. My assertion would, then, prevent you from acquiring a false belief, which is deception of type (IIa).

Now that you hopefully are convinced that you should be a manipulativist about deception, I will argue by way of examples that the manipulativist conception of interpersonal deception need not generate paradoxes when applied to self-deception. Although it does not solve all problems in understanding self-deception, manipulativism makes a giant step towards understanding this phenomenon better.

4.1. Here's Why You Should be a Manipulativist about *Self*-deception

The above cases mostly concerned the victim's epistemic situation, but the deceiver's epistemic situation is also much more diverse than is typically thought. In the main examples above – namely, *DP*, *TP*, *Stranger*, and *Superstitious Philosopher True* – D_R intentionally deceived D_D into believing what D_R believed to be true. However, D_R can intend to deceive D_D into believing that *p* (or to affect D_D's confidence in *p*) even if D_R has suspended belief concerning whether *p*. The following four cases are important because they further suggest that the widespread view that a deceiver must intend to make the deceived epistemically worse off offers too narrow an understanding of what it is to be a deceiver.

Presumably Triple Persecutors: Everything is as in *TP* except that I do not know whether the road you intend to take is besieged by persecutors of Christians but I know another road that you can take, which is longer yet absolutely safe. Not wanting you to take any chances and knowing that you would take some risks just to save some time, I say to you that the road you intend to take is besieged by persecutors (I triple bluff you), intending to make you take the safe, long road.

Presumably Stove: We are going on a fishing trip but neither you nor I can remember whether we left the stove on. I tend to panic about these things, and I know that you take me to be very trustworthy, so I tell you that I did leave it on, hoping that you will believe me and that we will then go back and check.

Guessing the Answer: My wife keeps bugging me how I'm terrible at maths and suddenly triumphantly says 'You don't even know how much 4 x 15 is!' 'It's 60!' – I fire the answer back at her, hoping that it's not 75. I never was good at these things, you know.

Guessing the Direction: My wife and I are going to visit old friends but I'm not really sure what the route is, so I ask her to guide me. Annoyed by the fact that I keep asking for directions even though we're on a part of the route where we've been many times, she begins

to nag vigorously: 'You have no idea where to go now, do you?' 'We should go left!' – I shout without thinking, hoping that I got it right. [14]

Here is another variation of a case of deception in which D_R has suspended belief as to whether *p*. I call this deceiving by bullshitting (bs-ing).[15] Deceivers who deceive by bs-ing do not have the intention to cause D_D to believe something false. They merely intend to make D_D believe what they say.

Fortune-tellers False: Suppose I believed a fortune-teller who told me that I will get married at 27. She randomly picked this number without any concern as to whether this might be true; she just wanted my money. Having turned 28 and still not married, I would feel the resentment of a duped person even if I learn that she did not believe to be false what she caused me to falsely believe as true. I would still see her as a deceiver.

Fortune-tellers True: everything is as in the *False* condition except that I really got married at 27. Nevertheless, I would feel the resentment of a duped person when I learn that she cared not whether the proposition she made me believe is true or false. I would still see her as a deceiver; she tricked me into believing this – my belief is luckily true.

These cases involve interpersonal deception even if it luckily turns out, in *Fortune-tellers True*, that the road is besieged by persecutors, that I did leave the stove on, that 4x15 is 60, and that we should go left. And even if I have not convinced you to be a manipulativist, I still may have a good argument based on these cases in favour of the view that self-deception is not paradoxical; that is, that self-deceivers need not intend to make themselves believe a falsehood (the Doxastic paradox) or to hide their intention to deceive themselves from themselves (the Strategy paradox). These cases show that D_R can *intend* to deceive D_D into believing that *p* even if D_R neither believes nor has an opinion as to whether *p* is true. That is, as against the aforesaid widely accepted view (Section 2.1), I need not believe that *p* is false in order to be able to form the intention to deceive you into believing that *p*.

Since I can intend to deceive you into believing that *p* even if I have no opinion as to whether *p*, I can equally intend to deceive myself into believing that *p* under the same circumstances. I offer two (among many) plausible cases of self-deception along these lines.

Presumably Stove Self-Deception: I just left for a fishing trip but I can't remember whether I left my stove on. I am too lazy to go back solely because of this gap in my memory but I am pessimistic enough to think I might regret this. Therefore, I start telling myself repeatedly how I must have left it on, even though I neither believe that it is on nor have an

[14] See also, Van Fraassen [1988: 124].
[15] On bullshitting, Frankfurt [2005], Sorensen [2011: 406], and Stokke and Fallis [2017].

opinion as to whether I should justifiably believe this. Repeating this to myself continuously makes me more anxious about not checking, eventually, I become sufficiently convinced that I left it on, hence, I go back.

I believe that this is an example of self-deception in which I intended to make myself more confident in *p* in a situation in which I initially had no opinion as to whether *p*. And no paradoxes arise. True, I did not tell myself that *p* with the intention to deceive myself (I wanted to make myself more convinced that *p*) but this does not break the analogy with interpersonal deception. Even in cases rightly described as interpersonal deception, D_R may believe that *p* is false and intend to make D_D believe that *p* under the description 'making D_D believe that *p*' rather than 'deceiving D_D into believing that *p*.' In pretending that I am dead, I may just want the enemy to think that I am dead; I need not have the further reflective thought that I am intentionally deceiving them into believing that I am dead. Many people intentionally deceive in the sense of 'cause to believe' and lie in the sense of 'assert something false.'

The following is a case of self-deception borrowed from Clifford [1918: 95] that I take to involve deceiving (oneself) by bullshitting.

My Ship is Safe: A ship owner suspects that his old ship carrying immigrants might not survive its next trip. The owner has nothing to lose. If the ship reaches its destination, he will reap the benefits, while if it does not, he will collect the insurance money. In short, he is indifferent as to whether the ship will really survive the trip. However, it is important to him to believe that the ship will survive; he sees himself as having good moral character. Therefore, by knowingly and willingly working himself into the relevant frame of mind (without any concern as to whether this frame is justified), he dispels his doubts, changes his mind about having it thoroughly repaired, and puts his faith in Providence. In this way, he acquires a comfortable conviction that his vessel can reach the destination. Fortuitously, the ship really can reach the destination and, in fact, it did.

The shipowner 'knowingly and willingly works himself into the relevant frame' simply because this frame delivers the required belief and thereby the needed relief. This situation, in which the owner deceives himself in the same way he might deceive you or me, in no sense entails that the owner believes contradictions and intends to make himself believe a falsehood. Even though he initially suspects that the ship might not survive, since he does not care about the truth, the owner does not consider anything relevant to this issue. Rather, he merely directs his reasoning towards the conclusion that will bring him relief.

All cases discussed so far suggest that you should be a manipulativist about interpersonal deception and that our other conceptions unnecessarily generate the paradoxes of self-deception. The manipulativist view yields a ground-breaking model of self-deception because it allows us to explain self-deception in a way we thought to be impossible: a unified and reasonably coherent self actually can intend to deceive itself and succeed in acting on this intention.

5. Conclusion

In this Subchapter (II-1), I proposed and defended the manipulativist theory of deception. I argued that it can explain cases of interpersonal deception other theories cannot explain and that it can explain all cases of deception as well as any other theory. This should have made you become a manipulativist about deception. In addition, I indicated that the manipulativist view generates a non-paradoxical account of self-deception. On this view, self-deception is an intentional and non-deviant self-manipulation by way of a trick of one's agential use of one's own cognitive capacities where the aim of this manipulation is affecting some relevant truth-evaluable mental state.

Self-deceivers need not believe contradictions or hide their intentions from themselves because they see their actions as manipulative tricks (something like Camus's tricking me into despising alienation), not as harming themselves. In fact, they may think that deceiving themselves is in their best interest (as in *Presumably Stove Self-Deception*). Notice, this view does not exclude the possibility that self-deceivers have some false beliefs; rather, it just says that we need not think that modelling self-deception on interpersonal deception requires that self-deceivers act with the intention to cause themselves to hold or retain what they take to be false beliefs.

All in all, the manipulativist theory of deception can tell us exactly what constitutes deception, it explains cases of deception other accounts cannot, it can explain human and non-human deception equally well, and it generates a non-paradoxical account of self-deception. And that's why we should be manipulativists about deception.

II-2. Self-Deception: Mapping the Terrain

1. The Manipulativist View and Self-Deception

Traditionalists see self-deception as an action by which the self intentionally deceives itself, whereas deflationists take it to be an intentional action by which the self deceives itself. The difference between the two is that, on the latter description, the intention was not to deceive oneself; rather, the person intentionally performed some other action that resulted in her ending up deceived – namely, with a false belief.

In this thesis, I give credit to both of these views (and I allow deception of the self to count as self-deception): most, if not the majority, of self-deceivers do not intend to deceive themselves, but a non-trivially large group of people does act on this intention. Nevertheless, this acceptance of deflationism comes with an important caveat: just like traditionalism, deflationism needs to be put on proper foundations. My analysis commits me to saying that the received deflationary view offers a theory of self-deception only in a figurative sense of deception ('having a false belief'), which, in addition to this, makes the theory too narrow, as it allows for self-deception to be intentional only in a very weak sense of intentional – since the person was intentionally performing an action that in no way violates her own norms of rationality or belief-formation. Strictly speaking, the current deflationary account is of unintentionally causing oneself to form a false belief, not of deceiving oneself.

The manipulativist view of deception is quite handy in deflating the intention to deceive – so that it does not involve the intention to end up with a false belief – while keeping features vital for deception (principally, the idea of deception as manipulation of cognitive capacities) and maintaining the plausible view that self-deceivers violate their own norms of rationality and belief-formation. According to the *manipulativist version* of the deflationary account,

> **Def**.: A is a self-deceiver if, motivated to get the intended φ-relevant response, A φ-s intentionally and thereby manipulates his agential use of his own cognitive capacities without realising that he is manipulating or deceiving himself by φ-ing; however, A may realise that he is obtaining the φ-relevant response in a way that is not typical for him.

I dub actions that satisfy the manipulativist traditional account of self-deception *simpliciter* and actions that satisfy this manipulativist deflationary description *self-deception secundum quid*. In self-deception *simpliciter* (simply, absolutely, unconditionally [*OED*]), the person intends to manipulate or deceive herself by φ-ing and, in self-deception *secundum quid* (in a particular respect [*OED*]), she merely φ-s intentionally. Self-deception *secundum quid*, then, counts as deception not because the product is a false belief – a condition that makes the deflationary account too narrow – but rather because the action that brought about the product involved a motivated, intentional manipulation, i.e., intentional violation of the person's own norms.

Notice that C&F classify deception relative to its product: deception is *simpliciter* when it causes a change (positive or negative) in belief and *secundum quid* if it sustains a state of belief that ought to be changed. In contrast, I classify self-deception considering the intention on which the self-deceiver acts. Self-deception is *simpliciter* if the intention is to manipulate or deceive and it is *secundum quid* if the intentional action caused self-deception but the intention was not to manipulate or deceive. These are the necessary and sufficient conditions for these types of self-deception.

A is a self-deceiver *simpliciter* if and only if A

(1) intentionally brings about φ-ing as the action of manipulating or deceiving himself into causing a certain φ-relevant response.
(2) φ-ing causally contributes towards manipulating A's agential use of his own cognitive capacities.
(3) the φ-relevant response was successfully generated in such way.

A is a self-deceiver *secundum quid* if and only if A

(1) intentionally brings about φ-ing as φ-ing intending to thereby affect his or her own truth-evaluable mental state.
(2) φ-ing causally contributes towards manipulating A's agential use of his or her own cognitive capacities in a way relevant for 1.
(3) A realises neither that he or she is deceiving themself by φ-ing nor that he or she is thereby manipulating themself (otherwise, A would not φ).
(4) A's truth-evaluable mental state is affected by way of φ-ing.

In the end, I propose that the general account of self-deception should be developed along these lines:

Def$_{SeD}$: self-deception is an action by which the person intentionally manipulates her agential use of her own cognitive capacities with the function of generating a certain response. However, this person may not understand that her acting in such a way counts as manipulating and/or deceiving herself.

Importantly, 'and/or' characterisation is there to allow a person to understand the action as manipulating herself while not seeing that this amounts to deceiving herself. I allow that self-deceivers *simpliciter* may understand their action as self-manipulation while not drawing a further conclusion that they are deceiving themselves. Many interpersonal deceivers see their action as 'making X believe what I want X to believe' rather than 'deceiving X,' so why should we deny that this may apply also to self-deceivers?

In what follows, I will offer a concise classification of self-deception to get a clearer grasp on various instances of this phenomenon. I first classify self-deception according to its product, i.e., whether the product is welcome; and then, in Section 3, according to why and how the action was performed, i.e., whether in the action that counts as self-deception a proposition was denied, promoted, or retained.

2. Welcome, Unwelcome, and Indifferent Self-Deception

The most commonly discussed cases of self-deception are those in which the person ends up with a belief or a belief-like thought she favours or welcomes, the so-called *welcome self-deception*. Some philosophers, e.g., Van Leeuwen [2007a: 423], call this *wishful self-deception* and, in the majority of cases – i.e., when self-deception is continuous with wishful thinking, as in the cases of Hillary, Maria, and Nicole – this name correctly portrays the phenomenon. However, there are cases in which the self-deceiver does not wish or favour the product but still welcomes it and we want to be able to capture these cases as well; hence, welcome self-deception. For example, given the preponderance of evidence of the affair, Hillary could deceive herself into thinking not that Bill is not having an affair, as this seems impossible, but rather that she does not really care that Bill is having an affair.

And yet, not all products of self-deception are welcome. People can deceive themselves into holding all kinds of beliefs or belief-like thoughts, even the unfavourable ones. I propose that we call those cases in which self-deceivers deceive themselves into generating a result they would rather avoid, e.g., believing something they would prefer not to be the case, *unwelcome self-deception* (see I/1, n. 3). For example, Zelda deceives herself into thinking that Scott is having an affair and I might deceive myself into thinking that the government is spying on me and so on. Unwelcome self-deception is problematic on two levels: it is not only that we need to explain why Zelda did this to herself; this explanation needs to be consistent with the way we understand welcome self-deception. That is, if the desire that

Scott is having an affair is not what motivates Zelda's action, then why should we take it that the desires that Bill or Arnold are not having affairs motivated Hillary or Maria to deceive themselves?

To this problem, I add another one: we may sometimes deceive ourselves into believing propositions with no particular personal significance. For instance, I can deceive myself into believing that the government is spying on *you* or that smoking is not bad for *your* health, or that 9/11 was an inside job, that Snowden is a Russian spy, that Trump will be a great president, that Bill did not mean to lie when he said that he did not have sexual relations with Monica, or even that Elvis is not dead, and many more. I dub this phenomenon *indifferent self-deception* – the deceiver does not draw any personal benefit from his action and thus has no reason to welcome or not welcome the proposition. This kind of self-deception has not received any notable attention in the literature but I find it quite important; it contradicts the widely accepted view of self-deception as a *self*-defensive strategy.

Some philosophers – e.g., Gardiner [1970: 242] and Patten [2003: 244] – have argued that the facts that (1) some self-deceivers are indifferent towards the proposition they end up believing and that (2) have no personal benefit or detriment from ending up this way entail that the action cannot be motivated at all. Gardiner [1970: 242], for instance, writes:

> It is … not clear what could be meant by, or what justification there could be for, speaking of somebody as deceiving himself if it were at the same time contended that what he was said to be deceiving himself about was a matter of total indifference to him, in no way related to his wants, fears, hopes, and so forth: could we, e.g., intelligibly talk about 'disinterested' or 'gratuitous' self-deception?

I would not say that indifferent self-deception justifies taking it that self-deception is not a motivated action. First, the premises in this argument are not plausible. A proposition in no way related to the person's wants, fears, hopes, and so forth will not be mentally entertained by that person – even mental tropisms, which are subintentional, are triggered in order to perform a function. More importantly, even if true, the thesis that the product of indifferent ('disinterested') the *person* sees no value in self-deception does not entail that it has no actual value. Therefore, the fact that the agent is indifferent towards the product of his deception does not entail that self-deception is unmotivated.

Having that said, I leave the issue of the motive aside; I will resolve it in subchapter V-4. Here, I just want to map the terrain and the next step is describing directions of the action by which we may deceive ourselves.

3. Denying, Promoting, and Retentive Self-deception

I have been able to identify three general types of actions that can count as self-deception: a person somehow *denies* something (available information, reasons for suspicion, belief, belief-like thought, etc.), she somehow *promotes* something (a state of affairs, value, belief, belief-like thought, etc.), and she somehow *retains* a thought (belief, belief-like thought, value, etc.). I will briefly sketch each type below.

Roughly, in *denying self-deception*, by negating what I do not welcome (a state of affairs, suspicion, belief, information, etc.), I form a belief or belief-like thought that I welcome.[16] Typically, denying self-deception is triggered by the desire that something is not the case. The Nicole case is a nice example: Nicole is obviously trying to deny a certain painful reality. Traditionalists also take the Hillary-type cases as representative of this family. Hillary obviously denies that Bill is having an affair but traditionalists also think that she would not be denying this had she not already been believing it to be true. On their view, by denying the unwelcome reality, she is revising, i.e., abandoning, her unwelcome belief. In response, I argue that we have no reason to think that Hillary actually formed the belief that Bill is having an affair. In fact, people typically deceive themselves in order to prevent themselves from believing something unfavourable, not in order to break out from believing it.

Hillary, we should note, may be sufficiently motivated to believe that Bill is not having an affair even if she did not believe that he is. She may (strongly) suspect that he is having an affair, without actually believing it.[17] Hillary's case is best described as exemplifying *retentive* self-deception, which I define as involving undermining the revision of an already held belief. It is not reasonable to assume that Hillary waited for herself to abandon her favourable belief about Bill's fidelity (she would not have married him had she thought he was going to cheat on her) before actually forming the unfavourable belief that he was having an affair in order, only then, to start defending herself from this unwelcome reality. Anyone

[16] For example, Davidson [1986, 1997], Gardner [1993: 16], Pataki [1997], Lockie [2003], and Von Hippel and Trivers [2011].

[17] Many philosophers offer this argument, e.g., Mele [1997: 94; 2001: 96], Van Leeuwen [2007a: 427–428]; Michel and Newen [2010: 734–735], von Hippel and Trivers [2011: 8–12], Archer [2013: 270–276].

would start defending herself from the unwelcome reality right from the start, namely, by preventing revision of their established favourable belief.

The Maria case is also of retentive self-deception. Maria undermines the revision of her belief that Arnold is not having an affair by undermining information that may become counterevidence to that belief; she sets some possibly damaging pieces of information aside. A retentive self-deceiver may prevent unwanted information from being encoded in the first place. For example, if one has some awareness that an upcoming piece of information may be inconsistent with what one already favourably believes, i.e. that an unwelcome piece of information is around the corner, one may decide to 'take a different route' (Von Hippel and Trivers [2011: 2, see 8–9]; see Michel and Newen [2010]). These are typically cases of keeping oneself in the dark by self-deceptively retaining a belief or a belief-like thought (see Mele 2010b: 747) and they do not involve believing contradictions.

In order to understand the third kind of self-deception, classified according to how the person ended up in the self-deceived state, we need to recall that, in *Presumably Triple Persecutors* (II-1/3.5), D_R caused D_D to believe that the road is besieged by persecutors while merely suspecting that this might be the case and that, in a different example, the fortune-teller deceived me into believing that I will get married at the age of 27 by bullshitting me. These cases show that, in interpersonal deception, the deceiver need have no prior beliefs concerning whether *p*. We may, then, assume that self-deception can start off without any prior beliefs concerning whether *p* as well. Therefore, we should posit *promoting self-deception*, in which the trigger simply is a desire that *p* be the case. The promoting self-deceiver is self-deceptively forming a thought (accepting a proposition, increasing her confidence in a proposition, etc.) by presenting the reality to herself as if the relevant proposition is true, namely, by manipulating herself in a relevant way, and this manipulation distinguishes the behaviour from wishful thinking. The person is simply promoting a certain picture of reality she welcomes; she is not running away from anything.

Heil [1993: 115] writes: 'forming beliefs is something we do voluntarily. I can, of course, elect to take steps that will result, predictably, in the formation or inhibition of particular beliefs' (similarly, Bermúdez 2000: 314). This proposal concerning promoting self-deception is built along similar lines. Suppose that I have just managed to make myself believe that my thesis is brilliant even though, in preparing its final form, I have only got to Subchapter II-2. I

have no reason to believe that my thesis is brilliant, since I have not had a look at the draft of chapter V in almost a year, but I have no reason to believe that the thesis is unremarkable either – so far, I like my work a lot! If this case involves self-deception, and I see no reason to think otherwise, it is promoting self-deception and the belief is welcome.

To summarise, in Subchapter II-1, I analysed deception and specifically intentional deception. I proposed and defended the manipulativist conception of deception. Then, in this subchapter (II-2), I presented some basic kinds of self-deception categorised according to the product (welcome, unwelcome and indifferent self-deception) and how the product was generated (denying, promoting, and retaining). I will now proceed to analyse some important received theories of self-deception, trying to get the best out of them.

III. RECEIVED THEORIES OF SELF-DECEPTION

1. Classification

Philosophers have two general perspectives on self-deception. On the one side, there is a group that argues that self-deception must be understood as an intentional action; that is, you must *do* something that causally contributes to your ending up being deceived if you are a self-deceiver. Accounts of this first type are typically called *intentionalist accounts* and they come in two sub-groups, depending on how this intentional action is understood. On the *traditionalist view*, the self-deceiver did what she did intending to deceive herself whereas, on the *deflationary view*, she intended to do something else but she ended up deceived instead/in addition.

As opposed to the intentionalist account, psychologists and some contemporary philosophers believe that self-deception is not an intentional action or not an action at all; rather, it may be a state in which the agent is, namely, a state of being deceived about something – typically, about oneself. Accounts of this second type are *non-intentionalist accounts*. On his view, being self-deceived just is being in a state in which you hold false beliefs about yourself – the aetiology of your state is irrelevant. These accounts are sometimes called *revisionist*, a name I will adopt here – given that this approach does seem to be revising the concept of self-deception (the idea of the self *deceiving* itself is missing).

Of course, these distinctions are never sharply drawn and many theories combine theses from different approaches. I call the views that combine some trademark thesis of intentionalist and non-intentionalist views, *hybrid* views. Hybrid views are typically based on the deflationary view, to which they add certain revisions of the concept of self-deception; such as that self-deceivers must hold false beliefs about themselves, typically second-order beliefs; or that the product of self-deception is not a belief but some belief-like thought (default thought, alief) or a hybrid mental state, such as besire (a hybrid of a belief and a desire) or avowed belief (a sort of a truncated belief, a belief lacking some vital features of beliefs), and the like. Nonetheless, hybrids of traditionalism and revisionism are also present: Sartre, for example, argued that in bad-faith, a kind of self-deception, we lie to ourselves

about who we really are; that is, that we are, in a traditional way, deceived by ourselves about ourselves.

The current state of the debate could be briefly summarised as follows. It has been a matter of extensive debate which of these two intentionalist descriptions, the traditional or the deflationary description, actually captures the essence of self-deception. The traditional account, which models it on interpersonal deception, has seemed a better candidate when it comes to the common views on intentional actions and on interpersonal deception. Nonetheless, since it is based on the soft-line view of deception according to which the deceiver, D_R, intends to deceive the deceived, D_D, into believing something D_R believes to be false (see II-1/2.2), it generates paradoxes. Because no acceptable solutions to these paradoxes have been provided (I would say 'yet'), traditionalism got slowly abandoned in favour of the account which seemed to be close enough to the intuitive notion of self-deception but which did not generate paradoxes, namely, the deflationary view. The problem with the deflationary view, however, is twofold: it relies completely on the hard-line view that takes the figurative sense of 'being deceived,' i.e., 'having a false belief,' as its primary sense and it seems to be too wide and thus unsatisfactory. Therefore, many contemporary philosophers turn to various revisionist accounts, both intentionalist and non-intentionalist in their nature, which are typically based on the idea that self-deception is or involves a deception about the self and that the product of self-deception is not a belief but rather some belief-like thought.

Below, I will first present the two main groups of intentionalist accounts (III-1 and III-2), and then the revisionist accounts (III-3), which can be intentionalist (hybrid) or non-intentionalist. This chapter will serve as a nice introduction to the stage in which I finalise my proposed account of self-deception. In brief, I am a revisionist about lying and interpersonal deception and a traditionalist about self-deception who also recognizes the value of other theories of self-deception, most notably, deflationism.

III-1. The Traditionalist View

1. Introduction

Considering the Davidsonian theory of action, a reasonable position is that, if my intentional action is to be rightly called an action of self-deception, it is not enough that it is intentional under any description, i.e., as any action, and that it resulted in me ending up being deceived about something. Rather, I must intend to deceive myself by performing it. Therefore, the first pillar of the traditionalist account is that I must intend to deceive myself in the same way in which I would intend to deceive you; that is, self-deception should be modelled on interpersonal deception.[18] Accordingly, because on the standard account of interpersonal deception (the soft-line approach), D_R intends to deceive D_D into believing something D_R believes to be false, 'deceiving myself' is naturally defined as 'intentionally causing that I end up believing as true what I believed to be false prior to my action.' Roughly, a self-deceiver intends to cause himself to have a belief he believes to be false.

Before presenting the traditionalist account in more detail, in order to avoid misleading the reader, I would like to highlight some points concerning my departure from traditionalism about self-deception. I hold the view that the product of deception need not be a false belief or even a belief believed by the deceiver to be false and I have argued that the right conception of deception should be manipulativist in its core (II-1). In short, I depart from traditionalism in my preferred account of interpersonal deception. On my version of traditionalism, I am deceiving myself if I am intentionally manipulating my agential use of my own cognitive capacities (II-2).

In contrast, the main characteristic of the standard version traditionalism is that self-deceivers intentionally get themselves to form beliefs they took to be false prior to their own deception. Moreover, traditionalists take it that believing something one does not want to believe is what motivates self-deception, a position I argued against in II-2/3. If we break this traditionalist view into parts, we get the following: if a subject S is self-deceived, then there is a proposition *p* such that:

(1) S believed that *p* prior to deception.
(2) (Because of this belief,) S formed the intention to get himself to believe that ~*p*.

[18] Hereafter just 'interpersonal deception' unless specified otherwise.

(3) S acted on his intention.
(4) S now believes that ~*p* (1 and 4 give the so-called the *dual-belief condition*).
(5) S's action is causally responsible for his forming the belief that ~*p*.

This account is meant to explain cases such as Hillary's: based on the evidence, she forms the belief that Bill is having an affair and she deceives herself into believing that he is not. The standard traditional analysis of the Hillary-type cases would go along the following lines. Hillary is a rational person who forms her beliefs according to available evidence and therefore she must have acquired the belief that Bill is having an affair. Yet, because she keeps sincerely saying that Bill is not having an affair and she is acting consistently with her avowals, it must be that, wanting to break out from believing the things she does not want to believe, she has deceived herself into believing that he is not.

However, if my analysis from II-2/3 is correct, then Hillary's case is best described as involving retentive self-deception. Hillary believed that her husband is faithful before she initiated deception and it is much more reasonable to assume that she initiated deception in order to prevent revision of this belief rather than to reacquire it. If I am right, then the standard version of traditionalism cannot explain this case – condition (1) is not satisfied – but this is not to say that this account is not applicable. The most plausible candidates for this account's explanandum are cases in which a person did not have prior beliefs on the topic of her self-deception: I may deceive myself that my argument is brilliant, a student may deceive herself that she will pass the exam, Donald may deceive himself that he won the debate, and so on. Let us, therefore, analyse traditionalism using the case in which, motivated by my unfavourable yet justified belief that my thesis is unremarkable, I deceive myself into believing (or forming a belief-like thought) that my thesis is brilliant – I dub this example *My Brilliance*.[19]

Now that we have a case that fits the traditional theory, we must recall that this understanding of self-deception raises two famous paradoxes. The first is the *Strategy* paradox or the *Dynamic* puzzle: intuitively, we cannot deceive someone who knows our intentions and we know our own intentions. Therefore, taking it that I could intentionally deceive myself that my argument is brilliant, in these circumstances, seems to be paradoxical because the success of self-deception entails that I concurrently know (I am D_R) and not

[19] This is not the classic *p* vs. ~*p* contradiction, in the sense that if *p* is false ~*p* must be true – brilliant is contrary to unremarkable but it could be that the thesis is neither of the two – but the inconsistency is still clearly there, thus I will treat it as *p* vs. ~*p*.

know (I am also D_D) my own intention to deceive myself. The second, *Doxastic* paradox or the *Static* puzzle arises because I should, as D_R, believe that my thesis is unremarkable and, as D_D, that it is brilliant, which should be impossible if my mind is reasonably coherent – since I must know that the thesis cannot be both (Mele 1987).

Furthermore, and this is important, the Doxastic paradox is not based on a simple case of believing contradictions. According to traditionalists, what motivates me into deceiving myself into believing that my argument is brilliant is, ultimately, my belief that my argument is unremarkable. That is, I intentionally deceived myself into believing that my argument is brilliant in order to escape the unwelcome belief that it is unremarkable; the former causes and sustains the latter (Davidson 1986). This excludes the interpretation according to which the unfavourable belief was somehow cancelled, denied, or repressed before I initiated the action of acquiring the belief that my argument is brilliant. The unfavourable belief must be present while I am forming the favourable belief so that the motivation for deception is operational. This feature makes self-deception extremely problematic on this view so that partitioning the mind seems unavoidable.[20] I call the need to posit the divided mind on the level of its intentional states in order to explain self-deception, the *Traditionalist Trap*. A scholar who has fallen into this trap has substituted a problematic conception of the mind for a problematic conception of self-deception.

I will not consider the psychological partitioning theories in detail in this thesis; this topic has already been widely discussed and a brief outline was given in the introduction. Besides, I do not intend to criticise or defend these theories; rather, I want to argue that traditionalism does not need them at all. Below I analyse some other ways in which traditionalists try to resolve these problems or in which they might have tried to resolve them.

2. Deceiving Your Future Self

One possible solution posits temporal divisions of the mind and says that one person-stage deceives another distant person-stage (e.g., Sorensen [1985]; Bermúdez [2000]) and that this is what may aptly enough be described as self-deception.

> **A *vs.* C:** Suppose that a person A transmits the information that they know to be false to the person B, and B then transfers the same false information to C. Now, B deceives C

[20] For example, Davidson [1982, 1986, 1997], Pears [1982, 1986, 1991], Pataki [1997], Lockie [2003], and many more.

> unintentionally but A's deception is intentional. What is more, if A intended that the false information communicated to B is then further transferred to C, we may say that A intentionally deceived C. Finally, if A, B, and C are all temporal stages of a single person, this is how one deceives oneself. For example, I may deceive my future self about the time of my distant appointment by writing the incorrect time in my calendar (Bermúdez 2000: 315).

To posit temporal partitioning of the mind is to posit logically distinct subjects, which in turn makes the action by which I deceive my future self non-problematic, but philosophers typically think that it falls short of self-deception (Davidson [1986: 208, n. 5]; Johnston [1988: 76–78]; Scott-Kakures [1996: 41–42]; Pataki [1997: 308–310]). Accordingly, not much attention has been given to this proposal. Calling this self-induced deception (McLaughlin [1988: 41]; see Davidson [1986: 208, ft. 5]; Barnes [1997: 24]) seems a good way to describe actions of this kind.

In his recent work, nevertheless, Scott-Kakures [2012: 22] argues that actions of this kind should count as self-deception. After all, he says, the traditional view on interpersonal deception requires only that agent intentionally cause themselves to have a belief they hold to be false, which is what happens in this case. While the definition of deception that Scott-Kakures assumes here is very problematic (see II-1; Mahon 2008), his position does have certain merits. For instance, one common objection, which says that 'at time t^2 [the future person] will be justified in believing that P although P is false, given the evidence then available to her, so that she will not be in a state of [self-deception]' (Galeotti 2012: 45), does not appear to be reasonable. Having an unjustified belief is not a necessary condition for being deceived: the victim of a double or a triple bluff is justified in what she believes and yet they were deceived into believing it (see II-1/3.2). Also, this objection excludes tampering with evidence from possible strategies of deception, which is unjustified.

In fact, that I can deceive my future self intentionally does seem to be a reasonable intuition: in the movie *Memento* [2000], Guy Pearce's character, Leonard, searches for his wife's murderer while overcoming obstacles posed by his short-term memory loss. Because he cannot remember anything that happened to him after a certain date, he tattoos clues and important memories on his body. By tattooing a piece of false information (spoiler alert!), he manages to deceive his future self about the identity of the murderer. We see that both in Bermúdez's *A vs. C* example and in *Memento*, it is the agent himself who generated the false evidence, a very common strategy of deception.

In particular, Leonard has intentionally caused his future self to form a false belief about the identity of the murderer; or as Scott-Kakures [2012: 37] puts it, he has 'intentionally caused himself to deceive himself unintentionally.' This rather convoluted formulation is incorrect and it needs to be reformulated. First, 'deceiving unintentionally' can only mean 'misleading unintentionally' (Bach [2009: 783]; Mahon [2008]) but this is the least of our concerns. The following comparison with cases of interpersonal deception may be useful here.

Grigory Potemkin, a Russian military leader, managed to impress Empress Catherine II during her journey to Crimea in 1787 by building portable fake settlements along the banks of the Dnieper River (thus the expression Potemkin village). Potemkin has deceived the Empress, but he has done so indirectly: the Empress was actually *deceived by appearance*, and being deceived here applies metaphorically; that is, this is not a real deception – since appearance cannot act intentionally (Barnes 1997: 4). Suppose my facial expression makes you think that I am angry but this is in fact just a face I make when I am deeply engaged in thinking. Seeing that my thinking face is the same as other people's angry face, the 'deception' seems to be the result of incoherence in the world rather than a purposeful action. Likewise, by writing the incorrect time in my calendar, I did not intentionally bring about conditions in which I unintentionally cause myself to have a false belief (*pace* Scott-Kakures 2012: 36); rather, I intentionally brought about conditions in which I got deceived by appearance. The same goes for Leonard and his tattoo strategy and, since both Leonard and I intentionally brought about conditions with the aim of causing ourselves to end up with a false belief, our behaviour should rightly count as deceiving.

In conclusion, Scott-Kakures is correct in arguing that deceiving your future self by tampering with evidence should rightly count as self-deception. However, he misdescribes self-deception of this kind. I did not intentionally cause myself to mislead myself unintentionally and I especially did not intentionally cause myself to deceive myself unintentionally. Rather, I intentionally deceived myself by causing it to be the case that the future-I gets deceived by appearances or, to be more precise, I deceived myself by causing it that my future self infers an incorrect belief from appearances.

The correct description, then, is that, as a result of the deceiver-I's manipulation, the deceived-I ended up with a justified false belief, which was a part of the deceptive intention. This kind of self-deception is a real phenomenon but it is not at all problematic or interesting

and, most importantly, 'what we call self-deception is rarely like that' (Darwall 1988: 412). Therefore, I leave it aside. Nevertheless, I do want to reject an interesting version of the temporal partitioning argument proposed in line with the notion of self-deception typically held by psychologists.

Psychologists characterise self-deception as a state in which the individual (i) simultaneously holds two contradictory beliefs, (ii) is not aware of holding one of the beliefs, and (iii) the lack of awareness is motivated (Gur and Sackeim 1979). This state, with some possible variations, is typically called 'literal self-deception.'[21] Operating within this theory, Quattrone and Tversky [1984: 248] explain self-handicapping strategies as involving self-deception. Here is their argument.

People deploy self-handicapping strategies in order to alter the circumstances of diagnostic performance with the aim of protecting the belief that they are competent (Jones and Berglas 1978). The point of self-handicapping strategies is to ascribe the failure to one's handicap: for example, 'by drinking or taking drugs [before the upcoming IQ test], any level of intellectual performance would not destroy the belief that one is basically bright, for even failure could be attributed to the debilitating effects of alcohol' (Quattrone and Tversky 1984: 248). And, on Quattrone and Tversky's argument, because they have acted with the intention to prevent themselves from confirming their subconscious belief that they are not bright, these people have intentionally deceived themselves; that is, the above conditions i–iii are satisfied. The idea is that, by drinking, I deceive my future self into believing that I failed the test due to an interfering factor, not because I was not smart enough.

However, this does not seem to be deception at all: I did fail the test because I was drunk and the manipulation was of my cognitive capacities (i.e., relevant cognitive processes) not of my agential use of them. True, I kept myself in the dark concerning what the outcome of the test *would have been*, had I been sober, but I did not do this by deceiving myself; rather, I avoided entering a situation that might prove me wrong, but avoiding *x*, i.e., willful ignorance of *x*, is not the same as deceiving oneself about what *x* may tell me (Lynch 2016). The importance of this example, and the reason I discuss it is that it demonstrates that the classic

[21] It is very interesting that an account that actually substantially revises our common folk-psychological conception of self-deception by reducing it to motivated self-knowledge with the dual belief condition is called an account of 'literal' self-deception. I do not see what is literal here.

psychologist's theory of self-deception is too wide. Psychologists are talking about a broader class of motivated irrationality.

3. Deceiving Yourself or Not Being Smart Enough

In his 1996 paper, Scott-Kakures offers an interesting attempt to avoid paradoxes. He proposes that actions that fit the traditionalist account of self-deception are to be found in the self-deceiver's effort to understand her own epistemic condition; namely, the self-deceiver deceives herself in evaluating her incorrectly acquired belief.

This account makes use of Johnston's [1988] proposal that subintentional purposeful processes causally and arationally may affect beliefs. I will consider Johnston's arguments in detail in III-2/3; here, I just make use of his proposed analysis. A Johnstonian self-deceiver has, due to a purposeful subintentional process, arationally 'jumped' from the belief that ~*p* (my work is unremarkable) to the belief that *p* (my work is brilliant). The former belief was repressed and the latter directly caused by the desire.

This is yet not self-deception, says Scott-Kakures; the processes of repression and belief-generation are sub-intentional whereas self-deception must be intentional. The irrational step relevant for self-deception is in the evaluation of the unjustified belief: in reflection, I see that I believe that *p* (my work is brilliant) but that I have reasons to believe that ~*p* (my work is unremarkable). That is to say, I do not know why I have lost the belief that ~*p* and why I have formed the belief that *p* – in fact, I believe that *p* while believing that ~*p* is best supported by evidence – and therefore, aiming to retain the belief that *p*, I invent reasons for keeping it. The evaluation is irrational for two reasons: it results in retention rather than in abandonment of the belief and I have realised that I have violated my own principles of irrationality.

This theory is very interesting – and it might explain cases such as Hillary's, she sure is smart enough – but it gives us an extremely complicated mechanism of self-deception. A person of average intelligence cannot deceive herself on this theory. Scott-Kakures [1996: 52] writes:

> The self-deceiver … thinks something to the effect of: Given my beliefs at t, I shouldn't now believe that *p* since no chain of reasoning (no rejection of the evidence in favor of ~*p*, etc.) connects these two states [these states were connected by Johnstonian subintentional processes] … To have this thought is not to think that one shouldn't … believe that *p*. It is rather to think that given the way one *was* epistemically constituted one shouldn't currently believe that *p*. The subject believes that she has *come* to believe that *p* in an epistemically

> unjustifiable manner… [Notice that] the subject does, by her current lights, have sufficient reason for believing that *p*. It is rather a case of her believing that she has come to believe that *p* in a way which makes for a violation of her principles of reasoning.

In order for me to be a self-deceiver on this account, I need to know the subtle difference between being justified *in believing* (at t) that *p* and being justified *in coming to believe* (at t) that *p*. But, considering how subtle this difference is, only a certain group of very smart or philosophically well-educated individuals could in principle be capable of deceiving themselves in this way. I, for instance, could not have ended up in the state of being deceived about the quality of my work by deploying this strategy. This point may be better illustrated on the following example, which nicely shows that many theories of self-deception ask too much of us.

The aforementioned Quattrone and Tversky [1984] conducted two experiments that supposedly demonstrate literal self-deception; their research and conclusions are extremely influential but I will present only the relevant bit. In their first experiment, subjects were told that the duration of one's being able to hold one's hand in cold water is associated with the desired heart type favouring longevity. The result, as interpreted by the two authors, was that the majority of subjects seem to have 'tried' to achieve the desired result and thus 'have' the desired heart type.

Their analysis of their experiment is based on the thesis that 'It should have been clear to subjects that shifts in tolerance would have no *causal* impact on their life expectancy. Shifts would be *merely diagnostic* of their life expectancy in that both shifts and life-expectancy were affected by an individual's heart type' (Quattrone and Tversky 1984: 243). Yet, to know this subtle difference is something we should expect only of abnormally intelligent or suitably well-educated individuals and we have no reason to think that the test subjects knew this difference. I, again, would never realise on my own that tolerance of cold is merely diagnostic of a heart type. Therefore, while it seems fairly obvious that the subjects' behaviour was motivated, it does not follow that they were deceiving themselves or even unintentionally misleading themselves.

Having said that, I note that I do think that Scott-Kakures's approach is valuable but that it needs substantial revision; I will, in subchapter V-2, propose an explanation of the Hillary case along similar lines. To conclude the present section, I will inspect the most famous traditionalist account, the one by Donald Davidson.

4. Donald Davidson

Davidson's main analysis of self-deception, as with most of his arguments, is very hard to follow: he often skips steps that connect his premises, or jumps back to some claims he made earlier, he repeatedly introduces examples that need careful analysis rather than clearly showing the point, and so forth. This style of writing makes it quite difficult to reconstruct the actual argument and so we are often forced to analyse its pieces separately. I believe that this is actually Davidson's tactics. The real motivation behind his papers on self-deception is not explaining this phenomenon but rather giving an argument in favour of positing divisions within the mind. The difficult writing style is just his way of hiding weak spots in the argument – Freud used similar tactics (see Derksen 2001). Therefore, I will treat his influential analysis of self-deception [1986, 1997], which motivated many philosophers to understand self-deception as necessarily involving internal incoherence, as an argument in favour of the conception of the mind as partitioned.

This argument can be briefly summarised as follows. In deceiving myself that my thesis is brilliant, I do not form this belief merely because I want it to be the case. Rather, the available information and the epistemic principle, which commands that I should believe only what the totality of evidence suggests, initially brought about the belief that the thesis is unremarkable. The belief that my thesis is unremarkable, then, triggers anxiety and I now wish either that my thesis is brilliant or that it is not that my thesis is unremarkable.[22] However, even though the unfavourable belief that my thesis is unremarkable and the attitude associated with it motivate me into believing that my thesis is brilliant, they are not sufficient reasons to believe it. That is, I will not believe that my thesis is brilliant just because I wish that it is brilliant – at least, not by way of deceiving myself. What prevents me from forming the belief that my thesis is brilliant is not just the evidence that suggests that my thesis is unremarkable and my belief that my thesis is unremarkable but equally, and more importantly, the epistemic principle that commands that I form my beliefs exclusively according to all available evidence.

In self-deception, then, compelled by my desire that it be true that my thesis is brilliant or to avoid the belief that my thesis is unremarkable, I (1) turn my attention away from the belief that my thesis is unremarkable and (2) ignore the epistemic principle that forces me to proportion my beliefs according to evidence. And this enables me to form the belief that my

[22] Let's suppose that 'my thesis is brilliant' being true equals my thesis is unremarkable being false.

thesis is brilliant *in the face* of evidence to the contrary. The problem that arises now is that, according to Davidson, this epistemic principle that I ignore is constitutive of rational agency; that is, no rational agent can ignore it on any considerations (Davidson 1986: 212). And yet, somehow, in forming the belief that my thesis is brilliant, I did ignore it. In ignoring this principle, I have exhibited a kind of inner inconsistency that is impossible [1986: 203] and this is why self-deception is conceptually problematic kind of irrationality – problematic for the conception of the mind as unified and reasonably coherent.

The solution must be, says Davidson, that there is a small and much simpler part of my mind, a part that need not abide by the principles of epistemic rationality, that at one point purposefully but not intentionally contributed to my epistemic deliberation and – governed by practical reasons (my motivation to believe that my thesis is brilliant) – walled off the belief that my thesis is unremarkable and the epistemic principle responsible for forming it, which in turn resulted in the formation of the belief that my thesis is brilliant solely based on my practical reasons (my wish that it is true that my thesis is brilliant) and in spite of the available evidence that suggests that my thesis is unremarkable.

In short, self-deception tells us that we should posit that our minds are partitioned because only this can explain it. In particular, my desire that my thesis is brilliant was a cause of, but not a reason for, ignoring the epistemic principle – *for a rational agent*, nothing is a reason to ignore it [1986: 212] – and therefore my action of ignoring it was performed without an agential reason. Because every action must be performed for a reason (Davidson 1963), and I did exercise my agency in ignoring it, the reason must come from a part of the agent's mind (similarly, Davidson 1982), a part that can ignore epistemic principles and use only practical considerations – e.g., a desire to believe that my thesis is brilliant and to ignore the epistemic principle – in epistemic deliberation. This is why, then, we should accept the conception of the mind as compartmentalised.

We see that, according to Davidson, self-deceivers are not only internally incoherent because they wall off a fundamental epistemic principle, they also believe contradictions; namely, I deceived myself into believing that my thesis is brilliant because I believed that it is unremarkable while I wanted it to be brilliant. Now, a person with a reasonably unified and coherent mind can believe that p and that $\sim p$ at the same time only if she does not realise that she actually believes contradictory propositions, which may happen non-problematically – she may fail to put two and two (p and $\sim p$) together for some reason (Davidson [1985: 353;

1999: 444]; Stroud [2013: 501]; see Nelson [2005: 8]; Bermúdez [2000: 313]). However, this does not explain Davidson's account of self-deception, according to which I deceive myself in order to avoid believing what I believe, and therefore the unwelcome belief must be operational while the action is being performed (Davidson 1997: 216). The idea is that, if my belief that my thesis is unremarkable initiates self-deception, then ceasing to believe it will take away the practical reason for performing the action in the first place. Therefore, I cannot unbelieve that my thesis is unremarkable if I am, in turn, to deceive myself into believing that it is brilliant.

Davidson [1986: 208] describes the relationship between two beliefs by saying that one causes the other (italics added).[23]

> The thought that *p*, or the thought that he ought rationally to believe *p* [e.g., Bill is having an affair], *motivates A to act in such a way as to cause himself to believe the negation of p.* The action involved may be no more than an intentional directing of attention away from the evidence in favour of *p*; or it may involve the active search for evidence against *p* [similarly, 1997: 229]. *All that self-deception demands of the action is that the motive originates in a belief that p is true* (or recognition that the evidence makes it more likely to be true than not),[24] and that the action be done with the intention of producing a belief in the negation of *p*. Finally, and *it is especially 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*.

In the strongest kind of self-deception, then, the belief that my thesis is unremarkable, by triggering the desire to avoid believing this, motivates me to believe that it is brilliant. Furthermore, because this desire was caused by my belief that it is unremarkable, this belief is, ultimately, the cause of my belief that it is brilliant and the former sustains the latter by keeping the motive operational. However, while my desire that the thesis is brilliant (or that it is not the case that it is unremarkable) causes the belief that it is brilliant, the desire is not the right kind of reason to believe this and, actually, it should not even be a causally efficacious kind of reason – the right kind of reason is epistemic. Davidson [1986: 205] writes (italics added):

> We must make an obvious distinction between having a reason to be a believer in a certain proposition, and having evidence in the light of which it is reasonable to think the proposition

[23] Davidson discusses a situation in which the epistemically justified belief '*p*' causes the belief that '*~p.*' In contrast, in the *My Brilliance* example, the belief that *~p* causes the belief that *p*.

[24] Please notice that the bracketed claim, 'recognition that the evidence makes it more likely to be true than not,' suggests that self-deceivers need not actually believe that *p* to be motivated to deceive themselves that *~p*. Nevertheless, Davidson focuses on the strongest case.

> true … A reason of the first sort is *evaluative*: it provides a motive for acting in such a way as to promote having a belief. A reason of the second kind is *cognitive*: it consists in evidence one has for the truth of a proposition.

What he means is this. The self-deceiver's motivation to believe a proposition, based on the desirability of the proposition's being true, causes the believing of the relevant proposition *as if* the motivation is evidence in favour of the truth of the proposition; namely, as if the motivation is a reason to believe it. However, although my motivation is a reason to believe that my thesis is brilliant (evaluative, practical), it is not a reason to believe that the proposition 'My thesis is brilliant' is *true* (Davidson 1997: 220).

Although beliefs are truth-directed and truth-regulated (Shah and Velleman 2005), forming beliefs on the basis of epistemically inadequate reasons happens all the time and this involves an epistemic failure but this is not the whole story of self-deception. Besides, the epistemic failure involved in self-deception is not that self-deceivers form beliefs on incorrect reasons (practical rather than cognitive) – this is how Davidson's position is being typically interpreted but this is not what he says (see below). Rather, and this thesis is crucial, their irrationality is conceptually problematic because self-deceivers do not have *any reason* to form their belief in the face of counterevidence and yet they *intentionally* act so to acquire it – I intentionally direct attention away from the evidence in favour of 'My thesis is unremarkable,' search for evidence against this proposition, etc. (Davidson 1986: 208; 1997: 229). This position needs to be explained in detail.

The *Principle* (*Requirement*) *of Total Evidence for Inductive Reasoning* (PTE henceforth), which commands 'give credence to the hypothesis most highly supported by all available relevant evidence,' as already noted, is held by Davidson [1985: 351; 1986: 201] to be constitutive of all rational agency. In self-deceptively forming the belief that my thesis is brilliant, I have sinned against PTE, which, then, allowed me to form the belief I wanted to form in the face of counterevidence. That is, because my *evaluative reason* (my wish that it is brilliant) hijacked the role of my cognitive reason ([counter]evidence), I did not accept the hypothesis I judged I should accept, and what made this possible is that I have ignored PTE (Davidson 1986: 201, 204).

Ignoring the PTE is the most important step in the argument: it makes self-deception internally irrational: by violating my own norms of belief-formation I have exhibited inner inconsistency (Davidson 1986: 203), in the way that requires positing divisions in the mind. In particular, because rational agents cannot decide not to accept fundamental attributes of

rationality, nothing can be a reason to ignore PTE (Davidson 1985: 352) and thus I cannot disregard it. Davidson makes this point very clear but, since he makes it quite late, it often goes unnoticed by the critics. He [1986: 212] writes (original italics)

> What causes it [PTE] to be thus temporarily exiled or isolated is, of course, the desire to avoid accepting what the requirement counsels. But this cannot be a *reason* for neglecting the requirement. Nothing can be viewed as a good reason for failing to reason according to one's best standards of rationality.

What he means is that a rational agent cannot ignore PTE but I did ignore it and I also made my belief that my thesis is unremarkable ineffective. The fact that I successfully ignored a principle constitutive of rational agency, the argument continues, should be explained by saying that a pseudo-rational part of my mind, a part that is internally coherent and operates mainly on practical reasons, contributed to my epistemic deliberation by walling off the undesirable belief and the PTE. I did not reject PTE; I sort of lost sight of it because it was walled off, hidden, in one of my mental compartments – i.e., 'sub-structures' (Cavell 1999: 416) or subsystems (Pears 1991). The PTE must be walled off because I cannot willingly disregard it (Davidson 1985: 351) and the unfavourable belief must be walled off because it needs to be inferentially, not physically (i.e., put in subconsciousness or dissociated in some other way), separated from the newly formed belief. However, and this is also crucial if the exhibited kind of irrationality is going to involve conceptually problematic internal incoherence, the 'deceiver' sub-structure will not house the evidence that led me to form the belief that my thesis is unremarkable.

The idea that the evidence is also walled off was proposed by Pears [1982: 273] and it is often attributed to Davidson (e.g., Heil [1989: 577]; Scott-Kakures [1996: 43]) but Davidson explicitly rejects it; he [1986: 209–210] says: 'the self-deceiver cannot afford to forget … the preponderance of evidence against the induced belief' and this is because, with the information out of the way, the self-deceiver would not be forced to ignore PTE, which makes self-deception a conceptually *un*problematic kind of irrationality and undermines Davidson's real conclusion – we now need not compartmentalise the mind. That the information must be available to the person is an important difference between Davidson and explanations offered by Freudian-style theories. According to Freudian-style views, one part of the mind uses various defence mechanisms to prevent the other part of the mind from accessing the (unpleasant) information or a mental state (unpleasant or humiliating desire); it defends the consciousness (vertical divisions) or the main agent (horizontal divisions) from

confronting this information or desire (see, e.g., Lockie [2003: 130]; Von Hippel and Trivers [2011: 6]). A Davidsonian self-deceiver, on the other hand, must intentionally deceive herself *while* having this information available. That said, I proceed to argue against this theory.

4.1. Challenging Davidson

Some philosophers argued that what Davidson describes is not self-deception but rather deception exercised by one rational centre of agency over another: one entity, the partition, deceives the other (Mele [1984: 138]; Foss [1980: 239]; Johnston [1988]). However, this is not what Davidson is saying: 'The image I wished to invite was not, then, that of two minds each somehow able to act like an independent agent; the image is rather that of a single mind not wholly integrated' [1997: 221]. The proposal is not that a partition intentionally initiates the action. Partitioning merely limits or diverts access to the usual array of deliberative resources, PTE being one of them.[25] In other words, had my mind been wholly unified, I would not be able to ignore PTE, but, since there is a part of me that is not wholly integrated with the rest of my mind, the principle and the unfavourable belief sort of 'got lost or misplaced' in it and the constitution of this part is such that PTE merely plays no guiding role, it was rendered ineffective.

Seeing that Davidson does not propose that the deceiver partition actually performs the deception, we must notice that many philosophers unfairly object to this view by assigning to it consequences it does not have. For example, Van Leeuwen [2013: 2] argues that a partition capable of deceiving must also be capable of: (a) knowing the goal of deception, (b) believing the contrary proposition such that what it counts as deception, (c) transmitting information to the main system, and (d) knowing what the main system already knows.

These concerns cannot be raised against Davidson's proposal: his argument only requires that the partition renders the unfavourable belief and the PTE causally inefficacious. Therefore, (a) and (d) straightforwardly do not hold; the partition need not know the goal of deception or what the main agent knows. Furthermore, because self-deception is an achievement of the whole agent, the partition need not house both beliefs; thus, (b) does not hold either. Then, the partition need not transmit any information; rather, it just harbours (figuratively speaking) the walled-off PTE and the unfavourable belief and so (c) does not hold. Notice that, and this is crucial, walling off is a logical operation, not physical.

[25] Pears [1991: 394] nicely explains the difference between Davidson's concept of the partition and rational centres of agency. Mijović-Prelec and Prelec [2010] offer a view similar to Pears's.

Therefore, the self-deceiver can and must be aware of the principle and of the belief (Pears 1991: 399–400); walling off merely makes them ineffective in practical reasoning.

Yet, none of this is to say that Davidson had me convinced. I have four objections to his argument. Principally, I see no good reason to accept his description of self-deception. Although we can imagine a situation such as this and although we can explain it by appealing to some kind of compartmentalisation of the mind that will show that laws of logic are not violated – believing *p* and ~*p* as described above violates the law of non-contradiction – we should avoid describing any case in this way unless no other description is applicable. If the self-deceiver's behaviour can be successfully explained on a less demanding hypothesis – and considering the complexity of this solution, that should not be hard to find – then this is of course what we should do.

Here is one rival explanation: according to Davidson, the self-deceiver exhibits inner inconsistency, which is done by ignoring the PTE. But, one may exhibit inner inconsistency in many ways that do not involve ignoring principles constitutive of rationality; one may simply, contrary to one's typical behaviour, ignore evidence. In fact, Davidson's solution also says that a self-deceiver simply ignores the evidence; he just describes this in a way that requires understanding the mind as partitioned.

Secondly, the concept of mental partition is too vague, even incoherent. Partitions are sufficiently like human agents yet are not rational centres of agency; how can these two claims be reconciled (see, e.g., Gardner [1993: 74]; Heil [1989])? Moreover, if walling off is logical, not physical, then we have no reason to posit the division of the mind – a physically wholly united mind can exhibit logical failures. Even worse, understanding walling off as a logical rather than a physical operation entails that the agent has failed to put two (PTE, the belief that ~*p*) and two together (the belief that *p*), which is exactly what Davidson [e.g., 1997: 216] denies to happen in self-deception. Partitions simply cannot explain Davidson's version of the phenomenon.

I find the second objection quite serious but objections three and four even more serious; they both say that the argument collapses. The third objection is that the argument does not justify positing partitions, i.e., it fails, and the fourth is that partitioning the mind leads into deflationism – self-deception becomes unintentional deception by the self. I start with the third.

It is said that a rational agent cannot have a reason to ignore PTE and that thus we must posit a separate part of the mind that can. However, the agent does not need a reason to wall PTE off. According to Davidson, the agent affects the division by intentionally doing things that will promote believing what he wants to believe, e.g., that my thesis is brilliant (see the long citation from Section 4). But, in that case, I did not intend to wall PTE off; rather, I intended to, e.g., avert my attention from things that suggest that my thesis is unremarkable and, by averting my attention, I have unintentionally walled PTE off (Barnes 1998: 21–22, 27). Because walling PTE off is not something I do intentionally, I do not need a reason to do it and, therefore, we need not posit partitions to explain my behaviour.

The second reason to abstain from positing partitions is even more pressing: positing partitions drives us into deflationism. Recall, Davidson modelled self-deception on the account of interpersonal deception that says that D_R intends to deceive D_D into believing what D_R believes to be false – this thesis led us to paradoxes in the first place. Diverting D_D's attention is a legitimate strategy of deception but it must be performed with the aim of getting D_D deceived, and this further requirement is missing in Davidson's theory. Specifically, I need to avert my attention with the aim of deceiving myself, but this is not what happens. In turn, seeing that the action is intentional under a different description takes away the only reason to posit partitions: they were supposed to 'save' the traditionalist view on self-deception. Also, positing the intention to deceive myself by averting my attention, the partitioning theory will face the same problems as the traditional conception of the mind: if I take it that the reason for my averting my intention is my ending up believing a false belief, then the paradoxes re-emerge.

The truth is that paradoxes of self-deception cannot be resolved by reconceptualising the mind. The theories of lying and interpersonal deception, not the theory of the mind, are what causes the paradoxes. Other traditional solutions that resolve paradoxes by appealing to mental divisions suffer from the same objection: the division will either route the theory into deflationism, which is the most common consequence, or will reintroduce paradoxes. With the thought that most traditionalist solutions end up in deflationism, I proceed to present the deflationary view.

III-2. The Deflationary View

1. Introduction

Philosophers belonging to the deflationary view – e.g., Mele [1987, 1997b, 2001, 2006, 2009; 2010b], Johnston [1988], Barnes [1997], Ainslie [1997], Baumeister and Leith [1997], Van Leeuwen [2007a, 2007b], Michel and Newen [2010], Galeotti [2012], Lynch [2012] – have conceptualised self-deception in the way the Maria case was described. Here is the Maria case again.

> **Maria** possesses much evidence that her husband, Arnold, is having an affair with one of his attractive female co-workers. Arnold has lost sexual interest in Maria, he has protected his phone with a password, and she sometimes picks up what appear to be subtle love signals her husband exchanges with this female co-worker. Yet, Maria intentionally sets these distressing pieces of information aside and, as a result, retains the false belief that Arnold is not having an affair. However, she does not intend to deceive herself (since she never suspected an affair, we think that she lacks the motive for deception). Rather, she retained her belief by what looks like explaining away pieces of information that cause her distress. In a sense, she seems to be acting just like scholars who only give weight to claims that support their own position and avoid those that may undermine it.

In a nutshell, the idea is that Maria's behaviour is best explained by claiming that she has intentionally set the evidence supporting the truth of the proposition that Arnold is having an affair aside and has instead focused on some evidence in favour of his fidelity thus maintaining the false belief (belief-like thought) that he is faithful. Vitally, Maria did not intend to deceive herself; she was merely intentionally avoiding taking account of some unpleasant information – the *process* (of avoiding certain pieces of information) is intentional but not the *outcome* (Galeotti 2012: 52).

This approach has some important upsides. It seems far more likely to think that Hillary deceived herself in the same manner we are here ascribing to Maria than to think that she ever formed the belief that Bill is having an affair and then intentionally caused herself to believe what she took to be false. This explanation works even better in the *My Brilliance* case, in which, following the traditionalist approach, we supposed that I formed the belief that my thesis is unremarkable and, wanting it to be the case that the thesis is brilliant, I intentionally deceived myself into thinking that it is brilliant. It sounds much better to think that, due to my desire that my thesis is brilliant, I have simply failed to appreciate the

counterevidence and that I instead focused on the information consistent with the content of my desire, which is a deflationary description.

The starting premise of the deflationary view is the idea that intentional self-deception is problematic and paradoxical because we model it on interpersonal deception. And it is simply too much to expect self-deceivers to completely internalise interpersonal deception. Deflationists typically, yet not exclusively (e.g., Johnston [1988], see Section 3), think that the self-deceiver performs an intentional action, or takes intentional steps, as a result of which she gets deceived by herself into believing something unsupported by the evidence; however, this person does not intend to deceive herself or to produce or sustain a particular belief in herself (e.g., Mele [1997b: 94]; Galeotti [2012: 41–43; 57–58]). This is what happens in the Maria example and, likewise, we may say that I do not aim intentionally to form the belief that my thesis is brilliant but that I nevertheless do focus intentionally on information that will bring it about.

Because 'focusing,' i.e., biased evidence gathering, is motivated and intentional as 'focusing' not as 'deceiving myself,' this kind of behaviour is sometimes called *weak self-deception*. But, why should it be called self-deception at all? As a description under which the action is intentional it is incorrect. After all, neither Maria nor I intended to deceive ourselves and neither of us was insincere towards him or herself (I/1.1). The standard answer lies in the hard-line understanding of 'being deceived' as 'being mistaken.'[26] Maria and I are self-deceived not because we intended to deceive ourselves but rather because we (1) ended up believing a falsehood, thus 'deceived,' and we did so (2) as a result of our own motivated intentional actions, thus 'self.' This move is aimed at preserving the intuition that the agent is blameworthy for self-deception (we should have known better) while avoiding the paradoxes by eliminating the intention to deceive.[27]

Deflationists hold that a self-deceiver forms the belief that *p* due to the causal influence of a motivational state. However, they depart from the traditionalist description in that the desire is not to break out from believing the unfavourable belief that ~*p*; hence, the Doxastic paradox does not arise. Rather, it is typically proposed that a desire that *p* is/be true is what motivates biases involved in the self-deceptive action that ends up with forming the belief

[26] Fallis [2015: 376] defined holding a false belief as being misled, and he did this in a way that suggests that this is the standard view; however, I am not sure that this is the case.

[27] In my earlier discussion (II-2), I argued that this could be self-deception only figuratively speaking. I will not repeat that argument here.

that *p* (e.g., Mele 1997b). However, there are proposals according to which self-deception is motivated by the desire to believe that *p* (Johnston [1988: 69]; Dalgleish [1997: 110]; Nelkin [2002]; Funkhouser [2005]) or solely by an emotion even if the desire is absent (Mele 2006: 114–115; 2009: 60): for instance, anger can play a biasing role that causes Zelda to falsely believe that Scott is having an affair even though she does not want it to be the case that he really is having an affair.

This, I think, is sufficient for a brief outline of the general deflationary position. I will now present and argue against two prominent deflationary theories, those proposed by Alfred Mele and Mark Johnston. Michel and Newen's [2010] deflationary theory will be presented in subchapter V-3 where I propose my explanation of the Maria case.

2. Alfred Mele

Mele bases his theory on four conditions sufficient for, as he specifies it, *entering* self-deception in acquiring the belief that *p*. These conditions are (Mele e.g., 1987: 127; 1997b: 95; 2010: 747).

1. The belief that *p* which *S* acquires is false,
2. *S* treats data (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 ~*p* than for *p*.

A supplement to condition 2 is recognizing that there are different ways in which the self-deceiver may exhibit this motivated epistemic negligence. These ways are (Mele 2009: 56–57): *negative misinterpretation* of data that count against the belief that *p* (caused by wanting that *p* is true); *positive misinterpretation* of data in favour of *p*; *selective focusing/attending*, i.e., a failure to focus attention on evidence against *p*; and *selective evidence gathering*, i.e., overlooking the evidence against *p* (due to the desire that *p* is true). In brief, S's motivation acts as a ~*p*-data filter and as a *p*-data amplifier (Michel and Newen 2010: 737). This is how Maria retains her belief that Arnold is not having an affair and I form the belief that my thesis is brilliant.

'Entering' self-deception is a nice way of saying that the self-deceiver did not intend to deceive herself but that she did in a sense walk into a trap she has set up for herself. Self-deception is not something you actually *do* (intentionally); rather, it is something you *enter* by doing something else, such as cherry-picking the evidence. Maria, for instance, intended

to avoid taking account of unsettling information but she did not to intend to avoid the conclusion ('Arnold is having an affair'), which this information warrants. This description, which draws its appeal from the opacity of intentions, should make the action non-problematic. Nevertheless, Lockie [2003: 140, 142] argues that intentionally avoiding taking account of evidence may equally be a conceptually problematic action. He writes:

> Why is this evidence 'threatening'? What it is for something to be *evidence* at all is that it contributes warrant towards some conclusion. The data as such are not intrinsically appetitive or aversive … The data only are aversive in as much as their *meaning* has at some level been attended to, and spelt out: that these events are *evidence for adultery* … [I]n most cases, one does not seek to *ignore evidence* as an end in itself. One avoids evidence *because* of what it *means* ('my [spouse] is unfaithful!') with the condition of *being deceived* as the necessary concomitant of the pursuit of that end.

Lockie's argument is not as successful as he takes it to be: two important claims involve a *non sequitur*. First, that Arnold's behaviour is evidence of adultery only justifies us in taking it that Maria may *suspect* that Arnold is having an affair – no more than that – and this is consistent with Mele's view. Second, even if she sees what the evidence 'means,' it does not follow that she will realise that she is deceiving herself by ignoring that evidence. Realising that you are deceiving yourself by ignoring the evidence requires reasoning and analysis of both your behaviour and of the relevant concepts (ignoring, deception, etc.). Thus, the claim that she must see that she will end up being deceived – again, understood as 'believing falsely' – as a result of her behaviour does not follow (see Bermúdez 2000).

Despite being a hard-core traditionalist, I agree with Mele [2006: 110] to the extent that self-deception need not involve intentionally deceiving oneself, or intending (or trying) to deceive oneself, or intending to make it easier for oneself to believe something. Nonetheless, I will argue that Mele has failed to offer a satisfactory account of ordinary instances of self-deception where such intentions are involved.

One often highlighted problem is that we cannot say with certainty that the person whose action satisfies Mele's four conditions is self-deceived; many other phenomena may satisfy them. Perhaps, unintentionally causing oneself to incorrectly believe that *p* is just an instance of improper belief formation, a sort of innocent error (Scott-Kakures [1996: 37]; Gergen [1997: 114]; Deweese-Boyd [2012]). Alternatively, this may just be unintentionally misleading oneself under the influence of a strong emotion, such as anger or fear (Pataki 1997: 311–312). Even better, we may call this *auto-manipulation*, a case where motivation directs acceptance of *p* – here, the manipulation is not intentional (Michel and Newen 2010:

732). And it might reasonably be described as *willful ignorance*, steering clear of evidence that ~*p* may be true (Lynch 2016). Roughly, the first concern with Mele's deflationary account is that it is too wide.

A further concern is that the account is inconsistent. This concern is, in short, that S cannot satisfy both 2 and 4 (Michel and Newen [2010: 737–739]; Galeotti [2012: 49–50]). According to condition 2, S's biased mechanisms of data-selection led to S's failure to correctly appraise the strength of the evidence; that is, S's motivated bias inhibited gathering of ~*p*-supporting data and amplified the power of *p*-supporting data. Because of this, the body of data actually possessed by S will not give greater warrant to ~*p* rather than *p* and 4 will not be satisfied. Even if S has access to all relevant *data*, it follows from condition 2 that S, as a biased subject, will not treat all data as *evidence* in favour of ~*p*. Consequently, since the body of data possessed by *S* does not provide greater warrant for ~*p* than for *p*, S's behaviour is not self-deception according to the proposed account.

Lynch [2012: 441] modified Mele's account by proposing that it is not that the body of data actually possessed by *S* at the time provides greater warrant for ~*p* than for *p*, it is rather that the body of data S *should have* possessed, had his motivated biased behaviour been absent, provides greater warrant for ~*p* than for *p*. This modification avoids Michel and Newen's objection but it opens the theory up for the worry that self-deception is not a first-person kind of irrationality. If the third-person perspective determines which evidence I should have possessed, then the third-person perspective is what determines whether I have entered the state of self-deception or not, which is not something we should readily accept. Self-deception must involve a kind of one's insincerity towards oneself.

It is worth pointing out, however, that, in defence of his position, Mele [e.g., 2001: 51–52] argues that condition 4 is sufficient, but not necessary, for self-deception; and that, if the body of data possessed by *S* does not provide greater warrant for ~*p* than for *p*, then this is because *S* was selectively gathering data. Accordingly, condition 4 could be understood as follows (see Nelkin [2002: 390], Lynch [2012: 441]):

> 4'. The body of data S should have possessed at the time provides greater warrant for ~*p* than for *p*; *or*, if it does not, then the explanation for that fact is selective data-gathering on the part of *S*.

Condition 4' is a disjunctive combination of Mele's condition 3 and how Lynch proposes condition 4 should be understood. However, the problem with disjunctive analyses is that, if

an analysis is based on two independent criteria, then we might actually be dealing with two separate phenomena (Kingsbury and McKeown-Green 2009: 578–581). In particular, the worry is that one phenomenon will have three sufficient conditions (since the second disjunct of 4' is condition 3) and the other will have four (the first disjunct of 4' is condition 4).

Mele defends his account in part on the basis of its consistency with, what he calls, the FTL theory of hypothesis testing.[28] The main idea of FTL is that a concern to minimize costly errors – i.e., having costly false beliefs – drives lay hypothesis testing. That is, how much we will test a hypothesis before it becomes a belief depends on how costly that belief might be if false. The cost of a false belief is the predicted severity with which it will affect the person's life, which is assessed by contrasting the cost of false acceptance (or rejection) of a proposition relative to the cost of the resources and effort invested in gathering and processing relevant information (Trope and Liberman 1996: 252). The predicted cost of believing falsely a proposition *p*, in turn, determines the confidence thresholds required for acceptance or rejection of a belief that *p*.

The acceptance threshold is the minimum confidence in the truth of a hypothesis *p* sufficient for acquiring a belief that *p* rather than continuing to test the given hypothesis *p*. Trope and Liberman [1996: 253] understand the rejection threshold as sufficient for acquiring a belief that ~*p*, but I would like to propose a modification here. Abstaining from accepting the hypothesis *p* as a belief, due to insufficient confidence, need not involve acquiring the belief that ~*p*; one may suspend belief or even judgement concerning whether *p*. Therefore, I suggest that we understand *the rejection threshold* as the minimum confidence in the untruth of a hypothesis *p* sufficient for not holding a belief that *p*, where 'not holding' stands for 'abstaining from acquiring' or 'abandoning.'

The greater the predicted damage caused by falsely believing that *p*, the greater the cost of falsely believing that *p*. And the greater aversion towards this specific costly error, the higher the confidence threshold required for acquiring the belief that *p* and the lower the confidence threshold for not holding it. The higher the acceptance threshold (i.e., the higher the expected cost of believing it falsely), the more evidence is needed for accepting a belief, whereas, the lower the acceptance threshold (i.e., the lower the expected cost), the less evidence is sufficient for accepting a belief.

[28] FTL is acronym of the names of the relevant authors; Friederich [1993] and Trope and Liberman [1996].

The FTL theory has two predictions relevant to Mele's view. One is likely to require less evidence to acquire a belief that cannot cause significant damage but could introduce an important benefit. Therefore, the first prediction is that because, other things being equal, lower thresholds are easier to reach than higher ones, the belief that *p* is a more likely outcome than the belief that ~*p*, other things being equal, if the person is more strongly averse towards falsely believing that ~*p* than to falsely believing that *p*; that is, if the assessed cost of falsely believing that *p* is lower than the assessed cost of falsely believing that ~*p* (Mele, e.g., 2006: 113). For example, my fear of my house burning down may readily cause a false belief that I have left the stove burner on (*p*) if I assess the cost of believing falsely that the stove is off (~*p*) as much greater than the cost of believing falsely that it is on (*p*): I assess an increased risk of my house burning down as more to my detriment than the time and effort required for going back and checking.

The second prediction is that aversions towards costly errors influence how we test hypotheses – e.g., whether we will exhibit confirmation bias – and when we stop testing them (having reached a relevant confidence threshold). Accordingly, if Maria is more strongly averse to falsely believing that *p*, 'Arnold is having an affair,' than to falsely believing that ~*p*, then she will primarily seek evidence for ~*p*, will be more attentive to such evidence than to evidence for *p*, and will interpret relatively neutral data as supporting ~*p* (Mele 2006: 113). And, on this account, Maria will be more strongly averse towards falsely believing that *p* than to falsely believing that ~*p* because, says Mele based on Friederich [1993: 314], falsely believing that *p*, 'Arnold is having an affair,' will lead to a situation in which Maria mistakenly criticizes herself or lowers her own self-esteem when in fact she has no reason for doing so.

While I generally agree with the idea that the belief that *p* is more costly than the belief that ~*p* (Maria may even endanger her marriage by falsely accusing Arnold of having an affair), I have two significant concerns regarding this proposal (not counting the important limitation that *p* needs to be false). First, there are cases of self-deception where the relative costs as assessed by the person concerned stand in the contrary relation to what is required by the theory. Consider the My Brilliance case, in which I deceive myself that my thesis is brilliant. The cost of falsely believing this is a much greater loss of my self-esteem than the cost of my falsely believing that it is unremarkable. In particular, it is far more important for me to make sure that my supervisor and the examiners see my work as brilliant than to give myself some relief at the cost of turning myself into a doofus at the oral defence. The cost of

falsely believing that my thesis is unremarkable is only in my having to deal with the fact that my supervisor finds my lack of confidence disappointing. The cost of the former … well, I don't want to think about it. This makes the My Brilliance case one where a person deceives himself while making the contrary judgment about the relative costs of falsely believing that *p* than is required on Mele's FTL reading of his theory.

My second concern relates to the standard of the self-deceiver's irrationality. According to this view, the self-deceiver is irrational only from the third-person perspective. We need to compare, and this is what Mele and deflationists following him suggest, the doxastic situation of the self-deceiver with the doxastic situation of her cognitive peers, and if the peers would form a different belief in the light of the evidence that is or should be available, then the person is self-deceived. Nevertheless, once we leave the philosopher's armchair, we see that self-deceivers are not characterised as such because they violate the norms of irrationality of their impartial cognitive peers or because they fail to achieve a certain level of responsiveness to evidence. Rather, they are so described because they violate *their own* norms of irrationality and this violation needs to be a result of the fact that the self-deceiving person was *insincere towards herself*. Only then, their impartial cognitive peers will accuse them of self-deception.

Placing the irrationality of self-deception into the third-person perspective is effectively breaking with the idea that self-deception is a first-person irrationality, which is a very dangerous implication I think – I call this the *Deflationary trap*. Being insincere towards yourself is the distinguishing feature of self-deception; ending up with an evidentially unjustified belief acquired under the influence of a desire is not.

I conclude, then, that Mele's theory is very problematic when analysed in detail. Nonetheless, the starting premise, that self-deceivers can act intentionally but not with the intention to deceive themselves, is seminal and, if combined with the manipulativist account of deception and the correct description of this manipulation that is consistent with the idea that self-deceivers violate their own norms, it will give us the right account of self-deception *secundum quid*. I do this in chapter V-3, where I offer my own account of the Maria case.

3. Mark Johnston

Johnston deflates the intention to deceive even more. Self-deception, he argues, is based on purposive but not intentional processes which serve some interest of the self-deceiver but

which are not necessarily initiated by the self-deceiver for the sake of those interests. That is to say, self-deceptive processes are initiated by the person and they do serve the purpose of protection from the painful truth, but the person did not initiate them with the intention of achieving this purpose. Because they serve a purpose but are not initiated with the intention to serve it, these processes – e.g., denial, repression, wishful perception, and so on – are *subintentional* (Johnston 1988: 65). Many rational actions performed by people, such as making deductive inferences, are also non-intentional yet purposive says Johnston ([1988: 87], similarly Fingarette [1998]).[29]

More specifically, what causes self-deception is a non-accidental but non-rational connection between a desire and a belief, where the former directly generates the latter. Johnston [1988: 67] calls this subintentional event, this purpose-serving mental mechanism in which the desire directly generates the belief, *mental tropism*. In self-deception, the anxiety that the person's desire that *p* will not be satisfied is reduced by her ceasing to acknowledge her own recognition of information that suggests that ~*p* (this is *repression*) while her subsequent acquisition of the belief that *p*, made possible by repressing the information that suggests that ~*p*, is based on her own desire to believe it (this is *wishful thinking*) (see Fingarette 1998: 295–298).

These processes serve the end of reducing anxiety (Johnston 1988: 86) or some other end (Pears 1991: 394). Repression directly alleviates the stress of believing the unfavourable proposition, while wishful thinking, since repressing the belief that ~*p* does not entail acquiring the belief that *p*, aims to acquire the desired belief. The desire that *p*, having no obstacles due to repression, directly, purposefully, non-rationally, and non-intentionally generates the belief that *p*.[30] According to this solution, then, I reduced the anxiety caused by the idea that my thesis is unremarkable by repressing my recognition of the evidence in favour of ~*p*, i.e., 'My thesis is unremarkable,' and then I generated the belief 'My thesis is brilliant' (*p*) by way of purposefully but not intentionally endorsing my wish that *p* as a belief. Hillary's case can be explained in the same way. She represses the information and the

[29] Lauria, Preissmann, and Clément [2016] offer a theory of self-deception that sits in between Mele's and Johnston's views. They see self-deception as involving subconscious, non-intentional affective bias: in short, the anticipated emotional impact on one's well-being determines how the particular piece of information will be interpreted.

[30] Since the unwanted evidence suggesting that ~*p* was already repressed, the desire that *p* need not alter my attention nor do anything similar; this is how it directly generates the belief (Scott-Kakures 1996: 45).

belief that Bill is having an affair and then generates the belief that he is not having an affair directly from her desire that this is the case.

The main problem with this solution may appear to be related to the question whether the transition from my wish to believe that *p* to my actually coming to believe that *p* is psychologically possible and Johnston resolves this problem by inviting us to consider an interesting example. A gambler who deceives himself into believing that he will get blackjack on the next hand and who bets accordingly does this because his desire to win has a potentiality to be rewarded by winning; namely, the event of the wishful belief's arising is reinforced and made possible by the fact that being rewarded is a real possibility – or, at least, believed to be so (Johnston 1988: 71–72). Likewise, in self-deception, repression, by eliminating evidence in favour of the undesired outcome, makes the possible reward operational and this possible reward facilitates relevant subintentional process. Therefore, my desire to believe that *p*, freed from the unwelcomed evidence supporting the contrary proposition and aided by the effective possibility that the thesis is brilliant, now directly generates the belief that the thesis is brilliant.

This proposal seems very compelling. However, I will give you good reasons to think that tropisms cannot explain self-deception or, at least, that they cannot explain it better than a traditionalist theory combined with the theory of compartmentalised mind.

The main reason for developing a non-traditional theory of self-deception is that traditionalism requires a specific architecture of the mind, the one of a functionally partitioned mind, and this architecture is very problematic. Johnston even introduces his argument by offering an extensive criticism of this architecture. The idea is that the tropisms solution is better than modelling self-deception on interpersonal deception because it can explain cases we understand as self-deception without the need for such theory of the mind and we do not need a theory of a divided mind because actions of self-deception are purposeful yet we did not intend them to have this purpose; that is, everything happened 'automatically,' i.e., purposefully but subintentionally. However, I will show you that tropisms do require partitioning of the mind and then argue that therefore we have no reasons to be Johnstonians about self-deception.

Recall, both the repression of unwanted beliefs and/or information and the motivated acquiring of desired beliefs are non-intentional outcomes of mental tropisms (Johnston 1988: 78). The desire that the person believes that *p*, supported by the reward that is still

operational, directly causes the belief by triggering the required subintentional processes. An example of this process, as proposed by Johnston, is when a person's attention suddenly and automatically shifts from a boring conversation she has been idly involved in a nearby debate about a topic she is deeply interested in. This person is not intentionally turning her attention to the debate she is interested in; rather, her automatic filtering process, which is inaccessible to introspection and which determines that what is salient in perception answers one's interests, initiates the purposeful action directly (Johnston [1988: 87]; see Fingarette [1998: 290–291, 295]).[31]

This analogy fails, however, because it disregards the substantial dissimilarity between the desire to avoid anxiety caused by evidence that the spouse is being unfaithful and the desire to avoid the uninteresting conversation one has been idly involved in. In avoiding anxiety caused by evidence that the spouse is having an affair, the evidence must be successfully repressed, which is very difficult given that what needs to be repressed are things that cause us distress and therefore preoccupy our attention; whereas, in focusing on an interesting debate, the person may still remain aware of the previous boring conversation in the background. Awareness of the boring conversation does not prevent you from focusing on the interesting debate; by contrast, the evidence that my thesis is unremarkable surely prevents me from observing evidence that the thesis is brilliant. We should not think that this evidence can be filtered out in the way in which a boring conversation or useless information can be filtered out.

In short, there are four vital differences between the two kinds of case: (1) the information or the belief that my thesis is unremarkable must be repressed, (2) the item that must be repressed causes me significant distress and (3) preoccupies my attention, and, vitally, (4) I have clear evidence that the interesting debate is taking place but no evidence that my thesis is brilliant. That is, in the uninteresting conversation situation, I have somewhere to turn my attention to (the interesting conversation), whereas, in the *My Brilliance* case, I do not – my thesis is not brilliant.

What I am saying is that I can non-problematically turn my attention away from the uninteresting conversation that annoys me to the interesting debate, because the interesting debate can non-problematically coexist in my mind with continued awareness of the

[31] It may be somewhat self-deceptive to regard this behaviour as tropic, i.e., purposeful but *uncontrollable*, a point owed to John Bishop.

uninteresting conversation – it would not be possible, of course, to pay focused attention to both conversations. Conversely, I cannot, if my mind is unified and reasonably coherent, non-problematically repress the evidence that my thesis is unremarkable – the evidence is in direct contradiction with my endeavour to form the desired belief. The same goes for Hillary and Maria. Because it is in direct contradiction with the desired belief, the evidence also makes the reward – namely, the thesis actually being brilliant, or Bill or Arnold not having an affair – is not a real possibility. Only an overall incoherent but functionally compartmentalised mind can successfully repress this kind of information by way of triggering mental tropisms.

My concern with this account, then, is whether we need it in the first place: while it does not require sub-agents (rational centres of agency within the mind), the theory based on mental tropisms does require some kind of partitioning of the mind. For a successful repression of evidence *E*, it is not enough that I turn my head away from *E*, *E itself* must be repressed – it needs to be functionally separated from my consciousness. Johnston argued that because repression is not intentional, it does not require the work of a censor, a sub-agent whose job is filtering information. This reply saves him only from the objection that tropisms require the Freudian conception of the mind. However, he fails to avoid positing functional compartmentalisation altogether and the reason is that, while repression need not be intentional and so we need not posit a switchman who will direct the content – evidence, beliefs, information, etc. – into sub-consciousness, the content still must be rendered inaccessible to the main agent and we need functional compartmentalisation for that.

In fact, tropisms seem to be essentially compartmentalising.[32] Let us amalgamate Pears's [1982] and Davidson's [1982, 1986] theories of the partitioned mind. Pears [1982] argued that in self-deception the information and the belief are being repressed, the walled-off content must be unconscious, and Davidson [1982: 171] described (italics added) 'parts of the mind [which] are in important respects *like* people, not only in having (or consisting of) beliefs, wants, and other psychological traits, but in that these factors can combine, *as in* intentional action, to cause further events in the mind or outside it.' We now get that, in self-deception, the information and the belief are non-intentionally walled off, but it looks *as if*

[32] Hales [1994: 283] offers an analogy with a 'tidal pool, with many different forces at work, and many different bits of flotsam within' but the 'forces at work' seem to be nothing other than subsystems.

everything was intentional, and then the favourable belief is generated as in wishful thinking. This is just like Johnston's theory of mental tropisms.[33]

The concern that tropisms require an incoherent mind is quite significant. The reason why we have deflationism in the first place was that we were unable to explain self-deception as involving the intent to deceive using the conception of a unified and reasonably coherent mind. Johnston himself says that the reason why he offers his account is to show that the traditional model, and the appeal to subsystems and compartmentalised mind, the *Traditionalist's Trap* as I call it, is unnecessary. And yet, he himself is unsuccessful in avoiding that trap.[34] Also, this kind of behaviour is not irrational in the way self-deception is supposed to be irrational: since my action is subintentional, it is difficult to see how *I* have violated my own norms of rationality and belief-formation.

4. Problems of Traditionalism and Deflationism

According to my analysis from II-1, both traditionalist and deflationary theories are based on incorrect accounts of deception. In my discussion in subchapters III-1 and III-2, I have uncovered more problems with two received perspectives on self-deception.

On the one hand, the traditionalists either explain trivial cases of self-deception or fall into the Traditionalist's Trap, i.e., they must divide the mind, and, by dividing the mind, they end up in deflationism. I intentionally acted, i.e., I averted my attention, and thereby affected a mental division but, by intentionally averting attention, I misled myself unintentionally. On the other hand, deflationists, who should have the upper hand of not being burdened by the demanding theory of interpersonal deception, typically either talk about innocent cognitive errors, having unjustified beliefs, that may trivially count as self-deception or they relapse and fall into the Traditionalist's Trap.

In short, either approach either offers a trivial account, or falls into a trap, or ends up in the rival camp. With these thoughts, which tell us why most contemporary philosophers are revisionists, I proceed to present revisionist theories.

[33] Pears [1991: 399–400, 402–403] argued against Johnston in a similar way.
[34] The same objection applies to Fingarette's [1998] view.

III-3 Revisionist Theories

1. Introduction

The first basic way in which self-deception can be reconceptualised is by positing that the relevant 'deception' is, or somehow involves, false self-knowledge. I dub this 'the deception *about* the self' approach. Three things need to be said about this approach. First, it is heavily dependent on the thesis that the belief I have about myself – my personal features or some other (first-order) belief of mine – is false. Second, seeing that self-deception is essentially deception about the self, some proponents of this approach are not intentionalists: on their view, being in the state of having false beliefs about yourself is self-deception. However, again seeing that self-deception is essentially deception about the self, some proponents of this approach take it that the self-deceiver has arrived at this state by drawing a valid, non-motivated, inference from false premises. The person is self-deceived because her action caused that she now has false beliefs about herself. I have dubbed such views 'hybrid.'

The second basic way of reconceptualising self-deception is by combining the deception about the self thesis with the intentionalist approach; that is, the person has arrived at the state of being deceived about herself as a result of a motivated action. Proponents of this view typically posit that the product of self-deception is not a first-order belief or even a belief at all. I dub this 'the failed attempt at deception *by* the self' approach – since the person acted so to deceive herself under one of the two intentionalist descriptions (traditional or deflationary) but she ended up with a mental state she was not aiming for. I will start by presenting some theories I categorise as belong to the deception about the self view.

III-3a. Deception *about* the self

1. Introduction

On this view, self-deception is, or necessarily involves, a failure of self-knowledge, which amounts to being deceived *about* the self. The focus is on the idea that the self is in some sense the object of self-deception, rather than an active subject that performs the action of deceiving. There are two main ways to understand the thesis that the deception is about the self, the *internalist view* and the *externalist view*, both of which rely on the hard-line conception of deception, according to which in deception you must hold a false belief.

Accounts belonging to the internalist view take deception about the self to be a matter of internal psychology. Most revisionist theories – even the hybrid ones that combine this revisionist thesis with deflationism or traditionalism – are internalist; these are the views of Jordi Fernández, David Patten, and Richard Holton I inspect below. This approach also seems to be predominant in psychiatry and psychology and, there, it would probably not be considered as revisionist but traditional. For psychiatrists and psychologists, the relevant failure of self-knowledge, as implied by Gur and Sackeim's influential Self-Deception Questionnaire (SEQ), concerns those mental states – desires, memories, preferences, etc. – that cause us some distress.[35] In short, I am self-deceived if I believe that I do not know that I have (or that I have had) a certain desire, memory, or preference, and my state of ignorance is motivated (see Gur and Sackeim 1979: 149).

According to the *externalist view*, self-deception is a social phenomenon that has to do not so much with the person's mental states as it has to do with her roles and relationships in society. Often, self-deception is forced by social pressure onto the individual who cowardly accepts it. In this camp, we find most prominently Sartre with his theory of bad faith (similarly, Solomon 2009).

[35] This questionnaire consists of 20 questions to which one should respond on the scale from 1 to 7, 1 being a strong negative response and 7 being strong positive – 1 and 2 are in the range of self-deception. Some of these questions are: 'Have you ever felt hatred toward either of your parents?,' 'Have you ever made a fool of yourself?,' 'Have you ever thought that your parents hated you?,' 'Have you ever been uncertain as to whether or not you are homosexual?,' 'Have you ever enjoyed your bowel movements?,' and so forth. I would treat with considerable caution the idea that this questionnaire is an indicator of self-deception (see 2.2).

2. Internalism

The first two accounts that I discuss understand self-deception as a state, not as an action. Self-deceivers are not motivated to end up in this state and the result is non-intentional. Both theories propose that the relevant deception is not in what we believe but in what we believe that we believe; a self-deceiver is deceived into thinking that she believes that ~*p*. The difference between these two theories is only in that, on the first account, the one by Jordi Fernández, one deceives oneself in believing that one believes that ~*p* by directly focusing on the state that should cause both the belief that *p* and the belief that one believes that *p*, while on the second, proposed by David Patten, one deceives oneself in holding that one believes that ~*p* by interpreting one's own behaviour. The third account, proposed by Richard Holton, is a hybrid one: it combines the deception about the self view with deflationism.

2.1 Jordi Fernández

Fernández [2013] presents a very interesting non-intentionalist theory of what he sees as a type of self-deception constituted by a specific failure of self-knowledge. In this kind of self-deception, the person violates the norm that prescribes that 'One should not form a belief in the face of evidence against it unless there are overriding considerations in support of that belief' (Fernández 2013: 396).[36] The relevant irrational step – i.e., the violation of the relevant norm – is not in forming a first-order belief (as in Davidsonian self-deception, e.g.) but rather in associating that belief to oneself; namely, a self-deceiver violates this norm and forms a belief that he believes that ~*p* even though he believes that *p*. I will explain this theory step by step.

We see that the theory relies on the concept of meta-beliefs, which are (second-order) beliefs about which beliefs we have. Before I proceed, I will briefly explain what meta-beliefs are. It is commonly thought that people do not only form their beliefs as a result of being in certain states; rather, being in certain states normally also results in forming meta-beliefs about those beliefs. We may thus say that we keep track of what we believe by way of having meta-beliefs; they constitute a database about our beliefs.[37]

[36] This norm is somewhat problematic: what counts as overriding considerations in support of the belief is left unspecified. Are these considerations non-evidential and, if they are, why would they override the evidence? And, if they are evidential, then why take it that the weight of the evidence is against the belief? (The point owed to John Bishop.) I leave this worry aside, however.

[37] The concept of meta-beliefs may be somewhat problematic. Hamilton [2000: 32], for instance, argued that proponents of this concept 'cannot explain what "I believe that I believe that *p*" is supposed to mean. It takes its

Fernández bases his account on the following aetiology of meta-beliefs. A person that forms a (first-order) belief as a result of being in a certain state is disposed to believe that she is in that state – this is what having the belief about that state just is. Not only that, the same state also causes the second-order belief, the belief that I have the belief that I am in that state. Accordingly, if I have the belief that *p* as a result of being in a state P, I typically also have the relevant meta-belief *Bp* ('I believe that *p*') about the belief that *p* by which I ascribe the belief that *p* to myself, and both beliefs are based on my being in the state P. Looking at an apple, for example, ideally causes both the belief that I see an apple and the belief that I believe that I see it. Nonetheless, P may cause the belief that *p* but fail to cause the belief that *Bp*. This is why people sometimes have to investigate what beliefs they have.

We can conduct this investigation from two perspectives: the third-person and the first-person perspective. The third-person perspective is nothing other than me observing my behaviour and performing some inferences from my observations (Patten [2003: 232–233]; 2.2). In the first-person perspective, I 'look past' my beliefs in order to self-ascribe them and observe the grounds that I have for those beliefs, i.e., the states that generated them (Fernández 2013: 390). In short, from the first-person perspective, I will form the belief that I believe that *p* by directly observing the state P. In particular, if asked whether I believe that I see an apple in front of me, I will not 'scan' my own mind or analyse my behaviour but rather just look at what is in front of me (similarly, Evans [1982: 230], Peacocke [2000: 71–72, 75]). This process of forming beliefs about beliefs Fernández names *bypass* and his account is of a kind of self-deception that occurs through bypass.

The important idea is that, if I have investigated my beliefs and I believe that I believe that *p*, then I have established through bypass that the state P triggers the belief that *p and* that I am in a relevant way in the state P (Fernández 2013: 393). Yet, meta-beliefs are beliefs and thus they can be false. Just as I could falsely believe that I am not sick (~*s*), despite the fact that the state S justifies the belief that I am sick (*s*), I may, based on S, form the belief that *s* but not the (meta-)belief that I believe that *s* (*Bs*). Because I have no relevant second-order beliefs (I just lack a relevant belief, I do not believe that I do not believe that *s*), this situation counts as a cognitive omission, but not as self-deception – as a hard-liner about deception (see II-1/2.1), Fernández thinks that you must have a false belief in order to be deceived. Nonetheless, you are one step closer to becoming self-deceived.

---

sense entirely from the third-person or mixed case ("I believe that Jonathan believes that I believe …" etc.), or from cases involving other tenses ("I believe that I believed that *p*").' I leave this issue aside as well.

Suppose that I do not have grounds to believe that I am sick. In fact, the state ~S gives me grounds to believe that ~*s*, and that I form this belief (~*s*) but not the relevant meta-belief. In this situation, I believe that ~*s*, 'I am *not* sick,' but I do not know or believe that I believe it. Subsequently, if, through the process of bypass, I form the false (meta-)belief that I believe that 'I *am* sick' (*Bs*) based on the state ~S, while believing that 'I am *not* sick (~*s*)' (also based on ~S), I have become a self-deceiver according to this view; I have a false belief, the (meta-)belief that *Bs*. In brief, a self-deceiver correctly believes that he is *not* sick and, based on the same state that causes this belief, incorrectly *(meta-)believes* that he believes that he *is* sick. Although I did not intentionally cause myself to mistakenly believe this, nor was I motivated to believe it, this is *deception* because the belief is false and I violated the principle of belief-formation and it is *self*-deception because 'the self [is] the subject matter, or the object, of self-deception' (Fernández 2013: 395).

Fernández proposes that his account has the potential to explain some of our cases. For instance, it should be able to easily explain the Nicole case in which Nicole sincerely asserts that Tony is not having an affair with Rachel but avoids driving by Rachel's house – even when it requires her to drive out of her way – when there is a chance of Tony's car being parked in front. Nicole's non-verbal behaviour contradicts sincere avowals and, on the proposed account, this is because she believes both that Tony is having an affair with Rachel (*a* for 'affair') and that she believes that she believes that they are not having an affair (~*Ba*) on the grounds of the state A (I try to identify this state below). She is a self-deceiver because she believes that *a* and that ~*Ba* (through bypass) on the grounds of A. Since Nicole should have formed the belief that *Ba*, rather than ~*Ba*, based on A in the same way she formed the belief that *a*, she has exhibited the relevant failure of self-knowledge. Therefore, her (meta-)belief is epistemically unjustified in the way that makes her a self-deceiver.

This is an elegant solution. It pinpoints Nicole's non-motivated cognitive error but it also tells us why she avoids driving past Rachel's house: she is acting on her belief that he is having an affair. Furthermore, because meta-beliefs are shallow, i.e., immediately accessible to consciousness, she will sincerely report on her meta-belief. Additionally, the tension is between two kinds of beliefs, not between two contradictory beliefs of the same kind, and thus the doxastic paradox does not arise. Finally, even though she does not intend to deceive herself and while the deception is not a motivated action, Nicole is blameworthy because she has sufficient grounds for believing that she believes that Tony is having an affair with Rachel.

However, notwithstanding its elegance, I have two considerable concerns about this interesting theory. The first is that, since 'having a false belief' is a figurative sense of 'being deceived,' it can only be proposed as a theory of a kind of self-deception, figuratively speaking. Incidentally, it is really hard to see how Nicole is not motivated – it is her husband we are talking about – but this is a minor concern. My second substantial concern is that this theory falls into the Traditionalist's trap. In particular, the state P can cause the belief that *p* but, at the same time, fail to cause the belief that I believe that *p* only in a highly incoherent, even dysfunctional mind. Let us take Nicole's case. First we need to identify the state that causes the belief that *a* ('Tony is having an affair with Rachel') but fails to cause the belief that *Ba* ('I believe that I believe that *a*').

Arguably, we should identify A in the same way we identify the state S in the analogical case of forming the belief that I believe that I am sick. It cannot be that, in normal cases, my perception of what people take to be symptoms of being sick generates the first-order and second-order beliefs directly in the way in which my perception of an apple renders the relevant beliefs directly. Recognising the symptoms of being sick is the result of an inference to the best explanation (I might be tired or suffering from a hangover), whereas recognising an apple is a result of simple matching cognitive predictions with perceptual input. Therefore, the relevant state S must be my judgement that I must be sick, or that I have the symptoms of being sick, rather than my mere perception of the symptoms. That is to say, the state S is the state of having the judgement that *s* effective in my mind. Bypass involves only accessing the judgement *s*.

Analogously, the state A ('affair') is the state of having the judgement that Tony is having an affair with Rachel. It follows, on the proposed theory, that self-deception through bypass is possible when the judgement causes the belief that *a* but not the belief that *Ba*. Self-deception will, in turn, occur when Nicole, by looking directly at the state A, forms the belief that *~Ba* while believing that *a*. However, if the state A is having the conscious judgement that *a*, then it is hard to see how the state A can cause the belief that 'Tony is having an affair with Rachel' (*a*) and fail to cause the belief that 'I believe that I believe that Tony is having an affair with Rachel' (*Ba*). And it is much harder to understand how Nicole's conscious judgement that *a* could ever cause both the belief that *a* and the false belief that *~Ba* in bypass (in which she is accessing A in order to see whether she believes that *a*), if Nicole's mind is not functionally severely divided.

In effect, we are back to the original problem: we have fallen into the Traditionalist Trap, i.e., we need to posit a substantial mental dissociation, and this takes away the principal reason for not being a traditionalist about self-deception. That said, I leave this argument and move on to a similar argument based on the first strategy of bypass: identifying what I believe by observing my behaviour.

2.2 David Patten

Patten's non-intentionalist account of self-deception differs from the one considered above only in that the meta-beliefs about which we are self-deceived on Patten's account are formed from the third-person perspective from which we form beliefs about our beliefs by analysing our behaviour. Additionally, Patten is principally concerned with people who are deceived (*fig*.) about their motives; i.e., they have false beliefs about what motivates their behaviour.[38] For example, if, in the case of smoking motivated by the fear of quitting (or failing to quit), a person makes an error in inference from the perceived lack of alternative explanations for her behaviour to the false assumption that she is smoking because she enjoys smoking, then this person meets the criteria for being self-deceived about her motive for continuing to smoke (Patten 2003: 233).

Self-deception, *qua* self-deception (not *qua* analysis of my behaviour), is not a motivated action, according to Patten. The smoker was not motivated to reach the conclusion she got. What causes the error in reasoning is, in addition to the failure to see her own fear of quitting, her stereotypical idea of herself as someone who, e.g., pursues only those activities that are enjoyable. Let's say that I am the smoker in question: I am self-deceived because I was wrong about myself. Specifically, I 'failed to notice' (Patten 2003: 242) the real reason, i.e., the real motive, for my continuing to smoke and, by analysing my behaviour, I have formed a false (meta-)belief about that motive.

As with many other accounts, this account is problematic principally because it relies on the figurative meaning of the term 'deceive' as 'having a false belief' and, by excluding the intention to deceive, it reduces self-deception to an innocent error of self-knowledge. In the above case, I have drawn my conclusion from an argument that involves false premises or lacks an important premise, but my inference was rational considering the information

[38] Around the same time Patten published this paper [2003], Borge [2003] proposed a view according to which a person's failing to see which mental states, specifically emotions, guide her behaviour should not be understood as self-deception. Borge does seem to have a point here, these are not cases of self-deception, but his further argument that the theoretical notion of self-deception should be abandoned is a *non sequitur*.

available to me. Besides, I am not sure that we can say that I am not motivated into believing that I am smoking because I enjoy it; this belief clearly preserves my self-esteem whereas the correct belief would undermine it severely. Hence, the claims that the inference is not motivated and that the action is non-intentional seem *ad hoc* and implausible.

Finally, and this is the reason why I discuss this proposal, we should ask ourselves what gives Patten the right to say that *I* do not know whether *I* am enjoying smoking or not. He cannot taste the flavour of the cigarette *I'm* smoking better than me and *I* obviously think that I enjoy it. This concern shows that explanations such as this – and, to a certain degree, *all* meta-beliefs solutions, including Fernández's – fall into what I call the *Freudian Trap*, namely, explaining a certain pattern of (possibly) problematic behaviour by offering a pseudoscientific, unfalsifiable theory. A relevant example of this trap can be found in a discussion of one of Freud's cases from his *Introductory Lectures on Psycho-analysis* [1916–1917]. While defending his position from an important objection that says that dreams cannot help us understand our mental life in a way that we should say that they reveal our unconscious mental life, he lists some interesting reactions from his patients who disproved his diagnosis of their behaviour. Consider the following (Freud 1929: 121):

> A woman dreamer says: 'Am I supposed to wish that my husband were dead? Really that is outrageous nonsense! Not only is our married life very happy, though perhaps you won't believe that, but if he died I should lose everything I possess in the world.'

Freud's answer to the woman's objection that his interpretation is unjustified is very interesting (italics added).

> *Assuming that* unconscious tendencies do exist in mental life, the fact that the opposite tendencies predominate in conscious life *goes to prove nothing*. Perhaps there is room in the mind for opposite tendencies, for contradictions, existing side by side … What does it matter if *you* [in this case, the woman in question] do find the results of dream-interpretation unpleasant, or even mortifying and repulsive? '*Ça n'empêche pas d'exister*' ['*It doesn't prevent things from existing*'] – as I, when a young doctor, heard my chief, Charcot, say in a similar case [1929: 122].

What Freud practically says is 'It does not matter what you say; I am right either way. If you deny my diagnosis, this is because my diagnosis got your unconscious mental state correctly.' However, this amounts to saying 'I know you better than you know yourself and I know this by analysing your behaviour *from within the perspective of my theory*.' There is a dangerous and obvious circularity in this. We do not know whether his theory is correct – in fact, this is exactly what is at stake. Bandura [2011: 16], for instance, nicely objects, and

rightly so, to Von Hippel and Trivers's [2011] theory that posits that the deceived mind is dissociated from the unconscious mind that knows the truth; he asks 'How does one know what the unconscious mind knows?' With his manoeuvre by which he says that conscious life cannot disprove the existence of unconscious purposes (granted that they exist!), Freud has transformed this case from a counterexample to his view to a case that supports it (see also Derksen 2001: 333–334). In the same way, anything I say in defence of my belief that I smoke because I enjoy it will be dismissed in view of Patten's theory; yet, I have no reason to trust Patten rather than myself and nor do you.

Now, I may think that, if my behaviour is a result of self-deception, then I am in no better position to interpret my behaviour than any other person (e.g., Borge 2003: 18). However, to think that is to engage in circular reasoning once more: thinking that I am wrong in saying that I smoke because I enjoy smoking is the reason why you say that I am self-deceived in the first place; hence, you cannot defend the hypothesis that my belief about my motive is wrong by claiming that I am self-deceived.

More importantly, even if this was the case and I was self-deceived, a theory that does not fall into a trap but explains my behaviour equally well should be preferred to the one that falls into a trap. The problem with falling into a trap, as I have been using this notion, is not that the account straightforwardly fails but rather that falling into a trap comes with a price: getting out of that trap minimally requires additional (typically, *ad hoc*) theoretical apparatus, which unnecessarily complicates the theory. Moreover, some theories cannot get out of their traps without being severely harmed: traditionalists, we saw, leave their trap as deflationists.

2.3 Richard Holton

Holton proposes that deception about the self is a necessary condition of self-deception that needs to be added to Mele's list of jointly sufficient conditions. Because it combines this revisionist thesis with the deflationary approach, the resulting account is hybrid; I call it the Holton-Mele account (H&M henceforth). H&M is an interesting hybrid account. Most contemporary philosophers are revisionists about self-deception and most of them say that some sort of error about the self is responsible for the internal conflict self-deceivers are experiencing (see III-3b): we saw that Nicole's sincere avowals contradict her non-verbal behaviour. By contrast, H&M is designed to explain the type of cases similar to the Maria case, which does not involve any internal tension. Here is one such often discussed case.

Mario: Suppose a man, let us call him Mario – contrary to all the evidence – deceives himself into believing that his wife, Arnoldine, is not having an affair.

Any explanation of cases such as this, says Holton [2001: 55–56], must start by noticing that a person is never deceived exclusively about one belief, but always about some complex state of affairs, i.e., a subject matter, a topic (also Szabados 1974: 57). That is to say, Mario cannot be deceived exclusively in believing that Arnoldine is not having an affair, i.e., that ~*p*; rather, his state of being deceived involves a whole set of beliefs concerning that subject matter and some of those beliefs must concern Mario himself – this is why Mario is also deceived *about* himself. For example, he likely believes that his wife is honest, that she is chaste, that she loves him, that he is a good husband, that he is basing his belief in her fidelity on correct standards, that the belief is justified, and so on. The belief that Arnoldine is not having an affair is only the one that is on the 'surface.'

However, not all false beliefs about the self need count as false self-knowledge of the kind relevant for self-deception. Aiming to identify the relevant errors, Holton distinguishes two kinds of errors of self-knowledge (italics added): 'those that result from the *application to oneself* of an erroneous belief that is *not* about oneself; and those that do not.' He names the latter kind of errors '*fundamentally de se* errors': thus, if Mario is self-deceived about ~*p*, his body of beliefs about *p* necessarily contains errors of this kind relevant for the question whether *p* (Holton 2001: 65).

I have been able to identify two key features of *fundamentally de se* errors and some that merely appear as such. One characterisation repeatedly used in the paper is that these errors are the kind of mistake made 'when people make mistakes about themselves in ways that, had they got it right, would have been self-knowledge' [2001: 65], but this cannot be a feature exclusive to *fundamentally de se* errors: any kind of error of self-knowledge is of the kind that, had they got it right, you would have self-knowledge. Rather, judging by the above-cited claim about what they are not, the first relevant feature of *fundamentally de se* errors is that they do *not* result from the application to oneself of an erroneous belief that is not about oneself. To this description, Holton [2001: 66] adds that these mistakes result from some intentional or at least culpable fault on the part of the self. When combined, these two theses suggest that *fundamentally de se* self-knowledge errors are those errors about the self that originate from how the self directly understands itself and it is the self's fault it ended up with these false beliefs.

Therefore, if Mario believes that Arnoldine is not having an affair because he is a mobster and, understandably, mobster wives are consistently faithful, then this topic is not a matter of self-deception despite the fact that the set of relevant beliefs contains false beliefs about oneself. The reason is that Mario's relevant erroneous belief about himself – that his profession somehow guarantees his wife's faithfulness – did not originate from how Mario directly understands himself. Rather, it is an error of self-knowledge caused by application of erroneous beliefs about mobster wives in general. Even Mario's false second-order belief that his belief 'Arnoldine is not having an affair' is true also involves a non-self-deceptive error of self-knowledge, says Holton [2001: 64–65], and the supposed reason is that the error does not 'begin at home,' namely, Mario is not culpable for it (one must believe that his belief is true).

So, what would be a genuinely *de se* error of self-knowledge, according to Holton? One kind is the error as to how justified his belief that ~*p* is; specifically, Mario's 'beliefs about the status of his own assessment of his wife's fidelity are part of his set of beliefs about his wife's fidelity' [2001: 56] and some of these beliefs are *de se* beliefs which are false. Then, there are Mario's erroneous meta-beliefs about what mental states he has; that is, he may be deceived about his motives, or desires, or beliefs, and so forth (Holton 2001: 55–56). Also, he may be mistaken about whether he is living up to his own belief-forming standards in forming the belief that Arnoldine is not having an affair [2001: 60, 64], and so on.

Having the relevant necessary *de se* failure of self-knowledge identified, Holton adds Mele's theory to his view. According to H&M (Holton 2001: 63, 65), S is self-deceived about a subject matter α if

1. S's body of beliefs about α contains fundamentally *de se* self-knowledge errors;
2. S treats data relevant, or at least seemingly relevant, to the truth value of these mistaken beliefs in a motivationally biased way;
3. This biased treatment is a nondeviant cause of S's acquiring these mistaken beliefs.
4. The body of data available to S at the time provides greater warrant for rejecting these beliefs than for accepting them.

The full H&M account, however, reveals some important concerns regarding Holton's proposal.[39] The first one is that Holton's contribution to H&M might be trivial: it follows from Mele's theory that the self-deceiver, who is a biased evidence gatherer, is going to form false beliefs about how justified the surface belief is and about how he lived up to his belief-forming standards in forming that belief. Furthermore, it is the agent's fault for having these

[39] I will not repeat here my earlier discussion of the shortcomings of Mele's analysis (see III-2/2).

false beliefs about himself: he obtained them by being biased (condition 3) and not by applying to himself erroneous beliefs that are not about himself. Therefore, Holton's added necessary condition seems in many cases to be implied by one of the already existing conditions.

The second concern is that Holton's examples of *fundamentally de se* errors of self-knowledge are not uncontroversial. He says that, if Mario is mistaken about whether he is living up to his own belief-forming standards in forming the belief that ~*p*, then this is a *de se* error of self-knowledge but that, if he falsely thinks that his own standards will lead him to true beliefs, then this is not a *de se* error [2001: 60, 64]. In *de se* errors, the explanation is, 'the mistake must be the result of some intentional or at least culpable fault on the part of the subject: the self deceives itself' [2001: 66, see 57–58]. But, what makes Mario *culpable* for his failing to see that he did not follow his own standards of belief-formation but not for his falsely believing that his own standards will lead him to true beliefs? The reason is that, says Holton [2001: 65], Mario first 'formed a general belief about which standards led to the truth, and then applied them to his own practice,' which resulted in forming the belief that his own standards will lead him to true beliefs. His mistake did not 'begin at home' and thus we should not blame Mario for it. This explanation is very problematic, as I show below.

Firstly, Mario need not form a general belief about which standards led to the truth and then apply it to his own practice. Mario could, due to having high self-esteem, think that he is an excellent reasoner and then apply this to this particular practice. In this case, he surely is culpable for forming the belief that his standards will lead him to the truth and the error 'began at home;' therefore, this should be the relevant *fundamentally de se* kind of error. Secondly, failing to see that he did not follow his own standards of belief-formation may not be his fault; it could be a result of absent-mindedness or of some other general belief about what it is to satisfy certain standards that he applied to his behaviour. Therefore, this error need not be a *fundamentally de se* kind of error. The examples do not seem to fit the theory.

Maybe we could try with different examples? Holton thinks that that one actual error is how justified the surface belief is [2001: 55], but it might be better to say that the *fundamentally de se* error is of the kind that *gives rise* to a mistaken estimate of the justificatory status of the surface belief. This suggestion is consistent with the main idea that the agent is culpable for these errors and that errors do not arise from the application to oneself of erroneous beliefs that are not about oneself. In welcome self-deception, examples

of such beliefs would be '*I* would have noticed the change in Arnoldine's behaviour,' or 'Bad things do not happen to *me*,' '*My* wife cannot be having an affair,' and the like.

Let us suppose that Mario's belief that *he* would have noticed Arnoldine's infidelity (*q*) will lead him into falsely supposing that his surface belief that ~*p* is justified. Because Mario is culpable for believing that *q*, this belief involves a relevant *fundamentally de se* error of self-knowledge. On this modified view (modifications in italics), S is self-deceived about a subject matter α if

1. S's body of beliefs about α contains *fundamentally de se* self-knowledge errors;
2. *These errors causally contribute to* S treating data relevant to the truth value of these mistaken beliefs in a motivationally biased way;
3. This biased treatment is a nondeviant cause of S's acquiring these mistaken beliefs.
4. The body of data available to S at the time provides greater warrant for rejecting these beliefs than for accepting them.

When we apply this account to Mario's case, it tells us the following. Because of (1) relevant *de se* errors of self-knowledge – namely, his belief that *q* – Mario (2) treats the data relevant to the truth value of the belief concerning Arnoldine's fidelity in a motivationally biased way (condition 2 makes him culpable for self-deception). This biased treatment of data, then, is (3) a nondeviant cause of his acquiring of the belief that Arnoldine is not having an affair (~*p*), (4) despite the fact that the body of data available to him at the time provides greater warrant for rejecting beliefs that *q* and that ~*p* than for accepting them. Therefore, since all four conditions are met, Mario has entered the state of self-deception.

Interestingly, this theory has the potential to explain cases in which the surface belief is true or unwelcomed. Mario is self-deceived as long as his surface belief is supported by a *de se* error of knowledge. It is irrelevant whether the surface belief is true or false, welcome or unwelcome. For example, due to the belief that *he* could not have done anything to jeopardise his marriage or that bad things always happen to *him*, Mario may selectively focus on evidence that suggests that Arnoldine *is* having an affair.

However, the main objection I have to the H&M theory is that it fails to avoid the same traps and problems as Mele's original theory.

For instance, it still falls into the Deflationary trap of making self-deception irrational only from the third perspective (see III-2/2). That is to say, we take Mario to be self-deceived about α (the subject matter of Arnoldine's fidelity) because we think that his *de se* beliefs are

false and that, therefore, the body of data available to Mario provides greater warrant for rejecting all the beliefs belonging to α than for accepting them (condition 4). However, most spouses are justified in believing that he or she would have noticed the change in his partner's behaviour. Thus, conditions 1 and 4 are problematic. Moreover, and this is vital, even if the beliefs '*I* would have noticed' or 'I couldn't have done anything wrong,' Mario does not seem to be a self-deceiver, a person that is being insincere towards himself, but rather a person who fell victim to his own egoism or naivety.

In conclusion, Holton is right to draw our attention to the fact that self-deception concerns a subject matter and yet his theory is problematic, and I was unable to modify it so that it does not raise any concerns. The main issue is that Mario does not violate his own norms of reasoning and he is not being insincere towards himself. It is just that he inadvertently accepts some implausible premises. With this thought, I proceed to Sartre's externalist theory.

3. Externalism

3.1 Jean-Paul Sartre

Self-deception, or *bad faith* (*mauvaise foi*) as Sartre calls it, can in brief be described as negating oneself and being what one is not, that is, existing in the mode of what one is not (similarly, Solomon 2009). On this view, I deceive myself when I do not accept myself as who I really am. Sartre [1978: 66] writes 'And what is the goal of bad faith? To cause me to be what I am, in the mode of "not being what one is," or not to be what I am in the mode of "being what one is".' The general method of deceiving in bad faith is as we traditionally understand deception: bad faith is a lie to myself (Sartre 1978: 148).

Sartre approaches bad faith from the perspective of phenomenological ontology: he is analysing the foundation of our being by analysing phenomena in which our being actualises itself. The rationale is that, by analysing how we actually are in the world, i.e., by understanding our being-in-the-world, we can understand the nature of our true being and by analysing bad faith, we discover what our being is not. Analytical accounts of self-deception typically analyse the ontic dimension of human existence: interpersonal or intrapersonal relations. In contrast, Sartre inspects the level of becoming a person, the ontological foundations of a person as herself. Accordingly, denying oneself should be understood not as the kind of denial that may be made in avowals (*pace* Lockie 2003: 137), which is how, for instance, Hillary denies Bill's infidelity, but as an *ontological stance* towards one's own

reality. It is refusing to accept who I 'am.' In order to understand bad faith, we first need to know what I am.

Human freedom is in being free to constitute oneself. I am a constant potentiality, becoming – this is why existence precedes essence on Sartre's view. The existence of others is necessary for my existence (his account is externalist because the self is socially established); however, becoming myself requires 'not allowing myself to accept the reality of others as *my* reality.' My consciousness, says Sartre [1978: 47], 'constitutes itself in its own flesh as the nihilation of a possibility which another human reality projects as *its* possibility.'

Allowing myself to accept the reality of others as my reality is the essence of bad faith: the self-deceiver negates his own reality by accepting another human reality, which this other human reality projects as *its* possibility. In a sense, in bad faith, I steal what is not mine and thereby become who I am not.[40] Bad faith, we may say, is a kind of an ontological schism, where two properties constitutive of the self fall apart. To be precise, humans have the double property of being at once *facticity* and *transcendence* (Sartre 1978: 56). These two aspects of human reality should be in synthesis, but bad faith wants to affirm their identity while preserving their differences, which leads to separating the self from itself by way of exercising a power (of transcendence) integral to the self as becoming.[41]

Suppose that a young woman is in the company of a young man who gently puts his hand over hers. This woman is in bad faith, says Sartre [1978: 55–56], when 'she *permits* herself to enjoy his desire' by leaving her hand there *without noticing* that she is leaving it. She does not notice this because she is at that moment 'all intellect.'[42] This woman wants to enjoy his desire for her without needing to take any responsibility for that enjoyment. And, by becoming all intellect, thus effectively separating her facticity from her transcendence, she has disarmed the significance of the actions of her companion by reducing them to being only what they are, mere actions. In this way, she makes it clear to herself that her encouraging

[40] The Other is my hell in the sense in which the Other – by actualising itself and projecting its reality into the world – gives me a possibility to accept its reality as mine and thereby negate myself.

[41] Golomb [1995: 108] notices that Sartre distinguishes three basic negative attitudes towards oneself: regarding oneself as pure transcendence, denying one's transcendence completely (ironically, denying transcendence is also transcendence), and treating oneself as other. I here discuss the first of these attitudes.

[42] Becoming 'all intellect' should be understood as redirecting one's attention or making oneself preoccupied with own thoughts in order to avoid thinking about a certain distressing memory or perception; basically, it is a strategy of denial.

response to his tentative 'pass' is no such thing at all. This attitude has fundamental ontological consequences: it degrades her into a lower, less authentic, mode of being.

Sartre distinguishes between two ontologically primitive types of being: *being-for-itself*, which is conscious being, and *being-in-itself*, which is an object of consciousness that has no consciousness of its own. Nothing can have, or be, both types of being at the same time (Golomb 1995: 105–106), which is to say that nothing can be both a conscious being and a thing. Consequently, seeing that the woman has disarmed this man's behaviour by reducing it to mere actions in order to *permit* herself to enjoy his desire, she has degraded her own mode of existence into existing as a being-in-itself, a thing. This thought may seem to be a *non sequitur* – after all, the woman has reduced the man's behaviour to mere actions by becoming 'all intellect,' not a thing – but this is not the case. For Sartre, pure consciousness is a being the nature of which is to be conscious of the nothingness of its being; therefore, intellect is nothing without facticity. In fact, intellect can never be mere intellect. It always has to be transcendence *within* facticity. And, since pure intellect is nothingness, the only reality left to be affirmed is the reality of a thing – a mere hand being held by another mere hand, a thing touching a thing.

The kind of transcendence involved in bad faith is a perversion of human nature. By negating the facticity of my reality in bad faith, 'I affirm that I *am* my transcendence in the mode of being of a thing' (Sartre 1978: 164). This is a denial of facticity, i.e., of the desire to continue enjoying the man desiring her; it is an ontological detachment from facticity in order for her to indulge her desire without authentically acknowledging it. Her hand can remain being held by the man's hand because, without her own desire acknowledged, this state of affairs is no longer a displeasing truth from her 'nice young woman's' perspective. However, by solving the problem of the desire that is unwanted from that perspective, she has perverted her own reality. Human reality is freedom – 'man is nothing other than what he makes of himself' (Sartre 2007: 22) – and, by not being able to be free frankly to enjoy the man's desire, to be the author of her enjoyment, the woman ceases to be herself.

The final question is why she did it. Why did she negate herself just so as to enjoy her desire being satisfied? Actually, why was her desire a displeasing truth in the first place? Why do people in bad faith accept someone else's reality rather than nihilate it? Sartre's [1978: 59] answer to all these questions is 'the public demands it.' The public's demands are the reason why it is difficult to live with freedom, which with itself brings responsibility. The

reality which people in bad faith embrace is the reality imposed by the public and this reality is wholly one of ceremony.

Self-deceivers are cowards; they back down under pressure. This woman is not free to enjoy her desire. Instead of creating herself, a bad faith self-deceiver 'has accepted the role his family and heritage have imposed on him … In deluding himself into believing that he is freely creating and choosing, [he] is playing a sophisticated game rather than actually proceeding towards authenticity' (Golomb 1995: 103). The society demands that we limit ourselves to our functions of a grocer, a well-mannered girl, a father, and so on. Sartre gives us an example of a waiter in a café. This man moves quickly and precisely, his voice and eyes express an interest a little too solicitous for the order of the customer. All of his behaviours seems to us as a game, his voice and gestures seem to be mechanisms, and this is because he is playing *at being* a waiter in a café (Sartre 1978: 59) just as the abovementioned woman is playing at being a woman who nonchalantly accepts a mere depersonalised gesture.

Finally, bad faith is 'bad' because, rather than being who I want to be, I am playing the role of someone who I do not want to be; the role imposed on me by the society, by my homeland, my ancestors, and so forth. And this is the core of self-deception as the bad faith type of deception about the self. Sartre [1978: 60] writes

> And it is precisely this person *who I have* to *be* (if I am the waiter in question) and who I am not … [T]here is no common measure between his being and mine. It is a 'representation' for others and for myself, which means that I can be he only in *representation*. But if I represent myself as him, I am not he … I cannot be he, I can only play *at being* him; that is, imagine to myself that I am he. … In vain do I fulfil the functions of a cafe waiter. I can be he only in the neutralized mode, as the actor is Hamlet … Yet there is no doubt that I *am* in a sense a cafe waiter … But if I am one, this cannot be in the mode of being in-itself [a thing]. I am a waiter in the mode of *being what I am not*.

I am not sure that this account deserves any kind of criticism as an account of bad faith but it must be noted that even if bad faith is a type of self-deception, the problem of how this action is possible re-emerges on the internalist level. Specifically, seeing that bad faith is a lie to oneself, and the liar is in complete possession of the truth, i.e., he knows who he should be (Sartre 1978: 148), this view falls into the Traditionalist trap, the need to posit partitions in the mind. However, Sartre did not accept the typical traditionalist solution that posits divisions in the mind. What is more, he was a fierce opponent of it and, I would say, his criticism of the concept of dynamic consciousness – specifically the idea of the censor – is quite compelling.

Sartre [1978: 60] argues that Freudians unjustifiably assume that the censor can register unwanted information as unwanted and then repress it without being conscious of (i.e., know) what he is doing. Wood [1988: 209–213] replied to Sartre by arguing that one can unconsciously know that *p* and unconsciously know that one believes that *p*. Accordingly, the censor could unconsciously repress a belief, i.e., repress a belief while not being conscious of what he is doing. I do not find this reply compelling: it requires of us to think of a censor itself as having a divided mind. Wood tries to avoid this implication by saying that censor's acts might be *self*-repressing, but why not then say that the self-deceiver's act is self-repressing? That way, we need not posit the censor at all. In any case, bad faith falls short of self-deception for the same reasons all received traditionalist approaches fail – they generate paradoxes. Even more, the woman and the waiter are not really deceived about who they are; they are merely afraid to be who they are. Therefore, it may be better to say that bad faith is a kind of pretence (see III-3b/7), not self-deception.

4. A Short Summary

The main feature of this particular revisionist approach is that self-deceivers must have some false beliefs about themselves and that this is what makes them self-deceivers. However, my analysis suggests that this does not represent much of a step closer to a better understanding of the phenomenon – even though Holton's idea that self-deception concerns a subject matter (of which false beliefs about the self are a part) is a significant contribution to the debate. Theories from this family either, on the one side, fall into the Traditionalist trap, i.e., dividing the mind (Fernández, Sartre), or into the Freudian trap, i.e., being pseudo-scientific, or even both (Patten, psychological theories); or, on the other side, fall into the deflationary trap by making self-deception irrational only from the third-person perspective (H&M). So, maybe self-deception may be understood instead as a failed attempt of deceiving by the self? I explore this type of account in the next subchapter.

III-3b. Failed Deception *By* the Self

1. Introduction and Overview

Accounts that may be classified as belonging to this group of theories are principally concerned with the Nicole-type cases. It may very well be these are the most problematic cases of self-deception: Nicole sincerely avows that ~*p*, i.e., 'Tony is not having an affair with Rachel,' but nonetheless acts as if *p* is true, e.g., avoids driving past Rachel's house when Tony might be there, which seems to be indicating that 'deep down' she still knows the truth. This conflict of verbal and non-verbal behaviour presumably causes tension, which, according to accounts presented in this section, is the hallmark of self-deception (e.g., Audi [1985: 173–177; 1988: 94; 1997b: 104]; also Martin [1997]; Perring [1997]). Because, unlike Hillary, Nicole did not end up completely separated from the truth, I classify intentionalist accounts that try to explain this case as belonging to 'the failed deception *by* the self' family. The hallmark thesis is that self-deceivers are motivated to believe that *p* but have somehow ended up with some other mental state, while some accounts also add that they incorrectly believe that they believe that *p*.

The following famous example by Amélie Rorty [1988: 11] puts the failed self-deceiver in a high-stakes context.

> **Androvna**: If anyone is ever self-deceived, Dr Laetitia Androvna is that person. A specialist in the diagnosis of cancer … she has begun to misdescribe and ignore symptoms that the most junior premedical student would recognize as the unmistakable symptoms of the late stages of a currently incurable form of cancer. Normally introspective, given to consulting friends on important matters, she now uncharacteristically deflects their questions and attempts to discuss her condition. Nevertheless, also uncharacteristically, she is bringing her practical and financial affairs into order: though young and by no means affluent, she is drawing up a detailed will. Never a serious correspondent, reticent about matters of affection, she has taken to writing effusive letters to distant friends and relatives, intimating farewells, and urging them to visit her soon. Let us suppose that … as far as she knows, she is hiding nothing. Of course, her critical condition may explain the surfacing of submerged aspects of her personality. … But let's suppose Laetitia Androvna's case is not like that. The best explanations of the specific changes in her behaviour require supposing that she has, on some level and in some sense, recognized her condition.

When a person's non-verbal behaviour conflicts with her avowals we typically conclude that the person has been lying to us. However, if it appears that the person's avowals to us are sincere but her behaviour is inconsistent with them, then it would seem fitting to say that the

person is lying to herself, not us. Unfortunately, this solution raises paradoxes and so it has not been seriously considered anywhere outside of the camp of scholars who hold a compartmental or even homuncular model of the mind (mind comprised of rational centres of agency). Notwithstanding this trend, in subchapter IV-1, I will take this road; in particular, I will argue that Nicole and Androvna are lying to themselves, arguing further that this explanation does not require partitioning the mind – on the theory of lying I will propose and defend, saying that one is lying to oneself raises no paradoxes.

In this section, I will examine some of the most common and most prominent explanations of Nicole-type cases. Then I will summarise the main points from the whole chapter III; I will highlight respective problems of the three main views. This will conclude my literature review. In chapter IV, I will proceed to offer what I argue is the correct theory of lying capable of explaining the Nicole and Androvna cases as cases of lying to oneself without appealing to any theory of the mind. This will lead the way to a non-problematic traditional account of the most peculiar cases of self-deception; in chapter V, I will argue that Nicole and Androvna could even deceive (on the manipulativist understanding) themselves with their lies. But, before that, let's see what other people have to say.

2. Meta-Beliefs vs. Beliefs

In his 2005 paper, Funkhouser offers an account of self-deception that combines revisionism with deflationism in a very innovative way. Because it is implausible to think that unwelcomed self-deceivers want that what they end up believing is true – e.g., Zelda does not want Scott to be having an affair and yet she believes it – Funkhouser [2005: 297], following Nelkin [2002], proposes that, 'whenever a person is self-deceived about *p*, that person's self-deception is motivated by a desire *to believe* that *p*' (i.e., rather than a desire that *p*). However, instead of getting what they want, i.e., the belief that *p*, self-deceivers such as Nicole and Androvna get the false (meta-)belief that they believe that *p*. That is to say, they merely believe that they got what they want, when in fact they failed and, since they do not get what they want, these cases may be described as failed deception by the self.

I will challenge the two main claims of this view: the one that people like Nicole merely want to believe that ~*p* rather than wanting that ~*p* is true and the one that they do not get what they wanted.

It is typically thought that self-deceived subjects want that a certain proposition is true – e.g., that Tony not be having an affair with Rachel – and this proposal seems quite plausible. However, Nelkin [2002] and then Funkhouser [2005: 297] correctly argue that this hypothesis cannot explain cases of unwelcome self-deception, as in the case of Zelda, where Zelda clearly does not want it to be true that Scott is having an affair with Ernest. Therefore, Funkhouser thinks, the correct view should be that unwelcome self-deceivers want to *believe* that *p* or, at least, to have some first-person qualities associated with such a belief, while not wanting that *p* is true. Being the jealous, insecure type, and having been cheated on before, Zelda wants that ~*p* (Scott is not having an affair) and, out of caution, wants that she believes that *p* (Funkhouser 2005: 298) – the rationale, I assume, is that believing what she does not want to be true will make her more vigilant. Welcome self-deceivers, then, merely want to believe the proposition whose truth they welcome.

Positing this kind of motivation is not only problematic and very complicated; it is also unnecessary and impractical.[43] Zelda can perform all the cautionary actions she needs in the attempt to make sure that Scott doesn't cheat on her without wanting to make herself believe that Scott is unfaithful. Also, suppose that a hypochondriac deceives herself into believing that she *has* cancer despite all the evidence that she is healthy. It cannot be that she wanted to believe that she has cancer in order to make it less likely that she indeed gets cancer; acting on her belief that she has cancer would involve taking chemotherapy, or undergoing surgery, etc., which is clearly harmful. Finally, unwelcome self-deceivers should not be motivated to believe that the unwelcomed state of affairs has already been attained. That would defeat the purpose of being more vigilant. In conclusion, it is very problematic to think that self-deceivers merely want to believe something.

The second main claim is that the person, let's say Nicole, ended up mistakenly believing that she believes that ~*p*. Accordingly, she sincerely reports on the belief she can access, and she can access the 'shallow' second-order belief, while her non-verbal behaviour is triggered by her belief that *p*. In turn, the tension is caused by the fact that she still believes that *p* while believing that she believes that ~*p*. The elegance of this solution is the reason why it became so popular so fast but, nonetheless, it falls into the Freudian trap, i.e., it is unfalsifiable, and more importantly, it is quite uncharitable toward Nicole.

[43] Against this proposal, see Fernández [2013: 387–388].

What I mean by being uncharitable is that avoiding driving past Rachel's house should alert Nicole not only to the fact that she does not believe that ~*p* ('Tony is not having an affair with Rachel') but equally that her belief that she believes that ~*p* is false. Insofar as believing that ~*p* disposes one to take it in practical reasoning that ~*p* is true, so should believing that one believes that ~*p*; hence, it is hard to see how she could keep this meta-belief in the face of obvious counterevidence against it. In short, even if she did form the false belief that she believes that ~*p*, Nicole should be quick in revising it. Finally, it is hard to see how she could, by performing an intentional action, deceive herself into believing that she believes that ~*p* while at the same time failing to deceive herself into believing that ~*p*.

In the end, if I may, I would say that positing meta-beliefs may just be a philosopher's trick to avoid paradoxes. This manoeuvre is elegant and ingenious but it is *ad hoc* and it generates some very important concerns.

3. Context-Dependent Beliefs/Credences

Another plausible explanation is that Nicole and Androvna have inconsistent first-order beliefs at one and the same time but that these have different triggering conditions. In particular, since there is evidence that our memory is context-dependent (e.g., Smith 1994), it could be that inconsistent beliefs (i) belong to different '(memory) search sets,' namely, sets of information in memory that are to be deliberately searched (Shiffrin 1970), or (ii) that they are stored in different corpuses of beliefs with each corpus having its own triggering conditions (Egan 2008).

The corpuses of beliefs hypothesis requires a subtle compartmentalisation of the mind that does not entail functional divisions but rather relies on the plausible intuition that similar beliefs, those related by a certain subject or similar subjects, are compartmentalised into different corpuses that get triggered by relevant sets of conditions. Beliefs within one corpus are mutually consistent but they need not be consistent with those in other corpuses; that is, the holism of the mental applies to corpuses internally but it does not apply across corpuses. The hypothesis of context-dependent memory does not require that beliefs are stored in corpuses while allowing for inconsistencies on the global level. In particular, on the memory *search set* model, only memories with contextual associations are included in the given search set (Smith, Glenberg, and Bjork 1978); and what determines the content of the set is not internal consistency of the content (as in the *corpuses* view) but contextual cues.

Accordingly, certain beliefs concerning the same topic may be inconsistent in virtue of the fact that they are not simultaneously accessible – due to different contextual cues.

Among many studies on context-dependent memory (see Smith 1994), the research conducted by Cialdini et al. [1991] indicates that people often have a number of different beliefs on the same topic. Which belief is activated depends on signals (cues) in the current context and so incompatible mental states can exist within different executive systems. In addition, evidence that suggests that we use different types of logic in different information-processing contexts (Tooby and Cosmides 1992: e.g. 24), which is an evolutionary adaptation, since different contexts introduce different priorities, but it sometimes delivers peculiar results. For example, 'we may be led to believe that "free love" is a splendid idea while sexually aroused in the presence of an attractive partner, and to believe precisely the opposite after viewing a film about AIDS' (Brown and Kendrick 1997: 109). Brown and Kendrick [1997: 109] use the concept of context-dependent memory to explain peculiar cases of self-deception. They argue the following (italics added).

> When I am talking to my physician during my annual check-up I may fully believe alcohol has all the toxicity of strychnine. Later that same day, when I am chatting with my friends in a pub, I may just as fully believe that a few drops of the spirits can have all the benefits of ambrosia. *It is a mistake, however, to assume that the most recently activated belief somehow erases the others*.

Accordingly, some situations – such as Nicole's girlfriends voicing their concerns about Tony's infidelity – would trigger the belief that Tony is not having an affair with Rachel, whereas some other situations – such as the thought of driving by Rachel's house – would trigger the belief that Tony is having an affair with Rachel. These two beliefs, the two authors reason, would be supported by two different types of 'logic' and would be activated in the memory by different respective contexts or would belong to two different corpuses of beliefs.

Theories positing context-dependent memory and, ultimately, beliefs, explain why it is easier to recall something in the same context in which it was learned or why I will recall a piece of information in one context but not in another. They also have the potential to explain how a belief acquired in one executive mode may exist simultaneously with a logically inconsistent belief acquired in another executive mode. This makes these views explanatorily very successful and thus compelling. Even so, it is not said *why* it is a mistake to think that the recently 'activated' belief erases and, importantly, this thesis sits uneasily with the view that we use different types of logic in different information-processing contexts. People can

change their minds and, in that case, the later belief originates from the former (I form it by re-evaluating my former belief) or is a qualified version of the former. Therefore, my concern is that some kinds of apparently inconsistent overall behaviour would unjustifiably be ascribed to context-dependent beliefs or self-deception; they involve people who have simply changed their minds.

For instance, on the context-dependent memory view, the sexual arousal context will activate the belief that free sex is a splendid idea whereas the HIV context will activate the opposite. However, these beliefs are products of judgements and, saying that I have two context-dependent beliefs about sex completely negates my rational power to change my mind when some reasons, such as the prospect of getting HIV, become more salient than others or when emotions or sexual arousal enter and skew my reasoning. In fact, there is no reason to posit different kinds of logic: I genuinely have different reasons in these different contexts. Consider the following situation.

> **Buridan's Bill**: Imagine that Hillary and Bill just met at a frat party, they immediately hit it off, and now they are kissing in Bill's room. But alas, in the moment of Bill's sexual arousal, his roommate, Madeleine, comes into the room and reminds Bill of the documentary about HIV he watched that morning – he swore he will never have sex with a stranger again.

Buridan's Bill does not give us a situation in which one belief ('free love is great') is triggered and the other ('no sex before marriage for me') is not. Rather, Bill will revaluate his position considering all the reasons he now has. And, depending on the intensity of arousal and the level of his current fear of HIV, Bill may easily decide that 'free love' may not be as dangerous as he thought, and return to his old judgement the very moment he satisfies his sexual desires. The arousal could be so high that it overshadows the information introduced by Madeleine completely, so that it never enters Bill's reasoning – overshadowing occurs when the subject's limited attentional capacity is taken up by the more salient cue (Smith 1994: 180). But we may imagine that this is a *Buridan's Bill* situation where the arousal and the inhibiting information are equally powerful, i.e., they give Bill equally strong reasons to act. In this situation, Bill's judgement will be paralyzed: Bill will suspend belief as to whether free love is great, and will just pick an option (see Tanney 1995: 112).

Another problem for the context-dependent memory and the corpuses of beliefs views is the emotional significance of the proposition in question. Memory is said to be context-dependent, in short, when contextual cues affect remembering – since contextual information

is stored together with memory targets (Smith 1994: 167). Hence, reinstantiating or even imagining the learning context contributes to recalling the events better – an internal *response* to a context, rather than its representation, leads to reinstatement of memories. For example, even though drugs, such as alcohol or marihuana, impair long-term memory, they can actually improve it in the cases in which the target material was learned while under the influence of the specific drug (Stillman et al. 1974). Similarly, on the corpuses of beliefs hypothesis, if I cannot remember Val Kilmer's character's call-sign in Top Gun but, to the question 'Was "Iceman" Val Kilmer's character's call-sign in Top Gun?', I immediately say 'Yes!', this is because the relevant belief was not in a corpus that was guiding my behaviour at the time I was trying to answer the question (Egan 2008: 51).

These are good theories and they explain a broad spectrum of our behaviour. Nevertheless, the personal significance of the issue whether Tony is having an affair with Rachel is a very strong reason to think that Nicole's thoughts about this will be relatively pervasive and context-*in*dependent. Emotionally laden stimuli are remembered better than those that are not. One cannot lose sight of one's belief that one's partner is having an affair in the same kind of way in which one may lose one's train of thought. Val Kilmer's character's call-sign in Top Gun is trivia. Finally, these proposals entail that Nicole is not being irrational and insincere towards herself, which does not fit the description of the case.

Perhaps, an explanation that posits different levels of confidence in a proposition captures the Nicole and Androvna cases – confidence in a proposition depends on reasons which may change with context changes (e.g., Lynch [2012]; Lauria, Preissmann, and Clément [2016: 122]). In some contexts, Nicole's reasons may justify having high conviction in *p*, such as the one in which she is having a conversation with her friends in a coffee shop, and some other context may introduce some reasons in favour of being more convinced that ~*p* rather than that *p* – for instance, realising that she needs to go by Rachel's house may bring some unfavourable memories to her attention. That need not make her abandon the proposition that ~*p* altogether. Rather, it may just significantly decrease the level of confidence she has in it.

However, without a clear description of Nicole's action as clearly involving deceiving herself, this explanation suffers from the same ache as all context-dependency views. We now have less reason to take this to be self-deception or any kind of insincerity towards oneself: Nicole updates her confidence according to available evidence (see Pettigrew 2013). If she is updating her confidence incorrectly, let us say that she is being biased in the Melean

way (Lynch 2012), it follows only that she is guilty of having unjustified confidence in *p* but not that she is being insincere towards herself and she does not seem to be violating her own norms of rationality either.

4. Half-Beliefs

Price [1969: 302–314] proposed that many situations can be explained by using the concept of 'half-belief,' a circumstance dependent belief. A person with a half-belief will systematically feel herself to be and act as if she were fully committed to *p* in one set of circumstances, while systematically feeling and acting as if the opposite were true in others. A supposed half-believing theist, for instance, would on Sundays exhibit all dispositions of believing in God, but on weekdays exhibit none of them. Similarly, Price thinks, many people only half-believe that exercise is conducive to health or that going to bed late will make waking up early difficult. That is, they believe this up to the moment in which they should actually start exercising or go to bed (Price 1969: 313).

The difference between context-dependent beliefs and half-beliefs is that, in the former case, the agent holds two beliefs that pop up depending on a context whereas, in the latter case, the person has only one 'half' of one belief – she seems to be believing a proposition only in certain circumstances and this may even be with high confidence in both cases. Since a half-belief gets abandoned when circumstances change, a half-believer is actually a part-time (mild, strong, etc.) believer. On the other hand, in believing that *p* with mild or high confidence, my attitude will remain the same: I have a full belief all the time but my confidence changes (Price 1969: 303, 306). Half-beliefs are context-dependent but in the sense that only a half of the belief will be triggered. So, it may be that Nicole and Androvna believe none of the relevant propositions fully. Rather, they half-believe that *p* (Funkhouser [2012: 92]; Gibbins [1997]).

Imagine again that Nicole, without intending to lie to them, avows to her friends that Tony is not having an affair with Rachel *while* avoiding the route that would force her to drive them past Rachel's house. Let's say that they had their drinks and she, the designated driver, is now driving her friends to their homes and she chooses a longer route that does not go near Rachel's house, which makes her friends suspicious and so they ask whether this route is because she suspects the affair. On this variation, Nicole asserts that ~*p* ('Tony is not having an affair with Rachel') while acting as if *p* is true.

The 'half-belief' proposal cannot explain this situation: it predicts that, just like people who stop believing that exercise is conducive to health at the moment in which they should actually start exercising, Nicole stops believing that Tony is not having an affair with Rachel at the moment in which she is supposed to drive by Rachel's house. So, if they are not having context-dependent beliefs or half-beliefs, maybe Nicole and Androvna have no beliefs on the question whether *p*?

5. No Beliefs

According to Erick Schwitzgebel's [2001, 2002] phenomenal, dispositional account of beliefs, Nicole is in an 'in-between' state of belief; i.e., there is no fact of the matter what she believes. This is because she has some stereotypical dispositions of the belief that Tony is having an affair with Rachel and some dispositions that he is not, but she does not have enough stereotypical dispositions for her to hold either of the two beliefs.[44] Cases of in-between believing are those in which agents, in a certain kind of mood, sincerely assert that *p* but, in another, genuinely recant such a confession. And, since there is no simple answer to the question of what, 'underneath it all,' they really believe, Schwitzgebel [2002: 261] argues that these are cases of in-between beliefs.

This position is very elegant. In cases such as Nicole's or Androvna's, rather than attributing conflicting beliefs or half-beliefs, we should merely attribute different manifested dispositions. We know how they are disposed to act in certain circumstances and that's it; there is nothing more that can be said here. The conflict is not between beliefs; rather, various components constitutive of the belief that *p* are in conflict and there is nothing paradoxical about that. One may object that this is not self-deception, since deception requires that the deceived's belief is in some sense affected, but this will not be a serious concern as long as the theory can actually explain the cases – quite possibly, we might be incorrectly describing this as self-deception.

However, here is again that very plausible situation in which Nicole, not intending to lie to them, avows to her friends that Tony is not having an affair while avoiding the route that would force her to drive them home past Rachel's house. By saying that Nicole simultaneously exhibits some stereotypical dispositions of believing that *p* and some of believing that ~*p*, we are not explaining her behaviour; rather, we are merely re-describing it

[44] See Porcher [2012]; similarly Hamilton [2000: 25]; Funkhouser [2009: 9].

in terms of dispositions rather than beliefs, and this is not *as such* an explanation of how this is possible. Therefore, the phenomenal, dispositional account cannot explain this variation of the case.

6. Avoiding the Thought that *p*

One very promising account of what may be denying welcome self-deception *secundum quid* – i.e., self-deceptively denying a painful reality without the aim of deceiving oneself – is offered by Kent Bach [1981, 2009]. This is a hybrid account of self-deception: it combines the deflationary claim that the intention to deceive oneself is not the intention on which a self-deceiver acts [1981: 368] with two revisionist theses; one that deception about the self is necessary for self-deception and the other, Bach's personal touch, that the product of self-deception is not a belief but rather an occurrent thought.

Bach's key move is distinguishing what a self-deceiver thinks from what she believes: the thought *that p* does not entail believing that *p* and, *vice versa*, believing that *p* does not entail occurrently thinking that *p*; that is, there is no mutual entailment between the occurrent thought that *p* and the belief that *p*. Furthermore, there is no entailment between the thought *of p* and the thought *that p*, i.e., the thought of *p* does not entail the thought that *p*, which is the thought that *p* is true, of course.[45] Accordingly, being self-deceived about *p* is compatible with (dispositionally) believing that *p* as long as one does not think *that p*, on a sustained or recurrent basis at least, when the thought *of p* occurs [1981: 364). Thinking that *p*, thinking that ~*p*, and not thinking either are all compatible with thinking *of p* (Bach 1981: 354).

It follows, then, that a self-deceiver may have the belief that *p* in her belief-box but, due to the desire that ~*p*, think that ~*p* or think neither that *p* nor that ~*p* while dispositionally believing that *p*. Bach later [2009: 786] elaborates this proposal by saying that being self-deceived involves a disposition to resist consciously thinking something that one believes and to affirmatively think something contrary to that proposition – for what matters is what occurs to us. A person achieves this state by deploying a strategy of self-deception by which she prevents herself from entertaining the thought *that p* when the thought *of p* occurs, thereby making herself unaware of her belief that *p* and of her own violation of rationality (Bach 1981: 359–360).

[45] Bach uses 'believe' only to refer to dispositional beliefs (a belief is dispositional if the person has it but it is not present in her working memory) and 'think' to refer to an occurent mental activity.

Three strategies of self-deception are discussed; rationalization, evasion, and jamming. Rationalisation covers any case of explaining away what the person would normally regard as adequate evidence for a certain proposition. The self-deceiving rationalizer need not believe that ~*p* but she needs to believe a generalization that is incompatible with *p* and to reason accordingly (Bach 1981: 360).[46] Evasion, the second strategy, is turning one's attention away from some touchy subject. Self-deceptive evaders avoid the thought of *p* in order to avoid the thought that *p*. Evaders do not access the strength of their reasons; they merely think of a single reason against thinking of *p* and turn their mind to something else.[47] Finally, even if a self-deceived person cannot avoid thinking of *p*, she can, whenever the thought of *p* occurs to her, think that ~*p* – this is what Bach calls jamming. For instance, she may think what it would be like if ~*p* were true and imagine consequences of ~*p*.

My version of Bach's [1981: 364–365] complete and final analysis of denying welcomed self-deception is this: Over t_1–t_2 S is self-deceived (into thinking) that ~*p* if and only if, over t_1–t_2,

(1) S desires that ~*p*,
(2) S believes that *p* (or that he has strong evidence for *p*), and
(3) (1) and (2) combine to motivate S to avoid the sustained or recurrent thought that *p*,
(4) By way of deploying one or more relevant strategies of self-deception, S succeeds in avoiding the sustained or recurrent thought that *p*,
(5) S occurrently thinks that ~*p*.

Bach's account contains another condition which was meant to include cases in which self-deceivers cannot avoid thinking *of p*. I excluded it because of conditions 1–3, as I have stated them, already cover with this situation. Condition 3 specifies that the person avoids only the thought *that p*; she may very well have recurrent thoughts *of p* while successfully avoiding the thought *that p*. Also, for the sake of clarity, I have separated Bach's condition 3 into two conditions (3 and 4) and added that avoiding the thought that *p* is done by way of deploying a relevant strategy of self-deception. This addition is important: one may avoid a thought by drowning it in alcohol, by engaging in reading serious philosophy, or by falling asleep (caused by reading philosophy perhaps?), none of which need amount to self-deception. Finally, I added condition (5), which requires that the self-deceiver thinks that ~*p*

[46] On rationalisation see D'Cruz [2015].
[47] Since evaders avoid thinking that *p* by evading thinking *of p*, the proper strategy should be described as thinking of reasons against thinking *of p*, not against *p* (*pace* Bach 1981: 361). Reasons against *p* are reasons against thinking that *p* is true, whereas 'I got better things to do' may serve as an excellent reason against thinking *of p*.

– below, I explain why I did this and Bach would agree with this addition (see Bach 2009: 786).

According to Bach's [1981: 364–365] account, Nicole, for example, is self-deceived while talking to her girlfriends exclusively because she manages to avoid thinking that *p* even though she thinks of *p* – after all, they are discussing whether *p*. This proposal is consistent with Nicole thinking that ~*p* but it is also consistent with thinking neither that *p* nor that ~*p* (Bach 1981: 354). Now, it may appear as if Bach's strategies of self-deception would count as deceptive on the manipulativist view even if the person does not think that ~*p* – this would be in the virtue of them being aimed at generating a certain response by way of manipulating Nicole's agential use of her own cognitive capacities. Yet, this response, i.e., neither thinking that *p* nor that ~*p*, is too weak to count as a product of human self-deception: not thinking is not a mental state. The agent who successfully avoids thinking that *p* in this way is better described as being in the state of denial.

In contrast, if the person who successfully avoids thinking that *p* does this by way of thinking that ~*p*, then her manipulation does result in her acquiring a truth-evaluable mental state and so we may rightly count her as self-deceived. This is why I added condition (5) to Bach's 1981 proposal. With condition (5) in place, we seem to have acquired a plausible account of a certain kind of self-deception – it explains Nicole's behaviour as involving self-deception – but it cannot be understood as an all-encompassing account of self-deception. Some self-deceivers do seem to end up with beliefs.

There is another, practical, concern related to this proposal. By appealing to dispositional beliefs the person would not ascribe to herself, e.g., that Tony *is* having an affair with Rachel, we are falling into the Freudian trap (see III-a/2.2); that is, the theory is unfalsifiable – since we are ascribing a belief to Nicole she denies having. This is not something that should avert us from adopting this proposal but, other things being equal, an explanation that does not fall into this trap is preferable.

7. Making-Believe that ~*p* vs. Believing that *p*

The proposal that a self-deceiver is acting on her pretence was initially proposed by Sartre [1978] (also, Darwall 1988: 413–15) but was recently convincingly reintroduced by Tamar

Gendler [2007, 2010], who also reinforces it using some insights provided by Bach [1981, 2009] and Funkhouser [2005].

Gendler argues that the lack of normal manifestation does not imply a lack of the associated belief; subjective access to the belief may be muted suppressing thereby its action-guiding role while leaving the belief intact. Furthermore, a cognitive attitude playing a typical belief-like role does not imply that this cognitive attitude is a belief. Because the lack of manifestation does not mean the lack of belief and since a different mental state can play a belief-like role, a person can non-problematically, for example, make-believe that ~*p* while believing that *p* without (sufficiently) manifesting the belief that *p*.[48] In fact, in paradigmatic instances of self-deception, says Gendler [2007: 237; 2010], a particular attitude – e.g., pretence, make-believe, imagining, or fantasy – plays a typical belief-like role; that is, it occupies many of the person's thoughts, guides many of her actions, and so on. The conflicting belief is present but it plays no role in some aspects of the person's behaviour.

In particular, because the person does not want that *p* is the case, she engages in a fantasy in which ~*p* is the case. And one of the consequences of attending to the thought that ~*p* in the context of fantasy is that it allows the person to figure out what the world would be like if ~*p* was true. Attending to what the world would be like if ~*p* was true, in turn, renders her particularly sensitive to the evidential basis for her belief that *p* and, if her discomfort with *p* is sufficient, she may make an (unconscious) effort to arrange her encounters with the world so as to minimize her interactions with evidence in favour of *p*. In short, she pretends that ~*p* and represents her reality to herself in a way in which she wants it to be.[49] This behaviour is typically followed by a decrease in frontal lobe activation, which increases unrealistic optimism and anosognosia, and a dopamine increase, which yields a pleasurable sensation (Lauria, Preissmann, and Clément 2016: 128–129). Therefore, eventually, her imaginative pretence that ~*p*, despite the awareness of her own actions, may gradually come to have the sort of vivacity normally associated with the belief that ~*p*. As a result, the pretence – which has now become a hybrid mental state with some features of a belief – will come to play a central role in guiding of her actions. At this point, while the person continues to believe that *p*, it may rightly be said that she is self-deceived that ~*p* (Gendler 2007: 241).

[48] A similar point, in relation to how emotions can cause belief-like behaviour, was made by Borge [2003: 15]. On pretence see Nichols and Stich [2000].

[49] Recall, Bach understands jamming as pretending that ~*p* as a means of avoiding the thought that *p*.

This ingenious explanation elegantly avoids paradoxes – pretending that ~*p* is consistent with believing that *p*, while my knowing that I intend to pretend that ~*p* does not undermine the success of my action – but some concerns could rightly be raised notwithstanding. The first is simple: although my knowing that I intend to pretend that ~*p* is consistent with my successfully performing this action, can it really be successful in the way in which I need it to be successful? This concern reveals how heavily Gendler's position relies on the very problematic hypothesis that Nicole is more concerned with wanting to believe that ~*p* or having an experience of ~*p* than with the issue whether ~*p* (Funkhouser [2005]; see III-3b/3). Only this hypothesis motivates pretending that ~*p*; however, we have much more reason to think that Nicole wants to avoid the painful reality itself, i.e., she wants *p* ('Tony is having an affair with Rachel') to be false, and successfully pretending that ~*p* will not satisfy that want. This concern does not invalidate the proposal but it limits it only to cases of benign self-deception, i.e., those in which *p* does not represent a distressing state of affairs. That is, it may explain Gendler making-believe that she will finish her paper today but it will hardly explain Nicole making-believe that Tony is not having an affair with Rachel – Nicole's behaviour seems to involve much more than making-believe.

Gendler further suggests that, if the person continues to act as if ~*p* is true, then 'it would likely seem to her that she *does not* believe that *p* (after all, its normal manifestations are absent), but that she *does* believe that ~*p* (after all, it is playing a typically belief-like role). That is, the self-deceived subject may well have a false belief about what she believes: she believes that she believes that ~*p*, while in fact she does not' (Gendler 2007: 253, n. 38; 2010: 170, n. 38; 2012: 808, n. 17). This proposal is analogous to Patten's [2003] account, according to which I deceive myself unintentionally by judging which beliefs I have from the third-person perspective, and therefore it suffers from the same weaknesses (see III-3a/2.2).

However, rejecting the idea that pretence may cause that you end up with false second-order beliefs does not entail that some people, realising that *p*, react to this by wanting to have an experience of ~*p* and thereby avoid the thought that *p*, which vindicates Gendler's view. This proposal is analogous to Bach's jamming and borrows some lines from Funkhouser [2005]. It says that Nicole and Androvna, helpless to do anything to change the actual state of affairs, engage in pretence that ~*p* because they want to have the experience of the world in which ~*p* is true, and they can do this even without believing that ~*p*. Gendler [2007: 244; 2010: 174] writes:

> [The subject has a] tacit recognition that her self-deceptive pretense is subject to evidential override. … [This is why] the self-deceived subject deliberately avoids putting herself in situations where she suspects she may come upon evidence wildly incompatible with ~*p* (since, given her general commitment to reality sensitivity, this would render it difficult for her *p*-related thoughts and actions to be governed by her fantasy that ~*p*). So, for example, a subject who is self-deceived about her husband's affair [Nicole] will be reluctant to drive past his purported lover's [Rachel's] driveway for fear of seeing his car parked there.

This explains only a part of Nicole's overall behaviour and, it explains it not as self-deception, but rather as willful ignorance (see, Lynch 2016). Please notice, Nicole is avoiding driving past Rachel's driveway, and most likely for fear of seeing Tony's car parked there, but her action is not intentional under this description. That is, once asked why she is taking the longer route, she will not say 'I fear seeing Tony's car there.' Rather, she will have some other reason and this is the reason on which she acted. Taking the other route is not intentional under the description 'avoiding driving past Rachel's house' and thus is not behaviour typical for pretence, as defined by Gendler.

Nicole's or Androvna's behaviour does not appear to be pretence that plays a belief-like role in the sense of involving insincerity towards themselves and, without insincerity, there is no deception. Rather, it is pretence in the 'real' sense of pretending, as in 'faking,' 'acting' (in a play). As interpreted by Gendler, Nicole's behaviour (where she intends to avoid driving past Rachel's house), is more like living in a very poetic denial or engaging in a childish game (Doggett 2012: 765–766) than self-deception. Nicole is just doing her best not to think about some distressing stuff in certain circumstances in the same way a hurt athlete may say to himself that his leg does not hurt (see Van Leeuwen 2007a: 431, n. 21). This behaviour may be irresponsible in some circumstances, if one pretends that his tooth is not aching, but it need not be irrational, if one cannot afford to go to the dentist before Monday (Fernández 2013: 387–8).[50]

None of this is to say that this suggested explanation is incorrect, and I am very sympathetic to it: undoubtedly, many people would rather pretend that ~*p* than face up to the fact that *p*. However, this kind of behaviour does not appear to have the insincerity-towards-oneself component necessary for self-deception and this component is present in the Nicole case. Gendler gives us a theory of one's wilfully living one's own life in fantasy, not a theory

[50] Some scholars have understood self-deceptive pretence as deception of others (Haight [1980, 1985]; Gergen [1985]).

of self-deception as pretence. I will now briefly address the general attempt to explain self-deception by positing hybrid mental states.

7.1. Hybrid Mental States

Many scholars try to explain self-deception by positing hybrid mental states. Lazar [1999: 286], for instance, says that some states involved in self-deception are somewhere in between desires and beliefs; they are 'strongly influenced by desires and emotions – they express them – yet they inform our behaviour more or less in the way that beliefs do.' Egan [2009: 275–276] calls these states 'besires,' intermediate attitudes between beliefs and desires. Audi [1982, 1985, 1988, 1989, 1997a, 1997b] and Rey [1988] posited a specific kind of belief, avowed belief, which lacks some important properties beliefs have, such as behavioural manifestation (against, Van Leeuwen [2007a] and Bach [2009]). Similarly, Mijović-Prelec and Prelec [2011: 227] distinguish three levels of belief: deep belief, stated belief, and experienced belief.

I object that these hybrid mental states and distinctions between kinds of beliefs are posited *ad hoc* and they are unfalsifiable. More importantly, all of these concepts are very problematic and they actually explain nothing; they just re-describe some behaviour in a non-paradoxical way but this is because they are hybrid and speculative – we have defined them by way of picking out things that we cannot explain.

One such proposal is not subjected to this kind of criticism. Tyler Doggett [2012] proposed a possible variation of Gendler's [2008] theory that involves introducing her concept of alief, an innate propensity to automatically and arationally respond to certain stimuli. Crying at movies would be an alief-driven behaviour. Accordingly, a self-deceiver would believe that p and alieve that $\sim p$. However, Gendler's [2012] reason for not going along these lines is that self-deceivers can be reassured while alievers cannot: you can convince Nicole that Tony is having an affair with Rachel but you cannot convince her not to cry in the movies – the latter is beyond her control.

III-4. Short Summary of the Three Main Views

In this chapter (III), I have argued that traditionalists – who understand self-deception as involving the intention to make yourself believe a falsehood – fall into the Traditionalist trap, the need to divide the mind, and/or slip into deflationism. That is to say, dividing the mind has the consequence that the person did not intend to deceive herself; rather, the distressing information or the belief got unintentionally lost in one of the subsystems. And deflationists, who should have the upper hand of not being constrained by the demanding standard understanding of the intention to deceive, either displace the irrationality of self-deception into the third-person perspective, thereby falling into the Deflationary Trap, or they even relapse and fall into the Traditionalist's Trap.

Revisionist theories from the family of solutions that understand self-deception as deception *about* the self typically fall either into the Traditionalist trap or into the Freudian trap (they are unfalsifiable), or even both. Some of these theories are internally inconsistent and most fail to avoid the Deflationary trap. The 'failed deception *by* the self' revisionist proposals are probably the most diversified of all. They postulate conflicting meta-beliefs and first-order beliefs, or beliefs triggered contextually, or half-beliefs, or even conflicts between some belief-components, or between beliefs and some hybrid mental states. These proposals are valiant but they are *ad hoc*, some are unfalsifiable, and/or fail to explain a plausible type of self-deception presented in a situation in which, not intending to lie to them, Nicole avows to her friends that Tony is not having an affair while avoiding the route that would force her to drive her and her friends past Rachel's house.

The manipulativist view, in contrast, deals with ordinary cases of deception better than any of the proposed theories; it generates no paradoxes and thus it does not require revising the concept of self-deception. However, it still struggles with the cases of Nicole and Androvna. In the above problematic situation in which Nicole tells her friends that Tony is not having an affair with Rachel, their most natural response would be 'Nicole, stop lying to yourself; they are having an affair. Face it!'. The manipulativist view cannot make sense of lying to oneself: it is thought that liars must assert what they believe is false and intend to deceive the hearer. Lying to herself would require of Nicole to form the intention to make herself believe a

falsehood, and we reasonably think that people cannot intend to do what they take to be impossible, such as making yourself believe as true what you believe as false.

So, what now? I think that the answer is simple: we need a better theory of lying. I will argue that Nicole and Androvna *are* lying to themselves, and the account of lying I propose and defend even allows both that they lie to themselves with the intention to deceive themselves and that their deception is successful. These cases are in no sense doxastically problematic, especially the version in which they are lying to themselves without the intention to deceive themselves. Please notice, in lying to herself, Nicole is *not* lying to her girlfriends – since she is actually not addressing them, but herself. And this is what they recognize in her behaviour: 'Nicole, stop lying to yourself … Face it!'. Nicole is being insincere towards them only in a sense in which she does not make it clear that the assertion is meant for her. However, she is not acting on her intention to be insincere towards them – she has instinctively blocked the distressing proposition with her lie. I will discuss this reply in more detail in IV-1/5.

In what follows, I will first offer and defend the view that some lies are non-deceptive and non-deceitful (shorthand for 'not intended to deceive'): this is the topic of the Section IV-1. In Section IV-2, I will argue that in some cases liars lie by asserting something they believe to be true and that the essence of lying is misrepresenting to your audience what you assert as a legitimate assertion, not in saying what you do not believe or in intending to deceive. Combining this theory of lying with the manipulativist conception of deception generates a perfectly unproblematic traditional theory of self-deception and allows us to explain the Nicole-type cases in any possible variation, which is what I do in V-1, where I close my account of these cases.

IV. LYING

1. Brief Summary

In chapter II, I argued that received theories of deception, especially those focused on intentional deception, are substantially narrow. I also argued that paradoxes of self-deception are not something that arises with respect to self-deception *qua* phenomenon of the self deceiving itself, which is why our ordinary talk of self-deception is not thought to be puzzling (as correctly noticed by Szabados 1974: 51), but are rather flaws of our theories of deception that merely become apparent when these are applied to self-deception. I demonstrated that these theories are flawed by introducing examples of interpersonal deception they incorrectly explain, cases of triple bluffs and the deceptive use of irony, most notably. In turn, I offered a novel theory of deception, the manipulativist view, which in its intentionalist reading, says that D_R (deceiver) intentionally deceived D_D (deceived) by way of φ-ing, if and only if

1) D_R φ-ed intending to thereby generate a specific φ-relevant response on the part of D_D,
2) D_R's φ-ing causally contributes towards a non-deviant manipulation of D_D's agential use of D_D's own cognitive capacities, and
3) Because of 2, D_D responded autonomously to φ-ing in a relevant way.

Gendler [2007: 234] recently wrote that, 'whatever else may be said of it, self-deception cannot involve deception of the self by the self in exactly the same way that interpersonal deception involves deception of one person by another.' My analysis indicates that this popular position is incorrect. On the manipulativist view, self-deception could be modelled on interpersonal deception without raising any paradoxes: forming the intention to deceive myself, on this account, need not involve forming the intention to make myself believe a falsehood or harm yourself in any sense. This was my first step toward rehabilitating traditionalism, and this step is a substantial one.

My next step was demonstrating that the manipulativist view should be a preferred theory of self-deception. Therefore, in the previous chapter, I argued that received theories of self-deception, while ingenious and insightful, are all too narrow (explaining only certain kinds of self-deception) and most of them cannot explain this phenomenon without raising some important concerns (i.e., without falling into various traps), including (for some views) the concern that they assume a very limited conception that identifies deception with holding a false belief.

Observing self-deception through the manipulativist's eyes is a much better approach because it does not require self-deceivers to intend to make themselves believe falsehoods or to intentionally harm themselves in any way (epistemic, doxastic, or practical) and so the need for developing ingenious solutions or for reconceptualising self-deception need not arise. In subchapter II-1, I argued, for instance, that I can intend to deceive you into believing what I believe is true (triple and double bluffs), or into believing something I merely guess is true (presumably stove), or even into believing a proposition the truth of which is not my concern (deceiving by bullshitting). By analogy with double and triple bluffers, let us say, some self-deceivers might intend to make themselves epistemically better off. And, if I think that I will not suffer any harm from my deceiving myself but rather received benefit, then I need not object to being deceived by myself. For example, I do not blame Camus for tricking me into despising alienation because I have reasons that are independent of Camus's action to think that alienation should be despised.

Similarly to the *Stanger* case, a double or triple bluffer may believe that she is doing a favour to her victim. Analogously, some self-deceivers may think that they are doing themselves a favour. Sometimes it is very difficult to proportion your belief to evidence and, if you are absolutely convinced that your belief is true despite the overwhelming evidence to the contrary and if this belief comes with great emotional significance (e.g., it concerns your intelligence or spouse or your child), you might decide to proportion (i.e., manipulate) your evidence to your belief realising that you are thereby effectively deceiving yourself but thinking nevertheless that the deception is justified (*Stranger*) or even necessary (*DP*, *TP*). I take this to be the correct explanation of the Hillary case (see V-2).

Taking it that some self-deceivers think that they are deceiving themselves for their own epistemic and overall benefit produces a non-paradoxical explanation of many cases of self-deception. In a case of unwelcome self-deception, a jealous husband who fears that his wife is having an affair – let us say that he has a gut feeling about this but no evidence – may decide to take steps to confirm his gut feeling realising that he is thereby effectively deceiving himself but continue notwithstanding that realisation; his gut has never been wrong. The same explanatory principle can be used to understand welcome self-deception in which the man retains his belief that the wife is not having an affair or that his child is not abusing drugs.

Nevertheless, while it successfully explains many cases of self-deception without the need for an additional theoretical apparatus, the manipulativist view cannot be non-problematically applied to all cases understood as involving self-deception. In particular, it cannot directly explain any of the four main cases I introduced. This failure is not a result of some inherent flaw of the manipulativist view; rather, I have deliberately devised those cases so that a manipulativist reading will not straightforwardly explain them. Manipulativism gives us a solid *basis* for explaining deception and self-deception, but human behaviour can never be simply explained on a single hypothesis, even if this hypothesis correctly identifies what all cases of self-deception have in common – namely, one's intentional manipulation of one's own agency that one tries to hide or conceal from oneself. Our behaviour is nuanced and complicated, hence, explaining it requires a diversified approach: we need to understand the motive, know something about the person's character, intellectual capabilities, personal significance of the proposition in question (if any), relevant context, and many more. Below, I highlight the features of my main cases that remain to be explained.

Maria seems to have deceived herself unintentionally; that is, she did not act with the intention to retain her belief. Because they do not count having a false belief as sufficient for being deceived, manipulativists need other reasons to take her behaviour as involving self-deception. They need to tell us whether and, if so, how Maria is being insincere towards herself in a way that rightly counts as deceiving oneself. In contrast, Nicole appears to have the intention to believe that Tony is not having an affair with Rachel and she vehemently affirms that this is the case, but she nonetheless behaves as if she believes the opposite. Her self-deception, if she is deceiving herself, seems to be unsuccessful. Finally, Zelda ends up believing what she does not want to be the case and the manipulativist view offers no new insights that could help us explain a case like this. Specifically, manipulativism cannot tell us where the jealous husband's *gut feeling* came from.

As we can see, each case comes with its own characteristic problems. This will give me the opportunity to solve these problems one at a time. In this chapter, I will argue in favour of a theory of lying that will enable us to explain the Nicole type cases. My proposed theory of lying will enable us to say that people can successfully lie to themselves and sometimes even intentionally deceive themselves by lying – I make this additional step in Subchapter V-1 based on the argument from IV-2. I start by showing that one can lie to others (and thus also to oneself) without the intention to deceive the hearer, Subchapter IV-1. On this theory of lying, the Nicole-like person lies to herself *without* intending to deceive herself. She just

cannot legitimise the distressing truth by affirming it in her avowals; she cannot bring herself to face up to this truth. I think that this description captures this kind of behaviour perfectly: the insincerity towards oneself is not the *deceiving oneself* but rather *lying to oneself* kind of insincerity.

In short, in this subchapter, I will argue in favour of the account of lying that allows us to say that some Nicole-type cases need not involve self-deception at all, even though they involve a very similar kind of behaviour – lying to oneself. In Subchapter V-1, based on the final account of lying that I develop in IV-2, will argue that some liars may even intend to deceive themselves by lying and succeed in acting on this intention.

2. Refining the Theory of Lying

The task of showing that one can lie to oneself successfully and even deceive oneself by lying is important because it takes a giant step towards rehabilitating the traditionalist account of self-deception without the need to partition the mind. On the standard account of lying, traditionalism about self-deception does not fit well with the idea that I can successfully lie to myself, let alone deceive myself by a self-addressed lie. I will argue, however, that the right kind of the account of lying will not generate any problems in explaining these otherwise puzzling cases of self-deception or those thought to involve self-deception but which are rather cases of lying to oneself.

Let us begin by distinguishing two variations of the standard account of lying. According to the traditional version, the so-called *traditional account* of lying, L_T, I lie to you if and only if I assert what I believe is false and I intend to deceive you (e.g., Augustine [1952/395]; Sartre [1978]; Davidson [1986, 1997]; Williams [2002]; Derrida [2002]; Meibauer [2011, 2014a, 2014b, 2016a, 2016b]; and Faulkner [2007, 2013]).

L_T: I lie to you if and only if

1. I assert that *p* to you.
2. I believe that *p* is false.[51]
3. By asserting that *p*, I intend to deceive you.

My intending to deceive you by lying is typically understood as 'making you believe that *p*,' 'making you more confident in *p*,' or 'making you believe (or more confident in thinking)

[51] I call this the *Belief Condition for lying* because it requires a belief that what you assert is false.

that I believe that *p*.'[52] When the same agent is the liar and the addressee, we get that lying to myself generates paradoxes of self-deception if the mind is not partitioned: in lying to myself, according to L_T, I intend to make myself believe as true what I already believe is false. In turn, while manipulativists may hold that the paradoxes of self-deception need not arise, L_T forces them to admit that they must arise if the person is lying to herself. But, if the intention to deceive is not integral to the intention to lie, then I may intend to lie to myself simply because this does not require intending to make myself believe my own lie.

In this thesis, I argue that lying should not be understood as necessarily involving deception or the intention to deceive at all. In other words, I will defend the *non-traditional* version of the standard account of lying ($L_{N\text{-}T}$) (e.g., Aquinas [1922, article 1]; Johnson [1983/1755]; Carson [2006, 2010]; Sorensen [2007, 2010]; Fallis [2009, 2012, 2013, 2014]; Saul [2013]; Stokke [2013, 2016a, 2016b]). According to $L_{N\text{-}T}$, I lie to you if and only if I assert what I believe is false.

$L_{N\text{-}T}$, I lie to you if and only if

1. I assert that *p* to you.
2. I believe that *p* is false.

$L_{N\text{-}T}$ evidently avoids the problem L_T raises when the liar is also the one who is being lied to. That is, if I can lie to myself and *not* intend to deceive myself with my lie, then I do not intend to make myself believe as true what I already believe is false. Therefore, the question whether liars must intend to deceive becomes vital for understanding some very problematic cases thought to involve self-deception.

[52] It is pertinent to note that there are other versions of the general view that lying involves intending to deceive. For example, according to Lackey [2013: 246], liars merely need to intend to *be deceptive* towards their hearer in stating that *p*, where 'being deceptive' may involve concealing information from the hearer regarding whether *p*, not making the hearer more confident in *p* (see Section 2.2). On this view, I lie to you if I assert what I believe is false and I intend to conceal information regarding whether *p* from you.

IV-1 Lying Without the Intention to Deceive

1. Lying and Deception

The following is one of the many proposed counterexamples to the traditional account of lying (L_T).

> **Artie**: The police do not have enough evidence to convict Tony of a vicious murder without eyewitness testimony. They know that Artie witnessed the crime, Artie knows that they know, they know that he knows that they know, etc. However, when the police question him about the crime, he says that he cannot help them because he saw nothing. Artie does not say this because he expects or intends them to believe what he says. Rather, he says it because Tony has threatened to harm him.[53]

The one who did not intend to deceive did not intend lie according to L_T. However, Artie seems to be lying to the police even though he does not intend to deceive them. And he does not intend to deceive them because he knows that they know that what he asserts is false; that is, he knows that he cannot make them believe his lie and so he does not even try to do so. Persuaded by Artie and similar cases, many contemporary philosophers became non-traditionalists. Nevertheless, many prominent philosophers still hold the traditional view. In fact, several studies suggest that people, both children and adults, think that the intention to deceive is necessary for lying (Lindskold and Han [1986]; Peterson [1995]; Lee and Ross [1997]; Taylor, Lussier, and Maring [2003]; see Turri and Turri [2015: 161–162]). Most importantly, traditionalists about lying have come up with many innovative arguments that make the interpretation according to which Artie lies without intending to deceive much less convincing – e.g. Kenyon [2003]; Staffel [2011]; Meibauer [2011, 2014a, 2014b, 2016a, 2016b], Lackey [2013]; Tollefsen [2014]; Dynel's [2015]; Keiser [2016]; and Leland [2015] – and, with some replies from Fallis [2014] and Stokke [2016a], this debate has become very interesting.

A range of examples of lies that are not intended to deceive has been discussed by philosophers. These include so-called bald-faced lies, coercion lies, and knowledge lies. A bald-faced lie is a knowingly undisguised lie, or at least, the liar thinks that his lie is undisguised (Fallis 2010: 2, n. 8). We may say that Artie's lie is bald-faced because he knows

[53] From Fallis [2013: 342–343]; see Carson [2006; 2010: 20].

that he cannot and that he, therefore, does not, hide the fact that he is lying (see Carson [2006; 2010]; Sorensen [2007]; Fallis [2009]). And, because he does not hide his lie, he does not intend to deceive either.

Artie is an interesting case because it fails to raise some concerns raised by similar cases. For instance, unlike some similar cases involving witnesses (e.g., Carson [2006, 2010]; Keiser [2016]), Artie is not lying-on-the-stand and thus he is not committing perjury. The lying-on-the-stand scenario also generates the worry that the witness might actually harm the defendant with an obviously false testimony, i.e., bald-faced lie. The jury could infer from this testimony that the defendant did commit the crime, which conflicts with the witness's best interest: the witness wants to avoid the retribution received by the defendant on account of being found guilty for the crime. The fact that he might harm himself with a bald-faced lying-on-the-stand might be an incentive for the witness not to testify at all or to even try to deceive the jury with his testimony (Meibauer 2016b: 259–260). Because Artie cannot harm Tony with his lie – the police cannot arrest Tony without Artie's testimony – his lie is less problematic than other *Witness*-kind bald-faced lies

Artie's lie is also a coercion-lie on account of Artie being forced to say that he saw nothing: Artie lies to the police because he is afraid of Tony, not because he wants to deceive them (see Fallis [2013]; Lackey [2013]; and Leland [2015]). Finally, according to Sorensen [2010], a knowledge-lie is a statement intended to block knowledge that the very statement is false but which is not meant to convince the hearer that it is true. This is a rather problematic conception that immediately provoked strong challenges and the proposed examples are contested (see Staffel [2011; forthcoming]; Fallis [2011]; Lackey [2013]); hence, I will not discuss it in detail here (however, see Krstić 2017).

Because bald-faced lies seem to be an interpersonal analogue of lying to yourself (the liar's addressee knows that the person is lying), I will focus on bald-faced lies. I will first present the traditionalists' arguments to the effect that bald-faced lies do not constitute counterexamples to their view (Section 2). Then (Section 3), I will introduce an example to show that it is indisputably a lie that is not deceptive in any relevant sense. Finally (Section 3.2), I will modify this example so that the agent involved not only *cannot deceive* in any possible sense, but in fact *lies intending to 'tell' the truth* to his addressee. A liar who lies in order to tell his addressee the truth by way of the addressee recognizing the liar's intention to tell the truth by lying can in no sense be understood as intending to deceive. From this, I

conclude that lying need not involve any intention to deceive: it follows that I may non-problematically lie to myself because this intention need not be self-defeating (Section 4); some cases thought to be self-deception are actually cases of people bald-facedly lying to themselves.

2. Bald-faced Lies

Genuine instances of bald-faced lying involve situations in which liars do not intend to deceive because they think that everybody knows that they are lying. This interpretation is based on a plausible idea that people will φ only if they think that they can φ; that is, only if they think that φ-ing is possible. Accordingly, thinking that I must fail at φ-ing, makes it very hard for me to form the intention to φ and, therefore, some lies should be conceived as not aimed at deceiving their audience. It is thought that people do intend to deceive themselves for the same reason.

Traditionalists about lying – the *deceptionists* – attempt to defend their view that lying does require the intention to deceive using two general arguments. The first says that bald-faced lies not aimed at deceiving are not genuine assertions and thus are not lies, while the second says that those that do involve genuine assertions are in some sense necessarily aimed at deceiving. I discuss these two arguments in turn.

2.1. They are not Lies

Acting on my own intention to lie to you need not make me a liar: one can act on an intention, and yet fail to act as intended. For example, when George Bush uttered the sentence 'Iraq has weapons of mass production,' this plausibly came about via his acting on his intention to lie (Saul [2013: 15–17]; see Derrida [2002: 29]; Sorensen [2011]). Even so, because he misspoke, he did not act the way he intended.

One may act on an intention and yet fail to act as intended for various other reasons. I will list two relevant to my discussion. Suppose that I mistakenly believe that winking while uttering *x* (a sentence meaning that *p*) counts as asserting that *p*, and that I act on my intention to lie by asserting that *p* by winking as I utter the sentence *x*. My attempt to lie will be unsuccessful because (under the actual prevailing conventions) my wink cancels the assertion status of my utterance. By winking, I have unknowingly performed the wrong (sub-)action and thus have failed to lie.

Suppose now that, having only driven cars with automatic gearboxes, I want to start a car with manual transmission. Not knowing that I also need to press the clutch, I will only turn the key, which will not yield the desired effect. In this example, I did not fail because I performed an incorrect (sub-)action but rather because I did not perform all the necessary sub-actions. This is a 'not performing a necessary (sub-)action' kind of error and the first argument against the possibility of lies not intended to deceive appeals to this kind of failure. The argument, in short, goes as follows. Just as starting a car with a manual gearbox involves pressing the clutch, asserting involves trying to make the audience believe what one says. Therefore, because he does not intend to make his audience believe what he says (he does not perform a necessary sub-action), Artie is not asserting what he says and those who do not assert do not lie either. Although he acts on his intention, Artie fails to act as intended.

According to the first argument, then, cases regarded as cases of bald-faced lying that do not involve the intention to deceive (in the sense of affecting the hearer's relevant epistemic condition) are not genuine instances of lying; they do not involve genuine assertions (e.g., Faulkner [2007, 2013]; Meibauer [2014a, 2014b, 2016a]; Dynel [2015]; and Keiser [2016]). Proponents of this argument mainly belong to the Gricean perspective on assertion, where the idea is that asserting requires the intention to affect the interlocutor's beliefs in some relevant way (against this view on assertion, see Fallis 2010: 3). Because he knows that the police know that he witnessed the crime, Artie may not intend to make them believe what he says and, as a result, he will not assert what he says.

I now proceed to present the second argument offered by deceptionists against the claim that bald-faced lies do not involve the intention to deceive.

2.2. They Involve Deception

The second argument says that genuine cases of bald-faced lying *do* involve some sort of deception, just not the one that requires that the hearer ends up with a false belief – placing this requirement on deception is too restrictive anyway. Keiser [2016: 470–471], for instance, proposes that, by asserting that *p*, the liar intentionally worsens the audience's epistemic condition concerning whether *p*, which should probably be understood as making them more confident in a falsehood, and that this is sufficient for deception. In more detail, since asserting that *p* involves providing evidence in favour of *p*, the liar's assertion that *p* gives the audience *a* reason to believe that *p* thus making them more confident in *p*, which is a falsehood (see Staffel 2011, forthcoming). Therefore, Artie can worsen the police's epistemic

state with respect to the proposition that he saw nothing by giving them a reason to believe it, even though he does not and could not actually make them believe it.

Intending to deceive by asserting may also be understood in the sense of giving you a reason to believe that I believe what I assert (Chisholm and Feehan [1977: 149]; Faulkner [2007: 543, 2013: 2]; Mahon [2016/2008]; Fallis [2010: 8–9]; Tollefsen [2014: 23]; Meibauer [2014a: 132]; and Keiser [2016: 470]). Hence, while Artie is not intending to make the police believe that he saw nothing, he may have intended to make them more confident in thinking that he believes this. This indeed is a common tactic used by liars. Derrida [2002: 34] writes that, even if it is demonstrated that a person's assertion is false, she can always pretend to have believed it. Furthermore, it is not just that many bald-faced lies might easily be intended as deceptive, many purported examples of bald-faced lies may actually be involving sincere assertions. That is, some people accused of being bald-faced liars may have actually asserted what they genuinely believed to be true at the time.

For example, Sorensen [2007: 252] argues that Asne Seierstad's Iraqi guide, Takhlef, who proudly asserted to Seierstad 'Everything [President Saddam Hussein] did in the past was good and everything he will do in the future is good' is a bald-faced liar. And Sorensen seems to be on the same page with Seierstad on this; she [2003: 30] writes 'He knows he is lying, he knows I am lying, he knows that I know that he knows that I am lying. I keep my mouth shut. To report my questions and attitudes is one of Takhlef's duties.' However, Meibauer [2016b: 267] correctly objects to this specific interpretation of Takhlef's behaviour by saying that it dismisses the possibility that Takhlef believes what he says. And not just Meibauer, the participants of his [2016b: 258] study did not think that Takhlef was lying either. The cases I offer are specifically designed to avoid this concern: there, the liar cannot be taken as believing what he asserts.

So far, these deceptionist replies have all argued that lying makes the hearer epistemically worse off in some relevant sense, i.e., with respect to *p*. The final traditional response, offered by Jennifer Lackey [2013], understands deception as including preventing the hearer from becoming epistemically better off, or at least, as well off as they could have been.

Lackey [2013: 241] distinguishes between *withholding* and *concealing* information and between being *deceptive* and being *deceitful*. Concealing information is one instance of being deceptive while being deceitful is another (withholding information is not sufficient for

deception). I conceal information by acting so as to hide it, while I withhold information simply by not revealing it; that is, by refraining from any action that would disclose it. Importantly, concealing need not involve keeping the information a secret or making the hearer more confident in a falsehood – 'one can be deceptive in the relevant sense, even if the information that one is aiming to conceal is common knowledge' [2013: 242]. Accordingly, Artie is deceptive because he is concealing his testimony so that the police cannot access it. This understanding of deception may be too wide but I will not dispute it (however, see Fallis 2015); instead, I will first offer cases involving a bald-faced liar who cannot conceal anything with his lie (Section 3) and then offer cases of a liar who does not even *intend* to conceal anything with his lie; rather, this person intends to reveal information by lying (3.2 and 3.2.1).

I will now proceed to test these deceptionist arguments in succession.

3. Pinocchio

> **Pinocchio**: In *Shrek 3* [DreamWorks 2007], while chasing Shrek, Prince Charming ran into Shrek's friend Pinocchio and, wanting to seize the opportunity, asked him 'Tell me puppet, where is Shrek?' We all know that the prince has an excellent reason to confront Pinocchio with his question: Pinocchio's nose shows that he is lying while he is lying.
>
> Let us now modify this situation a bit and say that the nose starts to grow at the very instant the intention is formed (it is a reaction to an intended misbehaviour) and that therefore an informed hearer can know that what follows is a lie. Let us also suppose that Pinocchio thinks that all lies are deceptive in some sense, even when it is obvious that they are lies. As a result, he naively decides to lie to Charming with the intention of deceiving him by asserting 'He went east.'[54]

Please notice the following important features of this case. First, this first Pinocchio case is only building up to the counterexamples that actually does the work and shows that L_T is too narrow (*Pinartio*, *Artocchio*, and *Spoonocchio*). In this example, Pinocchio lies to Prince Charming by uttering 'He went east' with the intention of deceiving the prince; thus, this is not a counterexample to L_T. The traditional definition counts this as lying because Pinocchio asserts what he believes to be false and intends to deceive the prince (Arico and Fallis 2013: 791); the non-traditional definition counts his behaviour as lying principally because he

[54] By positing that Pinocchio willingly lied, I avoid the objection that coerced lies are not lies (Kenyon [2003: 245; 2010]; Leland [2015]). The objection that coercion lies are not lies because the speaker is addressing not the hearer but rather the coercer, Artie is saying what Tony wants him to say, will become important in explaining the Nicole case (see Section 4).

asserts what he believes to be false, allowing for an additional intention to deceive the addressee; and I count it as lying because Pinocchio's nose grows as a result of his forming the intention to lie to the prince.

Second, genuine instances of bald-faced lying involve situations in which liars do not intend to deceive because they think that everybody knows that they are lying. Pinocchio does not have this intuition: he *is* aware of how his nose is behaving (i.e., he knows that the nose will tell the prince that he is lying), but he thinks that deception is still possible. Third, the description of Pinocchio's lie detector as flashing the 'lie' signal at the very instant the intention to lie is formed turns the case into an *inter*personal analogue of an *intra*personal lie. A person who is close to forming the intention to lie to herself will surely give away her intention to lie to herself at the very moment in which that intention is formed (maybe even before that very intention is formed) – and this is exactly how I imagine Pinocchio's nose is behaving. Finally, and vitally, the nose is responsive to forming the intention to *lie*, not to forming the intention to say or utter something he believes is false.

My argument will proceed in the following way. I will first analyse whether Pinocchio's lie really can involve deception in any sense and argue that it cannot; Pinocchio was wrong to think that he could deceive (Section 3.1). In Section 3.2, I discuss the 'intention to deceive' condition of L_T – since one may form this intention even if deception is impossible (one need not know that it is impossible, for instance). I end this discussion by proposing a case in which Pinocchio lies in order to tell the truth: he knows that he cannot deceive, he does not want to deceive, but rather wants to tell his addressee the truth by lying. I, then, defend this case from the objection that Pinocchio failed to lie (Section 3.2.1) and conclude by showing that, because Pinocchio lied even though he did not intend to deceive his addressee in any way, this case is the most convincing counterexample to L_T offered thus far, and I offer two other cases of the same kind (Sections 3.2.1 and 3.3).

3.1 Deception

Because Pinocchio intends to deceive the prince, L_T will count his behaviour as lying, but let us see whether Pinocchio actually can *be* deceptive by lying. This debate is relevant because it is meant to establish what could Pinocchio reasonably intend to achieve with his deceptive lie.

Lies that are deceptive make the hearers epistemically worse off (e.g., they make them more confident in a falsehood) or, at least, they prevent the hearers from becoming epistemically better off (e.g., they conceal something). Artie's lie, for example, might be considered as deceptive because he is making the police epistemically worse off with respect to whether he witnessed the crime or with respect to whether he believes this, and it could be said that he is preventing them from becoming epistemically better off by concealing his testimony behind his lie.

However, Pinocchio's case is in an important way different from Artie's case and all other cases of bald-faced lies. The main feature of all other cases is the relevant background knowledge. The police know that Artie saw the murder. Similarly, in Carson's [2006, 2010: 21] *cheating student* example, for instance, the Dean knows that the student cheated. Likewise, in the case in which an Iraqi doctor lies to Seierstad by saying that there are no soldiers in the hospital, Seierstad can actually see many soldiers admitted to the hospital (Seierstad [2003: 262]; Sorensen [2007: 253]). In short, that victims of bald-faced lies know the truth seems to be a prerequisite of the bald-faced kind of lying and the reason why we think these liars do not intend to deceive.

But, whereas what Artie and other bald-faced liars deny is common knowledge, in *Pinocchio*, the prince does not know where Shrek is and, accordingly, he does not know whether Shrek went east. It follows that, since Charming knows nothing about Shrek's whereabouts, Pinocchio is actually being helpful: *by lying, Pinocchio tells Charming where Shrek is not*! This feature has substantial consequences for the interpretation of the example.

For one, we cannot say that Pinocchio is worsening the prince's epistemic condition with respect to Shrek's whereabouts. The prince at first knows nothing. Therefore, by asserting 'He went east,' Pinocchio is making Charming epistemically slightly better off: he can now warrantedly believe that Shrek did *not* go east. In addition, because the nose shows that Pinocchio is lying, the prince has also learned that Pinocchio believes that Shrek did not go east – insofar as Charming believes that a liar must assert what he believes is false. Finally, Lackey's argument would be that, because he prevented the police from accessing his testimony, Artie concealed the testimony which makes him deceptive. Yet, because Pinocchio obviously cannot prevent the prince from using the information obtained from his lie, Pinocchio cannot be fully deceptive by lying.

However, Pinocchio can still be sufficiently deceptive. By asserting 'He went east,' he is both concealing and revealing information: he is concealing Shrek's exact location by revealing where he is not. Thus, he is making the prince slightly better off but, surely, the prince would like to become even better off. He wants to know where Shrek *actually* is. So, maybe this could be the intended product of deception: preventing Charming from getting the whole truth? In response, I propose two variations of the case.

Let us modify Carson's *cheating student* example and suppose that the student is Pinocchio and that the Dean, who abandoned his old policy of not punishing those who deny cheating, does not know whether Pinocchio cheated. By asserting that he did not cheat, Pinocchio will reveal to the Dean that he did cheat. The Dean now knows all that he wants to know. Let this be *Cheating Pinocchio* case. Similarly, let us suppose that Shrek hid in Pinocchio's apartment and that the prince asked 'Is he in your apartment?' If Pinocchio lies by asserting 'No,' he will give away the whole truth. Let this be *Shrek in the House* case. On these two variations, the prince and the Dean get exactly what they want and thereby they have become as epistemically well off as they possibly could; in these cases, Pinocchio simply cannot conceal anything. Therefore, we must say, not only that Pinocchio is not being deceptive in *Cheating Pinocchio* and *Shrek in the House* but that he is being perfectly helpful. *Cheating Pinocchio* and *Shrek in the House* show that some lies simply cannot involve deception in any sense.

3.2 Intention to Deceive

The *Cheating Pinocchio* and *Shrek in the House* cases show that some lies cannot involve deception and that thus deception is not necessary for lying. However, they do not show that the *intention* to deceive is not necessary for lying: Pinocchio lied because he thought that he could be deceptive. Therefore, L_T, which requires the intention to deceive rather than actual deception, correctly counts these cases as lies. This result forces me to say that, although Nicole and Androvna can non-deceptively lie to themselves, they cannot form the relevant necessary intention, which comes down to conceding that lying to oneself is conceptually impossible. Let us, thus, consider what we could expect of Pinocchio if he knew that he cannot be deceptive with his lies.

There are three intuitions one might get while thinking about the situation of a person who knows that she cannot deceive by lying. The first two are consistent with the traditional view and thus I start with them.

Surely, Pinocchio would not form the intention to lie in any of these examples had he known that he cannot deceive by lying: in the original *Pinocchio* example as well as in the *Shrek in the House* variation, he wanted to help Shrek and, in the *Cheating Pinocchio* variation, he wanted to get away with cheating. Therefore, the first intuition is that those who cannot deceive by lying will not even try to lie. Alternatively, one might think that even the person who believes that she cannot deceive by lying may form the intention to deceive by lying. Lackey [2013: 242] most notably argues that a liar may aim to conceal information, i.e., be deceptive, even if this is impossible (see Broncano-Berrocal 2013: 228–229). Specifically, even if he knew that he cannot conceal anything, Pinocchio could have formed the intention to conceal it. Lackey [2013: 242] explains her position in the following way.

> I can aim to win a marathon even if I know that I will ultimately fail to achieve this goal. In this sense, even if 'conceal' is a success term … 'aiming to conceal' is surely not – that is, A can aim to conceal *x* from B even if A fails to succeed in hiding *x* from B.

Lackey's position is controversial (intentions do appear to be bound by rational constraints), while the argument in support of it is very problematic and it involves an irrelevant premise. I start with the latter concern. The issue whether A actually fails to hide *x* is irrelevant for the issue whether A will aim to hide it. What is relevant is whether A *thinks* that he will fail to hide it, which brings me to the main concern. Lackey is not showing that I can intend to conceal *x* from you even if I think that I definitely cannot do this; rather, considering her marathon example, she is effectively rejecting the idea that bald-faced liars really think they cannot deceive. That is to say, her argument requires that bald-faced liars think that deception is highly unlikely but not impossible. When Lackey says that she knows that she will ultimately fail to win the marathon but that she still aims to win it, what she actually means by 'know' is 'know that it is highly unlikely' rather than 'know for certain.'

Many people aim to win marathons even though they know that this is highly unlikely but none of them would try to do something they know that they will certainly fail: no one in their right mind would try to run a marathon from Florida to Cuba. Lackey's argument rests on probabilities. A rational person who believes that she knows for certain that she cannot conceal any information from her hearer by lying to the hearer will not aim, hope, or intend

to conceal this piece of information by lying. Lackey 'knows' that she will fail to win the marathon in the same way you 'know' that my lottery ticket is not the winning ticket on the basis that its probability of winning is only one in a million. Even if we – *pace* Williamson [2000] – say that you do know that my ticket did not win, we must nonetheless admit that you do not know for certain that I did not win – namely, we need to wait for the results of the draw to be announced – and that therefore I still have a chance of winning, no matter how slim it is. Had I known for certain that the ticket would not win, I would not have bought it (Williamson 2000: 59).

Most importantly, even if I am wrong and you may aim to φ even when you know for certain that φ-ing is impossible, there is still one intuition left to be mentioned: People who know that they cannot deceive by lying may lie without intending to deceive. Some of them might even use this condition to communicate the truth by lying. Let me, therefore, finish this section by introducing my central counterexample, *Pinartio*, a case in which Pinocchio lies in order to tell the truth.

> **Pinartio**: In a moment of foolishness, Artie told the police that he witnessed the murder. Unfortunately for poor Artie, Tony found this out, and now Artie is sleeping with the fishes. In pursuit of revenge, Prince Charming chases Tony all the way to Pinocchio's neighbourhood, where Tony hides in Pinocchio's house. Charming locks down the whole neighbourhood. Eventually, he knocks on Pinocchio's door asking whether Tony is hiding in his house. Pinocchio wants to give Tony away but he is afraid that if he gives any indication of this to Tony he might end up like Artie, and so he comes up with a plan.
>
> Pinocchio knows that the prince knows about his peculiar nose and that he knows that Pinocchio knows that he knows, etc. But Pinocchio also knows that Tony does not know anything about this. Therefore, Pinocchio asserts to Charming: 'Tony definitely isn't in my house.' Pinocchio does this not because he wants to deceive Charming in any sense but rather because he wants to tell him that Tony is in his house by way of Charming's recognizing his intention to tell the truth by lying (he definitely does not want Charming to think that he is protecting a murderer).

In *Pinocchio*, Pinocchio lies to Prince Charming by uttering 'He went east' with the intention to deceive the prince. The traditional definition, L_T, counts this as lying because Pinocchio asserts what he believes is false and intends to deceive the prince; the non-traditional definition, $L_{N\text{-}T}$, counts it as lying because he asserts what he believes is false allowing for an additional intention to deceive the prince; and I count it as lying because Pinocchio's nose grows. In *Pinartio*, however, Pinocchio lies to Prince Charming by uttering 'Tony definitely isn't in my house' intending that Charming (who knows that the nose grows

when Pinocchio is lying) infers directly from the fact that the nose shows that Pinocchio is lying that Tony *is* in Pinocchio's house, which is true and believed by Pinocchio to be true. The non-traditional definition counts this as lying because Pinocchio asserts what he believes is false, I count this as lying because the nose grows, and the traditional definition does not count this as lying because Pinocchio does not intend to deceive his addressee notwithstanding the fact that the nose indicates that Pinocchio is lying – this is why *Pinartio* is a counterexample to the traditional definition.

In the next section, I resolve some concerns that might arise with respect to Pinartio, mainly the concern that Pinocchio failed to lie here, and then, in Section 3.3., I conclude my argument by giving a positive answer to my title question: You can lie without intending to deceive.

3.2.1 Not a lie

A natural reflex of a deceptionist is to argue that Pinocchio simply failed to lie in *Pinartio*. And indeed, this proposal is consistent with the intuitions some people might have with respect to this example. Even Prince Charming actually said in the cartoon '*You can't lie*; so tell me puppet, where is Shrek?' Furthermore, many theories of asserting and assertion entail that his nose makes lying in Pinocchio's position (and especially lying without the intent to deceive) impossible.

For example, Carson [2006, 2010] and Saul [2012] understand asserting roughly as saying that *p* and warranting that *p* is true; Davidson [1986, 1997] and Fallis [2013] see it as saying that *p* and representing oneself as believing that *p*; while Unger [1975: 250–270], DeRose [2002: 185], and Turri [2013: 280] define it as saying that *p* and representing oneself as knowing that *p* (see Black 1952: 31). Since the nose prevents Pinocchio from warranting that *p* is true and from representing himself as believing or knowing that *p*, Pinocchio did not assert what he said on any of these theories. According to these theories, Pinocchio cannot lie at all, since the nose will prevent him from asserting what he says.

However, this consequence only shows that these theories are incorrect. The nose prevents Pinocchio from warranting the truth of what he says and from representing himself as believing by indicating that he already is lying. And any theory that entails that the nose prevents him from asserting what he says generates contradictory results: it says that Pinocchio is not asserting when the nose shows that he is lying. Therefore, asserting that *p*

should not be understood as warranting that *p* is true or representing oneself as knowing or believing that *p*.

It is important to say that not every theory of asserting will generate contradictory results with respect to the *Pinocchio* cases. The *Pinocchio* cases are all consistent with the view that asserting that *p* is proposing that *p* be added to common ground, which is background information shared by participants of the conversation (Stalnaker 2002). Since you need not believe or know that *p* or warrant it as true to be able to propose that it be added to common ground in the sense in which this counts as asserting that *p* (e.g., Stokke 2013; 2016; 2017), Pinocchio can lie on this account. Analogously, since you need not represent yourself to yourself as believing that ~*p* (e.g., 'Tony is not having an affair with Rachel') nor warrant that ~*p* is true to be able to assert that ~*p* to yourself, you may non-paradoxically lie to yourself by asserting that *p* to yourself – insofar as you do not intend to make yourself believe your own lie. I now proceed to discuss theories that belong to the Gricean view on assertion. They all entail that Pinocchio cannot lie to Prince Charming in *Pinartio*.

The classic formulation of Grice's [1989: 219] account of assertion says that I asserted that *p* by uttering *x* if and only if I uttered *x* with the intention of inducing in you the belief that *p* by means of your recognition of my intention. Therefore, if I assert something I believe to be false (a supposed prerequisite of lying), then I want you to come to believe what I believe is false, which is nothing other than intending to deceive you. Conversely, if I do not intend to deceive you by uttering *x* in the meaning of *p*, then I equally do not intend to induce the (false) belief that *p* in you by uttering *x* and, as a result, I do not assert that *p* by uttering *x* either. And, since one cannot lie if one does not assert what one utters, I do not lie by uttering *x*. On this view, then, bald-faced 'lies' are either aimed at deceiving or else are not assertions at all.[55] According to the classic Gricean formulation, because Pinocchio does not intend to induce in the prince the belief that Tony *isn't* in his house, which is what he says, he does not assert this.

Even a moderate version of Griceanism, which does not require intending that your addressee actually comes to believe what you say, entails that liars either intend to deceive or else they do not assert. Recall that, according to Keiser [2016: 470], a real assertion must be intended to give the audience a reason to believe what is said, and some theories demand that

[55] Faulkner [2007, 2013] and Meibauer [2014] are proponents of this explanation while Mahon [2011] and Meibauer [2016a: 110] argue that this is an implication of Carson's [2006, 2010] account of lying

competent asserters must be in a position to inform their interlocutors or affirm something with their assertions (Hinchman [2013]; Tollefsen [2014]). Pinocchio is not giving the prince any reason to believe that Tony is *not* in his house (~*p*), he is not informing the prince that ~*p*, he is not giving him a reason to believe that ~*p*, and so forth. Rather, Pinocchio wants Tony out of there and his assertion that Tony is *not* in his house (~*p*) is intended to be clear evidence (and it is clear evidence) that Tony *is* in his house (*p*). These theories, therefore, also predict that Pinocchio cannot assert what he utters. However, this consequence is a sign that these theories are flawed: as long as Pinocchio can utter 'Tony definitely isn't in my house,' he is asserting that Tony definitely is *not* in his house. There is nothing in the context of Pinartio that may prevent Pinocchio from asserting what he says (or prevent his nose from growing when he lies) and the fact that Gricean views entail that he cannot assert it speaks against these views rather than *Pinartio*.

Also, the *Pinocchio* cases cannot be dismissed by saying that positing an impeccable lie-detector such as Pinocchio's nose is epistemically unjustified or that this case does not correspond to a real-life situation. Recall that the case is devised as an *inter*personal analogue of an *intra*personal lie. Just as Prince Charming will know that Pinocchio is about to lie every time Pinocchio forms the intention to lie to him, I will know that I intend to lie to myself every time I form the intention to lie to myself. Furthermore, many people feel quite uncomfortable while lying and then they blush, make awkward gestures, and so on. These indicators are not as reliable as Pinocchio's nose but this is not a reason to think that they are not lying-indicators. A real-life lying-indicator need not 'go off' every time the person is lying. In fact, it need not go off at all. All that is required is that the person *thinks* that it will go off and that this will make it possible for her to tell her audience the truth by lying. Here is one plausible situation that combines the main theses of *Artie* and *Pinartio* to get a plausible real-life situation.

> **Artocchio**: Suppose that Tony hid in Artie's house. The police come to question Artie about Tony's whereabouts but Artie is too afraid to tell them – Tony might hear him. Fortuitously, Artie has a tick that gets triggered when he lies: he stutters. Tony does not know about this tick because Artie never dared to lie to Tony but the police know; he lied to them many times. Artie *believes* that he always stutters when he lies and he thus believes that he will stutter in this situation if he lies to the police. Being the smart guy he is, Artie decides to use this to tell the police the truth without alarming Tony. Therefore, when the police ask him 'Is he in your house?' Artie decides to tell them that Tony is *not* in his house hoping that (1) he will stutter, that (2) this will show that he is lying, and (3) that the police will realise that Tony *is* in Artie's house inferring this directly from (1) and (2).

The relevant dissimilarity between *Artocchio* and the *Pinocchio* cases is that Artie merely *thinks* that his tick will show that he is lying. And, the relevant similarity with *Pinartio* is that Artie aims to use his tick to tell the truth to the police – since he knows that the police know about the tick and that Tony does not. The issue of whether Artie will succeed in telling the truth by lying is irrelevant for this discussion: we are testing whether liars must *intend* to deceive their audience and Artie in *Artocchio* surely does not act on this intention. Rather, he wants to tell the police the truth by having them recognizing his lie.

All in all, in *Pinartio*, Charming knows that Pinocchio knows that Charming knows that he is lying, and Pinocchio knows that Charming knows this, and so forth. Because Charming knows that Pinocchio is lying and that Pinocchio knows that he knows this, Charming also knows that Pinocchio does not believe what he says and that Pinocchio knows that he knows this, and Pinocchio knows that Charming knows this, and so on. Pinocchio's lie is perfectly transparent to everyone engaged in this linguistic interchange as a lie meant to communicate the truth. Therefore, it is a lie not aimed at deceiving. A similar analysis applies to *Artocchio*; there, Artie *aims* (hopes, intends) that everything will work out just as it worked out for Pinocchio in *Pinartio*.

3.3 You *Can* Lie Without Intending to Deceive

The *Pinartio* variation of the *Pinocchio* case and the *Artocchio* version of *Artie* and *Pinartio* are the most obvious counterexamples to any account that takes the intention to deceive as necessary for lying. Pinocchio lies intending that the prince realises that Tony is in Pinocchio's house by recognising both (i) that Pinocchio is lying to him by uttering 'Tony definitely *isn't* in my house' *and* (ii) that Pinocchio is lying because he wants Charming to deduce from the lie that Tony *is* in the house, but that (iii) Tony should not know this (Tony will not understand the behaviour of the nose whereas Charming will). This is nothing other than Pinocchio lying to Charming intending to tell him the whole truth by way of Charming recognizing the full description of Pinocchio's intention. This lie clearly does not involve the intention to deceive the addressee in any sense. The fact that Tony is being kept in the dark is irrelevant for the question whether lies must be aimed at deceiving; Pinocchio is addressing the prince, not Tony. Besides, we can even imagine a situation in which no one is being kept in the dark with respect to anything.

Spoonocchio: Pinocchio and Charming are friends enjoying a dinner in Pinocchio's house and Charming asks whether Pinocchio has a spoon, which is in a cupboard. Pinocchio is

going through a rough adolescent phase in which he does everything opposite to established norms and so he is insincere when he should be sincere and sincere when he should not be (fully) sincere (e.g., when he should avoid offending his interlocutor). By asserting 'It's not in the cupboard,' Pinocchio would, by lying, be telling Charming where to get it.

This lie was not aimed at deceiving: Pinocchio was merely utilising his specific condition to make his adolescent 'rebel' phase more efficient and even Artie could try to use his stuttering for the same purpose. *Pinartio* and *Spoonocchio* show that Pinocchio can intend to lie while not intending to deceive the person addressed in any sense of 'deceive' – Pinocchio intends to tell the truth by lying – and we see that even a 'real boy' (namely, Artie) can easily form the same intention. Therefore, these cases are clear counterexamples to the traditional definition of lying, L_T, according to which liars must intend to deceive their addressee, and they support, $L_{N\text{-}T}$, the non-traditional definition of lying that does not have the intention to deceive as a necessary condition of lying

4. Self-deception vs. Lying to Myself

$L_{N\text{-}T}$, in virtue of not requiring the intention to deceive, seems to be the right account of lying. It rightly counts all Pinocchio cases as lies but not the coach example. Even more, $L_{N\text{-}T}$ can be used to explain a very interesting kind of behaviour thought to involve self-deception, the Nicole and Androvna kind of cases. Recall, the hallmark of the Nicole case is that Nicole sincerely avows that $\sim p$, i.e., 'Tony is not having an affair with Rachel', but nonetheless acts as if p is true, i.e., she avoids driving past Rachel's house when Tony might be there, which seems to be indicating that 'deep down' she still knows the truth. Attempting to avoid saying that she believes both that p and that $\sim p$, scholars developed many ingenious solutions; however, I dismissed them all as unsatisfactory (III-3).

Even though it is most commonly understood as involving self-deception, Nicole's behaviour makes best sense on the hypothesis that Nicole is merely lying to herself with no intention of deceiving herself. I will offer a detailed account of this behaviour in subchapter V-1, once I have the full theory of lying in place. Here, I will just highlight some upsides of the proposed explanation. First and foremost, taking it that Nicole is lying to herself is consistent with our linguistic practices. It does seem quite fitting for her girlfriends to say 'Nicole, stop lying to yourself. He is cheating on you. Face it!'. And indeed, she is lying to herself precisely because she cannot 'face it (the truth)'. Her lie just is intentionally not 'facing it'. It is turning her head the other way.

Adding that she is lying to herself with no intention to deceive herself explains her behaviour without raising any concerns or the need to posit conflicting mental states. There is no tension between her avowals and her behaviour – she is lying to herself – and her behaviour is thus consistent with what she believes. Having that said, it is vitally important to notice that Nicole is not addressing her girlfriends at all but rather herself. Their questions acted as a trigger of distressing emotions and memories, to which Nicole reacts by lying to herself – lying is a form of damage control. And, since she is not addressing them, she is not lying to them either (see Kenyon [2003: 245]; Leland [2015]), hence, the intuition that she is being insincere towards herself but not towards them remains unharmed.

Furthermore, the 'lying to oneself' kind of self-insincerity (that may lead to self-deception) significantly contributes to some very successful theories that understand self-deception as involving living in a pretence, such as those offered by Sartre, Bach, and Gendler. Sartre conceptualised bad faith, i.e., deception about the self, as lying to oneself and living in pretence; $L_{N\text{-}T}$ makes perfect sense of this. A person in bad faith plays *at being* in a certain social role (a good parent, waiter, etc.) while knowing her own desires. By lying to herself, she refuses to acknowledge, and thus legitimize, her desires; this is how she keeps them subdued.

Similarly, Gendler's [2007; 2010] position makes best sense if taken as saying that pretending that ~*p* ('Tony is not having an affair') is a way of avoiding the thought that *p*. I objected that this seems to fall short of self-deception – since not all pretence involves insincerity, it is hard to see where the insincerity towards oneself is to be located. By saying that Nicole is lying to herself by asserting that ~*p*, we avoid this concern – lying is an obviously insincere kind of pretence. The same point may be used to reinforce Bach's [1981, 2009] theory: he says that self-deceivers avoid the thought that *p* and is there a more obvious way of avoiding it than lying to yourself by asserting that ~*p*?

IV-2. Must Liars Assert What They do not Believe?

1. Map of the Overall Argument

In chapter II, I argued that paradoxes of self-deception arise from flaws in received theories of interpersonal deception that merely become apparent when these flawed theories are applied to self-deception. As a replacement, I suggested the manipulativist view, according to which deceivers need not intend to produce any harmful consequences in their targets. On this view, the paradoxes of self-deception need not arise: I may intend to make myself believe what I merely guess is true (as in the case in which I deceive myself into believing that I left the stove on based on my irrational fear that it might be on); similarly, the shipowner makes himself believe that his ship will make it to the port by bs-ing himself; and so on.

Nevertheless, while it may be successfully used to non-paradoxically model self-deception on interpersonal deception, the manipulativist view cannot explain any of the four main cases I introduced at the outset. Therefore, in the previous Subchapter (IV-1), I resolved problems pertaining to cases such as Nicole's by arguing that liars need not intend to deceive, as what is prescribed by $L_{N\text{-}T}$. $L_{N\text{-}T}$ is not just an account of lying more successful than L_T but also a very useful tool for analysing human *intra*personal communication. Nicole vehemently asserts that Tony is not having an affair with Rachel, but nevertheless behaves as if she believes the opposite notwithstanding. In contrast to the majority of philosophers (see IV-1), I argued that this case may not involve self-deception but rather a kind of insincerity towards oneself that does not involve the intention to deceive oneself. Nicole is non-deceitfully lying to herself out of despair; she cannot face the truth and, by lying, she avoids facing it. She is 'living in a lie.'

However, taking it that Nicole is bald-faced lying to herself, in turn, raises the concern that 'non-deceitful lying to yourself' solution may leave some cases similar but not identical to Nicole's unresolved. Some Nicole-type people might actually be trying to deceive themselves by lying and $L_{N\text{-}T}$ cannot help us make sense of their behaviour. That is, since $L_{N\text{-}T}$ says that liars must assert what they believe is false, self-deceivers who intend to deceive themselves by lying will still be involved in believing clear contradictions. The problem that I cannot

intend to deceive myself by lying remains. Therefore, in this Subchapter, I will argue in favour of a theory of lying that does not contain the condition that liars must assert what they believe is false – as what generates the paradoxes of self-deception. This novel theory of lying will make many cases of self-deception non-paradoxical.

2. Introduction

When one thinks about lying, one's typical first thought is of people who intentionally assert something they believe to be false, or at least think that they believe to be false – I call this *the Belief Condition on lying*. Existing philosophical accounts of lying are all consistent with this intuition: e.g., Davidson [1986, 1997]; Derrida [2002]; Faulkner [2007, 2013]; Sorensen [2007]; Carson [2006, 2010]; Fallis [2009, 2010, 2012, 2013, 2014]; Saul [2012]; Lackey [2013]; Stokke [2013, 2016a, 2016b]; and so on. In fact, social scientists generally adopt the Belief Condition; Vrij [2008] is just one example. I, however, think that this intuition is incorrect. Even more, I will argue that liars can lie by intentionally asserting what they believe to be true.[56]

I base my argument on the case of a man, Peter, whose situation is such that he can sincerely assert that *p*, which is what he consciously believes to be false, to any addressee, including himself (Section 3.1), and who, equally, in that same situation, can lie by asserting that ~*p*, which is what he consciously believes to be true, to any addressee, including himself (Section 4). I will argue that, because Peter can lie by intentionally asserting what he believes, this case provides a clear counterexample to the Belief Condition. Furthermore, since he can sincerely assert what he consciously believes to be false, the case also shows that the *Belief Account* of sincerity, according to which I am sincere (if and) only if I assert what I believe (Searle [1969: 64–65]; Austin [1975: 50]; Lewis [1975: 167]; Grice [1989: 27, 42], Williams [2002: 74]; Owens [2006: 111]; Sorensen [2011: 408]) or what I think I believe (Ridge 2006), is incorrect.

While no one, to my knowledge, has argued against the Belief Condition on lying, many contemporary philosophers have opposed the Belief Account of sincerity (you are sincere in asserting that *p* if you believes that *p* or think you believe it): for instance, Douven [2006],

[56] Turri and Turri [2015] argued that people cannot be taken as liars if their statements – unbeknownst to them – turn out to be true, but Wiegmann, Samland, and Waldmann [2016] disproved their findings. My discussion in no sense hinges on this issue. I am interested in whether they must believe that what they assert is false; it is this thesis that causes explanatory problems I aim to resolve.

Chan and Kahane [2011], Pruss [2012], Stokke [2014], McKinnon [2015], and so forth. While I agree that saying what you believe or what you think you believe need not make you sincere in asserting, I am not sure that their examples justify abandoning the Belief Account altogether. Some of them have already been shown faulty: for example, Fallis [2012: 573] elegantly rebutted Pruss's [2012] case, while Stokke [2014: 503] challenged the proposed interpretations of some of Chan and Kahane's [2011] most important cases. With other examples, asserters are indeed justified in asserting what they do not believe, but only due to various factors – e.g., pragmatic factors (Douven 2006: 453, n. 6), or those arising from the addressee's particular doxastic context (Hinchman [2013]; McKinnon [2015: 61–2, 68–71]) – that make it unclear which norms are actually operational in the given context or whether these speakers have sacrificed their own sincerity for the hearer's benefit (i.e., there is a worry that they make the hearer epistemically better off by being insincere).

My case, I will argue, is much more convincing than any other supposed counterexample to the Belief Account offered thus far. Peter is able, for a certain proposition *p*, sincerely to assert that *p* while consciously believing that ~*p*. In fact, because he is quite aware that he believes that ~*p*, he is also able to sincerely assert that *p* but that he believes that ~*p*, giving us a case in which a claim of the form of the strong version of the so-called Moore's [1993: 207–11] paradox may sincerely be asserted, '*p* and I believe that ~*p*.' This is a feature other purported counterexamples to the Belief Account typically do not have. Most importantly, I will argue that Peter can lie by asserting the proposition he believes to be true, i.e., that ~*p*, and he knows that. The case is important because, by invalidating the Belief Condition on lying, it not only helps us with understanding some cases of self-deception but also because it invalidates all existing theories of lying – every single one.

The remainder of this Subchapter has two main parts. In Section 3.1., I set out my key example, the case of a delusional patient, Peter. Then (in Section 3.2), I refine my proposed interpretation of Peter's case, offer a novel two-factor theory of delusions, and defend both the interpretation of the case and the theory of delusions from possible objections. I also highlight the advantages of the proposed theory as compared to existing theories of delusion. Section 3.5 finalises the argument concerning the Peter case and introduces a surprise: the case may be deployed as a substantially novel counterexample to the 'JTB' ('Justified True Belief') account of knowledge – one may know that *p*, not just while not believing that *p*, but also while consciously believing that ~*p*.

In Section 4, I argue that by asserting that ~*p*, which is what he consciously believes to be true in the specific situation as described, Peter would be lying to any relevant addressee, thereby supporting a negative answer to the subchapter title question: it is not necessary for lying that the liar says what he does not believe. I conclude by briefly proposing a tentative account of assertion and asserting (4.1) and a plausible theory of lying that can explain Peter's lie (4.2) and be used to non-problematically explain cases of deceiving oneself by lying to oneself.

3. Sincerely Asserting What One Does Not Believe

3.1. The Peter Case

Dennett [1996: 111] writes that cases of Capgras delusion, the delusional thought that one's relative has been replaced by an identical-looking double, could be used as thought experiments for personal identity debates. By considering a case of a patient in the early stage of recovering from his Capgras delusion, I will demonstrate that this disorder is even more interesting for discussions about asserting and lying, and for epistemology generally.

The case I offer is based on a case of a male patient in the early stage of recovering from his Capgras delusion, a deluded belief that his wife has been replaced by an identical-looking imposter. Delusions are commonly explained using two-factor frameworks, which posit both impaired thought-formation and impaired thought-evaluation. One-factor frameworks – such as the explanation that delusions are reasonable hypotheses given the strangeness of the experience (e.g., Maher [1974; 1999]; cf. Davies et al. [2001]) – that do not require impaired thought-evaluation are not uncommon, however (see Bortolotti 2013).

As a starting point in explaining the Capgras delusion, I will use Turner and Coltheart's [2010] version of the two-factor framework and then offer some important refinements to this theory. Following others, they argue that the first factor, i.e., impaired thought-formation, starts with a relevant neurological deficiency – an impairment in the person's perceptual, affective, or mnemonic processing (Breen et al. [2000: 64–65, 68–70]; Ellis and Lewis [2001: 154]; Brighetti et al. [2007]) – which prompts the delusion and determines its content by generating abnormal data that affects the person's perceptual experience. The delusion arises when the perceptual cognitive module either explains this abnormal data using Bayesian abductive inference (Coltheart, Menzies, and Sutton [2010]; McKay [2012]; Davies and Egan [2013]), the so-called 'explanation' model, or simply endorses the input that presents the

spouse's identical lookalike (Davies and Coltheart [2000]; Davies and Egan [2013: 711–712]), the so-called 'endorsement' model (see Bayne and Pacherie [2004: 2–3]; Aimola-Davies and Davies [2009]).

Although this syndrome is sometimes exclusively visual – some patients recognize their spouse or a parent over the telephone (Hirstein and Ramachandran [1997: 438]; Stewart [2004: 65]) – blind patients may also suffer from this delusion (Rojo et al. [1991]; Reid, Young, and Hellawell [1993]; Hermanowicz [2002]). Therefore, impaired perception should not be understood strictly as visual perception. Having said that, I offer a brief account of the experience of patients whose delusion is based on visual perception.

Familiar faces have significant emotional value (positive or negative) and affective information is an integral part of familiar face recognition.[57] Therefore, the most common initiator of Capgras delusion, as initially correctly proposed by Capgras and Reboul-Lachaux [1923: 127–8], is the missing affective response, consciously experienced as 'the feeling of strangeness,' associated with the perception of the relevant familiar person. These patients suffered damage to neural pathways underpinning the emotional component of face recognition (the first factor) – i.e., their autonomic nervous system is disconnected from their facial recognition system (Ellis, Young, Quayle, and De Pauw [1997]; Ellis, Lewis, Moselhy, and Young [2000]; Brighetti et al. [2007]) – as a result of which they do not have affective reactions to perceiving people they know. In short, the delusion develops because the autonomic response to the specific familiar face is just as if the face was a face of a stranger. Here is one illuminating report: 'Mum's walked in the room … well, this picture of Mum … and started talking but it was only a picture of her but it didn't feel like her' (Turner and Coltheart 2010: 372).

The second factor, impaired thought-evaluation, has two levels: it consists of failures of unconscious and conscious processes that constitute two monitoring frameworks that check occurent thoughts. The *unconscious* checking system (*U*CS) either 'tags' suspect thoughts as requiring extra conscious evaluation, or 'passes' them thus conferring conviction (Turner and Coltheart 2010: 353). If the *U*CS 'passes' the delusion-generating explanation/endorsement of the input, it reaches the consciousness as a proper fully-fledged thought that need not be checked for consistency, which then leads to a '(preconscious) feeling of rightness.' That is,

[57] Tranel et al. [1985: 407]; Tranel et al. [1995]; Stone and Young [1997: 358]; Breen et al. [2000: 64–65]; Ellis and Lewis [2001].

it does not feel as if one knows the person one sees and this feeling is taken as correct. Because the feeling of rightness accompanies delusions, Turner and Coltheart – following Gilboa et al. [2006: 1412–1413] and Gilboa [2010: 154] – take it that the person's *U*CS is impaired. Nevertheless, the *conscious* checking system (*C*CS) still may examine the delusion. *C*CS is typically instigated by an experience of doubt, which accompanies 'tagging' of the delusion by the *U*CS as requiring conscious checking, but it can equally be triggered by the person herself or by her friends or therapists. From the fact that the patients' *C*CS was not initiated, Turner and Coltheart [2010: 372] inferred that this system is also impaired.

Our *C*CS monitoring framework consists of various abilities. One is the ability to make *plausibility judgements*,[58] namely, the ability to judge whether thoughts or cognitive inputs are plausible in the context of everyday knowledge about the world; another is the ability to conduct *reality monitoring*, namely, checking whether a thought represents an externally derived experienced event or an imagined event; and so on (Turner and Coltheart 2010: 362–8). If some of the Capgras delusion patient's relevant *C*CS capacities start recovering, the person will begin to evaluate her delusion and therefore question its truth. However, the delusion will not be abandoned or revised instantaneously; hence, the person's state becomes extremely interesting at this point.

The patient in the following case is recovering from Capgras delusion (Turner and Coltheart 2010: 371).

Case 1: (italics and comments added)

> **Examiner**: What has made you realise that they're just imaginations? People telling you that?
> **Patient**: No *it's just myself* that's done it. *I've started going through it* [self-initiating some *C*CS functions], and seeing what could possibly happen and what couldn't happen [making plausibility judgments]. … Mary couldn't suddenly disappear from the room, so there must be an explanation for it. So then I try and work out what. ... She [the 'double'] knows me way back. The lady knows me way back. She could say things that happened 40 years ago, and I wonder where she gets them from. And then *I worked it out* and I've wondered if it's Mary all the time. *It's nobody else* [conclusion of the conscious judgement].

I will now slightly modify this case in order to construct my counter-example: let us suppose that the patient is Peter, a brilliant philosopher and neuroscientist. This modification is important for the following reason. Recovering from a delusion is a very confusing and stressful state to be in and many patients do not realise what is happening to them; therefore,

[58] Parrott [2016] understands this as the ability to assess epistemic possibility of an input.

responses of these patients are misleading. My modification of case 1 is plausible, since I add only that the patient in question is well informed about his condition, and it avoids the concern that the testimony of the sufferer is not reliable.

The main points I want to draw from my modification, and which I defend below, are these: as the case says, Peter was able to realise that the lady is his wife by making a series of inferences; he 'worked it out.' Nonetheless, given the facts about how delusions are produced, I argue, it is very likely that, because the first factor is still operational and because the *U*CS and the *C*CS still pass the belief that the lady is not Mary as true, Peter still believes that the lady is not Mary. Finally, as a brilliant philosopher and neuroscientist, he correctly understands his overall cognitive situation. Therefore, on my variation, when the doctor, *right after* Peter's conscious judgement ('It's nobody else'), asks him 'So, do you now really believe it is Mary?' Peter sincerely asserts 'It's gotta be Mary but I believe it's not Mary.'

Please notice two things. The idea is that the brilliant neuroscientist-philosopher Peter realises that the lady is most definitely Mary but that he, due to his delusion, still has the full belief that she is not Mary. Moreover, Peter asserts '*p and* I believe that ~*p*' because (i) he wants to be informative by directly answering the question (thus '*p*') *and* because (ii) he fears that asserting only that *p* may mislead his colleague into thinking that he actually believes that *p* by way of implicating this information in the given context (thus 'but I believe that ~*p*'). Having that said, I add some further steps to the dialogue above (let the examiner's name be C):

Peter case (case 1 continues)**:**

Peter: And then *I worked it out* and I've wondered if it's Mary all the time. *It's nobody else*.
Examiner: So, you now believe it's Mary?
Peter: No, no – I honestly can't say that I do, C. It must be Mary, but it still doesn't feel right that it is.
Examiner: What are you saying, Peter?
Peter: C, *it's gotta be Mary but I believe it's not Mary*.
Examiner: You still believe that she's *not* Mary? This surely can't be if you realise that she's Mary.
Peter: That's the point, C. I am saying that, although my knowledge of Mary and of the nature of delusions clearly tell me that the lady is indeed Mary and that I'm delusional, the delusion hasn't lifted – I still believe that she's not Mary.
Examiner: Are you sure Peter? Maybe, you're just confused. Remember that guy we had last week?
Peter: C, *I'd be lying if I told you it isn't Mary, but this is what I nevertheless believe*!

Since believing that ~*p* is equivalent to believing that *p* is false, in this situation, Peter sincerely asserts what he consciously believes to be false and, I seek to show in Section 4, he is right when he says that he can lie by asserting what he consciously believes to be true – two features none of the cases so far raised in the literature have. If I am right, then, because Peter can lie by asserting what he believes to be true, the case invalidates both the Belief Account of sincerity and the Belief Condition on lying.

Please note this very carefully, the claim is *not* that Peter will never abandon his belief that the lady is not Mary, nor that he will not occasionally fluctuate between believing and not believing the same proposition (and I will not argue that all delusions are beliefs). The claim is that, *at the time of asserting* 'It's gotta be Mary but I believe it's not Mary,' Peter consciously believes that the lady is *not* Mary despite his judgement that she *is* Mary! Accordingly, my main task here is to give you good reasons to accept that Peter actually believes that the lady is not Mary while sincerely asserting that she is Mary. The following aetiology of Peter's delusion will be defended.

Despite judging that *p* (the lady is Mary) is true, Peter still fully believes that ~*p* based on four factors. The first two are (1) his abnormal experience – it doesn't feel like Mary – and (2) the endorsement of the perceptual input as not depicting Mary. What makes the thought that ~*p* effective as a belief when it reaches consciousness is that (3) the unconscious checking system (*U*CS) did not tag it as suspicious and that (4) *C*CS, who has detected that the thought is incorrect, cannot revise it. In short, factors (1) and (2) generate the belief by providing the content, (3) makes it operational by passing it, while (4) cannot revise it.

The following mechanism generates the delusion. Because it does not feel like her (step I), the visual perceptual processing module registers the perceptual input as incomplete (step II), and, because the input is recognized as incomplete for a representation of Mary, the perception becomes endorsed as a perception of Mary's identical look-alike (the 'lady') – step III. The delusional belief results from a normal cognitive response to the relevant input: because the input does not incite the affective response cognitively expected for a perception of Mary, the perceptual cognitive module understands it as not depicting Mary. Given that it corresponds to the input, the *U*CS passes this new thought (step IV), which then emerges into Peter's consciousness as a fully-fledged belief (see 3.3). Finally, due to the relevant fine-tuned impairment of the right hemisphere, *C*CS is powerless both to prevent this thought to become a belief and to revise the belief, once formed (step V) (see 3.4).

Under this aetiology, there is no implication that the content of the delusional belief was obtained incorrectly. Rather, the belief results from a normal cognitive response to the relevant input: because the input does not incite the affective response cognitively expected for a perception of Mary, the perceptual cognitive module understands it as not depicting Mary. Had Peter not been neuropsychologically impaired, he would have had the predicted affective response to his wife and, thus, he would have formed a non-pathological belief that the lady is Mary using the same sub-personal belief-forming mechanism, i.e., (I) + (II) + (III), approved by (IV). Therefore, in the case as it stands, the conscious belief that the lady is not Mary is a basic or non-inferential belief, sub-personally generated through Peter's perception of the lady.

I now proceed to try to resolve concerns that my interpretation of this case raises, starting with the one that it is implausible to think that Peter believes that the lady is not Mary.

3.1.1. A Problem

Many philosophers argue that delusions are not beliefs; rather, they might be imaginings (Currie [2000]; Currie and Jureidini [2001]; Currie and Ravenscroft [2002]), faulty perceptual inferences similar to illusions (Hohwy and Rajan 2012), default thoughts that occupy the role of beliefs (Gerrans 2014), or even bimaginations, states 'in between' belief and imagination (Egan [2009: 263–80]; see Bayne [2010: 332–4]). On these views, Peter does not really believe that the lady is not Mary; he merely believes that he believes it.

Apart from the obvious paradoxicality and bizarre nature of many delusions, the deluded patients' seeming disregard for the truth and the fact that they typically do not act on their delusions also support the view that delusions are not beliefs: Capgras patients, for instance, typically do not call the police or express concern about the missing person, sometimes they live with the 'lookalike.' People suffering from Cotard delusion – the conviction that one is dead or that one does not exist – sometimes try to commit suicide, and so forth. Finally, it appears as if even some delusional patients who know that they are delusional do not believe their delusions. Taking it that they suffer from a strong illusion seems as a better explanation. A good example is the case of Mr F., a 39-year-old male who suffered from delusions of grandeur. Here is an excerpt from an interview with Mr F. (italics added).

> I can be like Moses, or like the Satan. You understand, doctor?… I am a very sick man. … I would like to have a normal life, like everyone. … But I cannot manage with a simple life

because *I always have thoughts of grandeur in my head*. (Zislin, Kuperman, and Durst 2011: 116)

It seems quite implausible to suppose that Mr F., who committed a suicide two years after this interview, really believed that he was Moses or Satan. Nevertheless, cases like this do not support the generalisation that delusions are never believed, and since my present aim is to establish only that Peter's peculiar and temporally limited cognitive situation is possible, I will not involve myself in debates concerning the general nature of delusions (however, see Bayne and Pacherie [2005]; Bortolotti [2010]). Instead, I will propose that, even if delusions are not typically beliefs, we have every reason to regard *Peter's* bizarre propositional attitude (that the lady is not Mary) as a belief. This is because it satisfies three typically accepted features a propositional attitude needs to have in order to be counted as belief. Firstly, it is truth-directed and truth-regulated; i.e., it is generated in the way required for belief. Secondly, Peter exhibits necessary dispositions for his propositional attitude to be counted as a belief. Finally, Peter will eventually revise and abandon the delusion; it follows that his mental state is revisable, and being revisable is another feature typically ascribed to beliefs.

Because I acknowledge that Peter will recover from his delusion (my argument requires only that he will not abandon the belief instantaneously), I proceed to sketch my main argument, which I then defend in Sections 3.3–3.5.

3.2. Sketching the Argument

Shah and Velleman [2005] write that to answer the question *whether to believe that p* is to answer the question *whether p* (similarly, Evans 1982: 255). Obviously, Peter consciously sees that his propositional attitude that the lady is not Mary is somehow flawed. Nonetheless, it still satisfies their criteria of being truth-directed and truth-regulated. For one, the truth, as the standard of correctness, was applied sub-personally – the truth-tracking mechanism presented in steps I–III (see 3.1) was initiated by the perceptual cognitive module. However, for a mental state to count as a belief, on this view, the normative standard of being correct if and only if true must also be applied; i.e., the mental state must also be truth-regulated. On the Shah-Velleman theory [2005: 502–506, 516], if one judges that *p* and affirms that *p*, if the affirmation produces a standing representation of *p* as true, then this affirmation typically becomes the belief that *p*. This standard, in Peter's case, is also met sub-personally: the *U*CS's passing the thought as correct, step IV, is a sub-personal application of the normative standard of correctness.

In brief, steps I–IV involve the Shah-Velleman type of doxastic reasoning that generates beliefs on a sub-personal level: passing the belief, step IV, is applying the normative standard, whereas endorsing the input as veridical, steps I–III, is applying the descriptive standard of belief. The truth-directedness is owed to the work of the perceptual cognitive module while truth-regulation was performed by the *U*CS. Therefore, Peter's propositional attitude that the lady is not Mary satisfies both standards of belief-formation to count as belief and so it has what we may call the first typical feature of beliefs.[59] The next step is to show that Peter will act on his belief.

Because of the work of his cognitive systems, it seems reasonable to say that Peter will, despite his conscious judgement, for some time be unable to escape the feeling that the belief is credible and will be strongly disposed to act on it, which explains why Peter exhibits dispositions relevant for believing that the person is not Mary. Not exhibiting some other dispositions – he will not call the police, for instance – is not evidence that he does not believe that the lady is not Mary. True, Peter exhibits all kinds of dispositions, but we must discriminate those relevant to the question what he believes and those that are not relevant. And, since Peter is aware that he is delusional, dispositions consistent with his conscious judgement are not relevant for what he believes. After all, the judgment says that his belief is false. But, how to know which particular dispositions are relevant or not? There is no general answer to this question, yet we may still say something informative about the case.

For example, we must take it that Peter does not have the affective or emotional reactions that would correspond to reactions toward his wife: he does not feel affection towards her, he feels uncomfortable and strange in her presence, he likely struggles to be frank in front of her, to be intimate with her, and so forth. These are not dispositions we could easily disregard and say that he believes that the lady is Mary (p). The lack of some other dispositions that would normally be implicated in believing that the lady is not his wife ($\sim p$), such as the disposition to find out what has become of her, etc., can be satisfactorily excused by acknowledging that Peter's judgement that she is Mary must have affected his behaviour in some way (see Schwitzgebel [2002: 253–7]; Bayne and Pacherie [2005: 184]; Bortolotti [2011: 80]).

It is vital for the correct understanding of delusions to bear in mind that acting on a thought does not require adopting the thought as a belief. Peter may exhibit some dispositions

[59] A belief being generated subpersonally is consistent with Shah and Velleman's [2005: 516] theory.

that are consistent with the belief that the lady is Mary – he may treat her just as he would treat Mary – while actively believing that she is *not* Mary. And *vice versa*, not acting on a thought does not entail not believing it. A person who akratically believes that the plane will crash will judge that her belief is irrational and typically remain seated. Likewise, Peter may have an incredible urge to call the police or scream in pain calling for his wife but abstain from acting on his urges because he judges that the lady is Mary.

Coltheart [2007: 1053–4] describes a case of G.R. that closely resembles Peter's case (G.R., however, thinks that his wife has multiple doubles, which he calls 'apparitions'). G.R. asserts that *p* but *immediately* proceeds to act as if he believes that *~p*, which is explained by claiming that the man fluctuates between contradictory beliefs (also Coltheart, Langdon, and McKay 2007). However, I will challenge this interpretation in Section 3.2 by arguing (1) that belief is not merely a matter of judgement (Scanlon 1998: 25, 35) and G.R. asserts what he judges to be true, (2) that acting *as if* a proposition is true, i.e., asserting what one judges, need not entail that one believes that proposition (Tumulty [2012]; Gerrans [2014: 137–143]), and (3) that Coltheart's explanation incorrectly presupposes that believing is a binary (on/off) matter (see Subchapter V-1). In fact, (4) many cases understood as people abandoning their delusions only to have them rapidly return, I propose, are in fact misdescribed. These people never abandoned their delusion; they just judged the opposite of what they believed and reported on their judgement rather than on their beliefs, which is something even non-delusional people do quite often when speaking sincerely (notably, Lackey [2007]; Stokke [2014], see Section 2).

In conclusion, the main argument is this: because it was caused in the proper way, because it is revisable, and because Peter exhibits relevant dispositions, we may say that Peter is correct in saying that he believes that the lady is not Mary. Nonetheless, this still leaves three concerns open. The first is, if the thought was passed by the impaired *U*CS, then we may say that the normative standard of being correct if and only if true was not applied correctly – the unimpaired *U*CS would not have passed it – and thus the thought that the lady is not Mary is not a belief. The second concern is that, even if the thought reached consciousness as a fully-fledged thought, this thought is not a belief because beliefs are personal-level phenomena and some person-level features of a belief are missing. Finally, even if the thought reached consciousness as a fully-fledged belief, it seems natural to think that Peter's conscious judgement will override the sub-personal mechanism and the belief will be revised

instantaneously; if not, then the thought was never a belief. I will address each of these concerns in succession.

3.3. Refining Step IV

Following Gilboa et al. [2006] and Gilboa [2010], Turner and Coltheart [2010: 360–2] posited that an impairment of the unconscious checking system (*U*CS) caused the sub-personally generated thought to emerge in the patient's consciousness as a fully-fledged delusion – step IV. Had this system been working reliably, their rationale is, the thought would have been tagged as suspicious, which would then trigger a sense of doubt or unease about the thought and thereby alarm the *C*CS. This explanation, in turn, raises the concern that Peter's delusion is not a belief because the normative standard – step IV – was not applied correctly. I resolve it by rejecting the view that *U*CS needs to be impaired for a delusion to arise.

Indeed, the fact that delusional patients are not surprised by their passed-as-proper delusions when these emerge in their consciousness indicates that a mismatch between expectations and reality was not detected. Turner and Coltheart [2010: 355] find the lack of surprise and the inappropriate feeling of conviction indicative of the impaired *U*CS. I, however, wonder why anyone would be surprised by not having an affective reaction to an unknown person that looks identical to their spouse, since this is what Capgras patients actually perceive.

To be precise, I do not deny that these people should be surprised that what they perceive is not their spouse but rather a person that looks identical to the spouse, i.e., the spouse's 'picture' (*picture* henceforth). Considering the low probability of an event such as seeing the *picture* rather than the spouse, their conscious checking system (*C*CS) should signal that this thought should be reconsidered. The lack of this kind of surprise indicates that the patients' conscious ability to make plausibility judgements is impaired (Parrott 2016) and studies do suggest that delusional patients overestimate the probability of certain events; they 'jump to conclusions' on probability tasks.[60] Furthermore, I do not deny that these patients should feel surprised by the fact that everybody else recognizes the *picture* as their spouse (Parrott and Koralus 2015: 401, 405). This is also a failure of *C*CS, not of *U*CS.

[60] Huq et al. [1988]; Garety et al. [1991]; Garety and Freeman [1999]; Garety et al. [2005]; Fine et al. [2007]; see Langdon, Ward, and Coltheart [2010] and So et al., [2012].

What I do deny are two related thoughts. Firstly, I deny that these people should be surprised by their not having an affective reaction to the person they perceive: they do not perceive the spouse but rather the *picture* – perceiving familiar *faces* does not require affective reactions but perceiving familiar *people* does. Capgras patients are not aware of the lack of autonomic response to the person they see (Davies et al. [2001: 140]; Coltheart, Langdon, and Breen [2001: 140]; Coltheart [2005: 155–6]); rather, they just have the experience of seeing the *picture* and thus they cannot be surprised by the lack of autonomic response – *perceiving the *picture* is not supposed to have one*. Secondly and vitally, given their experience, I deny that these people should be surprised by their perception of the *picture* (*pace* McKay [2012: 340]; Parrott [2016]). I defend these two claims in turn.

Perception of familiar people is based on the process of systematically matching inputs with relevant stored 'schemas,' which are structured internal representations of familiar items acquired through perception.[61] Schemas act as top-down predictions of future inputs based on what we know about the world, and matching predictions of familiar people is highly dependent on expected affective reactions (Tranel et al. 1995: 407), as opposed to hair colour, hairstyle, a type of garment, or similar – changes in the latter may even go unnoticed (Friston 2002b: 125–6). This is because neuronal activity in sensory areas biases perceptual decisions toward correct inference and not toward a specific percept (Feldman and Friston 2010: 17). Therefore, when a vital ingredient such as the affective reaction is missing, the correct perceptual decision is 'This is not my spouse,' and when the hairstyle is different, the correct perceptual decision is 'This is my spouse, but not how I remember her.'

Because the input does not match the vital ingredient of the spouse-schema when Capgras patients are looking at their spouses, these people simply do not perceive their spouses but rather their *pictures*. This is something like seeing your wife's identical twin – you know which is which but the mechanism of discrimination is not available to your introspection (Bayne and Pacherie [2004: 5]; see Gregory [1997]) – and thus *U*CS, which checks whether the input is consistent with the thought, has absolutely no reason to prompt surprise.

I am not saying that delusional people – Peter, in particular – will not initially experience *a moment* of surprise characteristic to that of seeing a person looking just like someone you know. They may initially be surprised *to* see the *picture* (cognitive expectations) but they

[61] Ellis and Lewis [2001]; Friston [2002a, 2002b, 2005, 2009]; Friston and Stephan [2007: e.g., 437–438, 443–444]; Hohwy [2007: 322].

will not be surprised *that* they see the *picture* (veridicality of perception). That is to say, they may exhibit a kind of (*C*CS) surprise identical to you being surprised *to* see an old friend in a coffee shop (or a snowflake in a sauna), but they will not exhibit a kind of (*U*CS) surprise identical to you doubting *that* you are really seeing your old friend (see Friston and Stephan 2007: 425–6, 435–6). My overall argument, then, is that Capgras patients – having no reason to doubt their senses – are not surprised *by* their perception or the resulting delusion although, at first, they may be surprised *to* see the *pictures*.

Importantly, on the refinement of step IV I proposed, the first delusional thought is neither 'This is an imposter/stranger' nor 'This perceived person is not how I remember X,' as proposed by some other two-factor theories (e.g., Young [2008]; Coltheart, Menzies, and Sutton [2009]; Davies and Egan [2013]; Parrott [2016]). Rather, the first delusional thought is 'This is *not* X.' This refinement has many advantages. For example, it is consistent with how we understand the closely related delusion of mirrored-self misidentification: there, the initial delusional thought is 'The person I am looking at in the mirror *isn't me*' (McKay 2012: 333) and not 'The person I am looking at in the mirror is an imposter/stranger.' In addition, this refinement captures Peter's case nicely, it coheres with prominent theories of perception, and it explains similarities between delusions and illusions (see Hohwy and Rajan [2012]; Hohwy [2013: 61–4]): in the Müller-Lyer illusion, for instance, I see lines of different length (Gregory 1997: 1125) while Peter sees the *picture*. Finally, it is consistent with the main idea of this section.

This main idea is that Peter's delusional belief 'This is not Mary' was not generated by a bad inference based on sensory information and that, therefore, the *U*CS need not be impaired. Our sub-personal cognitive abilities are based on the work of domain-specific peripheral cognitive modules (Evans 2003: 454). The perceptive cognitive module mainly operates with quite limited information, while the *U*CS checks whether a specific module has performed its function successfully (Turner and Coltheart 2010: 357–362). *U*CS has access to some higher-order analytic processes that are built into cognitive top-down predictions (schemes) but it does not have access to all of them, and we should assume that *U*CS does not have access to those processes that assess epistemic possibilities – as these require access to working memory and background knowledge (Evans [2003]; Parrott [2016]). In fact, taking it that *U*CS should, on the count of low probability, tag the perception of the *picture* as suspicious commits us to taking it that it will also tag the perception of an old friend as

suspicions, which does not seem reasonable. It follows that, because Capgras patients do not perceive their loved ones but rather their *pictures*, the *U*CS should not find anything surprising in the 'This is not Mary' thought. Therefore, the thought is rational considering the perceptual input, the normative standard of belief-formation was correctly applied by the *U*CS, and the second factor in Capgras delusion may consist solely of an impaired *C*CS.

Importantly, the hypothesis that the *U*CS passes the thought that the lady is not Mary because the thought is perfectly rational considering the perceptive input is compatible with both the 'explanation' and 'endorsement' models of the first; hence, it can explain all kinds of delusion (I adopted the endorsement model to explain Peter's case). Additionally, it explains why deluded patients do not develop a wide range of odd beliefs – i.e., why they are not taken in by every visual illusion – but rather hold only a certain type of beliefs, for instance, those concerning familiar people. The explanation is that the unimpaired *U*CS passes only one type of false thoughts because those thoughts are based only on one kind of (abnormal) data caused by a specific neurological impairment. The fact that, if it persists, the delusion may extend to other people who are close to the patient so that a patient holds more than one delusion (Stone and Young 1997: 329) is consistent with this answer: as the condition worsens, the neurological impairment will cause a lack of affective reactions to other familiar people (or objects) and, in turn, a greater variety of incorrect thoughts will go undetected and passed by the *U*CS.

Still, accepting this argument may not be sufficient for establishing that Peter really believes that the lady is not Mary at the time of assertion. It may be that, since Peter has – due to the impaired *C*CS – disregarded consciously accessible information relevant to whether the lady is Mary, this thought is not sufficiently truth-directed and, thus, it is not a belief – the second concern presented at the end of the previous section.[62] Plausibly, Peter has an impression that he sees Mary's picture and incorrectly takes this to be a belief. This objection assumes that some functions of *C*CS, while not constitutive of perception, are nonetheless constitutive of the processes whereby we form perceptive beliefs (Campbell 2001) and that therefore they must be integral to how beliefs are truth-regulated.

However, this account, which requires conscious effort for a perception to count as a belief, sits uneasily with some important studies that indicate that perceptual beliefs are taken on board right away and then quickly 'unbelieved' if found as being at odds with established

[62] I am grateful to Neil Levy for raising this important concern.

facts (Gilbert, Malone, Krull [1990]; Gilbert, Tafadori, Malone [1993]; Egan [2008: 55–58]). This model of belief-formation, initially proposed by Spinoza and then used in their theory of delusion by Aimola-Davies and Davies [2009] and Davies and Egan [2013], by definition, excludes assessment or evaluation of hypotheses before they are adopted as beliefs. On the Spinozian model, both *U*CS and *C*CS initially pass all perceptual thoughts as beliefs only to have them subsequently re-evaluated. Accordingly, Peter's failure would be in his *C*CS inability to 'unbelieve' that the lady is not Mary.

This theory of belief-formation can be resisted (see Sperber et al. 2010) but I need not endorse it in order to make it useful for my argument. Because I have already demonstrated that this thought goes through unimpaired *U*CS monitoring, I need not hold that *U*CS does not monitor this process; all I need is to show that the functions of *C*CS are not integral to the formation of perceptual beliefs. I use the rationale behind the Spinozian model to do this: Thought-assessment is demanding of cognitive resources; hence, consciously monitoring perceptive information would easily lead to resource depletion (Gilbert et al. 1990: 610). And, because having *C*CS participating in the formation of perceptual beliefs would be evolutionarily disadvantageous, the work of *C*CS is not integral to the formation of perceptual beliefs.

This concludes the second stage of my argument in favour of the proposal that Peter's thought that the lady is not Mary is a fully-fledged belief. The main argument was that the mental state is truth-directed and truth-regulated (it was formed correctly), that it is revisable, and that Peter possesses enough of the relevant dispositions to count as believing that the lady is not Mary. His lacking some other dispositions can be excused because he judges that she is Mary. I defended this view in three steps. In the first step, I argued that Peter's mental state was truth-directed and truth-regulated sub-personally. The second step involved refining the claim that Peter's mental state is truth-regulated and I did this by arguing that *U*CS was not impaired (the first concern resolved) and that *C*CS need not participate in the formation of perceptual beliefs (the second concern resolved). Specifically, the *U*CS passed the thought that the lady is not Mary because it was consistent with the perceptive input, while the work of *C*CS is not required for applying the standard of truth-regulation to beliefs based on perception.

In the third and the most important step, I address the third significant concern, according to which the belief was quickly or even immediately abandoned, 'unbelieved,' as a result of

Peter 'working it out' that the lady is Mary – this significant concern corresponds to Coltheart's [2007] explanation of the parallel case of G.R. In reply, I will not deny that Peter will eventually abandon his transparent delusion, nor that he may occasionally fluctuate between believing that *p* and that ~*p*; however, I will maintain that his making the judgement that *p* cannot result in his belief that ~*p* being abandoned instantaneously. Then, I will argue that this interpretation also applies to the case of G.R.

My reply, in short, is this: True, some of Peter's *C*CS abilities are recovering but the most important *C*CS ability is the one that enables other *C*CS abilities – and 'working it out' is just one of many *C*CS abilities – to be applied to a particular thought. Because this ability is still impaired, Peter's judgement is ineffective when it comes to resolving the delusion and it will continue until this ability recovers – which may initially happen only intermittently. I will defend this view in three stages: first, I will offer a plausible reconstruction of the relevant failure of the *C*CS (3.4), then I will present some cases that are best explained on that hypothesis (3.4.1–3.4.2), and finally, I will argue that the judgement 'it's gotta be Mary' corresponds to a mental state that is not a belief (3.5).

3.4. Step V: Misidentification as the Second Factor

Thus far I have argued that Peter's delusional belief that the lady is not Mary lacks overall consistency – it is inconsistent with some of his other beliefs and knowledge of the world – but that it does not lack the relevant domain-specific consistency – Peter really does not perceive his wife – and that therefore only a failure of *C*CS is necessary for the second factor (the impaired thought-evaluation) in the aetiology of the delusion. In this section, I resolve the third concern (the belief should be revised instantaneously) by offering an account of this *C*CS failure, step V, in which Peter's *C*CS reluctantly leaves the belief unrevised.

Using the theory of hemispheric specialisation, Ramachandran proposed a convincing neurological explanation of the relevant thought-evaluation failure. On this view, the left hemisphere is primarily responsible for imposing consistency on the model of reality constructed from the variety of sensory inputs we are regularly exposed to; it confabulates and rationalises atypical or scarce information. When anomalous information reaches a certain threshold, the right hemisphere, which monitors abnormalities, forces the left to revise

the entire model and start from scratch, but this does not happen in delusions.[63] Therefore, it is the right hemisphere's fault, concludes Ramachandran. The relevant impairment is, probably, damage in the frontal lobe of the right hemisphere – plausibly, the right dorsolateral prefrontal cortex (Coltheart [2007]; Coltheart, Langdon, and McKay [2011]). I will slightly modify Ramachandran's proposal and say that the right hemisphere actually initiates the revision but is unable to conduct it in full.

My proposal, in short, is this. The final test of a thought or a belief is when the outputs of relevant heuristic and analytic reasoning processes are put together and evaluated in the light of one's available background knowledge. This process is typically called decontextualized processing; i.e., analysing the thought through different domains of reality. However, if the right hemisphere is unable to conduct the decontextualized revision thoroughly, it will fail to integrate into decontextualisation outputs of some vital (typically, analytic) reasoning processes. And, if this happens, the perceptual module's domain-specific explanation or the endorsement of the input will be misidentified as the result of decontextualization, namely, as the overall and final consistency test. I dub this second-factor failure the "Misidentification step," in step V.

One implication of the Misidentification view is that, if the right hemisphere cannot correctly integrate all relevant outputs when decontextualizing a particular belief, then having some of these processes working reliably will not result in the revision of the delusions. Those outputs that signal that the belief is false will not be included in the evaluative process and, as a result, the output of the domain-specific heuristic responsible for the content of the delusion – i.e., the cognitive module's explanation/endorsement of the abnormal input – will be misinterpreted as the result of the overall consistency test – step V. Consequently, Peter's conscious judgement, as an output that was not used in decontextualization, will not be able to override the belief, which is why Peter believed that ~*p* at the time of asserting that *p*.

The Misidentification hypothesis explains not only how and why a bizarre delusional belief is successfully incorporated into one's pre-existing system of beliefs or why it persists despite obviously conflicting with other beliefs, which is the most significant problem for the doxastic conception of delusions and my interpretation of the Peter case, but also why

[63] Ramachandran [1995; 1996]; Hirstein and Ramachandran [1997]; see Levy [2009]; McKay and Cipolotti [2007]; also Kinsbourne [1989: 251]. On the thesis of hemispheric specialisation, see, e.g., Gazzaniga [1967, 1970, 1998]; LeDoux, Wilson, and Gazzaniga [1977]; Sperry [1961, 1986].

patients hold these thoughts with such vigour, even while recognizing their bizarre nature (see 3.3.1) – because decontextualisation is faulty, the thought is passed as correct. It also explains why patients fluctuate between beliefs: the decontextualisation is intermittently complete – the ability to integrate is only weakened, not absent. Polythematic delusions and pathological disorders, such as schizophrenia, I maintain, involve a (more) general impairment of the ability to integrate and even to initiate certain reasoning processes in evaluating beliefs; thus, the proposed solution can explain both monothematic and polythematic delusions equally well.

Notice, the claim is not that some relevant questions are not being raised or that the contradictory data is not being considered or given sufficient weight (*pace* Fletcher and Frith [2009: 55]; Coltheart, Menzies, and Sutton [2010: 280]; Parrott and Koralus [2015]): some delusional patients, Mr F. for example, are not indifferent to dissonance in their beliefs; they just have no means of resolving it. Rather, the claim is that the relevant failure is such that patients may realise that some information is inconsistent with their delusion without being able to use that to re-evaluate the delusion: impaired integration makes relevant *C*CS abilities ineffective with respect to the delusion.

On the Misidentification view, the ability to integrate relevant reasoning processes, i.e., the ability to decontextualize thoughts thoroughly, is the most important of all of our abilities relevant for thought-evaluation. Accordingly, the only impairment necessary and sufficient for some, if not most, delusions is the one that causes incomplete decontextualized processing (some relevant reasoning processes were not integrated into decontextualisation). It follows that, at least in some of these patients, other *C*CS abilities, such as the ability to make plausibility judgements (assessing whether something is epistemically possible), need not be impaired in order for these abilities to be ineffective with respect to this specific delusion. This prediction is met in Peter's behaviour: he judges that *p* while believing that ~*p*. Another prediction is that some delusional patients – since their thought has been tagged as correct – should be able to recognize the same delusional symptoms in others as delusions while failing to recognise them in themselves. This prediction is met in certain schizophrenic patients who had little difficulty in recognizing psychotic symptoms as long as those were the symptoms ascribed to other people (Startup [1997]; see Young [1998: 37]).

In fact, the delusional patients' behaviour is particularly indicative of this interesting failure of integration. For instance, they seem to restrict the information used in reaching a

conclusion and have a seeming disregard for truth. A common explanation is that, due to some bias (McKay 2012), they deny or ignore data from other experiences that contradict the delusional belief even if they are well-established facts (Garety 1991: 195). However, the formal reasoning (applying rules of logic) of deluded subjects is not much worse than that of normal subjects. Actually, they may even perform better than non-delusional subjects (Gerrans [2002: 48]; Parrott and Koralus [2015: 405–7]). They are typically irrational solely with respect to their delusion, and sometimes they do not even deny the contradictory information or even admit that this information makes their position ludicrous, yet without abandoning it (see 3.4.2). Acknowledging that their position is ludicrous indicates that these people are unable to apply formal reasoning only to a specific delusional belief – namely, that some of their *C*CS abilities are unimpaired yet ineffective exclusively with respect to the delusion. This fits nicely with the proposed Misidentification theory: although they recognize the data that contradicts their belief (their specific reasoning processes are intact), they cannot, even when they would like to (*pace* Bayne and Pacherie 2004: 6), use this data when evaluating their belief.

This is not to say that impairments of other *C*CS abilities are not important for the overall delusional experience. Impaired abilities to conduct reality checks, to make probability judgements, and so on seem to be behind secondary rationalisations and delusion-related confabulations. Because the patients cannot revise their delusions (misidentification) and because they cannot properly evaluate reality (some other *C*CS inability is impaired), they modify their perception of reality to make it consistent with their delusion. In conclusion, the faulty integration makes the delusion both possible (the second factor) and pathological, while failures of other *C*CS abilities typically contribute to the overall delusional experience. Patients who do not rationalise or confabulate most likely do not have other *C*CS abilities impaired, while those who do seem to have other *C*CS abilities also impaired.

The failure of integration is not such that it follows that delusions are not beliefs. Doxasticists about delusions think that cases such as superstition – in which people consciously refuse to integrate their superstitious beliefs with their other beliefs – suggest that good integration with the person's overall network of beliefs is not a feature essentially constitutive of beliefs and that therefore this is not a reason to think delusions are not beliefs (Bortolotti [2010: 85–8, 93]; Radden [2014: 45]). This view can be resisted (Van Leeuwen 2014) but notice that delusions, on the view I propose, are not analogical to superstitions.

Unlike superstitions, delusions *are* integrated, just *not fully* integrated. In contrast to superstitious people, some delusional patients, such as Mr F., desperately want to abandon their delusions. Therefore, even those who believe that religious and superstitious credences are not beliefs may have good reasons to take some transparent delusions as beliefs.

Finally, the third concern – to the effect that Peter's delusional belief would surely be revised as soon as he judges that the lady is Mary – is resolved by positing that the misidentification step may be the full description of the second-factor failure. Peter's *C*CS is recovering and he is now capable of conducting relevant analytic processes, such as judging based on available evidence. Nevertheless, he is still unable to apply the result of his 'working it out' to his belief by initiating a thorough decontextualized revision: the judgement is not being used in this process. Thus, the delusional belief that the lady is not Mary did not get immediately revised or rejected. Rather, it will continue to be operational despite his bewilderment until his ability to incorporate newly gained insights into the belief's decontextualized processing sufficiently recovers. This gives him plenty of time in which he may correctly assert 'It's gotta be Mary but I believe it's not Mary.' I will now apply the proposed theory to the case of G.R.

3.4.1. Peter and G.R.

I am not saying that Peter may not, while recovering, occasionally fluctuate between believing and not believing that *p*. The claim is that Peter's correct judgement will not immediately suspend the belief that ~*p* and, since he asserts that *p* immediately after the judgement, he sincerely asserts what he believes to be false thus showing that the Belief Account of sincerity is incorrect: one can sincerely assert what one believes to be false. That being said, I will now briefly discuss the aforementioned G.R. case (Coltheart 2007: 1053–4) that closely resembles the Peter case, except that this patient was, according to Coltheart, fluctuating between believing *p* and ~*p*. I present only the relevant section of the case (italics and comments added):

> **G.R.**: 'Yes, I figured that out, professor, *it couldn't be anybody else*.' [behaviour that supposedly indicates that he believes that *p*.] But having expressed a rejection of the impostor belief, G.R. continued *immediately* with … [non-verbal behaviour that seems to indicate that he believes that ~*p*].

Abandoning a delusional belief only to have it immediately return does not strike me as the correct description of the G.R. case, or any case actually. Firstly, believing is not an

on/off matter: we believe propositions with different levels of confidence; we do not just turn them on or off (Eriksson and Hájek [2007]; Lynch [2012: 448, n. 2]; Davies and Egan [2013: 705], Pettigrew [2013]; Subchapter V-1). Therefore, one cannot just abandon one's delusion and revert to it in a split-second. Furthermore, judging that *p* does not entail believing that *p* (Scanlon 1998: 25, 35). Also, acting *as if* a proposition is true, i.e., asserting what one judges, need not entail that one believes that proposition (Tumulty [2012]; Gerrans [2014: 137–143]) and G.R. asserts that he figured it out that this is his wife but then he proceeds to act as if he immediately forgot what he said a moment ago. This behaviour suggests that his *C*CS is not applying his judgment to the belief, misidentifying as overall correct the belief that continues to be 'passed' by the *U*CS – as predicted by the misidentification step.

What I argue is that the same cognitive condition reflects differently on G.R. when compared to Peter. Because he is not a brilliant neuroscientist, G.R. does not understand what is happening to him and, unable to resolve the problem (for the time being), his mind accepts *both* his belief and his judgment as correct, which is why his behaviour is misleading. Fineberg and Corlett [2016] argue that delusions act as doxastic shear pins, namely, safety mechanisms that allow continued functioning of the mechanism, although at a less reliable level. They are only partially correct: delusions do not have this feature but some mechanisms involved in delusions have it. In particular, without accepting both his judgment and his belief, G.R.'s mind would be paralyzed and this is when the shear pin breaks – in the moment in which the patients is starting to recover from his delusion.

Another possibility is that Peter and G.R. are 'in-between' believing that *p* at the time of asserting, so that the question of what they believe has no determinate answer (Schwitzgebel [2002, 2012]; Tumulty [2012]). Schwitzgebel [2002: 261] considered cases in which agents, in a certain kind of mood, sincerely assert that *p*, but in another genuinely recant such a confession and, since there is no simple answer to the question of what, 'underneath it all,' they really believe, he argues that these are cases of in-between beliefs. While I generally agree with this argument, it does not explain Peter's or G.R's behaviour: it takes time for a person's kind of mood or her context to change, i.e., a person *becomes* paranoid or anxious (see Gerrans 2014: 124–128), and this vital condition is not satisfied in these cases.

In conclusion, the description of G.R.'s case does not justify taking it that he actually abandoned his delusional belief when he said 'Yes, I figured that out, professor, it couldn't be anybody else' and the misidentification step predicts that the impaired ability of integrating

relevant reasoning processes will not allow the belief to be abandoned instantaneously. What happened instead is that G.R. realised that the 'apparition' is his wife without being able to break out from believing that she is not and while not realising what exactly is happening; the shear pin broke and both the judgement and the belief were accepted as valid. The only real difference between Peter and G.R. on the other is that the former knows what is going on and this is because Peter has the required knowledge of delusions.

This version of the second-factor failure vindicates the proposed interpretation of the case. The final screw in the theoretical device able to explain Peter's behaviour is pinpointing the exact mental state prompted by the relevant judgement. I come back to this in section 3.5, where I finally close off the Peter case, but, before that, I want to offer some other cases that support the theory I have proposed and highlight some of its advantages.

3.4.2 Other Interesting Cases

The Misidentification view is not designed specifically to explain the Peter case; it successfully explains other interesting cases as well. One patient, Mr S., admitted that it is not merely unusual, but unbelievable and virtually impossible that he has two families – in fact, he could not believe that he believes that – and yet he could not utilise this awareness to revise this delusion (Alexander, Stuss, and Benson 1979: 335–6). Likewise, another patient admitted that it was 'logically impossible' that he has multiple heads and bodies but that 'the feeling was too real to have been a dream' (Weinstein, Kahn, Malitz, and Rozanski 1954). Then there is LU, a 24-year-old woman with Cotard delusion (McKay and Cipolotti 2007: 353).[64] To the question how she knew whether someone was dead, she answered that her dead grandma was motionless and did not speak. Then the doctors asked her to explain how she could be dead, given that she can move and speak. Although LU acknowledged the tension between her concept of being dead and her delusion, she resolved the tension by abandoning the delusion only after a whole week had elapsed. These people's behaviour makes perfect sense on the view that they were, at least for a while, unable to decontextualize their delusions thoroughly while nevertheless being generally able to understand reality.

[64] Cotard delusion is caused by perceptual anomalies similar to ones causing Capgras delusion. Whilst the first factor in Capgras delusion involves a disruption to pathways underpinning the emotional component of face recognition, Cotard patients have a more global disconnection of all sensory areas from the limbic system leading to a complete lack of emotional contact with the world. They believe that they are dead possibly, the first factor, because they cannot experience anything; they 'feel nothing inside' (Young and Leafhead [1996: 148, 160]; Ramachandran and Blakeslee [1998: 167]; Gerrans [2002: 50–51]; McKay and Cipolotti [2007]; Coltheart, Menzies, and Sutton [2010: 265]).

In contrast, Young and Leafhead [1996: 156–160] describe the case of JK, also suffering from Cotard delusion. She admitted that she could feel her heartbeat and when her bladder was full and that she could sense hot and cold. However, when the doctors pointed out to her that these sensations surely provided evidence that she was alive, she responded that they clearly did not, otherwise, a dead person, such as herself, would not have them. She also believed that her doctors would not feel their heartbeats were they dead but that she can. JK's behaviour can nicely be explained on the thesis that both the impairments of integrating reasoning processes and of making plausibility judgements completely isolated her delusion from reality. The difference between Mr F., Mr S., the man with multiple heads, and LU on the one side, and JK on the other, where the former group had the correct appraisal of reality whereas JK did not, seems to be determined by whether, in addition to the impaired ability of decontextualizing thoughts, some other *C*CS ability is impaired.

This concludes my discussion on the nature of Peter's delusional thought. There is one final question concerning the effect of Peter's conscious judgement. I address it below.

3.5. 'The Lady is Mary'

In sincerely asserting that the lady is Mary, Peter acted on his judgement. Now, seeing that he already believes that the lady is *not* Mary and that he realises that his belief is false, and since he cannot simultaneously and consciously believe both that she is not Mary, that his belief that she is not Mary is false, and that she *is* Mary, which mental state was expressed by his judgment? I seek to answer that question in this section. This answer is very important, I note; it makes many (supposed) cases of self-deception, such as *Nicole* and *Androvna*, intelligible.

People guide their actions not only in the light of what they believe to be true but also in the light of what they *accept* as true; in fact, accepting rather than believing truth-claims is often involved in developing scientific theories. It might be suggested, then, that Peter accepts that *p* and believes that ~*p*. This suggestion seems to misdescribe the case, however. Acceptance, as commonly defined, is a voluntary, context-relative assent to a proposition; it is 'taking it for granted,' 'going along with it,' 'treating it as true for some reason,' or similar (Van Fraassen [1980: 12]; Cohen [1989: 384]; Stalnaker [2002: 716]). Accordingly, practical pressures or prudential reasons can make it reasonable for you to accept that *p* in a certain context even if you do not find it reasonable to assign *p* a high probability. For example, in planning your day-trip, you might take it for granted that it will not rain even though you are

not certain about this (Bratman 1999: 22). However, it is not that Peter believes that the lady is not Mary and acts as if he believes that she is. Rather, he is certain that she is Mary; '*It's nobody else*,' he said (Turner and Coltheart 2010: 371).

One plausible way of interpreting my case is as suggesting that there is knowledge without belief. And indeed, I will argue that Peter *knows* that the lady is Mary despite his belief that she is not Mary. I argue this because the mental state associated with his judgement meets several commonly recognised marks of knowledge. Peter inferred the proposition 'The lady is Mary' from what he knows about Mary and what he knows about the world; therefore, his inference makes competent use of reasons (Truncellito 2007) and the basis for his inference is safe – it is constituted by what Peter knows about Mary and about the world.[65] Therefore, apart from being true, the conclusion of the judgement, 'The lady is Mary,' is (1) justified, (2) based on the success of Peter's cognitive abilities, and it is (3) not a product of luck (see, Pritchard 2012). Furthermore, seeing that (4) Peter recognized the conclusion as true – in fact, he knows that he knows that *p* – the mental state that results from this judgement meets commonly recognised marks of knowledge both on internalist and externalist theories of knowledge (see Ichikawa and Matthias 2012). Therefore, it must be that Peter *knows* that the lady is Mary – in particular, he knows that he knows this (he has reflective knowledge) – and that he asserts that the lady is Mary because he reports on his knowledge.

Because he consciously believes that ~*p* and thus does not believe that *p*, Peter's knowledge that *p* does not have the JTB structure. Hence, unlike Gettier cases – which can establish only that JTB is not sufficient for knowledge – Peter's case shows that believing that *p* is not necessary for knowing that *p*. In addition, my case is not susceptible to criticism applicable to other counterexamples that aim at demonstrating that believing is not necessary for knowledge. For example, in Radford's [1970: 171–85] famous cases, people supposedly know that *p* – e.g., the agent in question answered the question correctly – but they do not believe that they know that *p* until someone confirms to them that *p* is true – she believed that she does not know the answer. A common objection to Radford's argument, which limits the persuasiveness of the case only to philosophers who are externalists about knowledge, is that these agents did not really know what they asserted *before* it was confirmed to them that their assertions were true (Lehrer 1983: 173, 182). This concern, however, does not arise in my

[65] Both Sosa-Safety [2002: 264–86], 'If S knows that *p* on basis B, then S could not have easily formed the false belief that *p* on basis B,' and Pritchard-Safety [2005: 163; 2012: 250, n.3], 'If S knows that *p* on basis B, then S could not have easily formed a false belief on basis B,' are satisfied.

case: Peter needs no external confirmation in order to know that the lady is Mary: 'What has made you realise … People telling you that? No *it's just myself* that's done it ... *I* worked it out' (Turner and Coltheart 2010: 371).

Many prominent philosophers (e.g., Radford [1970]; Lehrer [1989: 136]; Williamson [2000: 46–47]; Myers-Schulz and Schwitzgebel [2013]) have already argued that we may know something is the case before we believe it. The Peter case, thus, is just further evidence for this view. In fact, the case shows that an even more drastic situation is possible: Peter knows that *p* while actively believing that ~*p*, thus showing that it is not just that knowledge is conceptually prior to belief; rather, it can conflict with firmly held conscious belief. Moreover, the fact that Peter consciously believes the opposite of what he knows substantially differentiates this case from recent examples given by Myers-Schulz and Schwitzgebel [2013] (who offer a Nicole-type case of self-deception) or Lehrer [1983: 173–174] that are aimed at showing that belief is not necessary for knowledge. Therefore, insofar as there is a good case for taking it that Peter knows that the lady is Mary while believing that she is not, we have a particularly clear counterexample to the JTB account of what it is for a person to know that *p*.

Interpreting Peter's case in this way (he knows that *p*, but believes that ~*p*) fits nicely with the so-called *capacity-tendency account*, based on Ryle's [1943: 133–134] view that 'knowledge' is a capacity verb while 'belief' is a tendency verb. Philosophers who agree that knowledge does not entail belief have used this distinction to propose that one can have a capacity without a tendency, namely, know without believing. Margolis [1973: 78], for instance, proposed that knowledge involves the capacity to provide the right information in the right way while belief involves the disposition to act appropriately. A person who knows that *p* but does not believe it has the *capacity* to act on well-grounded information that *p* but lacks the *tendency* to do so (Myers-Schulz and Schwitzgebel 2013: 381). Peter has the capacity to act as if the lady is Mary and the tendency to act as if she is not, which is quite consistent with my analysis in Section 3.2.

Of course, much more needs to be said about what the right account of knowledge is but this thesis is not a place for such a discussion. For the purposes of this argument, I think, it is sufficient to acknowledge that the JTB account is problematic, not just because it entails that Peter cannot recognize that the lady is Mary without actually coming to believe this, but also

because there are other counterexamples (e.g., Radford [1970]; Lehrer [1983, 1989]; Myers-Schulz and Schwitzgebel [2013]) and counterarguments (e.g., Williamson 2000) to this view.

I now turn to arguing that a variation in which Peter asserts 'This lady is *not* Mary' invalidates the Belief Condition on lying, showing thereby that all existing theories of lying are too restrictive. The theory of lying capable of explaining Peter's behaviour will not generate paradoxes of self-deception – this theory does not require of liars that they assert what they believe to be false.[66]

4. Lying

The Peter case shows that belief is not the norm of sincerity and that the Belief Account thus should be abandoned. More interestingly, it also shows that not all sincere Moorean assertions are, *qua* assertions, absurd (*pace* Moore 1993: 207–11) or misleading (*pace* Lackey 2007: 615–616). In fact, it is the opposite: by saying 'It's gotta be Mary' without a further qualification, Peter would be implicating that he believes what he asserts, he would be misleading, and this is what Peter aims to avoid. Moorean assertions seem to be appropriate, and even required, responses to some unusual situations in which the topic requires full disclosure – and what Peter believes is quite relevant in this context. Finally and most importantly, Peter can lie by asserting that the lady is not Mary, which is what he consciously believes, in any context and to any addressee, including himself – a point vital for understanding self-deception as 'lying to myself.' Consider the following variation.

> **Peter Breaking Bad**: Suppose that Peter hates Mary whom he had married only for her money and that he realises that the lady is Mary one night while they were alone in the house watching *Breaking Bad* (Capgras patients often live with the 'imposters'). Knowing that no one knows that he is recovering, Peter quickly decides to use his fortuitous condition to kill Mary and get away with it; he knows that Capgras patients can be extremely violent towards 'imposters.'[67] Upon their arrival, Peter tells the police that he has killed a burglar who looks just like his wife. When they tell him that the lady he had shot is actually Mary, he replies 'No, that lady definitely is *not* Mary.' Obviously, Peter lies (even though he asserts what he believes to be true) because he wants to get away with the murder of his wife, Mary.

[66] I am tremendously grateful to Neil Levy for his help with this subchapter.

[67] On some of these violent cases, see De Pauw and Szulecka [1988]; Silva et al. [1989: 7–9]; Förstl et al. [1991]; Bourget and Whitehurst [2004: 721–722]; also Howard, Hepburn, and Khalifa [2015].

Peter Breaking the IRS: We may imagine a less violent example in which Peter everything is the same as in *Peter Breaking Bad* except that Peter lies in order to protect his wife from the IRS agent knocking on their door.[68]

No theory of lying can correctly explain this case but the intuition is clear: Peter lies by asserting what he knows to be false, i.e., that ~*p*, he knows that he is lying, and he even intended to deceive the police – this was all part of his plan. Indeed, by asserting that ~*p*, which is what he believes to be true, he can lie to the police, his friends, even to himself – should he suddenly be struck by guilt. Interestingly, Peter is lying not just despite the fact that the Belief Condition on lying is not satisfied, he does not assert what he believes to be false, but rather *because* it is not satisfied; namely, by asserting what he believes to be false, Peter would assert what he knows to be true and you cannot lie by asserting what you know to be true.

My *Peter* cases are fictional and contentious but they are not the only kind of cases that could be used against the *Belief Condition* on lying. I suspect that there is abundance of cases in which liars assert what they believe is true. We just need to notice them. Consider the following case I modify from Lackey [2007: 598; 2008:110].

Martin: a racist, Martin, was called to serve on the jury of a case involving an African American on trial for raping a white woman. Even though Martin accepts that the evidence that the defendant is innocent is compelling (and on the capacity-tendency account, we may even say that he knows the guy did not do it), he cannot help believing that the defendant is guilty. However, he suspects that it is his racism that leads him to this belief, which he nevertheless still holds. Suppose now that his friend asks him 'Did the guy do it?' and that Martin replies 'Yes, he did it' intending to make his friend believe that the guy did it. Even if the capacity-tendency account is incorrect and Martin does not know that the guy did not do it, there is a clear sense in which Martin is lying to his friend, and even trying to make him believe his lie, even though he asserts what he believes is true.

If the above analysis is correct, our theories of lying will have to be seriously rethought; a simple modification will not do the trick. However, this analysis of lying is already too long for me to engage in a serious rethinking of lying. Therefore, I would just briefly sketch some theses that seem to follow from it. I suggest that liars intentionally make illegitimate assertions in contexts in which they expect them to be taken as legitimate. In this way, they violate the norm of assertion and other relevant norms of communication. Thus, I dub this account the *Violation Account of Lying* (VAL).

[68] I thank John Bishop for this non-violent variation of the case.

VAL: I lie to you by asserting that *p* in context C if and only if

1. I say that *p* to you in C,
2. I propose that what I say be taken as a legitimate assertion in C, and
3. I judge that my asserting that *p* in context C is not legitimate.

According to VAL, Peter may easily lie to you, the police, or even myself by asserting that *p* even if he actually believes that *p*. He lies because he is intentionally making an illegitimate assertion, 'This is not Mary,' in a context in which he expects it to be taken as legitimate. The same analysis applies to Martin and all other liars. Now, it would be instructive here to propose what makes an assertion legitimate but I cannot argue in favour of any specific norm of assertion. Instead, I will note that, even though people might disagree on what exactly this norm is, most of us know what can and what cannot be legitimately asserted. Below I just briefly sketch how VAL and the Peter cases fit with some already proposed norms of assertion.

The Peter cases are consistent with the knowledge norm of assertion. According to some versions of this norm, for instance, assertions are legitimate only if they are governed by knowledge (Williamson B. [2000]; Adler [2002: 275]; DeRose [2002]; Turri [2010b]), or if they express knowledge (Turri 2011), or if they transmit knowledge (Hinchman 2013).[69] However, there are considerations against accepting this norm (Douven [2006: 476–81]; Lackey [2008: 103–24; 2011: esp. 271–272]; McKinnon [2015: 53–57], and so forth); hence, one might reasonably reject the knowledge norm of assertion. According to the truth norm, only true assertions are legitimate (Weiner 2005), and the Peter cases fit this rule, but most considerations raised against the knowledge norm apply to this norm as well. Recently, Hawthorne, Rothschild, and Spectre [2016] argued that belief cannot be a norm of assertion and *Peter* cases are consistent with this position: it cannot be that Peter can legitimately assert that the lady is not Mary, whereby he would be reporting on his belief, while knowing that she is Mary.

Finally, Lackey's [2007, 2008] Reasonable to Believe Norm of Assertion or Douven's [2006, 2009] Rationally Credible Rule seem as good candidates for the norm of assertion. Nevertheless, while they do not suffer from objections raised against the knowledge or the truth norms and while they can explain the *Peter Braking Bad* case, they too have had some

[69] On the norm of assertion, see Pagin [2016]. For two prominent views on what we say by what we utter see, Wilson and Sperber [2002, 2012] on the (radical) contextualists side, and Schoubye and Stokke [2016] and Stokke [2016b], who follow Roberts [2012], on the compositional contextualists side.

important objections raised against them (Turri [2013]; possibly McKinnon [2015: 61]). All in all, VAL and the Peter cases are consistent with any norm of assertion except the belief norm, and this is very important because being fixed to only one norm would reduce the appeal of the case and of the suggested theory of lying, but they cannot help us decide which of them is better.

In the end, I briefly reflect on the Nicole and Androvna cases, which I explain in more detail in Subchapter V-1. VAL allows us to interpret their behaviour in the following way. Nicole believes that Tony is not having an affair with Rachel. But be that as it may, she recognizes the preponderance of evidence to the contrary, which makes her insecure and nervous, and it may lower her confidence in her belief that Tony is not having an affair with Rachel. In this situation, Nicole could keep saying to herself 'Tony is not having an affair with Rachel,' which is what she still believes to be true, while realising that she is effectively lying to herself. And Nicole would have to think that she is lying to herself if she judges that the amount of evidence to the contrary makes her assertion illegitimate despite her (firm) belief that it is true.

Finally, if Nicole is lying to herself in order to give credence to her belief that Tony is not having an affair with Rachel and thus prevent or postpone its revision, then she is trying to deceive herself by lying to herself. In short, I will defend the view that a person is deceiving herself by lying to herself if she realises that her belief that *p* that cannot legitimize asserting that *p* but nonetheless keeps telling herself that *p* in order to maintain her belief that *p*.

V. EXPLAINING THE CASES

1. Summary

The manipulativist conception of deception (*manipulation* henceforth) yields a non-paradoxical theory of self-deception. According to *manipulation*, self-deceivers need not intend to make themselves believe a falsehood; they may intend to deceive themselves into something they suspect is false or even into something they already believe as true. In addition, the Violation Account of Lying (*violation* henceforth) is an account of lying which, when applied to self-deception, i.e., as involving lying to oneself with the intention of deceiving oneself, also does not generate paradoxes. According to *violation*, a liar is a person who misrepresents her assertion as legitimate, not a person who asserts something she believes to be false. On *violation*, Nicole or Hillary are lying to themselves by asserting that ~*p*, e.g., 'Tony is not having an affair with Rachel,' when they assert that ~*p* while judging that ~*p* cannot be legitimately asserted, and they may actually believe their assertion.

Intending to deceive myself by lying, then, the so-called 'lying model' of self-deception, on the *violation–manipulation* combination is not paradoxical. With *violation* and *manipulation* in place, then, any scholar analysing self-deception should breathe a sigh of relief: this phenomenon is not conceptually problematic and it does not require a theory of the mind as compartmentalised. Therefore, I take it that the traditional account of self-deception has been fully vindicated and rehabilitated. Nonetheless, this is not to say that all that is to be said about self-deception has been said and, since my aim is not just to defend traditionalism but also to provide a clearer understanding of the phenomenon itself, the following chapter is relevant.

In this chapter, I will apply *violation–manipulation* combination, given that they do not depart from each other in crucial ways, to my main four cases and introduce additional theoretical refinements required for explaining their possible variations. Most importantly, in Subchapter V-4, I will propose a plausible candidate for the motive for self-deception. However, I will try to be maximally concise. I start with the Nicole-type cases.

V-1. The Nicole-Type Cases

1. Introduction

Davidson [1997: 215], among many others, cautioned us against taking the notion of lying to oneself too literally, since that would 'require that one perform an act with the intention both that that intention be recognized (by oneself) and not recognized (since to recognize it would defeat its purpose).' However, I first argued that cases such as Nicole and Androvna (N&A henceforth) are best explained by saying that these people are lying to themselves while not intending to deceive themselves; the non-traditional account of lying renders Davidson's worry unjustified. Here, I want to do two things. First, I want to give an account of why Nicole would lie to herself if not with the intention to deceive herself. Second, I want to push my argument further and propose that Nicole may even intend to deceive herself by lying to herself without having her purpose defeated.

1.1 Why Would You Lie to Yourself

It seems as if Nicole has nothing to gain by lying to herself if the lie is not meant to deceive her in some sense, but this is only in the appearance. Nicole lies in order to avoid legitimizing the unfavourable state of affairs. Her bald-faced lie to herself does not make her better off – namely, it does not reduce anxiety by lowering her confidence in the unfavourable proposition – but it does prevent ending up worse off. Openly acknowledging the truth to herself would require Nicole to react accordingly and do something she obviously does not want to do, e.g., to leave Tony or break down in tears, and this would just increase anxiety. This point becomes even more obvious in the high-stakes cases: Androvna 'refuses' to admit that she is in the terminal stage of cancer – since that would require of her to give up fighting. (The self-deceiver's refusing to legitimize *p* is not like refusing to accept Donald as your president; it is a weaker kind of refusal that involves a kind of inability to acknowledge the facticity of *p*.)

Being unable to legitimize the unfavourable state of affairs is a very common motive, Sartre [1978], Bach [1981, 2009] and Gendler [2007, 2010] seem to have it in mind, but it did not get the deserved attention in the literature. A heavy smoker may bald-faced lie to himself by repeating that he does not want to quit and that he enjoys smoking due to his fear of quitting; I may lie to myself lie to myself saying that I am not afraid of flying but that flying is extremely dangerous; due to the feeling of guilt, Gendler may lie to herself repeating that

she will finish her paper today; and so forth. I distinguish two kinds of reasons to lie to yourself.

1.1.1. Failing or Refusing to Admit the Truth

The most common reason to lie to yourself is failing or refusing to admit the truth. Explanations of the Androvna case typically pay insufficient attention to the fact that she exhibits 'unmistakable symptoms of the late stages of a currently incurable form of cancer' (Rorty 1988: 11). This piece of information is vital for the correct understanding of her behaviour. Openly admitting to herself that her condition is terminal would probably require of Androvna to lie down and helplessly await for death to come. Therefore, in high-stakes contexts, lying to oneself is unavoidable for those not ready to give and, while epistemically unjustified, it is biologically adaptive and functional. It acts as a way of blocking the distressing thought, or to use Bach's [1981] terminology, it is a method of jamming. Allow me to explain.

The Androvna-kind cases make best sense on the tendency-capacity account (knowledge is capacity to provide information and belief is tendency to feel and behave), which allows us to say that, since belief and knowledge are two separate mental states, some self-deceivers know the truth but, due to some emotional factors, cannot bring themselves to believe it (they acquire one mental state but not the other). Dr Androvna is a specialist in the diagnosis of cancer, an excellent reason to think that she knows the truth. The tendency-capacity account allows Androvna to know that *p* ('I am terminally ill') not just while not believing that *p* but also while consciously believing that ~*p*. She was simply unable to bring herself to come to believe what she learned; the stress, most likely, hindered the thorough decontextualisation of her belief.

The combination of *manipulation–violation* and the capacity-tendency account has many advantages. Principally, it can tell us why some N&A-type self-deceivers have their dispositions all over the place. For instance, we may say that some of Androvna's behaviour, such as writing the will, is connected to her knowing that she is terminally ill, while some other dispositions are connected to her inability to come to believe that something this horrible is really happening to her. Accordingly, some kinds of self-deception involve an important evolutionary adaption: the fact that Androvna knows the reality makes it possible for her to seek treatment and thus prolong her life or increase its quality, while her not being able to make herself believe the reality may serve the function of preserving her mental

harmony or even sanity. Her knowing that *p* and believing that ~*p* play the same evolutionary role.

Therefore, having an inability – in this case, the inability to thoroughly decontextualize your belief under a significant amount of stress – can sometimes actually turn out to be evolutionarily advantageous. Other have already made this point. Fineberg and Corlett [2016] argue that delusions act as doxastic shear pins, safety mechanisms that allow continued functioning of the mechanism, although at a less reliable level. They borrowed this analogy from McKay and Dennett [2009] who argue that misbeliefs protect their authors; they allow continued engagement with the world rather than paralysis. The main idea here is that some doxastic failures involve biological adaptations; they protect the individual. I propose that some failures to decontextualize beliefs in the light of newly acquired knowledge serve the same function. Those who, in turn, lie to themselves thereby by affirming the un-thoroughly decontextualized belief are the Androvna-type self-deceivers.

1.1.2. Intending to Deceive Yourself

The second reason to lie to yourself is to deceive yourself. My proposal appears similar to but is substantially different from the one made by Mijović-Prelec and Prelec [2010]. These authors distinguish three levels of belief – deep belief, stated belief, and experienced belief – and argue that, because discounting information affects positive and negative statements asymmetrically, your knowing the truth will not take away the motive to lie to yourself. Suppose that deep down I know that my thesis is unremarkable (deep belief) but that I desperately need it to be brilliant and that, on the surface, I believe that it is brilliant (experienced belief). Lying to myself is the most obvious strategy in this situation because, although asserting 'My thesis is brilliant' in the face of counterevidence may provide only a little or no reassurance that my thesis really is brilliant, asserting 'My thesis is unremarkable' (stated belief) would involve clear evidence that the thesis really is unremarkable and thus would support my unconscious deep belief.

In such situation, lying, while bringing almost no positive payoffs, prevents one ending up worse off: asserting 'My thesis is unremarkable' may help my unconscious belief to rise to the surface. And, since payoffs are vastly disproportionate, in a situation in which I know the truth but the truth hurts too much, the 'positive statement becomes mandatory not because it will be believed, but because of fear of the all-too-credible power of negative statement' (Mijović-Prelec and Prelec 2010: 231). That is to say, the lie is meant not to increase

confidence the person has in believing a falsehood but to preserve the uncertainty about the unconscious deep belief that my thesis is brilliant.

While all three of us see lying to ourselves as mandatory in some situations, the Mijović-Prelec and Prelec's proposal differs from mine in three crucial aspects. First, their kind of lying is necessarily deceptive – it prevents Nicole from becoming epistemically better off, i.e., less convinced in a falsehood – whereas I argue that these lies are typically not aimed at deceiving. According to my view, the purpose of self-addressed lies is not allowing things to go from bad to worse in a practical, biological sense: these lies play a biological function of preventing stress (anxiety, etc.) to reach levels that would bring only harm to the organism, they are dispositional shear pins, and they need not affect our beliefs to perform this purpose.

Second, the reason why Mijović-Prelec and Prelec [2010] describe these lies as deceptive is that they see the assertion that *p* as evidence that *p*, which is the view I disputed in IV-1. However, this is not to say that my position is that self-addressed lies cannot be deceptive, which brings me to the third point of departure. I am a proponent of *violation* and *manipulation* and the beauty of this pair is that it requires of liars only to assert what they judge cannot be legitimately asserted and of deceivers only to manipulate the target and their truth-evaluable mental states in a relevant way. This allows me to move away from the idea that I could lie to myself only if the lie is a bald-faced lie and come closer to the view that some self-addressed lies may indeed be deceptive without being forced to posit exotic mental states. By asserting what I believe to be true, perhaps, I may intend to affect my confidence in that proposition and thus effectively manipulate myself in the way that rightly counts as deception.[70]

The binary idea of belief (believing is a binary on-off matter) implies that any epistemic reason to doubt a belief *ipso facto* undermines the truth of that belief and that therefore the belief will require revision. Adler [2002: 249–277] convincingly argued against this assumption by appealing to the distinction between belief and confidence. This distinction introduces two important theses. First, a person can be entitled to a full belief without having unqualified confidence in that belief. And second, negative evidence is not evidence against the belief itself but rather against having a certain degree of confidence in it – you need

[70] Some scholars talk about degrees of conviction in a proposition or credence rather than beliefs or confidence by which we hold them. Locke [1690/2001: chapters 15, 19, esp. 16] calls confidence levels 'degrees of assent' and people also speak about 'degrees of belief' (see Eriksson and Hájek [2007]; Pettigrew [2013]). This debate is tangential to my argument and, for the sake of simplicity, I will use these terms interchangeably.

counterevidence to undermine a belief. Accordingly, and adding *violation-manipulation* to the picture, we may understand Nicole and Androvna in some of the following ways.

It could be that they both believe that *~p* (no affair, no cancer) but, due to unsettling negative evidence, the level of confidence they assign to their belief is insufficiently high to guide their behaviour in certain contexts. This explanation does not exclude the possibility that they know the truth; the tension could be between appropriating their degree of belief to what they know. If they are lying to themselves in order to boost their confidence, then their behaviour involves self-deception *simpliciter*; they are intentionally deceiving themselves on the 'lying model' of self-deception. Not realising that what they are doing amounts to self-deception does not entail that they are unintentionally misleading themselves – deceivers need not entertain the thought that they are actually engaged in deception.

An alternative plausible explanation would involve people who lied to themselves without the intention to deceive themselves but whose lies affected confidence level of their belief. We may say, for example, that Nicole has lied to herself simply because she did not want to admit the truth to herself – she was turning her head the other way – but that she has, nonetheless, unintentionally affected her confidence in *p*, which would qualify the action of lying to herself as self-deception *secundum quid*.

2. Concluding Remarks

All in all, refusing to openly legitimise the unfavourable state of affairs is a good reason to tell a bald-faced lie to yourself. However, you may also lie to yourself with the intention to increase, decrease, or prevent a (more drastic) change to, your confidence in a proposition (see Lynch 2012, 2016), all of which rightly counts as intending to deceive yourself. Any of these solutions, I believe, can serve as a plausible non-problematic explanation of the N&A-type cases. Furthermore, these proposals can be generalised to a much broader group of people who lie to themselves, they do not fall into any traps, they are consistent with our linguistic practices and our intuitions about the cases, they fit any theory of the mind, and they do not require self-deception to be reconceptualised. With these thoughts, I end my analysis of N&A cases.

V-2 The Hillary-Type Cases

1. Introduction

As I argued in II-2/3, the Hillary-kind of self-deception makes best sense on the interpretation that Hillary prevented herself from revising – and ultimately abandoning – her belief that Bill is *not* having an affair, which is what she believed before being confronted with particular negative evidence. In this Subchapter, I propose one way in which a Hillaryan self-deceiver could intentionally deceive herself *simpliciter*, a proposal which relies on the manipulativist conception of deception and the thought that I may intend to deceive you into believing a proposition I believe to be true. The relevant account of self-deception is, in short, this.

Hillary recognizes information consistent with *p* ('Bill is having an affair with Monica') but does not make the transition from judging 'this piece of information suggests that *p*' to judging 'it is *p*'. That is, she judges that 'the evidence suggests that *p*' but does not judge that '*p*,' which is how her belief that ~*p* is being maintained. This maintenance will involve self-deception *simpliciter* if Hillary, motivated to keep her belief that ~*p*, intentionally prevents this transition by way of manipulating her own normal way off appreciating the relevant information.

Hillary's self-deception is *simpliciter* because she recognizes her behaviour as deceiving herself concerning the meaning of the relevant piece of information but, and this is crucial for the success of the action, she does not see her action as deceiving herself as to whether *p*. Rather, she thinks that her behaviour is justified given that it is aimed at maintaining a true belief, i.e., that Bill is not having an affair (~*p*). Recall, a double or triple bluffer deceive their victim aiming to produce/maintain a true belief and thus, while they see their behaviour as deceiving, they do not see it as 'bad' deceiving. The same line of reasoning is responsible for the success of Hillary's self-deception and, I maintain, cases of this sort are fairly common.

2. Self-Deception *Simpliciter* in Retaining a Belief

My analysis of this case comes in four steps. I start with the plausible assumption that Hillary already believed that ~*p* ('no affair') and that her desire to continue believing it and

avoid believing that *p* motivated her action of deceiving herself (see Archer 2013: 278). In the second step, I propose the following three epistemological refinements.

First, negative evidence is not identical to counterevidence. Counterevidence undermines reasons for a belief whereas negative evidence undermines confidence with which I may hold a belief. Counterevidence to the view that Bill is the president of New Zealand is the fact that New Zealand does not have the institution of the presidency. Negative evidence to the view that Bill is *not* having an affair with Monica is a reliable testimony that they were seen together in an exclusive restaurant. The former makes it irrational for Hillary to think that Bill is the president of New Zealand, whereas the latter makes it sensible, but not necessarily epistemically justified, for her to lower her confidence in *~p*, and the next refinement explains why.

Second, the weight of evidence does not have a direction; that is, more negative evidence does not mean less confidence in a belief (Adler 2002). Whether there is one or five reliable testimonies that Bill and Monica were seen together in an exclusive restaurant makes little difference in terms of affecting your confidence in your belief that they are not having an affair. This refinement can be reformulated to say that the weight of evidence should be distinguished from the power of evidence. A greater weight of evidence does not entail a greater force of the evidence – and all this because new evidence could undermine the evidence we already have. For instance, one reliable testimony that Bill and Monica are working on an important business project undermines the power of five reliable testimonies that they were seen together in an exclusive restaurant, and this comes notwithstanding that the latter is 'weightier' (i.e., five testimonies outweigh one). Even though it weighs less, the last piece of information has more force.

According to the third refinement, a piece of data or information may fail to become evidence if it is undermined by the person's background beliefs. In particular, a piece of information that suggests that *p* ('affair') may fail to become negative evidence or counterevidence to the belief that *~p if* it is undermined by some evidence in favour of *~p* already possessed by Hillary – and maybe possessed exclusively by her.[71] It could be that Bill is exceptionally prudish or that he is a good Christian. He really is a great guy; he even runs a

[71] Van Inwagen [1998] argues that justification people sometimes find for their belief is 'incommunicable.' Therefore, possession of adequate evidence need not guarantee a formation of the relevant belief.

humanitarian organization. And this is where the trick lies: Pieces of information that suggest that *p* do not undermine pieces of information or beliefs that suggest that ~*p*. Specifically, a reliable testimony that Bill and Monica were seen together in an exclusive restaurant is not evidence against the view that Bill is prudish or a good Christian, or that he is a great humanitarian. However, the belief that Bill is prudish or that he is a good Christian is a good reason not to make the transition from 'This piece of information (i.e., given reliable testimony) suggests that *p*' to 'It seems to be that *p*.' Our pre-existing web of beliefs may make the new information less convincing or relevant, especially if you are motivated to keep your belief that is threatened by this new information – then, you may use a strategy to make the particular information even less convincing or relevant.

In the third step of my analysis, I apply this epistemology to Hillary's case. Let us say that P is any piece of information that suggests that *p* ('Bill is having an affair with Monica'). On the refined view introduced above, P may (1) fail to become evidence that *p* or (2) fail to become evidence powerful enough to justify believing that *p* or revising the belief that ~*p*. Let us now analyse the former scenario, which I take to be more common.

The transition from 'P is a piece of information consistent with *p*' to 'P is evidence that suggests that *p*' also requires approval based on background knowledge. For a person sufficiently convinced that Bill is a good Christian and a devoted father, the fact that 'this piece of information P is consistent with *p*' does not imply that *p* is true. Consistency is so weak that on its own 'P is true and P is consistent with *p*' is never a reason to think *p* is true, not even a very slight reason. That is different with 'P is true and P suggests that *p*.' Note that 'P is true and P is inconsistent with ~*p*' does entail that *p* is true. Also, 'P is consistent with *p*' should not be conflated with 'P is inconsistent with ~*p*'. Therefore, the transition from 'P is consistent with *p*' to '*p* is the case' is not something that just follows; P may be outweighed by some future piece of information ~P and, in Hillary's case, it is already inconsistent with the available information. (Even if P becomes evidence that suggests that *p*, it will not entail that *p* by default – since circumstantial evidence is also evidence and P might be interpreted as or transformed into circumstantial evidence.)

Even though they are responsive to evidence, beliefs are not automatic responses to evidence. Evidence needs power to affect beliefs. Hillary needs reasons that are independent of P to make the transition from 'P suggests that *p*' to '*p* is the case.' To the same effect,

Hinchman [2013: 633] writes: '"My evidence meets the epistemic standard for *p*!" is not yet an assertion that *p*, since staking a claim about one's evidence falls short of staking a claim about what that evidence *shows*.'[72] The gap between the judgement 'P suggests that *p*' and the judgement '*p* is true' leaves a lot of space for self-deception.

In the fourth step, I propose a plausible reconstruction of an action of self-deception *simpliciter*. A self-deceiving person may deploy various strategies to prevent herself from making the transition from judging that 'P suggests that *p*' to judging that *p*. She may intentionally manipulate her interpretation of P in order to make the transition from judging 'P suggests that *p*' to judging that '*p* is true' unjustified. She can come up with various rationalisations; or avoid, obfuscate, or exaggerate the truth; she can cast doubt on the truth; focus on positive evidence or add 'alternative facts'; and whatever comes to her mind. Two things need to be said about these strategies.

First, they all involve the intention to deceive herself with respect to the *power* of P (if P is understood as evidence) or the *meaning* of P (if P is a piece of data) but not the intention to deceive herself into maintaining a false belief. That is to say, a self-deceiver may realise that she is manipulating a particular piece of information or evidence (P) in order to retain her belief and so she will recognize that she is being deceptive towards herself with respect to P. However, the deception will not involve her intending to continue to believe a falsehood – since she takes her belief that ~*p* (e.g., 'no affair') to be true.

Second, these are all strategies of interpersonal deception. Typically, interpersonal deception also involves manipulating the deceived person's (D_D) evidence. But, this is not where the similarity with interpersonal deception ends, and this similarity is crucial for the success of the action. The fact that P is inconsistent both with Hillary's totality of evidence and her belief that ~*p* makes this manipulation justified from Hillary's perspective. Her action of self-deception is analogical to cases of interpersonal deception in which the deceiver, D_R, intends to deceive D_D into a belief D_R believes to be true – often seen in double or triple bluffing. Both D_R and Hillary engage in deception because they find this move justified in the overall score – in benevolent double and triple bluffs, D_R helps D_D by making him believe a true proposition and Hillary helps herself by keeping her true belief safe.

[72] I here take it that 'P suggests that *p*' is interchangeable with 'P meets the epistemic standard for *p*' while 'P shows that *p*' means 'P entails that *p*.'

Somewhere along these lines, Gardner [1993: 18] writes that (italics added)

> It would be wrong . . . to see self-deception as resulting from a preference for trying to solve an internal or psychological problem over its external or real counterpart. Instead, self-deceivers should be seen as *mistakenly taking themselves to have solved their real problem* in solving their psychological problem; or, put another way, as failing to make a proper distinction between psychological and real problems.

The kind of a self-deceiver I have in mind recognizes the discrepancy between her belief that ~*p* (and beliefs relevant to it) and P but she incorrectly thinks that the problem is in P rather than in her belief (and other relevant beliefs). The irrational step is in that this thought is motivated by her desire that ~*p* is true rather than by thinking that there is something problematic about P. Because she thinks that the problem is P, and the problem with P is mainly P's being inconsistent with ~*p*, she thinks that this should be solved by way of manipulating P in some sense. In turn, even though Hillary will see her action as deceptive, she will not see it as 'bad' deceptive (with respect to P); she will judge that deception is not relevant for the question whether ~*p* is true.

A bold claim seems to follow: in cases in which they think that the deception, i.e., the relevant kind of manipulation, is executed in order to support a true belief, self-deceivers need not hide their intention to deceive themselves from themselves (see 2.2). These self-deceivers find their manipulation of P (information or evidence) justified because it serves the truth and because this was the only method of reconciling P with this truth available at the time.

3. Concluding Remarks

The main idea is simple: if Hillary already believes that ~*p*, has many beliefs in support of ~*p*, and takes that P is outweighed by her totality of evidence, she will think that her manipulation of P was performed with the intention of harmonizing the input with her true belief – thinking that time will show that P was misleading. There is nothing strange about this proposal: people do not have a standing intention not to be deceived; they have a standing intention not to believe falsehoods.

Suppose that you discover that I used a double bluff to make you believe that the Persecutors are on the road in order to help you. You will not abandon this belief just because I manipulated you into believing it. Likewise, the kind of self-deceiver described above

reasons in the following way ‘Yes, what I am doing basically is deceiving myself, but I am doing this in support of a true belief.’ Contrary to what was thought to be an irrefutable intuition on self-deception (e.g., Borge [2003: 11]; Lynch [2016: 516]), we see that it is possible to be fully aware that you are deceiving yourself and still be successful in it. Becoming fully aware that you are deceiving yourself is not a tantamount to realising that your belief is false.

V-3 The Maria-Type Cases

1. Introduction

In this thesis, I focused on defending the traditional conception of self-deception, the traditionalist view. The deflationary view is generally thought of as much less problematic. However, while I agree with the general deflationary proposal that some people rightly described as self-deceivers act intentionally but deceive themselves unintentionally and while I do think that this behaviour is very common, I argued that the vast majority of the received deflationary proposals are unsatisfactory (III-2).

As a replacement, I propose a theory of what I call self-deception *secundum quid*, which embraces the main deflationary thought of acting intentionally but deceiving oneself unintentionally but rejects the majority of deflationary proposals as to how this is actually achieved. Instead, it uses the manipulativist conception to generate the intuition that an action of this kind is rightly called self-deception. According to my manipulativist account (Section II-2/1), A is a self-deceiver *secundum quid* if and only if A

(1) intentionally brings about φ-ing as φ-ing intending to thereby affect his or her own truth-evaluable mental state.
(2) φ-ing causally contributes toward manipulation of A's agential use of his or her own cognitive capacities in a way relevant for 1.
(3) A realises neither that he or she is deceiving themself by φ-ing nor that he or she is thereby manipulating themself (otherwise, A would not φ).
(4) A's truth-evaluable mental state is affected by way of φ-ing.

My proposal can be illustrated on the following example. Suppose that you ask me to play American football with you and your friends, a game I know nothing of. As the game progresses, it happens that I am all open and so you throw a ball towards me, I catch it, and you shout 'Run behind the opponents' goal line and touch the ground!' At your command, I run towards the opponents' goal line and I score a touchdown. In this case, I ran towards the opponents' goal line and touched the ground intentionally (conditions 1 and 2 above) but I did not realise that I am doing this in order to score a touchdown (condition 3), which my touching the ground (condition 4) effectively is. My action may rightly be called scoring a touchdown even though I did not know I was doing that. Self-deceivers *secundum quid* perform actions of this sort.

In this subchapter, I will defend the view that the manipulation from condition 2 should be understood as involving *intentional pseudo-rational reasoning*, namely, the person's intentional deviation from her own standards of reasoning motivated not by her desire to deceive or manipulate herself (i.e., to score a touchdown) but rather by her desire to thereby bring her specific truth-evaluable mental state into harmony with the available evidence (i.e., to touch the ground with the ball).

I note here two things. First, it is absolutely essential that the pseudo-reasoning is intentional in the sense in which the agent realises that he or she is deviating from her own typical behaviour. That is, if A has deceived himself into thinking that *p* by engaging in biased evidence gathering his action must be intentional in the sense in which it is plain to A that the way he is gathering evidence is not how he typically does that but he is still intentionally deviating from his normal behaviour. Self-deceivers *secundum quid* realise that they are being insincere towards themselves without realising that they are in effect deceiving themselves – since realising that they are deceiving themselves would make them self-deceivers *simpliciter*. Second, A's *realising in retrospective* that he or she deviated from their own norms of reasoning in the above-prescribed way should also count as A intentionally engaging in pseudo-rational reasoning as long as A persists in this kind of reasoning. That is to say, if, after being pointed it out to A that he would have reasoned differently had it been someone else in question, A does not change his reasoning (i.e., A accepts in retrospective that he violated his own norms), then A's reasoning is pseudo-rational in the sense required for A to count as being self-deceiver *secundum quid*.

Any kind of reasoning can be pseudo-rational as long as the person is intentionally deviating from her own typical way of reasoning. I will focus on two most common kinds: pseudo-rational belief formation and pseudo-rational inference to the best explanation.

2. Self-Deception as Pseudo-Rational Reasoning

2.1 Pseudo-Rational Belief-Evaluation

Studies conducted by Wentura and Greve [2003, 2005] found that, when faced with undeniable challenges to a certain feature of their self-image, people will incorporate the data in a way that preserves the particular feature of this self-image. Specifically, participants of their study preserved their own conception of themselves as erudite against the evidence to the contrary by modifying what counts as evidence for being erudite. These findings suggest

that beliefs are not ranked only according to rational criteria, such as epistemic justification or inferential support, but also by their levels of subjective importance and how they affect your well-being (Michel and Newen [2010: 739]; Lauria, Preissmann, and Clément [2016]). Consistently with influential models of hypothesis-testing (e.g., Trope and Liberman 1996), these findings indicate that your motivation to believe that *p* normally determines both whether you will require more evidence to reject *p* rather than ~*p* and whether you will require less evidence to hold *p* rather than ~*p*.

Consistent with this hypothesis are also the results of the following two studies. Gal and Rucker [2010] found that subjects who were prompted to doubt their beliefs became stronger advocates of those beliefs than those subjects who were prompted to feel more confident in their beliefs. This trend became even more apparent when the beliefs in question were experienced as important. Furthermore, negative evidence was unable to reduce confidence in those beliefs. Therefore, aiming to avoid this kind of belief-immunization, Van der Leer and McKay [2017: 25] instructed their subjects to imagine that their first estimate is incorrect and, then, to think which assumptions or considerations could be wrong if this was the case. This prompt to an alternative estimate, which targets the subjects' line of reasoning rather than the available information, was successful in shaking the confidence in the initial estimate, a result suggesting that subjective importance of our beliefs significantly effects the kind of reasoning involved in their evaluation.

Requiring more evidence to revise a belief of significant importance or avoiding information too difficult to handle is not sufficient for self-deception but it may, nevertheless, involve self-deception, *provided that* the person also exhibits a relevant kind of insincerity towards herself. And there is a sense in which retaining your beliefs may involve this relevant kind of insincerity. The participants of the Wentura and Greve's studies immunized their conception of eruditeness against certain types of information by way of reorganizing belief-sets about the subject matter in question; they developed a personal conception of eruditeness. With this new conception of eruditeness, a particular piece of information, which would otherwise become counterevidence to the view of one as erudite, is rendered consistent with their self-referential belief.

In the same way, it is pertinent to note, people may immunize many other 'important' beliefs, such as the belief that one can control one's own drinking or smoking, or that one is a good driver or parent, or the belief that one's memory or health is good, and many more. I

may, for instance, admit that I forgot to buy the milk while I was in the grocery store and that I consistently forget to buy things that I wanted to buy, but dispute that this particular skill says much about my general memory, thus 'trivialising' my failure (Greve and Wentura 2010: 723).

Greve and Wentura – along many other scholars (e.g., Gur and Sackeim [1979]; Quattrone and Tversky [1984]; Mele [1997b, 2001, 2006]; Von Hippel and Trivers [2011]; Lauria, Preissmann, and Clément [2016]) – see this behaviour as self-deceptive but, since motivated biased evidence-gathering or evaluation is insufficient for self-deception, they fail to tell us why this is the case. Michel and Newen [2010], on the other hand, say that this kind of belief-revision is self-deceptive because it is pseudo-rational; the person is applying a double standard of reasoning: had this been someone else, the self-deceiver would have readily said that the person is not erudite, that her memory is bad, health is deteriorating, and so forth. This is a very good characterisation of the kind of insincerity and manipulation required for self-deception. If we apply this to the *My Brilliance* case, we get that I have immunized the belief that my thesis is brilliant against the prospective counterevidence by initiating a pseudo-rational process. I have the following doxastic profile (following Michel and Newen 2010).

(*p*) My thesis is brilliant.
(*q*) Containing no flawed arguments while offering a substantial contribution to science is necessary for a thesis to be brilliant.
(*r*) My thesis offers a substantial contribution to science and has no flawed arguments.

The belief that needs to be preserved is *p*; the belief that gets refuted by prospective counterevidence (E¬) is *r*; and the belief that inferentially connects *p* and *r* is *q*. The belief that my thesis is brilliant (*p*) needs to be preserved because undermining this belief undermines my self-esteem by way of undermining the belief that *I* am brilliant (p_{ME}) thus causing me distress. The belief that *q* is a trait definition: it defines what counts as evidence against believing that *p* and in favour of believing that ~*p*.

In self-deceptively retaining the belief that my thesis is brilliant, I could do some of the following things. I could change my theory of 'being brilliant,' which is largely contributed by *q*, so that a thesis can be brilliant even if *q* is false. Replacing or waiving *q*, which connects *r* and *p*, blocks the inference from ~*r* to ~*p* and so *p* and p_{ME} will not be threatened by E¬. Alternatively, I could revise *r* so that information that used to be counterevidence to *r* is now consistent with it. If *r* is not threatened by E¬, *p* and p_{ME} are not threatened by it

either. The rationale behind both of these strategies is the same: revise the relevant theory so that it is consistent with evidence; that is, so that I can be brilliant in the face of E¬. But, I could discredit E¬ in some way. I could perform a pseudo-rational information or evidence-assessment; I could think that E¬ is ambiguous, for example. However, this strategy could rightly count as self-deceptive only if it is intentional in the right sense. A pre-conscious dismissal or downplay of evidence (Lauria, Preissmann, and Clément 2016: 125) is a motivated kind of reality but it does not involve the required kind of insincerity that can be seen in pseudo-rational reasoning and thus it cannot be considered as sufficient for self-deception. Having said that, I offer some possible ways in which a relevant pseudo-rational theory-revision could be done.

In revising my theory of *q*, *Tq*, I may exclude the first conjunct ('Containing no flawed arguments') from *q*; this will allow me to admit that some of my arguments are flawed and continue believing that *p*. Alternatively, I may introduce supplemental beliefs *s* and *t* that justify changing or waiving *q*. I may believe that 'As long as only some arguments are flawed but the solutions are brilliant, the thesis is brilliant,' and the like. In another strategy, I could undermine E¬, the prospective counterevidence to *r*, by revising my theory of *r*, *Tr*, and thereby bring my belief that *p* into harmony with the evidence about my performance. For example, I could convince myself that my arguments were not refuted but have rather been overlooked by those who made the objections. Also, I could think that my approach is too much advanced for it to be accepted at this point, and so on.

These techniques of pseudo-rational T-revision, Michel and Newen [2010: 743] correctly maintain, are capable of transforming the evidential status of any information towards the desired result (see D'Cruz 2015: 326–327). However, manipulating my theories so that they are made consistent with the new information or manipulating my assessment of available evidence so that threatening information is discounted as ambiguous is not what makes my behaviour self-deception, and this is vital. Rather, it is my using a double standard of belief-formation that makes me a self-deceiver. I am intentionally violating my own rules, that is. If it was you in question, I would have revised none of these theories or discounted none of the available evidence but would straightforwardly say that your thesis is not brilliant.

It is because of this intentional pseudo-rationality that my behaviour rightly counts as self-deception *secundum quid*; in particular:

(1) I have intentionally revised *q* as 'revising *q*' intending to thereby retain my belief.
(2) This revision causally contributed toward manipulation of my agential use of my own cognitive capacities.
(3) I did not realise that I am deceiving myself by revising *q* (otherwise, I would not have revised *q*).
(4) Due to the revision of *q*, my belief was retained in the face of E¬.

However, this theory cannot be applied to the Maria case: it can explain only the cases of self-deception that happen in reflection to one's personal features but Maria is deceiving herself about Arnold's behaviour, not about his qualities.[73] Therefore, following Szabados [1974: 60–61, 64–66], who also calls this kind of behaviour 'quasi-rationality,' I propose a general account of self-deception *secundum quid* that sees self-deception as any kind of motivated pseudo-rational reasoning.

2.2 Pseudo-Rational Abductive Inference

2.2.1 Jumping to Conclusions

Let us recall the Maria case first.

> **Maria** possesses much evidence that her husband, Arnold, is having an affair with one of his attractive female co-workers. Arnold has lost sexual interest in Maria, he has protected his phone with a password, and she sometimes picks up what appear to be subtle love signals her husband exchanges with this female co-worker. Yet, Maria intentionally sets these distressing pieces of information aside and, as a result, retains the false belief that Arnold is not having an affair. However, she does not intend to deceive herself (since she never suspected an affair, we think that she lacks the motive for deception). Rather, she retained her belief by what looks like explaining away pieces of information that cause her distress. In a sense, she seems to be acting just like scholars who only give weight to claims that support their own position and avoid those that may undermine it.

In the Maria case, the belief that *q* is 'Arnold is not having an affair' and not 'I am a trusty cuckquean, a fool.' Because the belief is about explaining behaviour and not about someone's personal characteristics, the pseudo-rational belief-revision theory cannot explain it. I suggest the following explanation. Since we analyse behaviour by making inferences to the best explanation, i.e., abductive inferences, Maria is guilty of self-deception (*secundum quid*) if her abductive inference to the conclusion that 'Arnold is not having an affair' is pseudo-rational; that is, if, motivated to retain her belief that Arnold is not having an affair, Maria intentionally applied the kind of abductive reasoning that allows her to retain it – realising that she is being insincere towards herself but not that she is effectively deceiving herself thereby. She may further support her belief with other sorts of pseudo-rational reasoning,

[73] The same limitation applies to Galeotti's [2012] version of Michel and Newen's account.

such as rationalisation, reappraisal, doubting the source of threatening information, explaining threatening information or doubts away, appraising the threatening evidence as ambiguous, shifting blame to others, and so forth.[74]

The following is an example borrowed from the Massey University *Critical Thinking* course. This example will be of particular use in analysing Maria's behaviour.

> **Casey**: I've tried to call Casey today and she hasn't picked up, she didn't respond when I waved at her, and she hasn't liked my post on Instagram. I think she must be mad at me.

It seems fair to say that this person judged that Casey is mad at her too quickly; she jumped to a conclusion. Maybe Casey is just absent-minded? Or, maybe Casey accidentally left the phone at home, did not see when the speaker waved at her, and her mom took her computer to work? The main difference between these two rival explanations is that the former treats Casey's overall behaviour as explainable on a single hypothesis whereas the latter sees it as an unconnected series of events, each requiring its own explanation. Of course, both of these proposed explanations are mere speculations but their point was not to actually explain the case. Rather, I wanted to highlight a tendency to jump to conclusions when explaining other people's behaviour. While a single-hypothesis explanation might be appropriate, we actually do not have good reasons to think that these particular instances of Casey's behaviour should be explained using only one hypothesis. This is why the reasoner jumped to a conclusion. She did not bother to consider other possibilities.

Jumping to conclusion, I argue, is a mistake that underlies interpretations of many examples of supposed self-deception offered in the literature; the conclusion that these agents are deceiving themselves because the evidence 'points' to the right belief is hastened. If we take a closer look at the Maria case, in particular, it is not at all obvious that Arnold is having an affair with his attractive female co-worker. All that we know is that Maria *possesses enough evidence*, and we should read 'evidence' as 'information' or 'data,' to conclude that he is having an affair. But this same 'evidence' could also be enough for her to conclude something completely different. Maybe Arnold is 'being in' his character even when they are not filming – this is a common practice that contributes to the quality of one's performance. Having enough evidence to conclude that *p* makes *p* a candidate explanation; it does not

[74] On some of these techniques, see, e.g., Snyder and Higgins [1988]; Snyder et al. [1992: 282–287]; Lauria, Preissmann, and Clément [2016].

make it the best explanation. The best explanation is to be adjudicated among candidate explanations, a point left undiscussed by many received theories of self-deception.

Many scholars simply take it for granted that the self-deceiver must hold a false or epistemically unjustified belief but this view breaks up the analogy with interpersonal deception and deception in general in a very radical way. The hallmark of deception is not falsity (triple bluffs are a good counterexample, see II-1/3.2) or being unjustified in believing something (tampering with evidence is a strategy of deception but the deceived is not unjustified in forming her belief); rather, the hallmarks of interpersonal deception are *insincerity* and *manipulation*.

Failing to correctly identify the essence of interpersonal deception yields a very problematic approach to behaviour that might involve self-deception. Looking at the case from the third-person perspective, the impression is that Maria should believe that Arnold is having an affair because this hypothesis best explains his overall behaviour; that is, it seems natural to think that the affair is *the* reason why Arnold is behaving in such a way. But, I ask, why is this the best explanation and, even more, why should *this* explanation, even if it is the best one, be the criterion by which we judge whether Maria is deceiving herself? It could easily be that Arnold protected his phone with a password because he leaves it on the set while shooting his scenes, that he has lost sexual interest in Maria either because he stopped taking protein shakes (which increase sexual desire) or because he is tired, and so on? Also, exchanging subtle love signals – if those really are love signals – merely reveals flirting, not an affair. Many people flirt without having any further intentions. Flirting is a self-esteem thing for them. Finally, Arnold could be flirting in order to keep the on-screen chemistry with his co-worker. With so many available options, motivated not choosing the best one *per se* hardly makes one a self-deceiver.

In what follows, I will first offer an explanation to the question why people typically go for the single-hypothesis explanation of behaviours such as Arnold's and Casey's. Then, I will propose that Maria is a self-deceiver not because she is failing to appreciate evidence or judge which explanation is the best but rather because, being motivated to keep her belief in Arnold's fidelity, she is intentionally using a double standard of reasoning: in a situation in which she would normally use a single-hypothesis to explain the behaviour at hand, prompted by the desire to retain her belief, she intentionally adopts the multiple-hypothesis explanation, thus exhibiting relevant kinds of pseudo-rationality and insincerity.

2.2.2 Motivated not Raising Further Questions

According to the erotetic theory of reasoning (Koralus and Mascarenhas 2013), human reasoning proceeds by raising questions (even if they do not look like questions); specifically, we interpret premises in reasoning as questions and then try to answer them as quickly as possible. This is a part of a systematic strategy that allows us to reason in a classically valid way. The erotetic theory nicely explains both our reasoning failures and successes; however, I will not discuss it in detail. Rather, I will only discuss the bit that is relevant to my argument: the thought that reasoning proceeds by raising questions and answering them.

Parrott and Koralus [2016] use the erotetic theory of reasoning to explain delusional thoughts. On their view, delusional subjects have an impaired ability to endogenously raise further questions. A Capgras patient, for instance, reaching the conclusion that the person they see is nobody they know will *hold this conclusion fixed* (Parrott and Koralus 2016: 403), namely, they will ask no further questions, which then explains their condition: their delusion is unresponsive to evidence because evidence is not being treated as introducing further premises (i.e., questions) into the person's reasoning about the topic. The faulty explanation of the Casey case, I argue, that says that she is mad at the reasoner, involves this kind of a mistake.

Recently, I presented the Casey example to my Critical Thinking students at the University of Auckland before discussing abductive inferences with them. Unsurprisingly, a great number of students from both universities reached the same conclusion as the imagined person from the example; most of them thought that Casey is mad at the faulty reasoner. I then asked my University of Auckland students to give me their reasons for this explanation (so that we could identify the origin of their error). One reaction I find particularly interesting. Having me pointing it out to her that this is a weak inference and that, even if Casey is mad, we are not justified in inferring this from her behaviour, the student explained how she reacted to the example and what has made her reach this explanation (italics added).

> For some reason, the first thing that springs to mind is 'she is mad at me,' which makes me *over-look and start finding* reasons 'why' she is mad.

The first thing that, 'for some reason,' springs to mind, i.e., 'She is mad at me,' is the answer to the first question. *If we hold this answer fixed* – rather than asking further relevant questions, such as 'Is there a reason for Casey to be mad at me?' – any further analysis will actually involve finding reasons as to *why Casey is mad* – given that the question is already

answered and the answer is held fixed. This 'first thing that springs to mind,' therefore, which has made the student over-look and start finding reasons 'why' Casey is mad, is a very important response to the case at hand: holding the answer to the first question fixed makes the student's behaviour *appear* from the third-person perspective just like biased evidence gathering, but this does not involve biased evidence gathering at all. (The students were not primed into thinking that the conclusion is justified; we were investigating whether the inference is justified.) Rather, the student reached this conclusion by developing the most plausible single-hypothesis explanation, where the hypothesis held fixed is the answer to the first question, and she was then merely interpreting information in view of this answer. This answer, in turn, acted as the relevant interpretative paradigm that shaped the analysis of the incoming information.

Even though this kind of reasoning falls short of the desired ideal of rational reasoning, it is not irrational *per se*. The dynamics of ordinary life often leaves no time for raising all relevant questions; hence, jumping to conclusions is biologically functional. Nevertheless, this tendency sometimes needs to be resisted. In such circumstances, we need to answer further questions or abstain from passing judgements.

Maria approaches the issue of analysing Arnold's behaviour in the way we should be analysing Casey's behaviour. Arnold's behaviour is suspicious and implies affair only when interpreted under 'one hypothesis' paradigm, that is, only if the answer to the first question is being held fixed. However, holding this answer fixed without a good reason (i.e., without a plausible answer to a further question) involves a reasoning error. Strictly speaking, Maria did not make this reasoning error. She is endogenously raising further questions and this is the correct approach: a sound reasoner will either treat Arnold's behaviour as a set of unrelated occurrences or avoid passing judgement until she acquires more information. Interestingly, we are jumping to conclusions in explaining Arnold's behaviour, not Maria.

But, even more interestingly, the fact that she is being closer to the reasoning ideal does not entail that Maria is not deceiving herself. In fact, we can rightly say that Maria is deceiving herself even if it turns out that her belief is true. While we have been jumping to conclusions because we thought that this is how Arnold's behaviour should be explained, i.e., we were just reasoning the way we normally reason, Maria was raising further questions because she was thereby intentionally avoiding the conclusion you and I have reached. In other words, she was not raising further questions as part of her general reasoning strategy

but to avoid making the unfavourable conclusion. She was running away from the truth by asking additional questions and she will probably continue to raise further questions as long as this is required to avoid the unwelcome conclusion. Had she been in our shoes, Maria too would judge both that Casey is mad at her and that Arnold *is* having an affair.

In short, Maria is not deceiving herself because her belief is false and/or unjustified (i.e., because she applied an incorrect analysis to Arnold's behaviour). Rather, she is deceiving herself because she is intentionally using double standards in reasoning due to a motivational effect of her desire to escape a certain conclusion.

3. Concluding Remarks

Maria is deceiving herself not because she is doing the wrong thing but rather because she is doing the right thing for the wrong reason (this is where her insincerity lies). She has chosen the reasoning strategy based on the desirability of the envisaged conclusion. While she normally raises no further questions when analysing other people's behaviour, confronted with the new information about Arnold, Maria, compelled by the need to protect the cherished belief, started asking questions until she asked enough questions to avoid the dreaded explanation.

Maria did not intend to deceive herself with her behaviour (she was not going for the touchdown) but, vitally, she knew exactly what she was doing (she knew that she was running like crazy towards the opponents' goal line). She knew that she is not being completely sincere towards herself in evaluating Arnold's behaviour. In contrast, biased evidence gatherers do not know that they are biased and they are in no sense being insincere towards themselves.

Maria exercises a dual standard of abductive reasoning: she switches abductive methods depending on the topic and this switching is motivated and intentional (it is not a preconscious response); she will still use the typically used abductive method when analysing other aspects of Arnold's behaviour or the behaviour of other people. For instance, seeing that he left his car keys at home and that his gym bag is not in the house, she will infer that he decided to jog or walk to the gym even though this is not what Arnold typically does. And, vitally, she is violating her own preferred method of abductive reasoning and not the objective observer's method of abductive reasoning.

In conclusion of this Subchapter, self-deception *secundum quid* is best understood as intentional pseudo-rational reasoning. In particular, any kind of intentional deployment of a specific reasoning method with the aim of affecting (acquiring, retaining, supporting) a certain belief should be understood as pseudo-rational self-deceptive kind of reasoning.

V-4. Unwelcome and Indifferent Self-Deception

1. Introduction

Roughly, unwelcome, or twisted, self-deceivers end up believing what they would rather like not to be the case.[75] It is a matter of consensus that the challenge posed by unwelcome self-deception is twofold. One problem is that it is hard to see why a person would deceive herself into believing something she does not want to be true – and self-deception is supposed to be a motivated kind of irrationality. Specifically, it is very hard to say why Zelda would deceive herself into believing that Scott *is* having an affair, why a person would deceive herself into believing that the government is spying on her or that she was abducted by aliens, why a prospective athlete would deceive herself into believing that she is injured a day before the competition, and so forth.

In short, the first problem is 'Where does the motivation for unwelcome self-deception come from?' The second problem is that the answer to this question needs to be a part of a general account of self-deception that can explain both welcome and unwelcome kind of self-deception. The kind of things that explains why Zelda deceives herself into believing that Scott is having an affair should also explain why Nicole, Hillary, and Maria deceive themselves into believing that their spouses are not having affairs.

To these two problems, I add another one: this account needs to be able to explain cases of what I call indifferent self-deception in which people end up believing propositions with no personal significance or value and so they have no reason to want them to be true or not (see II-2/2). Nothing in their life hinges on the truth of their belief and yet they still make herself believe it.

This is an important addition to the problem of what motivates self-deceivers for several reasons. First, thus far, we have only been concerned with cases in which *my* husband is (not) having an affair (Hilary, Maria, Nicole, and Zelda), *my* child is (not) abusing drugs, or in which *I* am not brilliant, *I* am not a moral person, *my* memory is not good, in which *I* enjoy smoking, the government is spying on *me*, and so on. But, I could easily deceive myself into

[75] Let us, for the sake of simplicity, take belief (degrees of belief) or levels of credences as unproblematic and non-rival concepts. This issue is tangential to my discussion and nothing in my discussion hinges on it.

believing that the government is spying on *you* or that smoking is not bad for *your* health, that *you* are brilliant, or that Donald is brilliant or moral, etc. Because the received theories are concerned only with motives that directly pertain to the self and the self's interest – namely, they see self-deception as a self-defensive, reality-coping strategy – the received theories are all too narrow. We need a theory capable of explaining cases which are rightfully described as cases of self-deception, a motivated kind of intentional insincerity towards oneself that involves self-manipulation, but which do not produce beliefs that are directly associated to self-deceivers themselves, their self-esteem, or their personal practical benefit.

Suppose that I have somehow (the method is irrelevant) managed to deceive myself into believing that Trump's hair is beautiful and his temper is subtle, and that Bush really believed that Iraq had weapons of mass destruction. We could sometimes ascribe these beliefs to my political preference, but if we were to do this every time a person deceives herself about Trump or Bush we would be wrong. Furthermore, to use this explanation is to leave other analogous cases unexplained. I could be deceiving myself into believing that Bill never meant to lie to the public, for example. Moreover, imagine that I deceive myself into believing that Elvis is not dead. Notice that these beliefs are fairly common and most interesting: Elvis has been spotted everywhere from Graceland to Russia and there is even a Facebook page 'Evidence Elvis Presley is Alive' with 25,349 likes and 25,303 followers, as assessed on 19 December 2017. Also, and this example is quite important, suppose that I have evidence – and I see it as evidence – both that my student cheated on the exam and that she is bald-faced lying to me that she did not cheat. Suppose further that somehow, and to the student's utmost surprise, I allow myself to be fooled by her lie; let us say that I allow the lie to make me (sufficiently) less confident in truth. Why would I do this to myself?

It seems obvious that self-deceivers get nothing from any these beliefs and that thus their benefit does not appear to be the purpose of their self-deception – in most of these cases, at least. This thought does not fit the idea of self-deception as a self-defensive strategy. And yet, we cannot deny that some people may allow themselves to believe these things and even intentionally make themselves believe them or exercise these thoughts in a way relevant for self-deception.

So, the problem we get is this: it seems quite unproblematic to think that people are motivated to deceive themselves about the matters of personal concern – and, if the result is something they would rather avoid, maybe the mechanism backfired? – but why would they

deceive themselves about issues that bear no significance on their self-esteem, self-understanding, or themselves in any salient sense? In fact, in some situations, as in the benevolent professor case, the action benefits the third party at the self-deceiver's expense. This is an important concern.

Nevertheless, while indifferent self-deception might appear as making things worse, it may actually help us explain the motivation of self-deceivers easier. Thus far, we have been focusing on the idea that self-deceivers need to 'get something' from their self-deception – even though, typically, there is a greater payoff in truth. Indifferent self-deception tells us that self-deception need not have any personal payoff, since it need not be for the sake of the self and it need not be motivated by any concern related to the self at all. Rather, it may be motivated by some other interest the person finds worthy, sometimes even more worthy than her own benefit. The answer that indifferent self-deception seems to be suggesting is that received theories are flawed because they were trying to answer the wrong question, namely, 'What does the person get from deceiving herself?' There need not be anything in it for the self-deceiver. And, it is important to note, this is a clear parallel with interpersonal deception: there, a deceiver might act to his detriment and for some 'greater good.'

In what follows, I will first address the problem of unwelcome self-deception and argue that the received solutions fail to explain it. Then, I will introduce my proposal, the idea that what motivates welcome and unwelcome self-deception is our incorrect understanding of ourselves, argue that this hypothesis can explain these two kinds of self-deception better than its rivals but that applying it to indifferent self-deception requires that we extend this incorrect understanding to our understanding of the world. Self-deception, I conclude, is a defensive strategy but not self-defensive. Self-deceivers are defending their own overvalued idea of the world.

2. Unwelcome Self-Deception

Let us first recall the Zelda case and then try to explain it.

> **Zelda** mistakenly believes that she possesses plenty of evidence that her husband Scott is having an affair with his new close friend Ernest. Although Scott denies vigorously that they are having an affair, they are getting along great and that can't be right – she is quite sure of that. In fact, the more she thinks about it, the more evidence she finds. Thinking about that one time they were in Paris and met Ernest, she remembered how Scott went to the bathroom in the middle of the night and then she became convinced that he actually went to Ernest. Cheating fits Scott, she reasons, because he's such an unlikely person to do it and their

marriage has been falling apart for months now. Zelda does not exhibit any signs of delusion or schizophrenia: she does not, for instance, think that Ernest is leaving Scott secret messages in newspapers, etc. Therefore, we cannot say that she is deluded but she does seem to be self-deceived. However, she certainly does not want Scott to be having an affair or that she believes that. Even worse, her belief makes her desperate.

Zelda deceived herself into believing that Scott is having an affair, which is something she does not want to be true. Furthermore, Zelda deceived herself in the face of an abundance of counterevidence and she even construes evidence out of counterevidence: she thinks that cheating fits Scott because he is such an *un*likely person to do it. She is surely trying hard to reach the dreaded conclusion. Many ingenious proposals are put forward to explain this kind of behaviour. I will address them in turn and argue that they are unsuccessful.

One very influential explanation is based on the hypothesis that emotions often affect belief-formation processes. When I am angry, writes Lazar [1999: 280], what my friend did to upset me may seem large and her kindness insignificant, and *vice versa* when I am in a good mood. On this view, the self-deceiver's emotional state is priming the relevant processing systems to gather evidence in a biased fashion, which, depending on the emotional state, sometimes produces favourable and sometimes unfavourable beliefs (Dalgleish [1997: 110]; Mele [2001, 2006]; Lauria, Preissmann, and Clément [2016]). Mele [2006: 114], for example, prominently argues that, whereas for many people it is more important to avoid acquiring the false belief that their spouses are having affairs, the converse may well be true of some insecure, jealous people. Mele's influential proposal is based on two theses.

The first is taken from Sharpsteen and Kirkpatrick [1997: 627] who write that jealousy complex is a mechanism for maintaining close relationships which appears to be 'triggered by separation, or the threat of separation, from attachment figures.' Let this be the *threat thesis*, according to which, the jealous desire to preserve her marriage is what triggers Zelda's self-deception by priming her cognitive mechanisms towards the unwelcome conclusion.

The second thesis is supplied by the FTL theory of hypothesis testing (see III-2/2). According to FTL (Friederich [1993]; Trope and Liberman [1996]), the concern for minimizing having costly false beliefs drives lay hypothesis testing and the predicted cost determines the confidence threshold required for acceptance or rejection of a belief. The greater the predicted damage caused by falsely accepting that *p*, the greater the cost of falsely accepting that *p* and, accordingly, the greater aversion towards falsely believing it. The level of aversion towards falsely believing a proposition determines the threshold for believing or

rejecting it. The lower the threshold, the less evidence is sufficient for accepting a belief, and the threshold is low if the expected cost of believing it falsely is low (Mele 2006, 2009).

According to the relevant combination of the *threat thesis* and the FTL, Zelda's case is explained by saying that her strong jealous desire to maintain her relationship with Scott plays a role in rendering the false belief that Scott is *not* having an affair a 'costly' error. Specifically, since believing erroneously that Scott is *not* having an affair may cause her not to take steps to protect the relationship against an intruder, Zelda would have a higher aversion towards falsely believing this proposition than towards falsely believing the one that causes her suffering (Pears [1986: 42–43]; Mele [2001: Ch. 5; 2006: 114]). And this higher aversion tells us why she made herself believe what she does not want to be true.

While undoubtedly innovative and interesting, this explanation simply strikes me as false. The first thing to notice is that the *threat thesis* limits the application of the view only to cases of unwelcome self-deception that involve jealousy, which makes it too narrow (similarly, Lazar 1999: 275). In addition, the main idea is that believing that Scott is having an affair will help Zelda protect the relationship against the intruder, but the problem is that this belief will actually endanger the relationship: Zelda might confront Scott concerning his 'affair' and, if persistent, she will drive him away. Besides, there is no threat of separation that could trigger jealousy. The threat of separation is actually caused by self-deception: Zelda invents evidence.

I note here that Zelda's inventing evidence is taken from a real case of mad jealousy followed by self-deception that eventually led to a divorce – it is not something I invented so that the case fits my view and undermine other accounts. This kind of behaviour, which can be observed in general population quite often, sits uneasily not only with the *threat thesis* but with the FTL as well: FTL does not predict inventing evidence with the purpose of accepting a belief. Inventing evidence defeats the purpose of minimizing errors. Furthermore, if Zelda believes that Scott is having an affair when in fact he is not, and acts on this belief, she may reveal herself to Scott as untrusting of their relationship in a way she would regret; this, in turn, makes the unwelcome belief the more costly one and, according to FTL, the confidence threshold for acquiring it should be higher than the threshold sufficient for acquiring the belief that Arnold is not having an affair. Therefore, the Melean proposal does not seem to be able to explain a large class of cases of unwelcome self-deception.

Another proposal is that the unwelcome belief serves the function of reducing anxiety (Johnston [1988]; Barnes [1997]). Barnes [1997: 41] argues that deceiving oneself into believing something one does not welcome reduces anxiety associated with some other proposition. She introduces an example of a jealous husband, John, a less dramatic version of Zelda, to illustrate her point.

> **John**: John's friend George agrees to do something when asked by John's wife, Mary, but had refused to do it when asked by John. John, who cares about George's regard, is taken aback by George's refusal of his request, and he begins to suspect and, ultimately, believe that George and Mary are having an affair. John, by self-deceptively believing that his wife has been unfaithful avoids concluding, as he otherwise might have concluded, that George's reaction is a consequence of his having a higher regard for Mary than for John.

Barnes's explanation is that, by misinterpreting evidence, John trades higher anxiety, which would be caused by believing the proposition supported by George's behaviour, for lesser anxiety, which is caused by believing the proposition that resulted from self-deception. The more powerful desire biases John into accepting the unwelcome proposition. However, this does not seem right: the unwelcome belief that his wife is unfaithful does not reduce John's anxiety; rather it increases it (Scott-Kakures 2001: 321). Believing the unwelcome proposition would serve the purpose of reducing anxiety only if John would care more about George's regard then about Mary's fidelity and, even then, this would not be 'unwelcome' but rather 'less-than-fully-welcome' self-deception – seeing that John gets what he wants, namely, George's high regard. In addition, this explanation cannot explain Zelda's case: her self-deception is driving her mad. She gets nothing from the trade-off simply because she has nothing to trade: the anxiety is generated exclusively by her jealousy. And finally, while John needs to explain George's behaviour, Zelda has nothing to explain: Scott is not acting unusually, she is inventing evidence.

Because unwelcome self-deception often causes or increases anxiety – Van Leeuwen [2007a: 425] calls it *dreadful* self-deception but this name seems too strong – some scholars believe that its main purpose is not reducing anxiety. Rather, and this is consistent with one of the main features of the discussed Melean [e.g., 2006: 114] proposal, it serves the purpose of making the agent's goals and interests, such as avoiding danger, more likely to be fulfilled than not (Scott-Kakures 2000, 2001). On this hypothesis, the self-deceiver's unwelcome belief, e.g. 'Scott is having an affair,' motivates her to take steps to put an end to what she believes mistakenly, i.e., that 'Scott is having an affair.' That is, by acquiring the unwelcome belief, self-deceivers make sure that they will take all necessary steps to avoid the unwelcome

state of affairs. The anxiety reduction is the anticipated upshot of acting to avoid, ameliorate, or delay the feared possibility, but it is not the sole reason for the behaviour (Scott-Kakures 2000; 2001: 323).

There are several problems with this view. One is that we have no reason to think that Mary and George will ever end up in a situation that may lead to an affair. George merely has a higher regard for Mary than for John. This is even clearer in the case of Scott and Ernest; they are just great buddies. The unwelcome state of affairs does not seem to be a possibility in either of these cases. It follows, then, that John and Zelda are in effect causing themselves unnecessary, unprovoked distress. Their actions in no sense serve their interests. Actually, they undermine their best interests.

Finally, according to the reasoning that supports the idea that unwelcome self-deception will help the person make necessary steps to avoid the unwelcome state of affairs, the belief 'Mary is having an affair' should help John realize that she is not having an affair but, in real life, this rarely happens. Rather, the belief 'Mary is having an affair' actually prevents John from realising that Mary is not having an affair by making him misinterpret the evidence. This is even more obvious in the Zelda case. The more evidence she gets that Scott is *not* having an affair (*~p*), the more convinced she becomes that he *is* having an affair (*p*); she interprets the evidence in favour of *~p* as evidence that *p*. Similarly, in the hurt athlete case, acquiring the unwelcome belief will prevent the person from competing rather than serving her best interest. Therefore, if this kind of behaviour is a result of a mechanism meant to serve the purpose of making the agent's goals and interests more likely to be fulfilled than not, then we must say that this mechanism is malfunctioned in the cases I just mentioned.

However, saying that cases such as Zelda and John involve a malfunctioned protective mechanism assumes that this hypothesis is correct, but whether it is correct is exactly what is at stake here. Cases such as Zelda, John, and the hurt athlete are counterexamples to this hypothesis; to deny them this status is to beg the question. Yet, there are cases in which unwelcome self-deception might be serving some of the agent's best interests. Consider the following case (Scott-Kakures 2001: 314).

> **Barbara**, a busy young attorney, leaves home on her daily commute, late for an important meeting. As she nears the freeway entrance, she is suddenly taken with worry that she has left her gas stove on. Barbara does not remember whether she had left it on. Due to a disagreement she had with her husband, she was absent-minded that day, which could have distracted her from her routine. She desperately tries to visualize turning off the burner, but she cannot. And

> then her fear of her house burning down becomes so real that it causes her to believe that she did leave it on. Convinced now that she has left the stove burning, she calls her husband but he has already left home. At great inconvenience, she returns home only to discover that the burner is off.

The unwelcome belief increases Barbara's chances of preserving her house; hence, if the case is of self-deception, the example fits nicely with the suggested theory and with the FTL. In particular, if, on the one hand, Barbara falsely believes that she has left the stove on, the cost is relatively low but if, on the other hand, Barbara falsely believes that she has not left the stove on, the cost is extremely high – Barbara's house will be destroyed by the fire (Scott-Kakures 2001: 323). Therefore, this specific mechanism serves her goals and interests in an obvious way, which gives us a reason to think that the cases above might also involve the same mechanism, only malfunctioned.

Nevertheless, even though it is obvious that this mechanism serves the agent's best interests, it is not clear that Barbara's behaviour involves self-deception. Principally, a sense in which Barbara is being insincere towards herself, a feature self-deception should have, is missing. Furthermore, it is not clear that Barbra's belief, though false, is epistemically unjustified: after all, she cannot remember turning the stove off and she was absent-minded that day. Finally, Barbara is not epistemically negligent and she is not violating her own rules of belief-formation. Rather, she is acting in her best interest. Therefore, her behaviour lacks the most basic marks of self-deception.

True, Barbara did end up with a false belief but identifying holding a false belief with being deceived misuses the metaphorical meaning of the term (see II-1) and, more importantly for the case at hand, what actually caused the belief is the combination of Barbara's fear and her inability to clearly remember turning the stove off. This does not seem to be what we may rightly call self-deception. She did not invent evidence like Zelda or misinterpret information like John. Rather, the belief seems to have been generated sub-personally. That is to say, granted that Barbara really believes that she left the stove on – and she could be mistaking her fear or some other mental state for a belief – what seems to have happened is that her fear of having the stove on has directly generated the belief, which is an interpretation Scott-Kakures [2001: 322] accepts.

I see three possible interpretations of the case. According to the first, this is not a belief but rather a default thought, a subpersonally generated simulation similar to imagination (see Gerrans 2014). The default network is a mechanism of testing thoughts for consistency and

adequacy, which in turn may qualify them for representational states such as beliefs. When unsupervised, default thoughts may occupy the role of representational states, i.e., beliefs. However, they may equally become beliefs if the mind judges that the thought is reasonable, which is the second interpretation. Accordingly, Barbara's fear triggered the relevant default thought, 'My house may burn down,' which, judged as reasonable according to prior beliefs and available evidence (she was absent-minded), became a belief. Finally, it could easily be that the fear is caused by Barbara's representation of the stove as on; this fear caused anxiety, which then caused behaviour (Levy 2016). Importantly, on neither of these explanations, the case involves self-deception: even though the belief is false, it is based on a reasonable hypothesis.[76]

So, the main problems are that Barbara's thought does not appear to be unjustified and that behaviour does not seem to involve any kind of insincerity towards herself. But, what is more interesting about this case is that, in these circumstances, the thought 'The stove is on' is not unwelcome even though the state of affairs in which the stove really is on is unwelcome. To be precise, while it is causing anxiety, this is not something Barbara does not want to believe in the situation as described, but it is, notwithstanding, something Barbara does not want to happen. She wants to go home and check but she hopes not to see the stove on once she is there. This makes the case consistent with another prominent account of self-deception.

It is typically thought that self-deceived subjects want that what they believe is true. However, seeing that this cannot explain unwelcome self-deception, Nelkin [2002] and Funkhouser [2005] argue that self-deceivers actually want to *believe* the unwelcome proposition or, at least, to have some first-person qualities associated with such a belief, but do not want that what they believe is true. Being the jealous, insecure type, Zelda wants that Scott is *not* having an affair and, out of caution, wants that she believes that Scott *is* having an affair (Funkhouser 2005: 298) – since believing what she does not want to be true will make her more vigilant. Similarly, believing that the stove is on will make Barbara go back and check; hence, Barbara has a good rationale to want to believe this while wanting that what she believes is not true.

This is a nice way to put forward the thesis that unwelcome self-deception is a protective mechanism in a way that allows its application to welcome self-deception as well, which

[76] Even on a variation in which Barbara is confident that the stove is off but the anxiety is still present, it would be sensible on her part to go back to check whether the stove is off: she would do that to assuage her anxiety (Levy 2016: 8).

resolves both problems posed by unwelcome self-deception. Both welcome and unwelcome self-deceivers want to believe the proposition they end up with; the only difference is that the former want this proposition to reflect the state of affairs while the latter do not; they rather want their believing this proposition to aid preventing this state of affairs from occurring. Still, postulating this kind of motivation is not only problematic and very complicated; it is also unnecessary and impractical.[77] Zelda and Barbara can perform all cautionary actions even without wanting to make themselves believe the unwelcome proposition. Actually, Zelda should not want to believe that Scott is having an affair because this belief will only cause her problems.

Suppose that a hypochondriac deceives herself into believing that she *has* a very serious illness despite all the evidence that she is healthy. It cannot be that she wanted to believe that she has (let's say) cancer in order to make it less likely that she indeed gets cancer. Acting on her belief that she has cancer would involve taking chemotherapy, or undergoing surgery, etc., which is clearly harmful. Unwelcome self-deceivers need not – and often should not – believe that the unwelcomed state of affairs has already occurred in order to become more vigilant (*pace* Scott-Kakures [2001: 323–324]; Funkhouser [2005]). This kind of belief may defeat the purpose of being more vigilant.

Maybe the right kind of a protective mechanism should be based on the wish to believe that one is *at risk* of getting cancer, or, analogously, that one is at risk of losing her husband's affection? This proposal is a step forward and it will explain the Zelda and the hypochondriac kind of cases, but it cannot be generalised across all cases of unwelcome self-deception. Many cases of unwelcome self-deception do not serve the agent's goals and interests.

Take, for example, the case of an athlete who deceived herself into thinking that she is injured a day before the important competition. It is not in this person's interest to wish to believe that she is hurt or that she is at risk of getting hurt right before the big competition: this belief might prevent her from competing and it will definitely reduce her chances of success by making her underestimate her own capabilities, which defeats the purpose of the supposed protective mechanism. Then, there is the benevolent professor. Making yourself believe the student's lie in no sense serves your interest, since you are at no risk. With these cases in mind, we must admit that the proposal that unwelcome self-deception serves the agents' best interest requires of us to posit a very large number of cases in which the

[77] Against this proposal, see Fernández [2013: 387–388].

mechanism of unwelcome self-deception malfunctioned, and having so many exceptions to the rule speaks against positing that rule in the first place.

In fact, the only clear case in which believing what one does not want to be the case but in which the belief is serving the person's interests is the Barbara case but this is also the only case of those discussed that does not appear to be involving self-deception. Therefore, seeing that the cases such as Zelda, John, the hypochondriac, and the hurt athlete should be understood as counterexamples to the view that unwelcome self-deception may serve some of the agent's interests by helping the person to make necessary steps to avoid unwelcome state of affairs and that the only situation in which the false belief serves the agents interest does not seem to involve self-deception, this view should be abandoned.

The final explanation, the least demanding version of the Melean proposal, is that the agent's inability to avoid thinking about the painful truth is actually responsible for unwelcome self-deception. The welcome self-deceiver cannot face some unwelcome reality and so she turns her head away, whereas the unwelcome self-deceiver cannot get her mind off some unwelcome reality and so she treats some data as evidence that what she fears of is true (Bach [2009]; similarly Lazar [1999]). Specifically, the person's fear, or whatever the anxious emotion, makes her incapable of getting her mind off the subject and this keeping her mind on the subject makes the dreadful possibility more realistic than it really is. The subject dismisses information she can easily put out of her mind and concentrates on information she cannot put out of her mind. As a result, the likelihood of the relevant unwelcome state of affairs obtaining might seem greater than is suggested by the evidence. Let us call this the *preoccupation thesis*.

The *preoccupation thesis* does not seem right when applied to the Zelda and the hurt athlete cases. In both cases, we seem to have self-deceivers who actually invent evidence rather than merely focus on information in support of their own belief. Zelda never really had 'evidence' or anything remotely resembling it to think that Scott is having an affair with Ernest. We may grant that the self-deceiving subject dismisses information she can easily put out her mind and concentrates on information she cannot put out of her mind but this does not explain why Zelda interprets information that should act as counterevidence to her belief as evidence in favour of the belief.

In addition to this problem, there is a concern that the preoccupation thesis, as favouring a plausible explanation of a particular case, cannot support the best explanation of many

important cases. For example, it seems much more probable that the athlete deceived herself into believing that she is hurt because of her fear of failure at the upcoming competition – the injury gives her an excuse not to compete – and not because she fears that she really is hurt. Most importantly, in cases in which the *preoccupation thesis* appears to be supporting the best explanation, the essential features of self-deception, insincerity towards oneself and the agent's intentional action that resulted in deceiving oneself, are missing; the relevant mental state is caused by subpersonal mechanisms. Therefore, while very plausible, this thesis cannot explain unwelcome self-deception.

I now turn to developing my own explanation.

2.1. The Superself

Studies suggest that most people have unrealistically positive views of themselves and that they hold unrealistic optimism about the future (Taylor and Brown [1988]; Taylor et al. [1989]). For instance, most people think that they are in more control of their life than they in fact are and they take more credit for success while denying responsibility for failure (Langer and Rodin [1976]; Snyder et al. [1992]). The presence of these unrealistically positive views is observable at all levels of consciousness: a study conducted by Epley and Whitchurch [2008] shows that, in an array of faces, people will identify their more attractive self than their actual self faster and, because unconscious processes are faster than conscious ones (e.g., Neely 1977), Von Hippel and Trivers [2011: 13] infer that the enhanced version of the self is likely to be represented in memory below conscious awareness.

This kind of one's failure to understand oneself is most likely motivated. People often see their close friends in an overly positive light (Kenny and Kashy 1994), which suggests that the misperception is not a result of a simple cognitive deficiency or of a lack of information. Rather, it seems to be resulting from a specific preconscious bias towards ourselves and our close others. Since the bias is motivated, many scholars understand this kind of behaviour – i.e., exaggerating your qualities and minimising your flaws – as self-deception rather than as having a false understanding of oneself (most notably, Gur and Sackeim [1979], Quattrone and Tversky [1984]). However, I argued against this view by saying that a motivated cognitive error need not involve self-deception, especially if this error is a result of how you normally use your cognitive capacities (III-3a/2.1–2.2); that is, especially if you normally see yourself in a positive light. Similarly, Van Leeuwen [2007a: 425, n.15] writes (italics added).

> There is another prevalent type of mental phenomenon commonly called self-deception that I think would better be classified otherwise. The phenomenon I have in mind should be called *self-inflation bias*. This is simply a general tendency that many or most people have to see themselves in a comparatively more positive light than evidence would justify. This differs from the types of self-deception I discuss here in that it is the result of a more general habit of thinking positively about oneself ... Thus, when people think about themselves as better drivers than they actually are, this is probably the result of self-inflation bias. The self-inflation bias seems to me to involve a fixed frame for interpreting incoming information in a self-flattering way. Self-deception, on the other hand, involves motivated selection of different frames.

Having an incorrect self-understanding (having false beliefs about your personal characteristics) is much better conceived as a consequence of a natural self-inflation bias than of self-deception – having a false belief is 'deception' only in a metaphorical sense of the term. Nonetheless, we should not think that most people only have the tendency to see themselves in a comparatively more positive light than evidence would justify and that the error of self-understanding is always inflationary (see Snyder et al. 1992: 287–290).[78] There are studies that strongly indicate that people with negative self-views preferentially seek negative evaluations. Some of them even divorce spouses who perceive them in an overly positive manner (Swann 1983: 2012). These people seem to be biased towards negatively portraying themselves and, if so, then why not say that Zelda or some other unwelcome self-deceiver, such as an athlete who deceived herself into thinking that she is hurt, belong to this group of individuals? Some Zelda-like self-deceiver might be exhibiting a fear of happiness and the athlete-like ones may be having a fear of success.[79] This seems like a good explanation but, before testing it, I must note that the idea that people can exhibit both positive or negative self-bias comes with two important qualifications.

Firstly, we should not think that a type of a bias – positive or negative – is a hallmark of a certain type of personality. Rather, we have reasons to think that the bias is context-dependent. The results of a study conducted by John and Robins [1994] have very interesting implications. Some 35% of their participants showed a self-enhancement bias, i.e., positively inflating one's self-understanding, and yet, 15% exhibited self-diminishment bias, while the remaining 50% were fairly accurate. These results do not only suggest that the direction of

[78] The mother's mental state or her attitude towards the child (both in natal and prenatal periods) plays an important role on the child's mental development (e.g., Davis et al. 2004). This may contribute to the direction of the bias, I speculate.

[79] We should be cautioned here not to take all individuals seeking negative information of themselves as necessarily having negative self-views. Sometimes, bolstering a person's self-image may result in the person becoming more willing to search out negative information about herself (Von Hippel and Trivers 2011: 2).

bias is not a constant, i.e., negative or positive (Funder 2011: 23), but, I add, also that whether a person will exhibit a failure of self-understanding in the first place depends on numerous factors, such as relevant context or personal significance of the topic. Below I explain my hypothesis.

The participants of this study were master of business administration (MBA) students at the University of California, Berkeley, who were simulating the meeting of a compensation committee in a large company. The fact that half of the participants did not exhibit any bias is best explained on the thesis that performing better or worse than other subjects played no motivational role for a half of these students. Possibly, they did not see this as a competition but rather as a teamwork and this perspective required of them to understand their personal contribution to the team correctly. If I am right, this study, then, suggests that whether one will see oneself in a positive or negative light depends, not only on the type of one's personality, but equally on the kind of context one thinks that one is currently in and on what one takes to be at stake in that particular context. That is, we should think that the same person will exhibit different kinds of self-related biases in different contexts and with different things at stake. Notice, one's *perspective* on the issue in question does seem to be vital for self-deception. The perspective determines not only whether there will be self-deception but also what kind of deception will take place (e.g., welcome, unwelcome).

Secondly, and most importantly, we have reasons to think that neither people with positive nor people with negative shifts in their self-understanding realise the direction of their bias. People have a very vague understanding of what an average score means and therefore it is hard to see what they actually mean by saying that they are better or worse than average (Brooks and Swann 2011: 17). Specifically, while we see ourselves as somehow and somewhat 'special,' it is unclear whether we really understand what we mean by that. It is as if the idea of ourselves and of our close others as special is just embedded in our psyche.

Having the understanding of yourself as somewhat special need not be an irrational feature: being a person by definition entails being unlike anybody else, it just is being special. What is problematic in our self-understanding, however, is that our features that are supposedly special are not random, namely, our specialness is motivated. For instance, those who see no value in driving will not experience themselves as better drivers. I shall call this motivated perspective on ourselves as being special *Superself*. We think that our achievements are greater than those of the others, but also that our problems are greater than

those of the others ('[It's easy for you;] I'm too addicted to be able to quit smoking'), we may think that we can do things the majority of people cannot ('I can quit smoking whenever I want'), that bad things won't happen to us, and so forth. With Superself in place, I proceed to use this concept to explain unwelcome self-deception in a way that welcome self-deception is also explained thus solving the first two problems.

2.2. Zelda and Zelda-like Cases

Before I proceed, I want to note the following. All the cases introduced in this thesis were deliberately left underdescribed and the least specified features are relevant for determining the motive. The reason is that I think that most theories of self-deception offer an overly simplistic understanding of the motive. The aim of my discussion from this subchapter was to show the variety of factors that will have an effect on self-deception; in particular, whether the person will deceive herself, which method she will use, and in which direction the deception will go. The direction my analysis will now take is speculative but unavoidable. The motive can be explained only by filling out the cases under inspection in various relevant ways, an insufficiently appreciated point made fifteen years ago by Nelkin [2002: 388]. I now proceed to develop my solution.

One's desire to preserve beliefs integral to the Superself may explain many cases of self-deception. The most obvious case is when I deceive myself into believing that my thesis is brilliant but we could equally assign this motivation to Hillary, Maria, Nicole, and even Zelda. It could be that, in deceiving herself that Scott is having an affair, Zelda is shifting blame for her failed marriage on her cheating husband, but it could also be that Zelda is of those people who think that they do not deserve happiness or cannot achieve happiness. The important thought here is Zelda fears not that the marriage is in danger. Rather she fears that it is her fault the marriage is in danger – maybe she is avoiding the thought that Scott does not find her sexually attractive ('cheating fits Scott' would mean that it is his faulty character rather than her looks that caused the affair) – or that she is getting what she deserves, or something along these lines.

Once we know *why*, *how* becomes easy to explain. Zelda seems to be deceiving herself in the way Maria deceived herself; by exercising pseudo-rational abduction. She is interpreting Scott's behaviour in a pseudo-rational way, often switching between abductive methods. At some times, she seeks for one hypothesis explanation, like the idea that cheating fits Scott because he's such an unlikely person to do it, and, at other times, she treats particular

instances of Scott's behaviour as random occurrences that seek individualised explanation, like the case when Scott went to the bathroom and she saw this as evidence of the affair (the consistent strategy would be to analyse this from the perspective of his typical 'bathroom' behaviour combined with how much he had to drink that night).

A similar answer to the *why* question is most likely behind John's belief that George and Mary are having an affair, but before suggesting one, I must make a brief note. The John example, as offered by Barnes [1997: 41], is insufficiently described. It is very unlikely that anyone would deceive him or herself into believing that their partner is having an affair just because of the high regard of the third person. There must be much more below the surface in these cases and so filling out the details is a necessary requirement if the motive is to be identified. I suggest that the case should be filled out in the following way.

John is of those people who value their own self-image more than their marriage or friendships. John cannot imagine that anyone who knows him will not be able to appreciate his qualities. Accordingly, the best explanation for George's φ-ing when asked by Mary but not when asked by John is that φ-ing is not something George wants to do; rather, George was obliged to φ due to the emotional relationship he has with Mary. After all, John too has numerously times done stuff he did not want to do only to please his wife. This explanation will cause John anxiety and, considering his personality, is not something John wants to believe but it is the best explanation of the given situation from John's perspective. However, what makes this a case of self-deception is that John reached the conclusion by means of making a pseudo-rational abduction. This is the best explanation, according to John, not because he typically uses this abductive method but rather because this particular method preserves his Superself.

There seems to be a predominant view that men are afraid of being cheated on much more than women. Evolutionary biologists explain this by saying that this is because they want to make sure their genes get transferred to their posterity (e.g., Buss 2007: 336), but I would say that this behaviour is closely connected to male kind of Superselves. It is men who predominantly have, and at least think of having, affairs (Symons 1979: 27); in fact, evolutionary risks of extrapair copulation for women were enormous (Dixson 2009: 121–122). So, seeing that men indeed predominantly fantasize about having affairs, why not assume that they also think that their partners fantasize about having affairs too? Those who are willing to cheat or fantasise about it will be more vulnerable to recognizing their fantasies

in the behaviour of their partners. If I am right, then some Zelda and John-like self-deceivers are blaming their partners for what they have done or are fantasizing about doing.

The hurt athlete's Superself may also easily explain her self-deceptive belief that she is hurt: you cannot fail in a competition if you do not compete and you should not compete if you are hurt. By not competing, then, she is protecting her Superself. The person who thinks that the government is spying on him must be thinking that he is important enough for the government to see him as a threat. The paranoia is giving him a sense of value, it feeds his Superself. This fits nicely with Bentall's [1994: 353; 1995] thesis that says that persecutory delusions may have the function of protecting the individual against chronic feelings of low self-esteem (namely, they blame others when something goes wrong) and with the hypothesis that socially well-established people are unlikely to develop this kind of self-deceptive paranoid thoughts.

However, there is a class of self-deceptive behaviour that cannot be explained by appealing to the need to protect the Superself. It is the class of indifferent self-deception, which I discuss in the next section.

3. Indifferent Self-deception

Suppose that I have somehow managed to deceive myself into thinking that Donald is a great president and a well-behaved person, that George really thought that Iraq has weapons of mass destruction, and that Bill did not mean to have sexual relations with Monica or to publicly lie about it. Yes, I would be a very peculiar person if I really did believe these things, but if cognitively unimpaired people can believe that alien lizards rule us, then it being peculiar is not a reason to dismiss my proposed case.

What makes the case interesting is that we have reasons to think that I neither welcome nor unwelcome these propositions, at least not in the sense in which Hillary and Maria want that their husbands are having affairs and Zelda and John do not want their spouses to have affairs. Unlike Hillary, who gains something by living in a world in which Bill is not having an affair with Monica, I gain nothing by living in a world in which Bill is not having an affair with Monica, Donald is well-mannered, and George is honest. In analysing unwelcome self-deception, Scott-Kakures (2001: 316) asks 'Just what is in it for these motivated believers?' but this is the point of indifferent self-deception – there is nothing in it for them. So, if I get nothing from deception, then what could be my motive to deceive myself?

The phenomenon of indifferent self-deception strongly suggests that the motivation for self-deception should be looked for in a concept that is much broader than the agent's self-understanding or self-esteem. I hypothesise that the motivation comes from the desire or the need to protect the endorsed picture of reality. For example, suppose that I believe that Elvis is still alive and that I constantly rationalise evidence to the contrary in a way that involves pseudo-rational abduction. It seems reasonable to think that people susceptible to the kind of self-deception are those who are unable to bring themselves to believing that heroes simply die. Indifferent self-deceivers are not protecting themselves, their presidents, or their heroes; rather, they are protecting a valiant world in which good is rewarded, heroes do not simply die but rather ride into glory, and presidents are honest and well-behaved.

Indifferent self-deceiver is not completely indifferent: they do want to protect their valiant world. However, unlike in welcome and unwelcome self-deception, they do not have any personal interest in self-deception, namely, in protecting this world; they are indifferent in this sense. That is, I really get nothing by having defended the good moral character of Bill, George, or Donald – in fact, I might even recognize that I suffer detriment, as in the gullible professor example – whereas Hillary, Maria, and Zelda do get some personal benefit from their behaviour. In the next section, I spell out this theory in detail.

4. The Theory

Self-deception is a defensive strategy but not necessarily self-defensive: the person is defending a certain conception of reality of which her 'self' is only a part. We could say that self-deception is a concept-defensive strategy, given that self-deceivers are defending a specific conception of the world. I discriminate the following kinds of motives, depending on which aspect of the cherished conception is being protected.

Ego-centric motives, i.e., those based on the Superself, are those that have something to do with the self-deceiver's personal benefit. Appealing to these motives can explain most cases of welcome and unwelcome self-deception but not all motives related to the self are ego-centric. Defending a certain idealised picture of my close relatives or my group (my nation, religion, team members, etc.) would not strictly speaking be ego-centrically motivated actions, but they do involve a level of concern for the self. These would be *nos-centric* (nos [lat.] – us) motives. For instance, I could be deceiving myself into believing that my team or

my nation is much better than it really is, but I could also make myself believe that philosophers are better experts in psychiatry than psychiatrists, and so forth.

Similarly, deceiving oneself into believing that one's child is not abusing drugs or that one's child's chances of recovery from serious illness are higher than they actually are typically has more to do with the child's benefit than with the self-deceiver's own benefit. Accordingly, a parent who deceived herself into believing that her child's chances of recovery are good may allow herself to undergo a painful or possibly fatal medical procedure, e.g., donate a kidney, to save her child. In this case, the self-deception would be to the self-deceiver's detriment. For the same reason, a child may deceive itself into believing that its father's favourite team is better than it really is thereby maintaining the positive image of the father. Similarly, the gullible professor might be allowing himself to be deceived into thinking that his student did not cheat in order to maintain his positive image of this student. (However, the professor could equally allow himself to be deceived so that he need not act and punish the student, which would be an ego-centric motive). Let the other person's benefit kind of motives be named *tu-centric* (tu [lat.] – you).

Ego/tu/nos-centric motives are the most common motives but, quite often, the motive could directly be a noble idea, an ideal, an ideology, or anything one holds strongly and finds worth preserving. These would be *value-driven motives*. In defining value-driven motives, I partly rely on Veale's [2002] account of overvalued ideas, i.e., beliefs associated with idealised values. According to Veale [2002: 386], a value is something thought to be good or important to an individual. Values can be terminal (e.g. the importance of happiness) and instrumental (e.g. the importance of being honest). Strongly held values, says Veale, are the principles on which one will not yield and which are not subject to empirical testing. If these values come into conflict with reality, I suspect, their holders may resort to deploying self-deceptive strategies. In a sense, all motives for self-deception are based on overvalued ideas, i.e., beliefs associated with idealised values; the only difference is which misconstrued value needs protection.

In deceiving himself into believing that his student did not cheat, the gullible professor may be protecting the value of being honest or the ideal of academic citizenship; the professor need not be motivated by the student's personal benefit. Similarly, if Maria was motivated by her desire to keep her idealised picture of Arnold as a great guy (many people idealise their spouses) or by her desire to preserve the ideal of her family as a happy family,

then her motive is value-driven. This is an interesting scenario: while there is a benefit for Maria, she is not deceiving herself for that benefit. Therefore, even some cases of welcome or unwelcome self-deception need not involve *self*-defensive strategies.

Idealised values are responsible for many beliefs the believer's environment would recognize as unethical or superficial. My beliefs about Donald, George, and Bill are excellent examples, I think. Of course, a value can be negative, which generates relevant beliefs (overvalued ideas): I may believe that utility drives people or that you cannot be a good person if you are a communist or a politician, and so on. A negative overvalued idea, such as the thought that the movie industry is run by deviant people, may motivate me to deceive myself into believing that Scott is having an affair with Ernest.

The final category of motives I list here is what I call *naïve motives*. These motives come from some naïve conceptions self-deceivers adopted in various ways, such as by trusting news media, internet, movies, specific ethnographic misconceptions, religious ideals, and so on. The most common naïve motives are 'Bad things don't happen to good people,' or the simple 'It won't happen to me' principle, I suspect.

Motives can be combined. For instance, Maria, Nicole, or Hillary could be motivated both by a naïve belief that bad things won't happen to them – this would be a naïve-ego-centric motive – and by their desire that their spouses are valiant people – this would be a value-driven motive. Similarly, in deceiving myself that my thesis is brilliant, I could be, for example, acting on an ideal that hard work always pays out or on the desire not to disappoint my parents or my supervisor.

To conclude, whether there will be self-deception or not depends not only on one's type of personality but also on how one understands reality and on whether a certain part of that picture is under threat. Not every unremarkable PhD candidate will deceive him or herself that his or her thesis is brilliant, just as not all MBA students will think they did better than their colleagues. The fact that some PhD students correctly appraise their work entails only that this specific context was motivationally empty for these people, in the sense that no particular idea, overvalued or not, was under threat in a way that could trigger self-deception.

Finally, I do recognise that this proposal may seem superficial as an explanation of self-deception but this is the point: the reality is 'superficial.' People are not reasoning machines. They get overcome by emotions, prejudices, their own self-conceptions, and many other

things. We do need theories capable of explaining these sorts of behaviour. Sophisticated theories of self-deception, and those that think that there is some benefit for the self-deceiver are relevant here, read too much into this phenomenon.

VI. CONCLUDING REMARKS

The conclusion of this thesis is actually Chapter V, where I propose final explanations of the cases I introduced. In this chapter, therefore, I will briefly summarise the findings of my analysis.

1. The Three Main Views

According to my analysis, received accounts of self-deception either raise some important concerns or cannot be generalised to explain the cases discussed here, and mostly both.

Traditionalists either explain trivial cases of genuine self-deception *simpliciter* or fall into the Traditionalist trap, the need to partition the mind, and then slip into deflationism. The reason why partitioning the mind leads into deflationism is simple: the traditional conception of self-deception requires that I intend to deceive myself but, if a functionally semi-independent part of my mind is responsible for the self-deceptive though, then my intention to deceive myself is lacking. Rather, I have intentionally performed a certain action – e.g., I averted my attention and thereby affected the mental division – but have thereby deceived myself unintentionally. The key problem is that I did not avert my attention intending to deceive myself. By taking it that parts of the mind are rational centres of agency, we get a theory of a schizophrenic mind paired with a confusing theory of *inter*personal *intra*personal deception (one rational agent deceives another) and this fails to capture the traditionalist main idea of the self intentionally deceiving *itself*. The self did not intend to deceive itself.

So, if all versions of traditionalism end in deflationism one way or another, the most natural strategy is to abandon the main thesis of traditionalism and assume that self-deceivers act intentionally but deceive themselves unintentionally. To do this, nevertheless, is not to embrace a non-problematic theory of self-deception. In fact, deflationists typically either talk about motivated auto-manipulation, a flaw of Melean accounts, or relapse and fall into the Traditionalist trap, a feature of Johnstonian accounts.

The most charitable reading of Melean accounts, for example, must concede that the irrationality of self-deception is in the third-person perspective and this is quite problematic. Placing the irrationality into the third-person perspective is effectively breaking up with the idea that insincerity towards oneself is the distinguishing feature of self-deception – I dubbed

this the Deflationary trap. In addition, Melean accounts, despite their best efforts and many virtues, fall short of explaining cases where the product is a true belief. The only successful deflationary theories are those offered by Bach [2009] and Michel and Newen [2010]. These theories can be used to explain some cases of self-deception *secundum quid* but neither can serve as a general explanation of this kind of self-deception.

Revisionist theories from the family of solutions that understand self-deception as deception *about* the self, on the one side, fall either into the Traditionalist trap, i.e., diving the mind, or into the Freudian trap, i.e., being pseudo-scientific, or even both; or, on the other side, fail to avoid the Deflationary trap – the agents were not insincere towards themselves. Because of this, these theories fail to solve problems they were meant to solve.

Finally, the 'failed deception *by* the self' revisionist theories are probably the most diversified of all. Their proposals involve a supposed conflict between (i) meta-beliefs and first-order beliefs, or different (ii) beliefs triggered contextually, or different (iii) half-beliefs, or even different (iv) belief-components – most notably, dispositions. These ideas are valiant but some of them fall into the Freudian trap (the meta-belief proposals) or fail to explain a plausible situation in which Nicole avows to her friends that Tony is not having an affair while avoiding the route that would force her to drive her and her friends past Rachel's house. The pretence or alief solutions imply that the phenomenology of these cases may just be the result of some other mental states that generate results, i.e., the verbal behaviour, that can easily be confused with the results generated by beliefs. This hypothesis is ingenious but also dangerously close to saying that this is not self-deception but pretence in the sense of deception of others. Finally, theories from all three families introduced an abundance of *ad hoc* hypotheses, mainly used to get out of various traps.

Interestingly, respective failures of these views may be highlighting the importance of understanding self-deception principally as an action, which is the main point of Chapter II of my thesis. The product does not constitute self-deception. What constitutes deception, and thus self-deception as well, is what the agent has *done* to get this product.

2. My Approach

I argued that the paradoxes of self-deception result from incorrect theories of interpersonal deception and lying. With the correct theories of lying and deception in place, the paradoxes

need not arise. In particular, since the manipulativist conception of deception does not require the deceiver to form the intention to cause the target to believe what the deceiver believes to be false, the Doxastic paradox need not arise when this conception is applied to self-deception. Similarly, since the intention to deceive need not involve making epistemically worse off, the Strategy paradox need not arise either: I may know my own intention to deceive myself and, thinking that this is for my own good, allow myself to be deceived. The Hillary case should be explained along these lines, I argued in V-2. Finally, because on the Violation Account (*violation*), lying entails neither the intention to deceive nor asserting what the liar believes is false, it is possible to lie oneself and even deceive oneself by lying.

The resulting traditionalist account of self-deception, I argued, is better than its rivals, adding that some deflationary theories, when slightly modified, also have their merit. Deflationism offers a substantial insight by maintaining that some actions of self-deception, while intentionally performed, were not performed with the intention to deceive oneself. I named these actions self-deception *secundum quid* as opposed to self-deception *simpliciter* captured by the traditionalist approach. To the insight that some actions may rightly be called self-deception even though the intention was not to deceive oneself, I added that the relevant behaviour must involve a kind of insincerity towards oneself; otherwise, it cannot rightly be called self-deception. Self-deceivers *secundum quid* are insincere towards themselves by engaging in pseudo-rational reasoning.

The traditionalist conception of self-deception based on the *manipulation-violation* combination can successfully explain all cases of self-deception while being consistent both with our common linguistic practices and with any theory of the mind, and while not raising any problems and paradoxes, or falling into any traps. Furthermore, this conception need not posit any kind of exotic or hybrid mental states, it need not appeal to the distinction between first-order and second-order mental states, and it need not say that we know the self-deceiver's mental states better than the self-deceiver himself. Therefore, it does not fall into the Freudian trap either. It also need not posit some sort of divisions of the mind at the level of its intentional mental states; hence, it avoids the Traditionalist trap. Then, seeing that self-deceivers do violate their own norms of reasoning or belief-evaluation, due to the manipulative trick, and that they are insincere towards themselves in the relevant sense, my account avoids the Deflationary trap.

My main solution is so simple and elegant that it gives us too many interpretative options, which may raise a concern that the theory is suspiciously successful; that is, successful in a way Freudian theories are 'successful' by actually being unfalsifiable. This concern would be justified only if we would have reason to believe that this elegance is a direct consequence of the unjustified reconceptualization of knowledge, deception, and lying. However, while it requires of us to abandon deeply-entrenched beliefs, such as that knowledge is a justified true belief and that liars must assert what they believe is false, I believe that my suggestions are quite justified: these well-entrenched theories are false and they are the thing that causes problems in understanding self-deception.

In the end, I addressed the problem of unwelcome and indifferent self-deception. I resolved it by suggesting that, rather than being a self-defensive mechanism, self-deception is actually triggered by the need to protect the person's idealised conception of reality. Because the person is an integral part, she is typically also represented in this reality as idealised, as a Superself.

I conclude with another speculative thought. If correct, this explanation of the motive brings to the surface an interesting theological dimension of self-deception. This idealised conception of reality, including the idealised conception of ourselves, closely resembles the theological idea of Paradise and human perfect condition lost as a result of human sin. Self-deceivers seem to be defending the conception of the Pre-Fall world, the perfect world as they imagine God created it. Overcoming the tendency to deceive ourselves might be the key element in salvation.

VII. REFERENCES

1. Adler, Jonathan. 2002. *Belief's Own Ethics*. Cambridge, MA: MIT Press.
2. Ainslie, George. 1997. 'If belief is a behavior, what controls it?' *Behavioral and Brain Sciences* 20: 103.
3. Alexander, M. P., D. T. Stuss, and D. F. Benson. 1979. 'Capgras syndrome: A reduplicative phenomenon.' *Neurology* 29: 334–339.
4. Amador, Xavier F., Michael Flaum, Nancy C. Andreasen, David H. Strauss, Scott A. Yale, Scott C. Clark, and Jack M. Gorman. 1994. 'Awareness of Illness in Schizophrenia and Schizoaffective and Mood Disorders.' *Archives of General Psychiatry* 51: 826–836.
5. Aimola-Davies, A. and Martin Davies. 2009. 'Explaining Pathologies of Belief.' In *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, eds. Matthew Broome and Lisa Bortolotti, 285–323. Oxford: Oxford University Press.
6. Aquinas, Thomas. 1922. 'Of Lying.' In *Summa Theologica*, vol. 12, 85–98. London: Burns, Oates, & Washbourne.
7. Archer, Sophie. 2013. 'Nondoxasticism about Self-Deception.' *Dialectica* 67: 265–282.
8. Artiga, Marc and Cedric Paternotte. 2017. 'Deception: a functional account.' *Philosophical Studies* (Published Online) 16: 1–22.
9. Audi, Robert. 1982. 'Self-Deception, Action, and Will.' *Erkenntnis* 18: 133–158.
10. ———. 1985. 'Self-deception and rationality.' In *Self-deception and self-understanding*, ed. M. Martin, 169–194. Lawrence: University of Kansas Press.
11. ———. 1988. 'Self-deception, rationalization, and reasons for acting.' In *Perspectives on Self-Deception*, eds. B. McLaughlin and A. Rorty, 92–120. Berkeley: University of California Press.
12. ———. 1989. 'Self-Deception and Practical Reasoning.' *Canadian Journal of Philosophy* 19: 247–266.
13. ———. 1997a. 'Self-deception, rationalization, and the ethics of belief: An essay in moral psychology.' In Robert Audi, *Moral knowledge and ethical character*, 131–156. New York: Oxford University Press.
14. ———. 1997b. 'Self-deception vs. self-caused deception: A comment on professor Mele.' *Behavioral and Brain Sciences* 20: 104.
15. Augustine. 1952/395. *Treatises on Various Subjects*, ed. R. J. Deferrari. New York: Fathers of the Church.
16. Austin, John Langshaw. 1962. *How to do Things with Words*. Oxford/Clarendon.
17. Bach, Kent. 1981. 'An Analysis of Self-Deception.' *Philosophy and Phenomenological Research* 41: 351–370.

18. ———. 1997. ‘Thinking and believing in self-deception.’ *Behavioral and Brain Sciences* 20: 105.
19. ———. 2009. ‘Self-Deception.’ In eds. Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter, *The Oxford Handbook of Philosophy of Mind*, 781–796. Oxford: Oxford University Press.
20. Bandura, Albert. 2011. ‘Self-deception: A paradox revisited.’ *Behavioral and Brain Sciences* 34: 16–17.
21. Barnes, Annette. 1997. *Seeing through self-deception*. New York: Cambridge University Press.
22. Baumeister, Roy F. and Karen Pezza Leith. 1997. ‘Biased steps toward reasonable conclusions: How self-deception remains hidden.’ *Behavioral and Brain Sciences* 20: 106–107.
23. Bayne, Tim and Elisabeth Pacherie. 2004. ‘Bottom-Up or Top-Down? Campbell’s Rationalist Account of Monothematic Delusions’ *Philosophy, Psychiatry, & Psychology* 11: 1–11.
24. ———. 2005. ‘In Defence of the Doxastic Conception of Delusions.’ *Mind & Language* 20: 163–88.
25. Bayne, Tim and Jordi Fernández. (2009). ‘Delusions and Self-Deception: Mapping the Terrain.’ In eds. Tim Bayne and Jordi Fernández. *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*, 1–22. New York: Psychology Press.
26. ———. 2009. *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*. New York: Psychology Press.
27. Bayne, Tim. 2010. ‘Delusions as Doxastic States: Contexts, Compartments, and Commitments.’ *Philosophy, Psychiatry, & Psychology* 17: 329–336.
28. Beane, Jeffrey C., Sean P. Graham, Thomas J. Thorp, and L. Todd Pusser. 2014. ‘Natural History of the Southern Hognose Snake (Heterodon simus) in North Carolina, USA.’ *The American Society of Ichthyologists and Herpetologists* 1: 168–175.
29. Bell, J. B., and B. Whaley. 1991. *Cheating and deception*. New Brunswick: Transaction Publishers.
30. Bentall, Richard P. 1994. ‘Cognitive biases and abnormal beliefs: Towards a model of persecutory delusions.’ In eds. A. S. David and J. C. Cutting. *The neuropsychology of schizophrenia*, 337–360. Hove, East Sussex: Psychology Press.
31. ———. 1995. ‘Brains, biases, deficits and disorders.’ *British Journal of Psychiatry* 167:153–155.
32. Bentall, Richard P., and Peter Kinderman. 1998. ‘Psychological processes and delusional beliefs: Implications for the treatment of paranoid states.’ In eds. T. Wykes, N. Tarrier, and S. Lewis. *Outcome and innovation in the psychological treatment of schizophrenia*, 119–144. Chichester: John Wiley and Sons.
33. Bermúdez, José Luis. 1997 ‘Defending intentionalist accounts of self-deception.’ *Behavioral and Brain Sciences* 20: 107–108.

34. ———. 2000. ‘Self-deception, intentions, and contradictory beliefs.’ *Analysis* 60: 309–319.
35. Black, Max. 1952. ‘Saying and Disbelieving.’ *Analysis* 13: 25–33.
36. Bobonich, Christopher. 2002. *Plato’s Utopia Recast: His Later Ethics and Politics*. Oxford: Oxford University Press.
37. Bond, Charles and Michael Robinson. 1988. ‘The evolution of deception.’ *Journal of Nonverbal Behavior* 12: 295–307.
38. Borge, Steffen. 2003. ‘The Myth of Self-Deception.’ *The Southern Journal of Philosophy* XLI: 1–28;
39. Bortolotti, Lisa. 2010. *Delusions and Other Irrational Beliefs: International Perspectives in Philosophy and Psychiatry series*. Oxford: Oxford University Press.
40. ———. 2011. ‘Shaking the bedrock.’ *Philosophy, Psychiatry, & Psychology* 18: 77–87.
41. ———. 2013. ‘Delusion.’ *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), ed. Edward N. Zalta, URL = http://plato.stanford.edu/archives/spr2016/entries/delusion/.
42. Bourget, Dominique, and Laurie Whitehurst. 2004. ‘Capgras Syndrome: A Review of the Neurophysiological Correlates in Cases Involving Physical Violence.’ *Canadian Journal of Psychiatry* 49: 719–725.
43. Bratman, Michael E. 1999. ‘Practical Reasoning and Acceptance in a Context.’ In his *Faces of Intention: Selected Essays on Intention and Agency*, 15–34. Cambridge: Cambridge University Press.
44. Braude, Stephen. 1995. *First person plural*. London: Rowman & Littlefield.
45. Breen, Nora, Diana Caine, and Max Coltheart. 2000. ‘Models of face recognition and delusional misidentification: A critical review.’ *Cognitive Neuropsychology* 17: 55–71.
46. Brighetti, Gianni, Paola Bonifacci, Rosita Borlimi, and Cristina Ottaviani. 2007. ‘”Far from the heart far from the eye”: Evidence from the Capgras delusion.’ *Cognitive Neuropsychiatry* 12: 189–197.
47. Broncano-Berrocal, Fernando. 2013. ‘Lies and Deception: A Failed Reconciliation.’ *Logos and Episteme* 4: 227–230.
48. Brooks, Matthew L., and William B. Jr. Swann. 2011. ‘Is social interaction based on guile or honesty?’ *Behavioral and Brain Sciences* 34: 17–18.
49. Brown, Stephanie L. and Douglas T. Kenrick. 1997. ‘Paradoxical self-deception: Maybe not so paradoxical after all.’ *Behavioral and Brain Sciences* 20: 109–110.
50. Buss, David M. 2007. *Evolutionary psychology: the new science of the mind – 3rd ed*. Pearson Education.
51. Campbell, John. 2001. ‘Rationality, meaning and the analysis of delusion.’ *Philosophy, Psychiatry, & Psychology* 8: 89–100.
52. Capgras, J., and J. Reboul-Lachaux. 1994 [1923]. ‘L’Illusion des “sosies” dans un délire systématisé chronique.’ *History of Psychiatry* 5: 119–133.

53. Carson, Thomas L. 2006. ‘The Definition of Lying.’ *Noûs* 40: 284–306.
54. ———. 2010. *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press.
55. Cavell, Marcia. 1999. ‘Reason and the Gardener.’ In *The Philosophy of Donald Davidson*, ed. L. Hahn E., 407–421. Chicago: Open Court Publishing Company.
56. Cermolacce, M., L. Sass, and J. Parnas. 2010. ‘What is Bizarre in Bizarre Delusions? A Critical Review.’ *Schizophrenia Bulletin* 36: 667–679.
57. Chan, Timothy, and Guy Kahane. 2011. ‘The Trouble with Being Sincere.’ *Canadian Journal of Philosophy* 41: 215–234.
58. Chisholm, Roderick M., and Thomas D. Feehan. 1977. ‘The Intent to Deceive.’ *The Journal of Philosophy* 74: 143–159.
59. Cialdini, Robert B., Carl A. Kallgren, and Raymond R. Reno. 1991. ‘A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior.’ *Advances in Experimental Social Psychology* 24: 202–234.
60. Clifford, William Kingdon. 1918. ‘The Ethics of Belief,’ In *Lectures and Essays*, eds. Leslie Stephen and Frederick Pollock, 95–109. London: Watts and Co.
61. Cohen, L. Jonathan. 1989. ‘Belief and acceptance.’ *Mind* 98: 367–389.
62. Collodi, C. 1916. *Pinocchio; The Tale of a Puppet*. Racine: Whitman.
63. Coltheart, Max. 2005. ‘Conscious experience and delusional belief.’ *Philosophy, Psychiatry & Psychology* 12: 153–157.
64. ———. 2007. ‘The 33rd Bartlett Lecture: Cognitive neuropsychiatry and delusional belief.’ *Quarterly Journal of Experimental Psychology* 60A: 1041–1062.
65. Coltheart, Max, Peter Menzies, and John Sutton. 2010. ‘Abductive inference and delusional belief.’ *Cognitive Neuropsychiatry* 15: 261–287.
66. Coltheart, Max, Robyn Langdon, and Ryan McKay. 2007. *Schizophrenia Bulletin* 33: 642–647.
67. ———. 2011. ‘Delusional Belief.’ *The Annual Review of Psychology* 62: 271–298.
68. Cuffel Brian J, Joseph Alford, Ellen P. Fischer, Richard R. Owen. 1996. ‘Awareness of illness in schizophrenia and outpatient treatment adherence.’ *The Journal of nervous and mental disease* 184: 653–659.
69. Currie, Gregory, 2000. ‘Imagination, delusion and hallucinations.’ In *Pathologies of Belief*, eds. M. Coltheart and M. Davies, 167–182. Blackwell.
70. Currie, Gregory, and Jon Jureidini. 2001. ‘Delusion, rationality, empathy.’ *Philosophy, Psychiatry & Psychology* 8: 159–162.
71. Currie, Gregory, and Ian Ravenscroft. 2002. *Recreative Minds*. Oxford: Oxford University Press.
72. Dalgleish, Tim. 1997. ‘Once more with feeling: The role of emotion in self-deception.’ *Behavioral and Brain Sciences* 20: 110–111.
73. Darwall, Stephen. ‘Self-Deception, Autonomy, and Moral Constitution.’ In *Perspectives on Self-Deception*, eds. B. McLaughlin and A. Rorty, 407–460. Berkeley: University of California Press.

74. Davidson, Donald. 1963. ‘Actions, Reasons, and Causes.’ *The Journal of Philosophy* 60: 685–700.
75. ———. 2001 [1970]. ‘How is Weakness of the Will Possible?’ In *Essays on Actions and Events*, 25–42. Oxford: Oxford University Press.
76. ———. 2001 [1978]. ‘Intending.’ In *Essays on Actions and Events*, 75–91. Oxford: Oxford University Press.
77. ———. 2004 [1982]. ‘Paradoxes of Irrationality.’ In his *Problems of Rationality*, 169–188. Oxford: Clarendon.
78. ———. 1985. ‘Incoherence and Irrationality.’ *Dialectica* 39: 345–354.
79. ———. 2004 [1986] ‘Deception and Division.’ In his *Problems of Rationality*, 199–212. Oxford: Oxford University Press.
80. ———. 2004 [1997]. ‘Who is Fooled.’ In his *Problems of Rationality*, 213–230. Oxford: Oxford University Press.
81. ———. 1999. ‘Reply to John Elster.’ In *The Philosophy of Donald Davidson*, ed. Lewis Hahn E., 443–446. Chicago: Open Court Publishing Company.
82. Davies, Martin and Max Coltheart. 2000. ‘Introduction: Pathologies of Belief.’ *Mind & Language* 15: 1–46.
83. Davies, Martin, Max Coltheart, Robyn Langdon, and Nora Breen. 2001. ‘Monothematic Delusions: Towards a Two-Factor Account.’ *Philosophy, Psychiatry, & Psychology* 8: 133–58.
84. Davies, Martin and Andy Egan. 2013. ‘Delusion, Cognitive Approaches: Bayesian Inference and Compartmentalisation.’ In *The Oxford Handbook of Philosophy and Psychiatry*, eds. K. W. M. Fulford et al., 689–727. Oxford: Oxford University Press.
85. Davis, Elysia Poggi, Nancy Snidman, Pathik D. Wadhwa, Laura M. Glynn, Chris Dunkel Schetter, Curt A. Sandman. 2004. ‘Prenatal Maternal Anxiety and Depression Predict Negative Behavioral Reactivity in Infancy.’ Infancy 6: 319–331.
86. ‘Deceive.’ In Oxford English Dictionary, (March 2017 update) URL = http://www.oed.com.ezproxy.auckland.ac.nz/view/Entry/48096?redirectedFrom=deceive#eid
87. Demos, Raphael. 1960. ‘Lying to Oneself.’ *The Journal of Philosophy* 57: 588–595.
88. Dennett, Daniel. 1996. *Kinds of Minds: Towards an Understanding* of *Consciousness*. London: Weidenfeld and Nicolson.
89. De Pauw, Karel W., and Krystyna T. Szulecka. 1988. ‘Dangerous Delusions: Violence and the Misidentification Syndromes.’ *British Journal of Psychiatry* 152: 91–96.
90. Derksen, Anthony A. 2001. ‘The Seven Strategies of the Sophisticated Pseudo-Scientist: a look into Freud’s rhetorical tool box.’ *Journal for General Philosophy of Science* 32: 329–350.
91. Derrida, Jacques. 2002. *Without Alibi*. Stanford: Stanford University Press.
92. DeRose, Keith. 2002. ‘Assertion, Knowledge, and Context.’ *The Philosophical Review* 111: 167–203.

93. Deweese-Boyd, Ian. 2016. 'Self-Deception.' *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, URL = https://plato.stanford.edu/entries/self-deception/
94. *Diagnostic and Statistical Manual of Mental Disorders, 5th ed.*, DSM-V, 160–1, 2013. Washington: American Psychiatric Association.
95. Dixson, Alan F., 2009. *Sexual Selection and the Origins of Human Mating Systems*. Oxford: Oxford University Press.
96. D'Cruz, Jason. 2015. 'Rationalisation, Evidence, and Pretense.' *Ratio* 28: 318–331.
97. Doggett, Tyler. 2012. 'Some Questions for Tamar Szabo Gendler.' *Analysis* 72: 764–774.
98. Douven, Igor. 2006. 'Assertion, Knowledge, and Rational Credibility.' *The Philosophical Review* 115: 449–85.
99. ———. 2009. 'Assertion, Moore, and Bayes.' *Philosophical Studies* 144: 361–375.
100. Dynel, Marta. 2015. 'Intention to deceive, bald-faced lies, and deceptive implicature: Insights into Lying at the semantics-pragmatics interface.' *Intercultural Pragmatics* 12: 309–332.
101. Egan, Andy. 2008. 'Seeing and Believing: Perception, Belief Formation and the Divided Mind.' *Philosophical Studies* 140: 47–63.
102. ———. 2009. 'Imagination, delusion, and self-deception.' In eds. Tim Bayne and Jordi Fernández. *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*, 263–280. New York: Psychology Press.
103. Egan, Andy and Adam Elga. 2005. 'I Can't Believe I'm Stupid.' *Philosophical Perspectives* 19: 77–93.
104. Ellis, Hadyn D., Andrew W. Young, Angela H. Quayle, and Karel W. De Pauw. 1997. 'Reduced autonomic responses to faces in Capgras delusion.' *Proceedings of the Royal Society B: Biological Sciences* 264, 1384: 1085–1092.
105. Ellis, Hadyn D., Michael B. Lewis, Hamdy F. Moselhy, and Andrew W. Young. 2000. 'Automatic without autonomic responses to familiar faces: Differential components of covert face recognition in a case of Capgras delusion.' *Cognitive Neuropsychiatry* 5: 255–69.
106. Ellis, Hadyn D., and Michael B. Lewis. 2001. 'Capgras delusion: A window on face recognition.' *Trends in Cognitive Sciences* 5: 149–156.
107. Epley, N., and Whitchurch, E. 2008 'Mirror, mirror on the wall: Enhancement in self recognition.' *Personality and Social Psychology Bulletin* 34:1159–1170.
108. Erat, Sanjiv and Uri Gneezy. 2012. 'White Lies.' *Management Science* 58: 732–733.
109. Eriksson, Lina and Alan Hájek. 2007. 'What are degrees of belief?' *Studia Logica* 86: 183–213.
110. Evans, Gareth. 1982. *The Varieties of Reference*. New York: Oxford University Press.
111. Evans, Jonathan St B. T. 2003. 'In two minds: dual-process accounts of reasoning.' *Trends in Cognitive Sciences* 7: 454–459.
112. Fallis, Don. 2009. 'What is lying?' *The Journal of Philosophy* 61: 29–56.
113. ———. 2010. 'Lying and Deception.' *Philosophers' Imprint* 10: 1–22.

114. ———. 2011. ‘What Liars Can Tell Us About the Knowledge Norm of Practical Reasoning.’ *Southern Journal of Philosophy* 49: 347–367.
115. ———. 2012. ‘Lying as a Violation of Grice’s First Maxim of Quality.’ *Dialectica* 66: 563–81.
116. ———. 2013. ‘Davidson was Almost Right about Lying.’ *Australasian Journal of Philosophy* 91: 337–53.
117. ———. 2014. ‘Are Bald-Faced Lies Deceptive After All?’ *Ratio* 28: 81–96.
118. ———. 2015. ‘Skyrms on the possibility of universal deception.’ *Philosophical Studies* 172: 375–397.
119. ———. Forthcoming. ‘What is Deceptive Lying?’ In eds. A. Stokke and E. Michaelson. *Lying: Language, Knowledge, Ethics, Politics*. Oxford: Oxford University Press.
120. Fallis, Don, and Peter J. Lewis. 2017. ‘Toward a formal analysis of deceptive signalling.’ *Synthese* (Online): 1–25.
121. Faulkner, Paul. 2007. ‘What Is Wrong With Lying?’ *Philosophy and Phenomenological Research* 75: 535–557.
122. ———. 2013. ‘Lying and Deceit.’ In *The International Encyclopedia of Ethics*, ed. H. LaFollette, 3101–09. Oxford, Wiley-Blackwell.
123. Feldman, Harriet, and Karl J. Friston. 2010. ‘Attention, uncertainty and free-energy.’ *Frontiers in Human Neuroscience* 4: 215.
124. Fernández, Jordi. 2013. ‘Self-deception and self-knowledge.’ *Philosophical Studies* 162: 379–400.
125. Fingarette, Herbert. 1989. ‘Self-Deception Needs No Explaining.’ *The Philosophical Quarterly* 48: 289–301.
126. Fine, Cordelia, Mark Gardner, Jillian Craigie, and Ian Gold. 2007. ‘Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions.’ *Cognitive Neuropsychiatry* 12: 46–77.
127. Fineberg, Sarah K. and Philip R. Corlett. 2016. ‘The doxastic shear pin: delusions as errors of learning and memory.’ *Cognitive Neuropsychiatry* 21: 73–89.
128. Fletcher, Paul C. and Chris D. Frith. 2009. ‘Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia.’ *Nature Reviews Neuroscience* 10: 48–58.
129. Forbes, Frances Alice M. 2009. *St. Athanasius: The Father of Orthodoxy*. EBook #27707: Project Gutenberg.
130. Foss, Jeffrey. 1980. ‘Rethinking Self-Deception.’ *American Philosophical Quarterly* 17: 237–243.
131. ———. 1997. ‘How Many Beliefs Can Dance in the Head of the Self-Deceived?’ *Behavioral and Brain Sciences* 20: 111–112.
132. Frankfurt, Harry. 2005. *On Bullshit*. Princeton: Princeton University Press.

133. ———. ‘On Truth, Lies, and Bullshit.’ In ed. C. Martin. The philosophy of deception, 37–48. Oxford: Oxford University Press.
134. Freud, Sigmund. 1929. ‘Introductory Lectures on Psycho-analysis [1916–1917].’ transl. Joan Riviere. Unwin Brothers.
135. Frey, Ulrich and Voland Eckart. 2011. ‘The evolutionary route to self-deception: Why offensive versus defensive strategy might be a false alternative.’ *Behavioral and Brain Sciences* 34: 21–22.
136. Friederich, James. 1993. ‘Primary Error Detection and Minimization PEDMIN Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena.’ *Psychological Review* 100: 298–319.
137. Friston, Karl. J. 2002a. ‘Beyond phrenology: What can neuroimaging tell us about distributed circuitry?’ *Annual Review of Neuroscience* 25: 221–250.
138. ———. 2002b. ‘Functional integration and inference in the brain.’ *Progress in Neurobiology* 68: 113–43.
139. ———. 2005. ‘A theory of cortical responses.’ *Philosophical Transactions of the Royal Society B: Biological Sciences* 360: 815–36.
140. ———. 2009. ‘The free-energy principle: A rough guide to the brain?’ *Trends in Cognitive Sciences* 13: 293–301.
141. Friston, Karl J., and Klaas E. Stephan. 2007. ‘Free-energy and the brain.’ *Synthese* 159: 417–458.
142. Förstl, Hans, Osvaldo P. Almeida, Adrian M. Owen, Alistair Burns, and Robert Howard. 1991. ‘Psychiatric, neurological and medical aspects of misidentification syndromes: a review of 260 cases.’ *Psychological Medicine* 21: 905–910.
143. Fuller, Gary. 1976. ‘Other-Deception.’ *The Southwestern Journal of Philosophy* 7: 21–31.
144. Funder, David C. 2011. ‘Directions and beliefs of self-presentational bias.’ *Behavioral and Brain Sciences* 34: 23.
145. Funkhouser, Eric. 2005. ‘Do the Self-Deceived Get What They Want?’ *Pacific Philosophical Quarterly* 86: 295–312.
146. ———. 2009. ‘Self-Deception and the Limits of Folk Psychology.’ *Social Theory and Practice* 35: 1–13.
147. ———. 2012. ‘Practical Self-Deception.’ *Humana.Mente: Journal of Philosophical Studies* 20: 86–97.
148. Gal, David and Derek D. Rucker. 2010. ‘When in doubt, shout! Paradoxical influences of doubt on proselytizing.’ *Psychological Science* 21: 1701–1707.
149. Galeotti, Anna Elisabetta. 2012. ‘Self-Deception: Intentional Plan or Mental Event?’ *Humana.Mente: Journal of Philosophical Studies* 20: 41–66.
150. Gardiner, Patrick. 1970. ‘Error, faith, and self-deception.’ *Proceedings of the Aristotelian Society* 70: 221–243.

151. Gardner, Sebastian. 1993. *Irrationality and the Philosophy of Psychoanalysis*. Cambridge/New York: Cambridge University Press.
152. Garety, A. Philippa, D. R. Hemsley, and M. R. C. Wessely. 1991 'Reasoning in Deluded Schizophrenic and Paranoid Patients: Biases in Performance on a Probabilistic Inference Task.' *The Journal of Nervous and Mental Disease* 179: 194–201.
153. Garety, A. Philippa and Daniel Freeman. 1999. 'Cognitive approaches to delusions: A critical review of theories and evidence.' *British Journal of Clinical Psychology* 38: 113–54.
154. Garety, Philippa A., Daniel Freeman, Suzanne Jolley, Graham Dunn, Paul E. Bebbington, David G. Fowler, Elizabeth Kuipers, and Robert Dudley. 2005. 'Reasoning, emotions, and delusional conviction in psychosis.' *Journal of Abnormal Psychology* 114: 373–384.
155. Gazzaniga, Michael S. 1967. 'The Split Brain in Man.' *Scientific American* 217: 24–29.
156. ———. 1970. *The Bisected Brain*. New York: Appleton-Century-Crofts.
157. ———. 1998. The Mind's Past. Berkeley: University of California Press.
158. Gendler, Tamar Szabó. 2007. 'Self-deception as Pretense.' *Philosophical Perspectives* 21: 231–58.
159. ———. 2008. 'Alief in Action (and Reaction).' *Mind & Language* 23: 552–585.
160. ———. 2010. *Intuition, Imagination, and Philosophical Methodology*. Oxford: Oxford University Press.
161. ———. 2012. 'Between Reason and Reflex: Response to Commentators.' *Analysis* 72: 799–811.
162. Gergen, Kenneth J. 1997. 'Detecting deception.' *Behavioral and Brain Sciences* 20: 114–115.
163. ———. 1985. 'The Ethnopsychology of SD.' In ed. M.W. Martin. *Self-Deception and Self-Understanding*, 228–243. Lawrence: University Press of Kansas.
164. Gerrans, Philip. 2002. 'A One-Stage Explanation of the Cotard Delusion.' *Philosophy, Psychiatry, & Psychology* 9: 47–53.
165. ———. 2014. *The Measure of Madness: Philosophy of Mind, Cognitive Neuroscience, and Delusional Thought*. Cambridge US: The MIT Press.
166. Gibbins, Keith. 1997. 'Partial belief as a solution to the logical problem of holding simultaneous, contrary beliefs in self-deception research.' *Behavioral and Brain Sciences* 20: 115–116.
167. Gilbert, Daniel T., Patrick S. Malone, Douglas S. Krull. 1990. 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information.' J*ournal of Personality and Social Psychology* 59: 601–613.
168. Gilbert, Daniel T., Romin W. Tafadori, and Patrick S. Malone. 1993. 'You can't not believe everything you read.' *Journal of Personality and Social Psychology* 65: 221–233.
169. Gilboa, Asaf, Claude Alain, Donald T. Stuss, Brenda Melo, Sarah Miller, and Morris Moscovitch. 2006. 'Mechanisms of spontaneous confabulations: a strategic retrieval account.' *Brain* 129: 1399–1414.

170. Gilboa, Asaf. 2010. ‘Strategic retrieval, confabulations, and delusions: Theory and data.’ *Cognitive Neuropsychiatry* 15: 145–180.
171. Golomb, Jacob. *In Search of Authenticity*. New York: Routledge.
172. Gregory, Richard L. 1997. ‘Knowledge in perception and illusion.’ *Philosophical Transactions of the Royal Society B: Biological Sciences* 552: 1121–1127.
173. Greve, Werner and Dirk Wentura. 2010. ‘True lies: Self-stabilization without self-deception.’ *Consciousness and Cognition* 19: 721–730.
174. Grice, Paul. 1989. *Studies in the Way of Words*. Cambridge: Harvard University Press.
175. Gur, Ruben C. and Harold A. Sackeim 1979. ‘Self-Deception: A Concept in Search of a Phenomenon.’ *Journal of Personality and Social Psychology* 37: 147–169.
176. Haight, Mary Rowland. 1980. *A Study on Self-Deception*. Sussex: Harvester Press.
177. ———. 1985. ‘Tales from a Black Box.’ In ed. M.W. Martin, *Self-Deception and Self-Understanding*, 244–260. Lawrence: University Press of Kansas.
178. Hales, Steven D. 1994. ‘Self-deception and Belief Attribution.’ *Synthese* 101: 273–289.
179. Hamilton, Andy. 2000. ‘The Authority of Avowals and the Concept of Belief.’ *European Journal of Philosophy* 8: 20–39.
180. Hauser, M. D. 1997. Minding the behaviour of deception. In eds. A.Whiten and R.W. Byrne, *Machiavellian intelligence* II, 112–143. Cambridge: Cambridge University Press.
181. Hawthorne, John, Daniel Rothschild, and Levi Spectre. 2016. ‘Belief is weak.’ *Philosophical Studies* 173: 1393–404.
182. Heil, John. 1989. ‘Minds Divided.’ *Mind* 98: 571–583.
183. Heil, John. 1993. ‘Going to Pieces.’ In ed. George Graham. *Philosophical Psychopathology*, 111–133. Cambridge: MIT Press.
184. Henden, Edmund. 2004. ‘Weakness of Will and Divisions of the Mind.’ *European Journal of Philosophy* 12: 199–213.
185. Hermanowicz, Neal. 2002. ‘A Blind Man With Parkinson’s Disease, Visual Hallucinations, and Capgras Syndrome.’ *The Journal of Neuropsychiatry and Clinical Neurosciences* 14: 462–463.
186. Hinchman, Edward S. 2013. ‘Assertion, Sincerity, and Knowledge.’ *Noûs* 47: 613–646.
187. Hirstein, William, and Vilayanur, Subramanian Ramachandran. 1997. ‘Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons.’ *Proceedings of the Royal Society of London (B): Biological Science* 264: 437–44.
188. Hohwy, Jakob. 2007. ‘Functional Integration and the Mind.’ *Synthese* 159, 3, Functional Integration and the Mind: 315–328.
189. ———. 2013. ‘Delusions, Illusions and Inference under Uncertainty.’ *Mind and Language* 28: 57–71.
190. Hohwy, Jakob and Vivek Rajan. 2012. ‘Delusions as Forensically Disturbing Perceptual Inferences.’ *Neuroethics* 5: 5–11.

191. Holton, Richard. 2001. ‘What is the Role of Self in Self-deception.’ *Proceedings of the Aristotelian Society* 101: 53–69.
192. Howard, Richard C., Eve Hepburn, and Najat Khalifa. 2015. ‘Is delusional ideation a critical link in the nexus between personality disorder and violent offending?’ *The Journal of Forensic Psychiatry & Psychology* 26: 368–382.
193. Huq, S. F., P. A. Garety, and D. R. Hemsley. 1988. ‘Probabilistic judgements in deluded and non-deluded subjects.’ *The Quarterly Journal of Experimental Psychology Section A* 40: 801–12.
194. Ichikawa, Jonathan Jenkins, and Steup Matthias. 2012. ‘The Analysis of Knowledge.. In *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), ed. E. N. Zalta, URL = https://plato.stanford.edu/archives/spr2016/entries/knowledge-analysis/.
195. John, Oliver P. and Richard W. Robins. 1994. ‘Accuracy and bias in self-perception: Individual differences in self-enhancement and narcissism.’ *Journal of Personality and Social Psychology* 66: 206–219.
196. Johnson, Samuel. 1983/1755. *A Dictionary of the English Language*. London: Times Books.
197. Jones, Edward E., and Steven C. Berglas. 1978. ‘Control of attributions about the self through self-handicapping strategies: The appeal of alcohol and the role of underachievement.’ *Personality and Social Psychology Bulletin* 4: 200–206.
198. Johnston, Mark. 1988. ‘Self-Deception and the Nature of Mind.’ In *Perspectives on Self-Deception*, eds. B. McLaughlin and A. Rorty, 63–91. Berkeley: University of California Press.
199. Joseph, A. B. 1986. ‘Focal central nervous system abnormalities in patients with misidentification syndromes.’ *Bibliotheca Psychiatrica* 164: 68–69.
200. Kant, Immanuel. 1999 [1797]. ‘On a supposed right to lie from philanthropy.’ In eds. Paul Guyer and Allen W. Wood, *Immanuel Kant: Practical philosophy*, 611–615. Cambridge: Cambridge University Press.
201. ———. 1996 [1786]. *Groundwork of the Metaphysics of Morals (A German–English Edition)*. Cambridge, Cambridge University Press.
202. Keiser, Jessica. 2016. ‘Bald-faced lies: how to make a move in a language game without making a move in a conversation.’ *Philosophical Studies* 173: 461–477.
203. Kenny, David A. and Kashy, Deborah A. 1994. ‘Enhanced co-orientation in the perception of friends: A social relations analysis.’ *Journal of Personality and Social Psychology* 67: 1024–1033.
204. Kenyon, Tim. 2003. ‘Convention, pragmatics, and saying “uncle”.’ *American Philosophical Quarterly* 40: 241–248.
205. ———. 2010. ‘Assertion and capitulation.’ *Pacific Philosophical Quarterly* 91: 352–368.
206. Kingsbury, Justine and Jonathan McKeown-Green. 2009. ‘Definitions: Does Disjunction Mean Dysfunction?’ *The Journal of Philosophy* 106: 568–585.

207. Kinsbourne, Marcel. 1989. ‘A model of adaptive behavior related to cerebral participation in emotional control.’ In eds. G. Gainnotti and C. Caltagrione, *Emotions and the Dual Brain*, 48–260. Heidelberg: Springer Verlag.
208. Koralus, Philipp and Salvador Mascarenhas. 2013. ‘The erotetic theory of reasoning: Bridges between formal semantics and the psychology of deductive inference.’ *Philosophical Perspectives* 27: 312–365.
209. Krstić, Vladimir. 2017. ‘Knowledge-lies re-examined.’ *Ratio* 00: 1–9.
210. Kurzban, Robert. 2011. ‘Two problems with “self-deception”: No “self” and no “deception”.’ *Behavioral and Brain Sciences* 34: 32–33.
211. La Caze, Marguerite. 2016. ‘Pretending Peace: Provisional political trust and sincerity in Kant and Améry.’ In eds. Sorin Baiasu and Sylvie Loriaux. *Sincerity in Politics and International Relations*. London: Routledge.
212. Lackey, Jennifer. 2007. ‘Norms of Assertion.’ *Noûs* 41: 594–626.
213. ———. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford, Oxford University Press.
214. ———. 2011. ‘Assertion and Isolated Second-Hand Knowledge.’ In eds. J. Brown and C. Herman, *Assertion: New Philosophical Essays*, 251–75. Oxford: Oxford University Press.
215. ———. 2013. ‘Lies and Deception: An Unhappy Divorce.’ *Analysis* 73: 236–248.
216. Langdon, Robyn and Philip B. Ward. 2009. ‘Taking the Perspective of the Other Contributes to Awareness of Illness in Schizophrenia.’ *Schizophrenia Bulletin* 35: 1003–1011.
217. Langdon, Robyn, Philip B. Ward, and Max Coltheart. 2010. ‘Reasoning anomalies associated with delusions in schizophrenia.’ *Schizophrenia Bulletin* 36: 321–330.
218. Langer, Ellen J. and Judith Rodin. 1976. ‘The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting.’ *Journal of Personality and Social Psychology* 34: 191–198.
219. Lauria, Federico, Delphine Preissmann, and Fabrice Clément. 2016. ‘Self-deception as affective coping. An empirical perspective on philosophical issues.’ *Consciousness and Cognition* 41: 119–134.
220. Lazar, Ariela. 1999. ‘Deceiving oneself or self-deceived? On the formation of beliefs “under the influence”’ *Mind* 108: 263–290.
221. LeDoux, Joseph E., Donald H. Wilson, and Michael S. Gazzaniga. 1977. ‘A Divided Mind: Observations on the Conscious Properties of the Separated Hemispheres.’ *Annals of Neurology* 2: 417–421.
222. Lee, Kang, and Hollie J. Ross. 1997. ‘The concept of lying in adolescents and young adults: Testing Sweetser’s folkloristic model.’ *Merrill–Palmer Quarterly* 43: 255–270.
223. Lehrer, Keith. 1983 ‘Belief, Acceptance and Cognition.’ In ed. H. Parret, *On Believing*, 172–183. Berlin: Walter de Gruyter.

224. ———. 1989. ‘Knowledge Reconsidered,’ In eds. M. Clay and K. Lehrer, *Knowledge and Skepticism*, 131–154. Boulder: Westview.
225. Leland, Patrick R. 2015. ‘Rational responsibility and the assertoric character of bald-faced lies.’ *Analysis* 75: 550–554.
226. Levy, Neil. 2009. ‘Self-Deception Without Thought Experiments.’ In eds. T. Bayne and Fernández J., *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*, 227–42. New York: Psychology Press.
227. ———. 2016. ‘Have I Turned the Stove Off? Explaining Everyday Anxiety.’ *Philosophers’ Imprint* 16: 1–10.
228. Lewis, David. 1975. ‘Languages and Language.’ In his *Philosophical Papers Volume I*, 163–188. Oxford: Oxford University Press.
229. Lindskold, Svenn, and Gyuseog Han. 1986. ‘Intent and the judgment of lies.’ *The Journal of Social Psychology* 126: 129–130.
230. Locke, John. 1690/2001. *An Essay Concerning Human Understanding, Book IV*. Kitchener.
231. Lockie, Robert. 2003. ‘Depth psychology and self-deception.’ *Philosophical Psychology* 16: 127–148.
232. Lorenz, Hendrik. 2006. *The Brute Within: Appetitive Desire in Plato and Aristotle*. Oxford: Oxford University Press.
233. Lynch, Kevin. 2010. ‘Self-Deception, Religious Belief, and False Belief Condition.’ *The Heythrop Journal* 51: 1073–1074.
234. ———. 2012. ‘On the “tension” inherent in self-deception.’ *Philosophical Psychology* 25: 433–450.
235. ———. 2016. ‘Willful Ignorance and Self-Deception.’ *Philosophical Studies* 173: 505–523.
236. Lynch, Michael P. 2009. ‘Deception and the nature of truth.’ In ed. C. Martin. *The philosophy of deception*, 188–200. Oxford: Oxford University Press.
237. Maher, Brendan A. 1974. ‘Delusional thinking and perceptual disorder.’ *Journal of Individual Psychology* 30: 98–113.
238. ———. 1999. ‘Anomalous experience in everyday life: Its significance for psychopathology.’ *The Monist* 82: 547–570.
239. Mahon, James Edwin. 2007. ‘A definition of deceiving.’ *International Journal of Applied Philosophy* 21: 181–94.
240. ———. 2008. ‘The Definition of Lying and Deception.’ *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2016/entries/lying-definition.
241. ———. 2011. Review of Lying and Deception by T. L. Carson, Notre Dame Philosophical Reviews, URL = http://ndpr.nd.edu/news/24572-lying-and-deception-theory-and-practice.

242. ———. ‘Why Men of Action Don’t Lie.’ 2016. In *The Princess Bride and Philosophy: Inconceivable!* Eds. Rachel Robison-Greene and Richard Greene, 13–26, Chicago: Open Court.
243. Margolis, Joseph. 1973. *Knowledge and existence; an introduction to philosophical problems*. New York: Oxford University Press.
244. Martin, Mike W. 1997. ‘Self-Deceiving Intentions.’ *Behavioral and Brain Sciences* 20: 122–123.
245. Matthen, Mohan. 2010. ‘Two visual systems and the feeling of presence.’ In eds. Nivedita Gangopadhyay, Michael Madary, and Finn Spicer, *Perception, Action, and Consciousness: Sensorimotor Dynamics and Two Visual Systems*, 107–23. Oxford: Oxford University Press.
246. Macintyre, Alasdair. 2014. ‘Truthfulness, Lies, and Moral Philosophers: What Can We Learn from Mill and Kant?’ In *Tanner Lectures on Human Values 16*, 309–361 (delivered Princeton University, 1994) URL = http://tannerlectures.utah.edu/_documents/a-to-z/m/macintyre_1994.pdf
247. McKay, Ryan, and Lisa Cipolotti. 2007. ‘Attributional style in a case of Cotard delusion.’ *Consciousness and Cognition* 16: 349–359.
248. McKay, Ryan, and Daniel C. Dennett. 2009. ‘The evolution of misbelief.’ *The Behavioral and Brain Sciences* 32: 493–510.
249. McKay, Ryan. 2012. ‘Delusional Inference.’ *Mind and Language* 27: 330–355.
250. McKinnon, Rachel. 2015. *The Norms of Assertion: Truth, Lies, and Warrant*. Palgrave McMillan.
251. McLaughlin, Brian P. 1988. ‘Exploring the Possibility of Self-deception in Belief.’ In eds. Brian McLaughlin and Amelie Rorty. *Perspectives on Self-Deception*, 29–62. Berkeley: University of California Press.
252. McWhirter, G. 2016. ‘Behavioural deception and formal models of communication.’ *British Journal for the Philosophy of Science*. 67: 757–780.
253. Meibauer, Jörg. 2011. ‘On lying: intentionality, implicature, and imprecision.’ *Intercultural Pragmatics* 8: 277–292.
254. ———. 2014a. ‘Bald-Faced Lies as Acts of Verbal Aggression.’ *Journal of Language Aggression and Conflict* 2: 127–150.
255. ———. 2014b. *Lying and the semantics-pragmatics interface*. Berlin: Mouton de Gruyter.
256. ———. 2016a. ‘Topics in the linguistics of lying: A reply to Marta Dynel.’ *Intercultural Pragmatics* 13: 107–123.
257. ———. 2016b. ‘Understanding bald-faced lies. An experimental approach.’ *International Review of Pragmatics, Special Issue: Empirical Approaches to Lying and Deception* 8: 247–270.
258. Mele, Alfred. R. 1987. *Irrationality - An Essay on Akrasia, Self-deception, and Self-control*. Oxford: Oxford University Press.
259. ———. 1997a. ‘Introduction.’ In ed. Alfred R. Mele. *The Philosophy of Action*, 1–26. Oxford: Oxford University Press.

260. ———. 1997b. ‘Real Self-deception.’ *Behavioral and Brain Sciences* 20: 91–136.
261. ———. 1999. ‘Twisted self-deception.’ *Philosophical Psychology*, 12: 117–137
262. ———. 2001. *Self-Deception Unmasked*. Princeton: Princeton University Press.
263. ———. 2006. ‘Self-deception and Delusions.’ *European Journal of Analytic Philosophy* 2: 109–24.
264. ———. 2009. ‘Self-deception and Delusions.’ In eds. Tim Bayne and Jordi Fernández. *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation*, 55–70. New York: Psychology Press.
265. ———. 2010a. ‘Intention.’ In eds. T. O’Connor and C. Sandis. *A companion to the philosophy of action*, 108–113. Oxford: Wiley-Blackwell.
266. ———. 2010b. ‘Approaching self-deception: How Robert Audi and I part company.’ *Consciousness and Cognition* 19: 745–50.
267. Michel, Christoph and Newen Albert. 2010. ‘Self-deception as pseudo-rational regulation of belief.’ *Consciousness and Cognition* 19: 731–744.
268. Mijović-Prelec, Danica and Prelec Dražen. 2010. ‘Self-Deception as Self-Signalling: A Model and Experimental Evidence.’ *Philosophical Transactions: Biological Sciences* 365, 1538, Rationality and Emotions: 227–240.
269. Miyazono, Kengo. 2015. ‘Delusions as harmful malfunctioning beliefs.’ *Consciousness and Cognition* 33: 561–573.
270. Mishara, Aaron L., and Phil Corlett. 2009. ‘Are delusions biologically adaptive?: Salvaging the doxastic shear pin.’ *Behavioral and Brain Sciences* 32: 530–531.
271. Moore, George E. 1962. *Commonplace Book: 1919–1953*. London, Allen and Unwin.
272. ———. 1993. ‘Moore’s Paradox.’ In ed. Thomas Baldwin, *G. E. Moore: Selected Writings*, 207–12. London/New York: Routledge.
273. Myers-Schulz, Blake, and Eric Schwitzgebel. 2013. ‘Knowing that *P* without Believing that *P*.’ *Noûs* 47: 371–384.
274. Neely, James H. 1977. ‘Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention.’ *Journal of Experimental Psychology*: General 106: 226–254.
275. Nelkin, Dana. 2002. ‘Self-Deception, Motivation and the Desire to Believe.’ *Pacific Philosophical Quarterly* 83: 384–406.
276. Nelson, H. 2005. *Cognitive Behavioural Therapy with Delusions and Hallucinations. A Practice Manual*. Cheltenham: Nelson Thomes.
277. Nichols, Shaun and Stephen Stich. 2000. ‘A cognitive theory of pretense.’ *Cognition* 74: 115–47.
278. Owens, David. 2006. ‘Testimony and Assertion.’ *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 131: 487–510.

279. Pagin, Peter. 2014. ‘Assertion.’ In *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), ed. E. N. Zalta, URL = https://plato.stanford.edu/archives/spr2016/entries/assertion/.
280. Parrott, Matthew, and Philipp Koralus. 2015. ‘The erotetic theory of delusional thinking.’ *Cognitive Neuropsychiatry* 20: 398–415.
281. Parrott, Matthew. 2016. ‘Bayesian Models, Delusional Beliefs, and Epistemic Possibilities.’ *The British Journal for the Philosophy of Science* 67: 271–296.
282. Pascal, Blaise. 2004. ‘Faith as a Rational wager.’ In Louis P Pojman. *Introduction to philosophy: classical and contemporary readings*, 233–35. New York: Oxford University Press.
283. Pataki, Tamas. 1997. ‘Self-deception and wish-fulfilment.’ *Philosophia* 25: 297–322.
284. Patten, David. 2003. ‘How do we deceive ourselves?’ *Philosophical Psychology* 16: 229–246.
285. Peacocke, Christopher. 2000. ‘Conscious Attitudes, Attention, and Self-Knowledge.’ In eds. C. Wright, B. C. Smith, and C. Macdonald, *Knowing Our Own Minds*, 63–98. Oxford, Oxford University Press.
286. Pears, David. 1982. ‘Motivated irrationality, Freudian theory and cognitive dissonance.’ In ed. Richard Wollheim, *Philosophical Essays on Freud*, 264–288. Cambridge: Cambridge University Press.
287. ———. 1986. ‘The Goals and Strategies of Self-Deception.’ In ed. J. Elster, *The Multiple Self*, 59–78. Cambridge: Cambridge University Press.
288. ———. 1991. ‘Self-Deceptive Belief-Formation.’ *Synthese* 89: 393–405.
289. Perring, Christian. 1997. ‘Direct, Fully Intentional Self-Deception is also Real.’ Behavioral and Brain Sciences 20: 123–124.
290. Peterson, Candida C. 1995. ‘The role of perceived intention to deceive in children’s and adults’ concepts of lying.’ *British Journal of Developmental Psychology* 13: 237–260.
291. Pettigrew, Richard. 2013. ‘Epistemic Utility and Norms for Credences.’ *Philosophy Compass* 8: 897–908.
292. Platt, Dwight R. 1969. *Natural history of the hognose snakes, Heterodon platyrhinos and Heterodon nasicus*. University of Kansas Publications, Museum of Natural History 18: 253–420.
293. Porcher, José Eduardo. 2012. ‘Against the Deflationary Account of Self-Deception.’ *Humana.Mente: Journal of Philosophical Studies* 20: 67–84.
294. Pratchett, Terry, Ian Stewart, and Jack Cohen. 1999. *The Science of Discworld*. London, Ebury Press.
295. Price, H. H. 1969. *Belief: The Gifford Lectures Delivered at the University of Aberdeen in 1960*. Muirhead Library: George Allen and Unwin.
296. Price, A.W. 2009. ‘Are Plato’s Soul-Parts Psychological Subjects?’ *Ancient Philosophy* 29: 1–15.
297. Pritchard, Duncan. 2005. *Epistemic Luck*. New York, Oxford.
298. ———. 2012. ‘Anti-Luck Virtue Epistemology.’ *The Journal of Philosophy* 109: 247–279.

299. Pruss, Alexander R. 2012. ‘Sincerely Asserting What You Do Not Believe.’ *Australasian Journal of Philosophy* 90: 541–546.
300. Quattrone, George A. and Tversky Amos. 1984. ‘Causal Versus Diagnostic Contingencies: On Self-Deception and on the Voter’s Illusion.’ *Journal of Personality and Social Psychology* 46: 237–248.
301. Radden, Jennifer. 2014. ‘Belief as Delusional and Delusion as Belief.’ *Philosophy, Psychiatry, & Psychology* 21: 43–46.
302. Radford, Colin. 1970. ‘Knowledge–By Examples.’ In eds. M. D. Roth and L. Galis, *Knowing: Essays in the Analysis of Knowledge*, 171–85. New York: Random House; alternatively 1966. *Analysis* 27: 1–11.
303. Raikka, Juha. 2007. ‘Self-Deception and Religious Beliefs.’ *The Heythrop Journal* 48: 513–526.
304. Ramachandran, Vilayanur Subramanian. 1995. ‘Anosognosia in parietal lobe syndrome.’ *Consciousness and* Cognition 4: 22–51.
305. ———. 1996. ‘The Evolutionary Biology of Self-Deception, Laughter, Dreaming and Depression: Some Clues from Anosognosia.’ *Medical Hypotheses* 47: 347–362.
306. Ramachandran, Vilayanur Subramanian, and Sandra Blakeslee. 1998. *Phantoms in the brain: Human nature and the architecture of the mind*. London: Fourth Estate.
307. Reid, I., A. W. Young, and D. J. Hellawell. 1993. ‘Voice recognition impairment in a blind Capgras patient.’ *Behavioural Neurology* 6: 225–228.
308. Rey, Georges. 1988. ‘Towards a Computational Account of *akrasia* and Self-deception.’ In *Perspectives on Self-Deception*, eds. B. McLaughlin and A. Rorty, 264–296. Berkeley: University of California Press.
309. Ridge, Michael. 2006. ‘Sincerity and expressivism.’ *Philosophical Studies* 131: 478–510.
310. Roberts, Craige. 2012. ‘Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics.’ *Semantics & Pragmatics* 49: 1–69.
311. Rojo, Vincente I., Luis Caballero, Luis M. Iruela, and Enrique Baca. 1991. ‘Capgras’ syndrome in a blind patient.’ *American Journal of Psychiatry* 148: 1271b–1272.
312. Rorty Oksenberg, Amélie. 1988. ‘The Deceptive Self: Liars, Layers and Lairs.’ In *Perspectives on Self-Deception*, eds. B. McLaughlin and A. Rorty, 11–28. Berkeley: University of California Press
313. Ryle, Gilbert. 1949. *The concept of mind*. New York, Barnes & Noble.
314. Sartre, Jean-Paul. 1978. ‘Bad Faith.’ In his *Being and Nothingness*, 147–186. Pocket Books.
315. ———. 2007. *Existentialism Is a Humanism*. transl. Carol Macomber. New Haven: Yale University Press.
316. Sartwell, Crispin. 1991. ‘Knowledge Is Merely True Belief.’ *American Philosophical Quarterly* 28: 157–165.
317. ———. 1992. ‘Why Knowledge Is Merely True Belief.’ *The Journal of Philosophy* 89: 167–180.

318. Sass, Louis. 1994. *The Paradoxes of Delusion: Wittgenstein, Schreber, and the Schizophrenic Mind*. New York: Cornell University Press.
319. Saul, Jennifer Mather. 2012. *Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and in Ethics*. Oxford Scholarship Online.
320. Scanlon, Thomas. 1998. *What We Owe to Each Other*. Cambridge: Harvard University Press.
321. Schoubye, Anders J., and Andreas Stokke. 2016. 'What is Said?' *Noûs* 50: 759–793.
322. Schwitzgebel, Eric. 2001. In-between Believing. *Philosophical Quarterly* 51: 76–82.
323. ———. 2002. 'A phenomenal, dispositional account of belief.' *Noûs* 36: 249–75.
324. ———. 2010. 'Acting Contrary to Our Professed Beliefs, or the Gulf Between Occurrent Judgment and Dispositional Belief.' *Pacific Philosophical Quarterly* 91: 531–553.
325. ———. 2012. 'Mad Belief?' *Neuroethics* 5: 13–7.
326. ———. 2015. 'Belief.' *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), ed. Edward N. Zalta, URL = http://plato.stanford.edu/archives/sum2015/entries/belief/
327. Scott-Kakures, Dion. 1996. 'Self-Deception and Internal Irrationality.' *Philosophy and Phenomenological Research* 56: 31–56.
328. ———. 2000. 'Motivated Believing: Wishful and Unwelcome.' *Noûs* 34: 348–375.
329. ———. 2001. 'High anxiety: Barnes on what moves the unwelcome believer.' *Philosophical Psychology* 14: 313–326.
330. ———. 2002. 'At Permanent Risk: Reasoning and Self-knowledge in Self-deception.' *Philosophy and Phenomenological Research* 65: 576–603.
331. ———. 2012. 'Can You Succeed in Intentionally Deceiving Yourself?' *Humana.Mente Journal of Philosophical Studies* 20: 17–39.
332. Searcy, William A. and Nowicki Stephen. 2005. *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton: Princeton University Press.
333. Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. New York: Cambridge University Press.
334. Shah, Nishi, and David J. Velleman. 2005. 'Doxastic Deliberation.' *The Philosophical Review* 114: 497–534.
335. Sharpsteen, Don J. and Kirkpatrick Lee A. 1997. 'Romantic Jealousy and Adult Romantic Attachment.' *Journal of Personality and Social Psychology* 72: 627–640.
336. Shea, Nicholas, Peter Godfrey-Smith and Rosa Cao. 2017. 'Content in simple signalling systems.' *British Journal for the Philosophy of Science* 0: 1–27.
337. Shiffrin, R.M. 1970. 'Memory search.' In *Models of Human Memory*, ed. D. A. Norman, 375–447. New York: Academic Press.
338. Shoemaker, Sydney. 1996. *The first-person perspective and other essays*. New York: Cambridge University Press.

339. Seierstad, Asne. 2003. *A Hundred and One Days: A Baghdad Journal*, trans. Ingrid Christophersen. New York: Basic Books.
340. Silva, Arturo J., Gregory B. Leong, Robert Weinstock, and Catherine L. Boyer. 1989. 'Capgras Syndrome and Dangerousness.' *Bulletin of the American Academy of Psychiatry* 17: 5–14.
341. Skyrms, Brian. 2010. *Signals*. New York: Oxford University Press.
342. Smith, S. M., Glenberg, A. M. and Bjork, R. A. 1978. 'Environmental context and human memory.' *Memory and Cognition* 6: 342–353.
343. Smith, S. M. 1994. 'Theoretical Principles of Context-Dependent Memory.' In *Theoretical Aspects of Memory, 2nd ed*. eds. Peter Morris and Michael Gruneberg, 167–194. London: Taylor & Francis.
344. Snyder, Charles R., and Raymond L. Higgins. 1988. 'Excuses: Their role in the negotiation of reality.' Psychological Bulletin 104: 23–35.
345. Snyder, Charles R., Lori M. Irving, Sandra T. Sigmon, and Sharon Holleran. 1992. 'Reality negotiation and valence/linkage of self-theories: Psychic showdown at the "I'm OK" corral and beyond.' In *Life crises and experience of loss in adulthood*, eds. L. Montada, S.-H. Filipp, and M. J. Lerner, 275–297. Hillsdale, NJ: Erlbaum.
346. So, Suzanne H., Daniel Freeman, Graham Dunn, Shitij Kapur, Elizabeth Kuipers, Paul Bebbington, David Fowler, and Philippa A. Garety. 2012. 'Jumping to conclusions, a lack of belief flexibility and delusional conviction in psychosis: A longitudinal investigation of the structure, frequency, and relatedness of reasoning biases.' *Journal of Abnormal Psychology* 121: 129–130.
347. Solomon, Robert C. 2009. 'Self, Deception, and Self-Deception in Philosophy.' In ed. Martin Clancy, *The Philosophy of Deception*, 15–36. Oxford: Oxford University Press.
348. Sorensen, Roy. 1984. 'Self-Deception and Scattered Events.' *Mind* 94: 64–69.
349. ———. 2007. 'Bald-Faced Lies! Lying without the Intent to Deceive.' *Pacific Philosophical Quarterly* 88: 251–64.
350. ———. 2010. 'Knowledge-lies.' *Analysis* 70: 608–615.
351. ———. 2011. 'What Lies Behind Misspeaking.' *American Philosophical Quarterly* 48: 399–409.
352. Sosa, Ernest. 1999. 'How Must Knowledge Be Modally Related to What Is Known?' *Philosophical Topics* 26: 373–384.
353. ———. 2002. 'Tracking, Competence, and Knowledge.' In *The Oxford Handbook of Epistemology*. ed. P. Moser, 264–86. Oxford: Oxford University Press.
354. Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., and Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, 25, 359–393.

355. Sperry, R. W. 1961. ‘Cerebral Organization and Behavior: The split brain behaves in many respects like two separate brains, providing new research possibilities.’ *Science* 133, 3466: 1749–1757.
356. ———. 1986 ‘Hemisphere deconnection and unity in conscious awareness.’ *American Psychologist* 23: 723–33.
357. Staffel, Julia. 2011. ‘Reply to Roy Sorensen, “Knowledge-lies”.’ *Analysis* 71: 300–302.
358. ———. forthcoming. ‘Knowledge-lies and Group Lies.’ In ed. J. Meibauer, *The Oxford Handbook of Lying*.
359. Stalnaker, Robert. 2002. ‘Common Ground.’ *Linguistics and Philosophy* 25: 701–721.
360. Startup, Mike. 1997. ‘Awareness of own and others’ schizophrenic illness.’ *Schizophrenia Research* 26: 203–211.
361. Stewart, Jonathan T. 2004. ‘Capgras syndrome related to diazepam treatment.’ *Southern Medical Journal* 97: 65–66.
362. Stillman, Richard C., Herbert Weingartner, Richard Jed Wyatt, Christian J. Gillin, and James E. Eich. 1974. ‘State-dependent (dissociative) effects of marijuana on human memory.’ *Archives of General Psychiatry* 31: 81–85.
363. Stokke, Andreas. 2013. ‘Lying and asserting.’ *The Journal of Philosophy* 60: 33–60.
364. ———. 2014. ‘Insincerity.’ *Noûs* 48: 496–520.
365. ———. 2016. ‘Metaphors and Martinis: A Response to Jessica Keiser.’ *Philosophical Studies Online First*, 1–7.
366. ———. 2016. ‘Lying and Misleading in Discourse.’ *The Philosophical Review* 125: 83–134.
367. Stokke, Andreas, and Don Fallis. 2017. ‘Bullshitting, Lying, and Indifference toward Truth.’ *Ergo* 4: 277–309.
368. Stone, Tony, and Andrew W. Young. 1997. ‘Delusions and brain injury: The philosophy and psychology of belief.’ *Mind and Language* 12: 327–364.
369. Stroud, Sarah. 2013. ‘Irrationality.’ In *A Companion to Donald Davidson*. eds. Ernie Lepore and Kirk Ludwig, 489–505. West Sussex: Wiley-Blackwell.
370. Swann, William B. Jr. 1983. ‘Self-verification: Bringing social reality into harmony with the self.’ In *Social psychological perspectives on the self, vol. 2*. eds. J. Suls and A. G. Greenwald, 33–66. Erlbaum.
371. ———. 2012. ‘Self-verification theory.’ In *Handbook of theories of social psychology*, ed. Paul A. M. Van Lange, Arie W Kruglanski, and Edward Tory Higgins, ch. 27. Los Angeles: Sage.
372. Symons, Donald. 1979. *The Evolution of Human Sexuality*. Oxford: Oxford University Press.
373. Szabados, Béla. 1974. ‘Self-Deception.’ *Canadian Journal of Philosophy* 4: 51–68.
374. Tanney, Julia. 1995. ‘Why Reasons may Not be Causes.’ *Mind & Language* 10: 105–128.
375. Taylor, Shelley E. and Brown, Jonathon D. 1988. ‘Illusion and well-being: A social psychological perspective on mental health.’ *Psychological Bulletin* 103: 193—210.

376. Taylor, Collins Rebecca L., Skokan Laurie A., and Aspinwall Lisa G. 1989. 'Maintaining Positive Illusions in the Face of Negative Information: Getting the Facts Without Letting them Get to You.' *Journal of Social and Clinical Psychology* 8: 114–129.
377. Taylor, Marjorie, Gretchen L. Lussier, and Bayta L. Maring. 2003. 'The distinction between lying and pretending.' *Journal of Cognition and Development* 4: 299–323.
378. Tollefsen, C. O. 2014. *Lying and Christian Ethic*. Cambridge: Cambridge University Press.
379. Tooby, John and Cosmides Leda. 1992. 'The psychological foundations of culture.' In *The adapted mind: Evolutionary psychology and the generation of culture*. eds. Jerome H. Barkow, Leda Cosmides and John Tooby, 19–136. Oxford: Oxford University Press.
380. Tranel, Daniel, Don C. Fowles, and Antonio R. Damasio, 1985. 'Electrodermal Discrimination of Familiar and Unfamiliar Faces: A Methodology.' *Psychophysiology* 22: 403–408.
381. Tranel, Daniel, Hanna Damasio, and Antonio R. Damasio. 1995. 'Double dissociation between overt and covert recognition.' *Journal of Cognitive Neuroscience* 7: 425–432.
382. 'Trick.' In Oxford English Dictionary, (March 2017 update) URL = http://www.oed.com.ezproxy.auckland.ac.nz/view/Entry/205845?rskey=XO0Ngg&result=4#eid
383. Trope, Y., and Liberman, A. 1996. 'Social hypothesis testing: Cognitive and motivational mechanisms.' In *Social psychology: A handbook of basic principles*. eds. E. Higgins & E. Kruglanski, 239–270. New York: Guilford Press.
384. Truncellito, David. 2007. 'Epistemology.' In *Internet Encyclopedia of Philosophy*, eds. B. Dowden and J. Fieser, URL = http://www.iep.utm.edu/epistemo/.
385. Tumulty, Maura. 2012. 'Delusions and Not-Quite-Beliefs.' *Neuroethics* 5: 29–37.
386. Turri, John. 2010a. 'On the Relationship between Propositional and Doxastic Justification.' *Philosophy and Phenomenological Research* 80: 312–326.
387. ———. 2010b. 'Epistemic Invariantism and Speech Act Contextualism.' *The Philosophical Review* 119: 77–95.
388. ———. 2011. 'The Express Knowledge Account of Assertion.' *Australasian Journal of Philosophy* 89: 37–45.
389. ———. 2013. 'The test of truth: An experimental investigation of the norm of assertion.' *Cognition* 129: 279–291.
390. Turri, Angelo and John Turri. 2015. 'The truth about lying.' *Cognition* 138: 161–168.
391. Turner, Martha, and Max Coltheart. 2010. 'Confabulation and Delusion: A Common Monitoring Framework.' *Cognitive Neuropsychiatry* 15: 346–376.
392. Unger, Peter. 1975. *Ignorance: A Case for Scepticism*. Oxford: Oxford University Press.
393. Van der Hart, Onno, and Rutger Horst. 1989. 'The Dissociation Theory of Pierre Janet.' *Journal of Traumatic Stress* 2: 397–412.
394. Van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford: Oxford University Press.

395. Van Inwagen, Peter. 1998. ‘It Is Wrong, Everywhere, Always, and for Anyone, to Believe Anything upon Insufficient Evidence.’ In *The Possibility of Resurrection and Other Essays in Christian Apologetics*, 29–44. Boulder: Westview Press.
396. Van Leeuwen, Neil. 2007a. ‘The Product of Self-Deception.’ *Erkenntnis* 67: 419–437.
397. ———. 2013. ‘Self-Deception.’ In ed. Hugh LaFollette. *The International Encyclopedia of Ethics*, 4753–4762. Blackwell Publishing.
398. ———. 2014. ‘Religious credence is not factual belief.’ *Cognition* 133: 698–715.
399. Veale D. 2002. ‘Over-valued ideas: a conceptual analysis.’ *Behaviour Research and Therapy* 40: 383–400.
400. Velleman, David J. 2000. ‘On the aim of belief.’ In his *The Possibility of Practical Reason*, 244–281. Oxford: Oxford University Press.
401. Vincent, Jocelyne M. and Cristiano Castelfranchi. 1981. ‘On the Art of Deception: How to Lie while Saying the Truth.’ In eds. Herman Parret, Marina Sbisa, and Jef Verschueren. *Possibilities and Limitations of Pragmatics*, 749–777. Amsterdam: John Benjamins B. V.
402. Walsh, Clare R., and Johnson-Laird, P. N. 2004. ‘Co-Reference and Reasoning.’ *Memory and Cognition* 32: 96–106.
403. Waterman, Richard J. and Martin I. Bidartondo. 2008. ‘Deception above, deception below: linking pollination and mycorrhizal biology of orchids.’ *Journal of Experimental Botany* 59: 1085–1096.
404. Weiner, Matthew. 2005. ‘Must We Know What We Say?’ *The Philosophical Review* 114: 227–251.
405. Weinstein, Edwin A., Robert L. Kahn, Sidney Malitz, Jules Rozanski. 1954. ‘Delusional reduplication of parts of the body.’ *Brain: a journal of neurology* 77: 45–60.
406. Wentura, Dirk and Werner Greve. 2003. ‘Who want to be ... Erudite? Everyone! Evidence for automatic adaptation of trait definitions.’ *Social Cognition* 22: 30–53.
407. ———. 2005. ‘Assessing the structure of self-concept: Evidence for self-defensive processes by using a sentence priming task.’ Self and Identity 4: 193–211.
408. Wiegmann, Alex, Jana Samland, and Michael R. Waldmann. 2016. ‘Lying despite telling the truth.’ *Cognition* 150: 37–42.
409. Williams, Bernard. 2002. *Truth and Truthfulness*. Princeton: Princeton University Press.
410. Williams J. N. 1979. ‘Moore’s Paradox: One or Two?’ *Analysis* 39: 141–142.
411. Williamson, Timothy. 2000. *Knowledge and its Limits*. New York: Oxford University Press.
412. Wilson, Deirdre, and Dan Sperber. 2002. ‘Truthfulness and Relevance.’ *Mind* 111: 583–632.
413. ———. 2012. *Meaning and Relevance*. Cambridge: Cambridge University Press.
414. Wilson, Robbie S. and Michael J. Angilletta Jr. 2015. ‘Dishonest signaling during aggressive interactions: Theory and empirical evidence.’ In eds. Duncan J. Irschick, Mark Briffa, and

Jeffrey Podos, *Animal signaling and function: An integrative approach*, 205–227. Hoboken: Wiley.

415. Wood, Allen W. ‘Self-Deception and Bad Faith.’ In *Perspectives on Self-Deception*, eds. B. McLaughlin and A. Rorty, 207–227. Berkeley: University of California Press.
416. Van der Leer, Leslie and Ryan McKay. 2017. ‘The optimist within? Selective sampling and self-deception.’ *Consciousness and Cognition* 50: 23–29.
417. Van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford: Oxford University Press.
418. ———. 1988. ‘The Peculiar Effects of Love and Desire.’ In *Perspectives on Self-Deception*, eds. B. McLaughlin and A. Rorty, 124–156. Berkeley: University of California Press.
419. Van Leeuwen, Neil. 2007a. ‘The Product of Self-Deception.’ *Erkenntnis* 67: 419–437.
420. ———. 2007b. ‘The Spandrels of Self-Deception: Prospects for a Biological Theory of a Mental Phenomenon.’ *Philosophical Psychology* 20: 329–348
421. ———. 2013. ‘Self-Deception.’ In ed. Hugh LaFollette. *The International Encyclopedia of Ethics*, 4753–4762. Blackwell Publishing.
422. ———. 2014. ‘Religious credence is not factual belief.’ *Cognition* 133: 698–715.
423. Von Hippel, William and Robert Trivers. 2011. ‘The evolution and psychology of self-deception.’ *Behavioral and Brain Sciences* 34: 1–16.
424. Young, Andrew W. and Kate M. Leafhead. 1996. ‘Betwixt life and death: Case studies of the Cotard delusion.’ In eds. P. W. Halligan and J. C. Marshall, *Method in madness: Case studies in cognitive neuropsychiatry*, 147–171. Hove UK: Psychology Press.
425. Young, Andrew W. 1988. *Face and Mind.* Oxford: Oxford University Press.
426. ———. 2008. ‘Capgras delusion: An interactionist model.’ *Consciousness and Cognition* 17: 863–876.
427. Zislin, Joseph, Victor Kuperman, Rimona Durst. 2011. ‘”Ego-Dystonic” Delusions as a Predictor of Dangerous Behavior.’ *Psychiatric Quarterly* 82: 113–120.