Undeceiving Ourselves:
An Inquiry into Self-Undeception

Martina Orlandi

Department of Philosophy

McGill University, Montréal, Québec

October 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

TABLE OF CONTENTS

ABSTRACT

Decades of empirical research have shown something many people thought was obvious: we are often irrational. We rely on our gut feelings when we should instead turn to effortful deliberation, we believe against the evidence and refuse to admit it, we indulge in comfortable beliefs when we should instead face a difficult truth, and so on. Given this, it should not be surprising that phenomena such as wishful thinking or self-deception are so pervasive among otherwise rational agents.

Traditionally, the philosophical literature on irrationality has focused on trying to understand *why* we backslide in these ways. Yet sometimes we also seem to *know better*. We are able to overcome our biases, be self-controlled, and accept reality for what it is. It is possible to know better, and we sometimes do, so how can individuals come to do this? This is the central question that I examine in my dissertation. In particular, I provide an account of the underexplored phenomenon of self-*un*deception, that is, the phenomenon where self-deceived individuals *cease* to be self-deceived and reconcile with truth.

I begin by arguing that an account of self-undeception is dependent on what one takes self-deception to be and devote the first half of the dissertation to presenting and defending my preferred theory of self-deception which I call the *doxastic violation account*. According to the doxastic violation account, self-deception is a phenomenon where the self-deceived agent holds a belief that is at odds with the epistemic norms they currently hold. I defend the doxastic violation account by showing that, compared to its competing theories, it does a better job distinguishing self-deception from other similar biases; it better accounts for self-deception's irrationality; it rescues the role of the self in self-deception; and it better captures the experiential tension characteristic of self-deception.

With my preferred theory of self-deception in hand, I then devote the final part of the dissertation to advancing my account of self-undeception. I argue that a plausible account of self-

undeception should provide a metaphysical characterization of the phenomenon (i.e. what self-undeception is), a psychological characterization (i.e. how self-undeception occurs), and a normative characterization (i.e. a desirable path to self-undeception). This last criterion involves advancing an account of self-undeception that occurs out of self-reflection and that sets up the individual for what they see as an improved epistemic future. I argue that the doxastic violation account is the best candidate for self-undeception because, unlike competing theories, it is the only one that can provide a normatively proper characterization of the phenomenon.

RÉSUMÉ

Des décennies de recherche empirique ont démontré ce que plusieurs considéraient comme une évidence : nous sommes souvent irrationnels. Nous nous fions à nos meilleures intuitions dans des situations où nous devrions plutôt effectuer une réflexion en bonne et due forme; nous entretenons des croyances allant à l'encontre des faits et refusons de l'admettre; nous entretenons des croyances confortables plutôt que de faire face à des vérités difficiles, etc. À la lumière de ces exemples, il n'est pas surprenant que des phénomènes comme les vœux pieux ou la duperie de soi soient si répandus chez des agents autrement rationnels.

La littérature philosophique sur l'irrationalité s'est plus traditionnellement orientée vers la compréhension du *pourquoi* de nos dérives en la matière. Et pourtant, nous semblons souvent déjà être en mesure de *savoir* comment nous aurions pu faire mieux. Nous sommes en mesure de surmonter nos biais, manifester une certaine maîtrise de soi, et accepter la réalité telle qu'elle est. Il est possible de savoir comment faire mieux et on y arrive bien de temps en temps; comment peut-on donc parvenir à faire mieux? Il s'agit là de la question directrice de ma thèse. J'élabore ainsi une explication du phénomène jusqu'ici négligé de la sortie de la duperie de soi [*self-undeception*], c'est-à-dire le phénomène par lequel une personne qui se dupe elle-même est en mesure de mettre fin à ce phénomène de duperie et de se réconcilier avec la vérité.

Je débute en démontrant qu'une compréhension adéquate de la sortie de la duperie de soi dépend de ce qu'on comprend par duperie de soi et consacre la première moitié de ma thèse à présenter et défendre une conception particulière de la duperie de soi que j'appelle violation doxastique. Selon cette conception de la duperie de soi comme violation doxastique, la duperie de soi est un phénomène où l'agent qui se dupe lui- ou elle-même entretient une croyance qui entre en contradiction avec les normes épistémiques auxquelles il ou elle adhère autrement. Je démontre que

cette conception de la duperie de soi comme violation doxastique offre l'avantage, par rapport à des conceptions alternatives de la duperie de soi, de permettre une meilleure distinction entre la duperie de soi et d'autres biais similaires; de permettre une meilleure explication du caractère irrationnel de la duperie de soi; de remettre le rôle du soi au premier plan dans la duperie de soi; et de mieux rendre compte de la tension empirique caractérisant la duperie de soi.

Sur la base de cette théorie de la duperie de soi comme violation doxastique, la dernière partie de ma thèse vient ensuite consolider ma conception originale de la sortie de la duperie de soi. J'y avance qu'une conception plausible de la sortie de la duperie de soi devrait être en mesure de fournir une caractérisation métaphysique de ce phénomène (i.e. ce en quoi la duperie de soi consiste), une caractérisation psychologique (i.e. comment la sortie de la duperie de soi prend place), et une caractérisation normative (i.e. une voie souhaitable en vue de la sortie de la duperie de soi). Ce dernier critère implique de proposer une conception de la sortie de la duperie de soi qui origine d'une réflexion sur soi et qui prépare l'individu en vue de ce qu'il ou elle conçoit comme un avenir épistémique meilleur. Je démontre que l'approche de la violation doxastique est la mieux outillée pour la sortie de la duperie de soi car, contrairement aux théories concurrentes, elle est la seule à pouvoir proposer une caractérisation proprement normative du phénomène.

ACKNOWLEDGEMENTS

There are many people without the help of which not only I couldn't have completed this dissertation, but I couldn't have finished my PhD.

First and foremost, my advisors: to them goes a lot of my gratitude. Ian's hopeless optimism in my abilities, his constant encouragement, kindness, and empathy have often turned bad days into hopeful ones. His psychological insights have made this dissertation better than it ever could be. Sarah has not been just an advisor, she has been a mentor for me. She was nurturing, and she believed in me when I didn't. She pushed me to get out of academic comfort zone and taught me, sometimes the hard way, not to be afraid of challenges, but instead to be confident in my abilities to overcome them. She has been my strongest supporter and her dedication, love, and care continue to inspire me as a researcher and as a teacher. There are simply not enough words in this language to express the immense gratitude that I have towards her. She has made me a better philosopher and a better person.

Along the way, there have been many people that in small but crucial ways have contributed in making my PhD and this dissertation a memorable journey. Al Mele's help in reading the first drafts of my chapters was invaluable, and his collaboration continues to make me a better philosopher. To David Davies's encouragement and helpful insights go a lot of my gratitude. His enthusiasm about my work (and about Italian!) have brightened many of my days. One post-proseminar conversation, during my first year of the program, where he told me he had "no doubt" I would have succeeded in this program, has been a memory that I often went back to in moments of discomfort. In all honesty, those words have often made the difference between staying in or dropping out. I want to thank Marguerite Deslauriers, whose wisdom, bluntness, and enjoyable *conversazioni di Italiano* have meant more for me than she might imagine. Andrew Reisner's encouragement to overcome my imposter syndrome, Natalie Stoljar's constant availability and help on my work, Eran Tal's, Eric Lewis', Emily Carson's, Chris Howard's, and Stephanie Leary's incredible support during the job market, both professionally and psychologically, have been invaluable to me. To Angela Fotopoulos goes my gratitude for helping navigate the struggles of being an international student and for supporting me during these last difficult months, in a pandemic that has taken a toll on many of us. Finally, I want to thank Emily Carson for having been the *best* DGS a graduate student could ask for, and for having always supported and fought for the well-being of every student.

My colleagues have been invaluable friends. To Frédérick Armstrong, Sarah Clairmont, Celia Edell, Kosta Gligorijevic, Maiya Jordan, Éliot Litalien, Andre Martin, Sophie Osiecky, Sebastian Rodriguez Duque, Joey Van Weelden, and Emily Wash go my particular thanks. You've been phenomenal friends and have contributed in making this past six years a great experience. Without Alice Everly's and Michel Xhignesse's love, encouragement, and care I am not sure I would have been able to go through this journey.

Finally, an immense amount of gratitude goes to the two people that have supported me the most, and without whom I would not have been able to function: Charlotte Sabourin and Matthew Scarfone. I feel so fortunate and grateful to have you in my life that I am not even sure how to articulate it. The amount of care, patience, and time you have both put into encouraging me, supporting me, listening and believing in me has been, and still is, one of the things that I cherish the most. Without you, I wouldn't have been able to get to this point. And there are not enough words to thank you.

INTRODUCTION

My Long-Winded Answer

"It's better to confront these little sadnesses. Because sooner or later these things will arrive and if you've tried to hide from them the blow is even harder. If you don't have the moral courage to look *fino in fondo* – into the depths – you learn nothing. How can you teach anyone else anything if you don't know yourself?"

— Micaela Coletti, survivor of the Vajont disaster[1]

1. Opening Story

In May 2015 I embarked on a trip to the University of Amsterdam to give my first conference presentation. My paper, "How We Believe", examined whether the wishful beliefs involved in self-deception aim at truth, and if they do not, as I argued, whether it makes sense to classify them as beliefs at all. During the Q&A, which I handled with the typical nervousness of someone who is only one year into their PhD program, an undergraduate student, whose name now escapes me, asked a question that stood out: "How do we get out of self-deception?". The question left me dumbfounded, as I had never thought of it myself. Like me, many other philosophers had not either, preferring to instead focus on how we get *into* self-deception. I praised the originality of her question and admitted

[1] On October 9, 1963 a landslide caused a tsunami in a region near Venice (Italy) where the Vajont dam was situated and destroyed several towns. As a result, 1,917 people lost their life. Coletti's quote is from BBC article by Mark Duff (2013), "Italy Vajont anniversary: Night of the 'tsunami'".

that, since I had no answer, I would need to think about it (a reply-card that I later learned philosophers are allowed to play only sporadically).

After the Q&A, the student and I kept on discussing the issue. Though her asking such a great question validated my view of undergraduates – not merely as learners, but also as philosophical collaborators who can provide often genuine insights – I quickly realized that none of the literature on self-deception that I could think of had systematically examined the question. We both felt that this lack of attention was somewhat surprising. If self-deception is such an irrational and undesirable state, then why was no one interested in exploring how it can be abandoned? Even anecdotally, we noted, people seem to get out of self-deception. Certainly, this fortunate phenomenon is not as widespread as its counterpart, but why think that its perhaps rarefied nature disqualifies it from any attention?

Upon returning to Canada, a closer look at the literature confirmed that I was right: the phenomenon of coming out of self-deception, or self-*un*deception as I had begun to call it, was uncharted territory. I was however compelled to pursue the topic as I had left Amsterdam promising my interlocutor that I would have looked into her question. Five years later, this dissertation is my answer. Alas, it is not a very concise one.

2. Motivating the Project

As is the case with any philosophical issue that is not surrounded by a robust amount of literature, examining self-undeception revealed itself to be a methodologically challenging project. The first difficulty I faced was that since there had not been any examination of the topic, self-undeception required philosophical motivation that could not simply be exhausted by noting that the topic was inherently fascinating (though it is). To remedy this, two motivations seemed to support the project: the first is a conceptual or philosophical one, grounding why self-undeception is worthy as a topic of philosophical investigation; the second is a more practical motivation, which pointed to the concrete, cultural, and psychological implications that examining self-undeception can bring about.

2.2 A Philosophical Motivation

With respect to the philosophical motivation, I found that bringing self-undeception into the conversation seemed appropriate for completeness sake. An account of how individuals fall into the grips of self-deception simply seemed incomplete if not coupled with a view that at least gestures at how agents free themselves from this state. Since self-deception is a state that the agents are, even to some small degree, contributing to bring about, it made sense to also ask what individuals can do in order to unravel it.

Philosophers seem to naturally interrogate themselves about this very same question with respect to similar phenomena, such as *akrasia*. A vast philosophical literature has focused on what it means to embrace its opposite, that is, self-control.[2] This attention is not limited exclusively to the domain of action, but also extends to the domain of belief. A large amount of research has been devoted to the notion of doxastic control, or the ability to control one's own beliefs in the face of doxastic temptations.[3] In more empirically driven philosophical fields, scholars have also investigated how individuals can implement unbiasing techniques in order to overcome problematic prejudices.[4] Similarly, the metaethical literature on well-being has recently started investigating its flip-side, that is, ill-being.[5] Attention to parallels of well-known phenomena in other philosophical fields is a *pro tanto* reason for why the same attention should be directed to self-undeception. It is at least a reason for why the phenomenon should not be dismissed as irrelevant.

[2] See for example Mele 1987, Kennett 2013, Baumeister, Vohs, and Tice 2007, Henden 2008.
[3] See for example Holton and Shute 2007, Audi 2008, Paul 2015*a*, Paul 2015*b*, McHugh 2017.
[4] See for example Plant, Peruche, and Butz 2005 and Rudman, Ashmore, and Gary 2001.
[5] See for example Kagan 2015 and Tully 2017.

Inserting self-undeception into the philosophical conversation can also have implications that pertain to other related fields. For one, assessing the kind of epistemic resources that are needed on behalf of the agent to reach a state of self-undeception should be of interest to virtue epistemology. Undeceiving surely represents an improvement of the agent's epistemic life, but self-undeception is not an isolated event. It can provide the agent with an opportunity to reflect on their previous state as self-deceived and develop habits of thought that maintain a self-undeceived state in the future, thus steering clear from other instances of self-deception. More generally, anyone who is interested in issues that pertain to rationality should also be interested in exploring what insights an investigation into self-undeception can add to those debates. Plausibly, the process of self-undeception signals a shift *from* irrationality *to* rationality. So if one is committed to understanding how individuals, in general, come to be rational, certainly self-undeception can shed some much needed light on the issue.

Though this dissertation will not explore all these implications, it seems to be worth pursuing the project that quite easily gives rise to them.

2.3 A Practical Motivation

A second, more practical motivation for the relevance of examining self-undeception was one that I did not have to formulate, but that was naturally provided to me simply by observing the state of the world we have come to inhabit. Having written this dissertation over five years, I have witnessed our society becoming increasingly permeated by wild conspiracy theories, vaccine skepticism, Holocaust denial, denigration of expertise, and other sorts of pernicious reasoning. Even now, as I am typing this introduction in July 2020, many individuals refuse to believe that the current coronavirus pandemic is genuinely occurring, and as a result, are not implementing simple public health measures to prevent its spread, such as wearing masks.

Now, none of those malignant developments I cited perfectly map onto self-deception, but there are certain underlying worrying aspects that they all share. All of them involve belief perseverance where individuals do not believe evidence or warranted propositions. All of them involved a recalcitrance driven by one's pre-existing political, social, conative motivations that perniciously influence the way evidence is gathered and interpreted. If so, this means that examining how one comes to be self-deceived, or likewise how one comes to embrace conspiracy theories, is useful in order to understand the structure of these phenomena, but it is *not*, and cannot, be the only project worth pursuing. An additional, pressing task is to understand how these phenomena can be unraveled. That means that we, as philosophers, but also as persons, can contribute something towards isolating the dynamics that govern the dissolution of these phenomena. Debunking these phenomena at large then acquires a much more urgent practical flavour, because an investigation into how the same belief perseverance can be undermined in self-deception can pave the way for a unified methodology for how to proceed with respect to similarly pernicious phenomena. Predicting the positive implications that this unified approach can have further validates why examining self-undeception is so practically important, and why it not only warrants but demands our attention.

3. Connection to Self-Deception

With the motivation for the project now clear, I came to realize that the absence of literature on self-undeception raised a further challenge. In attempting to formulate an account of what it is to self-undeceive, I would often find myself spinning in the conceptual void. In order to remedy this philosophical disorientation, I made the decision to begin investigating such an unknown phenomenon by starting from one that I already knew – that is, by starting from self-deception.

While it was apparent that very little was known about self-undeception, its flip-side phenomenon of self-deception had been carefully analyzed.[6] Surely, what we know about how and

[6] See for example Davison 1985, Audi 1997, Lazar 1999, Funkhouser 2005, Levy 2004, Van Leeuwen 2008, Gendler 2007.

why we deceive ourselves can inform how self-deception can be unravelled, or so I came to think. However, the literature on self-deception comprises many different theories of what philosophers take the phenomenon to be, so the first step was for me to formulate an account of what I myself took self-deception to amount to. Based on the theory that I thought accurately captured what self-deception is, I could then build my characterization of self-undeception.

I want to stress that there are certainly different ways to tackle self-undeception, and in this dissertation I do not claim that my methodology is the only proper one. Unlike self-deception, I do not take self-undeception to be a stand-alone, self-contained phenomenon. Rather, self-undeception is dependent upon what self-deception is. Even intuitively, self-undeception seems to involve abandoning a state of self-deceit, and in this sense, it seems to unfold from self-deception. Understanding what abandoning a state of self-deception involves, raises in turn the preliminary question of what it is meant by "state of self-deception" in the first place. This is, in fact, what I set out to do in the first three chapters of this dissertation. Here I endorse and defend my preferred theory of self-deception, on which I then build and advance my account of self-undeception in Chapter 4.

4. The Structure of the Dissertation

It would be unclear why I endorsed a specific theory of self-deception if one was not familiar with the different theories hosted in the philosophical panorama. So, in Chapter 1, I provide the reader with a map of the different characterizations that philosophers have heretofore adopted. I start with an intuitive notion of self-deception as 'lying to oneself', commonly employed in non-philosophical language, to then highlight its limitations and show how a more sophisticated philosophical analysis of the phenomenon brings to life questions that escape its pre-theoretical conception. Do we deceive ourselves intentionally, that is, aiming at deceiving ourselves, where this aim is directly accessible to our introspection? And once we are self-deceived, do we truly *believe* that, say, our partner is being faithful, or do we merely *pretend* to?

The way philosophers have answered these questions reveals them to be endorsing different views. So-called *intentionalists* argue that self-deceivers deliberately engage in actions aimed at deceiving themselves.[7] According to such accounts, self-deception mirrors *inter*personal deception: we deceive ourselves in the same that we deceive others, with the only difference that in the *intra*personal version the deceiver and the deceived are the same person.

But intentionalist accounts quickly run into well-known problems. For example, how could one intentionally aim at deceiving themselves and have the deception be in any way successful? The worry is that, in an intentional self-deception, one would *see through* their own intentions.[8] Additionally, if the deceived and the deceiver are the same person, then this would imply that an individual simultaneously holds two contradictory beliefs, *p* and not-*p*. The two beliefs would need not neutralize each other for the self-deceit to succeed but this seems psychologically implausible, especially if the content of both believes involves *p*.

These issues have led to so-called *deflationist* accounts of self-deception. In these accounts, the aim has been to deflate the problematic intentional element in self-deception. All deflationist theories agree that self-deception occurs below the agent's awareness, but they diverge with respect to the doxastic attitude that the self-deceived individual holds towards the comfortable proposition they profess to endorse. *Pretense accounts* argue that the self-deceived person does not believe their comfortable proposition, but merely pretends to in the sense of make-believe.[9] *Belief-based deflationism* instead holds that the self-deceived person believes their self-deceptive proposition (e.g. that their

[7] See for example Rorty 1988, Davidson 1985, Pears 1982 and Bermúdez 2000, Talbott 1995 for a milder version of intentionalism.
[8] I borrow this elegant expression from Annette Barnes, who famously employs it in her book "Seeing Through Self-Deception". See Barnes 1997.
[9] See Gendler 2007.

partner is faithful). The two varieties of these views which I discuss – *minimalism* and *doxastic violation accounts* – disagree with respect to which conditions are sufficient for considering an agent self-deceived.[10]

As I will argue, I consider belief-based deflationist approaches to be the most plausible characterization of self-deception, and I prefer doxastic violation accounts over minimalism. I argue that the doxastic violation account is superior because it can meet minimalism's challenges and respond to its objections satisfyingly. In Chapter 2 I present the doxastic violation account in more detail, and in Chapter 3 I defend its superiority.

4.1 The Doxastic Violation Account

I begin Chapter 2 by noting that, like all belief-based deflationist accounts, the doxastic violation account claims that self-deception occurs non-intentionally and it results in a belief. However, in contrast to minimalism, the doxastic violation account claims that a motivationally biased hypothesis testing is not sufficient to declare an individual self-deceived, and claims that self-deception must involve the violation of the agent's epistemic norms.

Epistemic norms are, for example, norms that may involve deferring to experts, or trusting the clock at the post office. They are patterns of belief formation that individuals believe they have at least a *pro tanto* reason to follow in virtue of being their own norms. In this sense, epistemic norms *guide* agents' epistemic behavior with respect to how they think they ought to form beliefs.

Not all epistemic norms are the same though. There is certainly a difference between trusting the clock at the post office and deferring to experts. One difference is that while the former seems to merely play a doxastic function in helping the agent navigate the environment, the latter can possess additional, non-doxastic functions. If held by a scientist, for example, the epistemic norm of deferring

[10] See Mele 2001.

to experts may speak to the individual's identity in important ways or modulate access to the scientific community. Playing these non-doxastic functions is what typifies what I call *loaded* epistemic norms, in contrast to *bland* epistemic norms that are simply doxastic tools.

What happens then when an individual holds a belief that is at odds with the epistemic norms they actually hold? It depends: if the individual violates a bland epistemic norm, I suggest that little reaction is to be expected on behalf of that individual. We all hold epistemic norms that do not mean much to us (believing the clock at the post office is an example). Surely the fact that they are norms *we* hold may generate doubts or a feeling of unease when violating them, but these reactions will be rather bland (hence the name). However, if individuals violate norms that are loaded, then a more intense response is to be expected. Or so I'll argue.

This more acute response is in fact compatible with the phenomenology that self-deceived individuals often experience. As typically understood, the self-deceived person is not confident that the proposition they profess to endorse is true. Moreover, their self-deceived condition, far from making them happy, is ordinarily accompanied by feelings of unease, worry, and anxiety. I suggest that this phenomenology can be naturally explained by postulating that the self-deceived person holds a belief that is at odds with a loaded epistemic norm that they actually hold.

Consider my father, for example. My father is a doctor and as such he typically defers to experts. His epistemic norm however is not just a mere pattern of belief, but it has also a particular valence to him. Believing other physicians, and other experts in general, speaks to his identity in important ways. Being the kind of person who regularly defers to experts also allows him to qualify as integral part of the scientific community.

When my partner and I told him of our intention of getting married, my father reacted enthusiastically. Unfortunately, the happiness of the moment was interrupted by the realization that, due to the current coronavirus pandemic, the wedding would need to take place after the distribution of a vaccine. Since then, my father has repeatedly assured me that there will be a vaccine as early as September on the grounds that Donald Trump has guaranteed it will be the case. When I relayed to him information coming instead from Dr. Anthony Fauci, who instead believes the vaccine will be available only in 2021, my father embarked in effortful mental gymnastics to justify (read: rationalize) why Trump was actually right.

There is an expression commonly used in Italian that means "mirror climbing". Climbing a mirror is an obviously effortful activity and it is also bound to fail because the mirror's slippery surface makes it easy to fall from. Italians use the expression metaphorically to denote the struggle of someone attempting to justify a proposition that by their own lights is not justifiable. My father was certainly guilty of some mirror climbing, alternating between confidence and nagging doubts and anxiety. He eventually realized that he "couldn't believe" he for a moment trusted Trump over Fauci, an expert.

Now, my father is a clear example of someone who, driven by the strong desire to walk his daughter down the aisle, gathers and interprets evidence in a biased way and eventually finds himself holding a belief that goes against his loaded epistemic norm. The tension he experienced is also typical of the self-deceived person's phenomenology.

Certainly not all instances of self-deception are marked by this tension. Does that imply that instances that are phenomenology-free do not qualify as cases of self-deception? I will argue that they *do* count as instance of self-deception, and the doxastic violation account I prefer is well-equipped to explain why. The more acute the tension generated is, the more likely it is that the self-deceived person violated a loaded epistemic norm. A lower degree (or absence) of tension, does not indicate *absence* of self-deception, but rather, it indicates that the norm violated is of a blander variety.

That the doxastic violation account is well-suited to explain the phenomenological aspect of self-deception is not the only point in its favour. In Chapter 3 I argue that compared to minimalism, the doxastic violation account has several advantages. In addition to better capturing the experiential tension characteristic of self-deception, the emphasis on the agent's epistemic norms makes the theory well-equipped to explain how self-deception involves the *self*. It also does a better job at distinguishing self-deception from other similar biases, like confirmation bias or wishful thinking, which do not definitionally involve a violation of the agent's epistemic norms. The doxastic violation account also better explains the subjective irrationality of self-deception because according to the theory the self-deceived person violates norms that they themselves endorse.

4.2 Self-Undeception

In the fourth, final chapter I examine the core topic of my dissertation: self-undeception. Though most of my dissertation is on self-deception, I consider the work done in the first three chapters to provide a solid preparatory ground for introducing self-undeception. The way I frame my analysis of the heretofore unexplored phenomenon is by arguing that the doxastic violation account is superior to minimalism and pretense accounts not only with respect to its characterization of self-deception, but also with respect to its characterization of self-undeception.

As I said at the outset, there are many different ways to tackle self-undeception and I suspect that my approach might seem somewhat unusual because, once introduced to the novelty of self-undeception, most people probably expect a general analysis that aims at providing a unified account of what the phenomenon is. The expectation is not unreasonable because this is the methodology that is in fact employed in the literature on self-deception. Presenting a theory of what self-deception is and how it occurs equals advancing a theory that aims at settling the question of what the phenomenon amounts to *universally*. But as I already mentioned, I do not take self-undeception to be a stand-alone phenomenon (unlike self-deception). Instead, I take it to be one that derives, so to speak, from self-deception. Given this, it makes sense that an account of self-undeception *depends* on what one takes self-deception to be. To give an example, surely an intentionalist like Donald Davidson would not propose a theory of self-undeception that relies on unbiasing techniques, since intentionalism does not take self-deception to involve a motivational bias.

It would obviously be ideal to have an account of self-undeception that is compatible with the numerous characterizations of self-deception present in the literature, and I do not rule out that such an account can be formulated. But the primary aim of this dissertation is not to settle the question of what self-undeception is and how it occurs, universally. My aim here is to open the debate on self-undeception and introduce the phenomenon into the philosophical conversation. That is why I approach self-undeception by arguing, perhaps more modestly, that compared to minimalism and pretense accounts, the doxastic violation account provides a characterization of self-undeception that is superior.

Since minimalist accounts conceive self-deception as a species of motivationally biased reasoning, their account of self-undeception incorporates empirical research on belief revision in motivated reasoning. Minimalist accounts suggest that individuals abandon self-deception when counterevidence increasingly accumulates to the point that the mental tension normally experienced by the self-deceived person becomes too much for the individual to withstand. Pretense-based accounts instead advance a characterization of self-undeception framed in terms of costs adjustment. The suggestion is that when facing high-stake situations, the self-deceived person may abandon their self-deceit when the cost of maintaining it is *trumped* by the benefit of embracing reality.

I argue that both of proposals, minimalist and the pretense-based, are problematic insofar as they do not propose a way to undeceive that is, on reflection, desirable for the agent. What is a *desirable* way to undeceive oneself? A desirable way to undeceive oneself, I suggest, is one where the epistemic benefit of abandoning self-deception does not end at the very moment that the individual undeceives, but allows the agent to learn from their epistemic shortcomings and in doing so, it sets them up for what the individual sees as an improved epistemic future. This is a way that I argue is a *normatively* appropriate way to self-undeceive.

Consider the way minimalism characterizes self-undeception, for example. There the process of belief revision is somewhat forced on the individual who abandons their self-deceit to lessen the anxiety caused by counterevidence being too laborious to rationalize. In this sense, the self-deceived person must hit rock bottom, so to speak, in order to self-undeceive. Counterevidence must accumulate so dramatically and must be so incongruent with their prefixed beliefs that the individual cannot help but revise their beliefs. But is this a desirable way to update one's own beliefs?

I argue that it is not. A recent news story helps showing why. Richard Rose III was an Army veteran from Ohio who stubbornly did not believe in the "hype" of wearing masks because he was convinced that the current coronavirus pandemic was a hoax.[11] Rose later tested positive for covid-19 on July 1st, and while in quarantine he shared the diagnosis on social media, admitting that he had been having difficulties breathing. He passed away three days later, on July 4th. By then, Rose had started taking the coronavirus pandemic very seriously, but ideally, one would have preferred his change of mind to occur *before* contracting the virus because his future cannot certainly improve now. In this sense, a model that conceives belief revision as a function of the agent's hitting rock bottom is not a normatively desirable model because it cannot advance any helpful recommendations for the improvement of the self-undeceived's future.

Self-undeception through trumped incentives faces a similar issue. There the agent slides out of self-deception without necessarily gaining *awareness* that they were ever self-deceived. What seems to trigger self-undeception then does not seems to be the agent self-reflecting on their condition or

[11] See Rose's story: https://www.usatoday.com/story/news/nation/2020/07/17/covid-19-ohio-veteran-37-refused-wear-mask-died/5457283002/. Rose's case is sadly far from isolated. See here for a married couple who changed their mind about Covid-19 being a hoax after being hospitalized with symptoms: https://www.bbc.com/news/stories-52731624

on what might have possibly gone wrong in their epistemic practice. Rather, self-undeception through trumped incentives depicts the self-deceived person agent undeceiving out of *convenience* by gravitating towards the option that at the moment is perceived as less costly. The agent *does* reconcile with truth, yet this reconciliation occurs because the warranted proposition just so happens to be the most practically advantageous choice. This focus on advantageousness comes at the expense of an understanding of the agent's epistemic shortcomings.

These views' limitations become more apparent when faced with instances of self-undeception where an individual self-undeceives *in spite* of the subjective cost of both reality and self-deception remaining constant. How do we explain an individual who abandons their self-deceit when it is still the easier path to embrace? Both characterizations of self-undeception cannot explain this instance of self-undeception because both models rely on a shift of costs.

4.3 How the Doxastic Violation Account Explains Self-Undeception

The doxastic violation account is instead well-equipped to account for instances of self-undeception that are not characterized by any cost shifting. In the second half of Chapter 4 I propose two models of self-undeception that the doxastic violation account can advance. The first is an interpersonal model. This model involves the individual abandoning their self-deceit via reasoning with an interlocutor that they trust. Here the self-deceived person is offered both practical and epistemic reasons in favour of embracing reality and, through a process of self-reflection that occurs with their interlocutor, they slowly gain awareness of their own condition as self-deceived.

The second model is intrapersonal. This model portrays self-undeception as a state that the individual can attain by relying on their own resources. Here a self-deceived agent S recognizes an *isomorphism* between her own situation and the situation that another self-deceived agent B experiences. Self-undeception then occurs when S is able to transfer their own judgment of B to themselves and their own situation. Witnessing B's self-deceit can lead S to the realization that they have also been violating their epistemic standards which in turn may lead them to re-commit to their epistemic norms.

Successfully detecting an isomorphism between S and B is, however, unlikely to occur in purely rational ways. Though only anecdotally, self-deceived individuals are often able to identify self-deception in others without transferring the same judgment to themselves. Recognizing an isomorphism is also more challenging when the individual in question is someone that the self-deceived agent has little in common of with whom they do not easily identify. For these reasons, I suggest that *empathizing* with the self-deceived individual B can create an emotional dynamic able to bypass the obstacles that would be in place if the recognition of the isomorphism occurred in purely rational ways.

These two proposals allow the self-deceived person to gain awareness of their own condition, as well as come to an understanding of their previous epistemic shortcomings. This understanding is important because the epistemic benefits of undeception do not, and should not, end at the moment that the individual undeceives. The undeceived individual can reflect on their previous self-deceit and take their epistemic shortcomings as an opportunity to *grow*. The newly undeceived agent can benefit from the wisdom of their rational condition by committing themselves to practical strategies that will contribute to preventing future instances of self-deception. Epistemic growth is a diachronic process, one which can only be achieved by grasping one's epistemic weaknesses and learning from them.

I end the dissertation by extrapolating the main insights that an investigation into self-undeception has brought into the philosophical conversation. Conceptually, examining self-undeception has indirectly shown that among the theories present in the self-deception literature, the doxastic violation account is not only the best candidate for explaining self-deception but also the best candidate at characterizing self-undeception. Endorsing a particular theory of self-deception is of course not a new aim. But I defend the doxastic violation account in a non-standard way. By arguing that the theory is the most plausible candidate for explaining self-undeception, I have added *new* reasons for preferring this view to its competitors and, in doing so, I have also introduced a novel, underexplored phenomenon.

Practically, the examination of self-undeception can perhaps teach us about more than merely conceptual aspects. One of the upshots of my investigation of self-undeception is that it shows how a process that would intuitively seem to involve mainly epistemic factors instead heavily relies on non-doxastic features. As I initially conceived this dissertation, I expected self-undeception to hinge on a pure reiteration of evidence. Yet the more I worked on the project, the more I was forced to admit that a self-undeceptive process that occurs in purely epistemic ways does not seem psychologically plausible, given the motivational interests at stake. Empathizing with those who made our same epistemic mistakes as well as conversing with those who we trust, are not epistemic, but psychological features that can sometimes make the difference about whether we embrace or reject evidence.

This conclusion is important because it again sheds light on how other similarly dangerous phenomena such as conspiracy theories or antivaccine attitudes, which also involve a hostile attitude to evidence, can be debunked. In this sense, an investigation of self-undeception has relevance both philosophically and practically.

CHAPTER 1

A Map of Self-Deception

When I was a master's student in logic, I always had a hard time explaining what I worked on to non-philosophers who curiously asked me about it. My responses often came across as either abstruse or unclear (a result that I secretly thought was due to my own limitations rather than logic's). When my interests started shifting towards philosophy of action, and self-deception in particular, I felt hopeful that this time explaining my work to non-philosophers would have been easier. The reason being that everyone knows what self-deception is. Yet my optimism did not last because it quickly became apparent that what I conceived as self-deception was very different from what non-philosophers meant.

Non-philosophers have a pre-theoretical, intuitive conception of self-deception that is commonly known as 'lying to oneself', while philosophers take the intuitive characterization as a starting point from which to extrapolate puzzles that escape the pre-theoretical conception. Interestingly, both philosophers and non-philosophers are correct, in the sense that both intuitive and philosophical explanations of self-deception pick out aspects that are thought to be real of the phenomenon. Adverse attitude towards uncomfortable evidence and doing so because of an underlying interest or desire, are universally considered to be the drivers of self-deception.

In this chapter I begin by illustrating the transition from an intuitive, pre-theoretical conception of self-deception to a philosophical one. I do this to show not only the features that the two conceptions have in common, but also to do justice to the depth that the philosophical literature adds to the debate on self-deception. In detailing the philosophical characterization of self-deception, I canvass the different views that philosophers have formulated. I present their structure as well as assess their positive and negative traits. My aim is to provide the reader with a map of the various characterizations of self-deception. This map will serve as useful background against which to frame my preferred theory of self-deception, on which I will model, in the last chapter, my account of self-undeception.

1. An Intuitive Illustration of Self-Deception

Consider the following examples:

> *Crime*. A father refuses to believe that his son is a thief, despite compelling evidence that he in fact is.[12]
>
> *Cancer*. An oncologist denies she has cancer, even though the symptoms she is experiencing would normally lead her to believe otherwise.[13]
>
> *Affair*. A man believes that his partner is not having an affair, even though there is overwhelming evidence that points to the contrary.[14]

These examples are instances of self-deception. Self-deception is a phenomenon that is familiar to us all. It is this pervasiveness that perhaps explains how individuals already have a pre-theoretical notion of what self-deception is. Consider *Cancer*. There are two elements that can be detected here. The first is that the oncologist believes that she does not have cancer against compelling

[12] This example is from Canfield and Gustavson 1962: 35. See Sanford 1988: 165 for a detailed discussion of it.

[13] This example is from Rorty 1988: 142.

[14] This example is based on Lazar 1999: 265.

evidence. The second, implied by the first element, is that she believes she does not have cancer *because* she wishes that to be the case.[15]

The idea of believing against compelling evidence intuitively strikes us as unnatural. This is because it is generally thought that we only believe what we have good evidence for (Heil 1984: 61-62). When it comes to perceptual evidence, for example, if one notes that it is not raining, and they know they are not hallucinating (for example), then it seems impossible for them to avoid forming the belief that it is *not* raining. Evidence seems to control and guide belief-formation. Beliefs are instead tools agents use to navigate their environments. The way beliefs help us navigate is by picking out truths. So, evidence only points in the direction of what to believe, without agents playing any active role in believing (Williams 1973). Thus, the only good reasons for believing are *evidential* reasons. This so-called evidentialist line of thought is likely behind the oddity caused by hearing about agents believing against strong evidence. John Heil sums this idea up by saying that "beliefs seem most often to come to us, unsought and unbidden, *on the heels of* thought and investigation", thus "the notion that one might come to believe something simply by willing it has, if not exactly an air of contradiction, at least a strong whiff of implausibility" (Heil 1984: 59).

According to the evidentialist tradition, if it is evidence that almost automatically imprints beliefs upon us, then when one holds a belief in the face of stronger evidence to the contrary there is a meaningful sense in which the agent *themselves* must be the imprinter who causes themselves to believe. This is indeed how self-deception seems to operate. Instead of evidence controlling the ship's wheel, guiding one towards the most plausible belief, in self-deception it is agents themselves who take control of the ship's wheel, navigating towards what is convenient to believe. This last part of the

[15] This is not uncontroversial. There is a dispute about whether the self-deceived person believes their friendly proposition because they wish that to be the case or whether they merely wish to bring themselves to believe it, regardless of the proposition's truth values. As it will become clear later in this dissertation, the characterization of self-deception I endorse assumes the former. For more on the dispute see Funkhouser 2005 and Bilgrami 2006.

metaphor naturally leads to the second element of *Cancer*. The motivating factor that causes the oncologist to believe, against strong evidence, that she does not have cancer is a *desire* or *interest* that this belief be true which unduly shapes her epistemic investigation.

That said, it would be a mistake to overgeneralize this and say that all beliefs that are causally influenced by desires are epistemically problematic. Suppose for example that I fancy some gelato. At the gelateria I meet a colleague who tells me that a common acquaintance I dislike just lost their job.[16] As a consequence, I now believe that our acquaintance is unemployed. There is a sense in which the belief I formed is causally influenced by a desire, because if I had lacked the desire to eat gelato, I would not have formed the belief that the disliked acquaintance lost their job. Yet, as Yuval Avnur and Dion Scott-Kakures argue, this is not the sort of causal influence that is problematic, because the belief that the acquaintance got laid-off does not seem to involve desires in an epistemically problematic way in the sense that the causal influence here does not threaten the "justification of my belief" (Avnur & Scott-Kakures 2015: 11). But in self-deception the role of desires *does* seem to be particularly pernicious. This difference calls out for explanation, which I will address in the next section.

So far, I have said that an intuitive account of self-deception is able to isolate two key elements of self-deception:

(i) an agent S believes some comfortable proposition *p* in the face of strong evidence to the contrary;
(ii) S believes *p* because she wishes *p* to be true.

[16] This example is adapted from Avnur & Scott-Kakures who use it to explain the concept of "positional influences". See Avnur & Scott-Kakures 2015: 11.

However, (i) and (ii) do not alone articulate *what* self-deception is. A complete characterization requires supplying details that can be provided only by engaging in a deep philosophical analysis of the phenomenon. These details here include, but are not limited to, specifying why self-deception is irrational, and whether we need to postulate a particular metaphysical structure of the mind in order to explain how one can lie to themselves, so to speak. In this respect, appealing to the notion of believing against strong evidence does not turn out to be particularly illuminating because it does not capture the idea of lying to oneself (one can believe against the evidence unknowingly, while lying is by definition intentional) nor does it provide an explanation of what makes self-deception uniquely irrational. Similarly, the intuitive illustration I just provided does not identify the etiology of self-deception. That is, it does not explain *how* one comes to form the self-deceptive belief that *p* while under the influence of desires, and why such influence is problematic. In the light of this, philosophers have provided philosophical accounts of self-deception that aims at answering a *metaphysical* question of what self-deception is and a *psychological* question of how self-deception occurs.

2. Intentionalist Accounts of Self-Deception

Philosophical accounts of self-deception can be grouped into three broad categories: *robust intentionalist* accounts (Rorty 1988, Davidson 1985, Pears 1982), *weak intentionalist* accounts (Bermúdez 2000, Talbott 1995, Johnston 1988), and *deflationist* accounts (Mele 2001, Gendler 2007, Lazar 1999, Van Leeuwen 2008, Nelkin 2002, Lynch 2012). This section will consider both robust and weak intentionalist accounts. Section 3 considers deflationist accounts.

2.1 Robust Intentionalism

Robust intentionalist accounts, most notably that of Donald Davidson, mirror self-deception with its interpersonal analogue (i.e. deception), where an agent A, who believes *p,* deliberately deceives an agent B into believing not-*p*. In *inter*personal self-deception, Davidson says an agent A (deliberately) deceives another agent B about *p* under, I take it, what I take to be the following conditions:

- A knows or believes that *p*;
- A aims to bring it about that B believes that not-*p*, and intentionally takes steps to achieve this aim;
- Throughout the time that she is deceiving B about *p*, A is aware that she is deceiving B about *p*;
- Throughout the time that she is deceiving B about *p*, A continues to believe that *p*;
- If the deception succeeds, B comes to believe that not-*p*;
- A believes that *p*, B believes that not-*p*, but no one believes *p* & not-*p*.

This much is familiar to us about deception. But in self-deception, which for Davidson is the *intra*personal version of ordinary deception, A and B are the same person. This creates an interesting and difficult philosophical puzzle, because A (deliberately) deceiving *themselves* about *p*. This involves the following:

- $\text{A}_{\text{deceiver}}$ knows or believes that *p*;
- $\text{A}_{\text{deceiver}}$ aims to bring it about that $\text{A}_{\text{deceived}}$ believes that not-*p*, and intentionally takes steps to achieve this aim;
- Throughout the time that she is deceiving $\text{A}_{\text{deceived}}$ about *p*, $\text{A}_{\text{deceiver}}$ is aware that she is deceiving $\text{A}_{\text{deceived}}$ about *p*;
- Throughout the time that she is deceiving $\text{A}_{\text{deceived}}$ about *p*, $\text{A}_{\text{deceiver}}$ continues to believe that *p*;
- If the deception succeeds, $\text{A}_{\text{deceived}}$ comes to believe that not-*p*;
- $\text{A}_{\text{deceiver}}$ believes that *p*, $\text{A}_{\text{deceived}}$ believes that not-*p*, but no one believes *p* & not-*p*.

As a theory of *self-*deception, robust intentionalism adopts this model of deception. It falls out of robust intentionalism's picture that the theory is committed to two requirements:

> *Contradictory beliefs requirement:* the self-deceiver holds and maintains contradictory beliefs about the content of her deception, but does not believe their conjunction.

Intentionality requirement: the self-deceiver intentionally produces her deception (Levy 2004: 295, lightly altered).

The contradictory belief requirement implies the intentionality requirement. Since believing the conjunction of two contradictory beliefs would automatically lead to their own undoing, if an individual holds two contradictory beliefs about the same topic without juxtaposing them, she must do it intentionally.

2.2 The Two Paradoxes

A few clarifications about the requirements are needed. Consider an agent, Sally, and the following two contradictory belief statements:

(1) I have cancer

(2) I do not have cancer

First, the contradictory belief requirement commits Sally to believe that she has cancer and to believe she does not have cancer. The proposition that she has cancer, and the discomfort derived from such proposition, is a causal condition for Sally's belief that she does not have cancer. What this means is that Sally believes she does not have cancer *because* she aims at avoiding the discomfort triggered by the belief that she has cancer. However, Sally does not believe the conjunction of the two beliefs. The reason for this is that if Sally believed in the conjunction the two contradictory statements, she would believe a straightforward contradiction, which would likely result in the two beliefs undoing each other.[17] The contradictory belief requirement then raises the so-called *static paradox* where Sally believes that she does and does not have cancer at the same time. The challenge is then to show how

[17] Or at last, epistemic norms require this, though psychologically the story is likely to be more complex.

Sally can hold these two contradictory beliefs without her also believing their conjunction, even though she believes she does not have cancer *because* she believes she does (Mele 2001).

Second, since Davidson characterizes self-deception as the mirroring ordinary deception, the intentionality requirement requires Sally to intentionally deceive herself through activities that shall be discussed. I understand Davidson's commitment to the intentionality requirement as claiming that Sally intentionally deceives herself if and only if i) she aims at deceiving herself, and ii) it is immediately available to Sally's awareness that she aims at deceiving herself.

With respect to i), Davidson points out that Sally's self-deceived state is not simply the end result of having a false belief due to actions Sally performed, for then she "would be self-deceived even if [s]he read and believed a false report in a newspaper" (Davidson 1985: 207). The point is that, as Davidson puts it, the self-deceiver "must intend the 'deception" (Ibid.). Davidson is implying that not only does Sally actively perform actions that later result in her believing she does not have cancer, in the sense that she directly controls her bodily movements while performing those actions, but she also has the clear-sighted *aim* of directing her intentional behavior at a specific outcome that she has previously adopted. That is, Sally has the *goal* of coming to believe she does not have cancer. This is why Davidson says that "self-deception requires the agent to *do* something with the aim of changing [her] own view" (Ibid.). So not only does Sally aim at deceiving herself, but she is somehow also aware, per ii) above, of her self-deceptive aim.

A further, so-called *dynamic paradox* then arises from meeting the intentionality requirement: if Sally is aware that she is deploying a deceitful strategy, then her self-deception seems destined to fail because she would see through her own strategy. The challenge then is to explain how Sally can be aware of her intention to deceive herself and actually succeed in her self-deceit (Mele 2001).

Given these clarifications about the contradictory belief requirement and intentionality requirement, a typical case of self-deception can be construed as follows:

a. Sally has evidence on the basis of which she believes that the proposition "I have cancer" is more likely to be true than the proposition "I do not have cancer".

b. The thought that she may have cancer or that she ought to "rationally believe" that she may have cancer, generates discomfort, and Sally wishes that the proposition that she does have cancer were true, but fears it may not be.

c. The discomfort generated by the thought that she may have cancer "motivates" Sally to *intentionally* act in such ways as to "cause herself" to believe that she does not have cancer.

d. Sally is then more inclined to believe that the proposition "I do not have cancer" is true, even though the "totality of the evidence available to [her] does not support this attitude" and leads her to also believe that she does have cancer (Ibid., 200).

In the strongest case, the belief that Sally has cancer not only causes the belief that she does not, but it also "sustains it" (Davidson 1985: 208).[18] Furthermore, according to Davidson, Sally endorses the *requirement of total evidence for inductive reasoning*. The requirement states that "when we are deciding among a set of mutually exclusive hypotheses", we are required to "give credence to the hypothesis most highly supported by all available relevant evidence" (Ibid., 201).[19] A person accepts the principle of total evidence if her "pattern of thoughts" are in accordance with the requirement, and if that person is "disposed in the appropriate circumstances to conform to it" (Ibid., 204). Such is the case of Sally who "reasons and thinks in accordance with the requirement" when the appropriate circumstances require it (Ibid.).[20] If she accepts the requirement, then the requirement binds her to

[18] A possible misunderstanding to flag, regarding a), is to take a) as saying that the proposition that she has cancer need not to be a full-fledged belief, but merely a suspicion that this proposition may be true or an inclination to believe it without fully believing it. This, however, would not be an accurate reading of Davidson. Davidson himself, in fact, stresses that the idea of being "inclined to believe" is "too anodyne" (Davidson 1985: 200).

[19] Davidson takes this requirement from Hempel. For more on Hempel's formulation of the requirement, see Hempel 1965.

[20] It should be noted that are times when such requirement cannot be applied with absolute generality like, for example, cases where we do not yet feel we have enough evidence to decide at all. However, the cases Davidson is

believe the most supported hypothesis. From this it follows that she does not merely fear or suspect that the proposition that she has cancer might be true; rather she must fully believe it.

The thought that she may have cancer generates discomfort, while the thought that she does not have cancer is obviously accompanied by a feeling of relief. Sally's self-deception then originates also from her wish that the proposition that she does not have cancer be true. This wish is part of the motivation that brings Sally to behave in ways that aim to cause herself to believe that she does not have cancer. This may occur in several ways. Sally may, for example, purposefully "seek, favor or emphasize" evidence that undermines the proposition that she has cancer, *intentionally* directing her attention away from the evidence that points toward the proposition that she has cancer, or actively searching for evidence in favor of the opposite, more comfortable prospect (Ibid., 200). In particular, suppose Sally has compelling evidence that she has cancer. She has test results and scans that show where the cancer has spread. Suppose she is also an oncologist and would normally conclude from such test that she is most likely to have cancer. Additionally, she takes the evidence she has to be good evidence, and this leads her to believe that she has cancer. This belief, however, generates discomfort which motivates her to look for counter-evidence that supports the friendlier belief that she does not have cancer. She finds an article in a questionable medical journal that warns that one should always be careful when interpreting test results in medicine, particularly the ones related to cancer. Furthermore, she remembers an old friend of hers who had the same test results that later turned out to be indicative of a different illness, less serious than cancer. Here, throughout the self-deception, Sally remains aware that the totality of evidence points to the truth of "I have cancer" yet on the basis of the weak evidence she finds, she believes she does not.[21] If self-deception is successful Sally is

discussing are cases where the person possesses evidence that justifies her to go straight to an outright belief.

[21] It is worth mentioning, for the sake of avoiding possible misunderstandings, that the belief "I have cancer" *indirectly* causes (i.e. through intentional activities) Sally to believe "I do not have cancer". This clarification is important, for if "I have cancer" directly (and intentionally) caused "I do not have cancer", this would mean that Sally has intentionally decided to believe "I do not have cancer". This resembles the idea of believing at will, which has been highly criticized as being psychologically implausible (Williams 1973). What is not psychologically

deviating from her own requirement of total evidence. She knows or believes she does not have good reasons to endorse the belief that she does not have cancer, yet she believes she does not.[22] In conclusion, robust intentionalism characterizes self-deception as "a self-induced weakness of the warrant where the motive for inducing a belief is a contradictory belief" that occurs through intentional activities (Davidson 1985: 208).

2.3 Solving the Paradoxes

We are left now with the two tasks flagged above: the task of solving the static paradox and the task of solving the dynamic paradox. Let's start with the first one. Solving the static paradox involves showing how Sally may believe that she does and does not have cancer without believing their conjunction. For his part, Davidson attempts to solve the paradox by appealing to the compartmentalization of the mind. He suggests that on an intuitive level, people can and do sometimes hold "related but opposed beliefs apart" kept isolated by "boundaries between parts of the mind" (Davidson 1985: 211-212). The mind then, can be partitioned or divided, where such boundaries are not accessible to individuals' awareness. Partitioning is put forward as a psychological phenomenon that helps to make sense of agents' irrational attitudes that would be otherwise challenging to understand. On the one hand, the boundary does not necessarily define permanent and separate territories, as contradictory beliefs can belong to strongly overlapping domains, but on the other hand, they never happen to belong to the exact same territory. Erasing the line would result in neutralizing one of the two beliefs.

implausible for Sally to do, is to deliberately decide to select the evidence in favour of her preferred belief, which is what Davidson suggests. In this way, Sally does not believe at will, but believes "I do not have cancer" on the basis of the evidence that she intentionally selected.

[22] Sally's situation seems similar to a case of weakness of the warrant, where an agent judges better to believe *p* and yet believes not-*p*. In fact, Davidson specifies that self-deception is similar to weakness of the warrant. For one, they both require the existence of an irrational belief in the face of conflicting evidence as a necessary condition. But while in self-deception this conflict is always motivated, in weakness of the warrant it does not always need be.

With the divided mind metaphor in mind, let's look at how we can solve the static paradox. If Sally's mind is divided, then this means that in one part of her mind she believes she has cancer, while in a separate part of her mind she believes she does not. Thus, in virtue of being the deceiver, she believes she has cancer and, in virtue of being the deceived, she believes she does not. These two beliefs are kept separate by a 'boundary' which also allows for the two beliefs to be held at the same time without Sally believing their conjunction. What causes the isolation of the two parts is her avoiding accepting what the requirement of total evidence (that she endorses) advises. Self-deception's irrationality, then, not only consists in the beliefs' etiology (because Sally holds her beliefs not for epistemic reasons, but merely for motivational reasons), but it also consists in the structure of the mind of the self-deceived person. In particular, the irrationality consists in the very drawing of the boundary that allows Sally to believe and maintain two contradictory beliefs.

We can now also solve the dynamic paradox. The dynamic paradox consists in the challenge of having to explain how intentional self-deception can succeed, given that the deceiver and the deceived are the same person. As said, the partitioning of the mind allows Sally to be her own 'deceiver' and 'deceived' where these two reside in different territories of the mind that are kept separate. This allows for the deceiver's aim to be kept hidden from the deceived, who is not aware that she is being deceived. The deceiver makes sure, so to speak, through continuous and active effort that the deceived is kept ignorant of her condition. If the deception is successful, then Sally will not acknowledge, even to herself, that she intends to deceive herself. For, acknowledging it would force her to acknowledge the very belief she wishes to avoid.

2.4 For and Against Robust Intentionalism

There are good reasons why one may want to endorse robust intentionalism. To start, robust intentionalism's model of self-deception as the intrapersonal analogue of ordinary deception provides a straightforward account of what one may intuitively take self-deception to consist in. The idea that the mind of the self-deceived person is divided preserves the structure of interpersonal lying. Characterizing self-deception as paralleling interpersonal deception also has the advantage of allowing the distinction between genuine instances of self-deception and instances of beliefs held against strong evidence that merely result from a mistake. The latter occurs non-intentionally, while the former occurs through intentional active effort. Since the deceiver must intend their deception, the intentional element explains why self-deception is "selective" - that is, not everyone, when motivated by a wish that *p,* and challenged by strong evidence that not-*p*, deceives herself in believing that *p* (Bermúdez 2000: 310).[23]

Moreover, robust intentionalism characterizes self-deception in a way that allows us to distinguish it from other irrational phenomena, such as wishful thinking. Davidson points out that wishful thinking involves an individual believing *p* merely because they wish it to be true (Davidson 1982: 206). The weight of irrationality here is on the etiology of the belief, for the agent believes *p* for evaluative reasons and not evidential reasons, and, as Davidson puts it: "the desire to hold a belief does not constitute evidence of the truth of the belief" (Davidson 1982: 208). Self-deception also involves evaluative elements, yet it seems to go beyond wishful thinking because its irrationality involves not only the etiology of belief, but also the state of the self-deceived individual, in particular the divided mind that allows for a departure from one's own norms (i.e. the requirement of total evidence).

However, despite its positive aspects, robust intentionalism has also been heavily criticized, particularly with respect to its commitment to the divided mind. John Heil, for example, points out that the divided mind portraits a psychologically implausible picture.[24] He casts doubt on the claim

[23] See Bermúdez 2000 for a more detailed discussion on why self-deception is selective.
[24] For more on the criticism of the divided mind see Heil 1989, Talbott 1995, Levy 2004, Bermúdez 2000.

that Davidson's partitions of the mind never overlap, and questions whether self-deception conceived in such a way can successfully capture cases of irrationality that become increasingly complex over time. If the agent's contradictory beliefs are about the same domain (e.g. both about cancer, in Sally's case) they will presumably be tied to a single network of beliefs which inevitably leads to their domains overlapping in the long run (Heil 1989: 583). Heil points out that "we act on beliefs, employ them in our practical deliberations, form other beliefs on their basis" and it is far from obvious that the mental compartmentalization could accommodate these items (Ibid.).

Heil's criticism reveals how Davidson's picture, that initially seemed attractive, actually clashes with intuitive cases of self-deception where individuals are wrapped up in highly sophisticated instances of self-deception that can last a lifetime. In this sense, Davidson's account seems to flout Ockham's razor because it presupposes some heavy metaphysical baggage (i.e. a *partition* in the *mind*) that is not obviously necessary. Even if Davidson's view was able to successfully explain self-deception, all else being equal, one should not prefer it to other explanations of irrationality that are simpler and less convoluted. The question, then, is whether such alternatives exist.

It has also been argued that the divided mind is unable to capture behavioral and experiential aspects of self-deception. Robert Audi and Eric Funkhouser argue that while the self-deceived individual professes to hold a particular belief, their behavior seems to be inconsistent with that belief (Audi 1997, Funkhouser 2005). Funkhouser gives the example of Mitchell, a man who avows believing he is not going bald, yet poses at a certain angle when being photographed and does not allow his wife to "tussle" his hair (Funkhouser 2005: 296). This behavior, described as "avoidance behavior" (Ibid. 297), validates a friction or tension between the self-deceived individual's linguistic and non-linguistic behavior. Coupled with the behavioral tension, the self-deceived person also experiences a mental tension where they are not fully confident about the belief they profess to endorse. This is typically construed as an experience of "tension", "conflict", "instability", "discomfort" (Audi 1997: 104, Losonsky 1997: 122, Noordhof 2009, Graham 1986: 226). These two kinds of tension are not considered conceptually necessary but they are often regarded as characteristic of self-deception (Lynch 2012). Indeed, most core instances of the phenomenon are marked by these kinds of psychological conflict.

So, if robust intentionalism is correct in positing the divided mind, then the two contradictory beliefs the self-deceived person holds do not interact with each other but are instead safely kept in separate regions thanks to the boundary. This would have to be the case because if the two beliefs *did* come into contact, then they would neutralize each other. So, if the self-deception is successful and the boundary provides a stable separation, then this yields a portrait of the self-deceived person as an individual who sincerely believes their lie. But as the two kinds of tensions noted above reveal, this does not look like an accurate representation of self-deception, which is instead characterized by a more conflicted phenomenology that the divided mind picture is unable to account for.

Further, some have argued that the partitioning of the mind also undermines the agent's violation of the requirement of total evidence, thus casting doubt on self-deception as a phenomenon of internal irrationality. Dion Scott-Kakures 1996, for example points out that with the deceiver and deceived kept apart, the deceived does not violate the requirement of total evidence which is where the phenomenon's irrationality lies, because to the compartmentalized mind the violation has not been achieved with "open eyes", that is, in full awareness (Scott-Kakures 1996: 42-43). This calls into question the irrationality of self-deception; while the deceived holds their belief on the basis of objectively bad reasons, there is no irrationality because due to the mental boundaries they lack access to contrary evidence (Ibid.).

2.5 Attempts to Improve Robust Intentionalism: Weak Intentionalism

The drawbacks of robust intentionalism have motivated philosophers to either reject the contradictory belief requirement or the intentionality requirement, or both. *Weak intentionalists* as I call them, reject the contradictory belief requirement by deflating one of the two beliefs into a weaker doxastic attitude and attempting to maintain a somewhat mitigated version of the intentionality requirement by postulating either a *forgotten* or an *unconscious intention* (Johnston 1988, Bermúdez 2000, Talbott 1995).

José Luis Bermúdez argues that although believing the unfriendly proposition is not a necessary requirement for self-deception, self-deception nevertheless is intentional with the individual *diachronically* deceiving themselves through intentional behavior. That is, the self-deceived person intends to deceive themselves at time t_1 with this intention even transparent to them, and over time "los[ing] touch" with this intention (Bermúdez 2000: 314). Thus, while the project is still to self-deceive, the deception is carried out "unknowingly" (Ibid., 315). Bermúdez gives the example of someone who has the desire of advancing their career. This desire shapes their choices to the point that it seems proper to claim that they are acting on the intention of advancing their career. But this does not imply that their actions are carried out "knowing" that they are being performed with that intention (Ibid., 314). Mark Johnston suggests something similar when providing the example of someone who attempts to trick themselves into performing a mischievous action:

> One's memory of what one did in the hour or so before taking the pill is very indistinct and sometimes apparently erased completely. Knowing this, one could get up to mischief during such a period and avoid the guilt of the morning after by taking the precaution of rearranging things so that in the morning one will be misled about what one did the night before (Johnston 1988: 76).[25]

[25] This example is also discussed in Lazar 1999: 271.

Here the intention is fully transparent to the agent and the deception occurs diachronically. At t_1 the agent's intention is to deceive themselves into thinking they did not commit the mischief, but at t_2, following the action, the agent intentionally leads themselves to *forget* their original intention to ensure the success of the deception.

Along similar lines, W. J. Talbott rejects the contradictory beliefs requirement but still maintains that the individual intentionally brings about their self-deceit (Talbott 1995). According to this characterization, self-deception occurs through intentional biases that operate below the surface of the agent's awareness (Ibid. 33). Thus, what is intentional is the individual biasing their "cognitive

es to favor belief in *p*, regardless of whether *p* is true" (Ibid.).

While these are all interesting refinements of the initial intentionalist theory, there are good reasons to reject weak intentionalism. As it has been argued, the theory of 'forgotten intentions' that Bermúdez and Johnston put forward provides a portrait of self-deception that undermines its irrationality, both as a process and as a state. Bermúdez and Johnston suggest that the individual initially intends to self-deceive and then later, while acting on their intention, they either lose touch with it or forget it through intentional maneuvers. Their account has the advantage of being immune from the static and dynamic paradox, but the drawback is that there seems to be nothing *irrational* in intending to self-deceive, even less so in forgetting the intention to do it.[26] Consider the following example from Jenkins 2018. Suppose Mia is constantly late for appointments. Frustrated at herself, she intends to henceforth be on time, and decides to set her watch ten minutes fast to trick herself into leaving her house earlier (Jenkins 2018: 12, adapted). Of course, this stratagem only works if Mia actually forgets her initial intention, which she just so happens to do. Now, is there anything irrational here? I argue that there is not. There is nothing irrational in the content of Mia's intentions, nor there

[26] Davidson argues something similar with respect to Pears' view. See Davidson 1985: 210.

is anything irrational in losing touch with it, which seems to simply be a by-product of the fact that rational agents are cognitively limited and often tend to forget their intentions. Nor is there anything irrational in the state that Mia ends up in. Granted, she holds a false belief, but this alone does not qualify as irrational. Furthermore, it is hard to see how Bermúdez's and Johnson's view could account for the tense internal conflict characteristic of self-deception (Levy 2004). If Bermúdez and Johnson are correct that the self-deceived person genuinely forgets their original intention, then it is unclear why the self-deceived person should experience internal conflict at all since their self-deceptive state is stable.[27]

Talbott's view of self-deception fares better than that of Bermúdez and Johnston. This is because instead of appealing to forgotten intentions, it instead characterizes the phenomenon as a biased investigation of evidence. This investigation is intentionally initiated by the agent, and it eventually yields the endorsement of the friendly belief. This account has the advantage of preserving the irrationality of self-deception as a process because it is motivationally influenced by non-epistemic elements. According to Talbott the only intentional element in self-deception is the agent's action of gathering and interpreting evidence. Moreover, this aspect has the advantage of avoiding the dynamic paradox, but by minimizing the intentional element the account cannot clearly distinguish itself from more robust deflationary views that instead reject the intentionality requirement. As I explain in the next section, a strand of deflationary accounts also relies on the idea that self-deception consists in a motivationally biased examination of evidence thus making Talbott's account hard to classify as intentional, even if weak.

[27] Others have also pointed out that the Bermúdez and Johnston characterization cannot be considered accounts of self-deception because there is no rationalization of contrary evidence. With respect to Johnston's case, Ariela Lazar points out that "it seems that the subject's belief that she did not engage in the reprehensible activity (if it is indeed generated) is not formed in the presence of strong evidence to the contrary" thus his case does "not seem to generate the much debated puzzle of self-deception" (Lazar 1999: 271).

3. Beyond Intentionalist Accounts: Deflationism

Deflationist accounts deny both the contradictory belief requirement and the intentionality requirement. Despite all deflationist accounts rejecting both of these requirements, not all deflationist theories are the same. Some deflationist accounts argue that the self-deceived person does not believe the comfortable proposition they profess to endorse, but merely *pretends* to. Accounts of these sort are so-called *pretense* accounts (Gendler 2007). Other accounts, in contrast, claim that the product of self-deception is a belief. Accounts of this sort are what I call *belief-based* accounts (Mele 2001, Van Leeuwen 2008). Belief-based accounts can in turn be divided into what I call *minimalist* accounts and *doxastic violation* accounts (Mele 2001, Van Leeuwen 2008). The most common way of elaborating a belief-based theory is a minimalist one. Minimalism argues that performing motivationally biased hypothesis testing is sufficient condition for that agent to be considered self-deceived (hence the label *minimalist* accounts or *minimalism*). *Doxastic violation* accounts, in contrast, object that motivationally biased hypothesis testing is an insufficient condition for self-deception and instead suggest that self-deception be characterized as a violation of the agent's own epistemic norms (Van Leeuwen 2008). In what follows, while I will argue that a belief-based strategy is superior to pretense accounts, I will suggest that minimalism is not the best way to develop a belief-based strategy. In Chapters 2 and 3 respectively I will explain and defend the superiority of the doxastic violation account.

3.1 Pretense Accounts

Pretense accounts, most notably Tamar Gendler's, hold that the self-deceived person does not believe the friendly proposition, but merely pretends to, in the sense of make-believe, while often endorsing the unfriendly belief (Gendler 2007). However, the pretense plays a role normally played by beliefs both with respect to their "vivacity" and "motivation to action" (Ibid., 241). By contrast, belief-based accounts maintain that the self-deceived agent genuinely believes the friendly proposition.

Gendler's theory fits with the larger narrative that self-deception occurs when someone aims at avoiding a painful truth. The self-deceived person begins imagining what it would be like if a particularly friendly proposition *p* was the case. In doing so, there is still no self-deception occurring, but merely a wishful fantasy. Yet, entertaining the unfriendly belief not-*p* generates discomfort, so the individual may minimize interaction with evidence that favors it, instead directing their attention to those features of the world that nicely fit with *p* (Gendler 2007: 241). Over time, the pretense may increasingly intensify and shift from a *performative* kind of pretense, where individual performs to her peers, to an *imaginative* kind of pretense*,* where the agent pretends to themselves. Both pretenses are of a behavioral nature where the self-deceived person merely acts as if *p* was true. Self-deception then occurs when, as the pretense process develops, the individual's pretense comes to have the same "vivacity" that ordinary beliefs have (Ibid., 231). Likewise, the pretense provides the same motivation to act as ordinary beliefs, by guiding the individual's choices. Notice here that according to Gendler the transition occurs *non-intentionally*. The self-deceived person *slips* into their deception unknowingly, but still believes that not-*p* is true and the unfriendly belief is what motivates the pretense.

Gendler's view has several important things going for it. First, it captures both self-deception's *irrationality* and its typical *mental tension*. With respect to the latter, the self-deceived person allows a projective attitude (i.e. their pretense) to play a role that only receptive attitudes (i.e. beliefs) should play, and this makes the phenomenon unstable (Gendler 2007: 242). With respect to the former, the self-deceived person's 'use' of a "projective" attitude as if it was "receptive" is what renders the deception irrational, because the improper use here violates the individual's "commitment to rationality" (Ibid.). Additionally, Gendler's theory makes sense of our first-hand experience of self-deception. Recall that one advantage of robust intentionalism was that the theory nicely captures the ordinary notion of self-deception as 'lying to ourselves', and it does so primarily because of its commitment to the intentionality requirement. Now, by denying the intentionality requirement, deflationary accounts tend to lose their intuitive appeal because they reject characterizing self-deception as the intrapersonal analogue of interpersonal deception. Yet, in this respect Gendler's theory is an exception. It succeeds in maintaining intuitive appeal precisely by positing that self-deception is a form of pretense.

The idea of the self-deceived individual as someone who lives in a realm of make-believe fits well with our pre-theoretical image of the self-deceived person as someone *living in their own fantasy*. Conceiving of self-deception in this way also seems to fit with the opposite phenomenon of *ceasing* to be self-deceived. Anecdotal evidence shows that those who managed at some point in their life to free themselves from self-deception report feeling as if they 'woke up from a dream', which evokes the idea that the deception had them living in a fantasy. Gendler is perhaps the only philosopher who also attempts to explain how the process of self-*un*deception, as I call it, may occur. She argues that individuals cease to be self-deceived when maintaining the self-deceit reveals it to be too detrimental for the well-being of the subject. When this occurs, the self-deceived person abandons their pretense and elects to act consistently with their unfriendly belief.

Consider the example that Gendler provides, of someone who is self-deceived that they do not have a terrible illness. Doctors, she says, present them with the following choice:

> [T]aking medicine M gives great health benefits to those with disease D and causes terrible pain to those without disease D, and not taking medicine M has precisely the opposite effects. If he actually holds the delusional belief that he is not suffering from D (not-P), then he may elect not to take medicine M. But if he is self-deceived in the sense I have been considering, then while he allows his pretense that not-P to guide his actions and thoughts when his fantasy-maintenance desire is dominant, there will be situations where he will instead allow his (tacit or explicit) belief that P to serve as the basis for his actions. In particular,

> when confronted with this high-stakes forced choice, the self-deceived subject will likely elect to take the medicine. (Gendler 2007: 245)

This example suggests that when presented with high-stakes choices, the self-deceived person will likely abandon their pretense. This is controversial, however. Some have argued that high-stakes choices may not lead to the self-deceived person abandoning their pretense (Porcher 2014). The reason why is that self-deception is often self-destructive. It is common for the self-deceived person to exhibit behavior consistent with their own self-deceptive belief, even when this consistency comes at their own cost. For these reasons, it has been suggested that high-stakes situations may not always be sufficient for self-deceived people to trump their pretense. Consider for example the case of someone who, despite being highly exhausted, believes that they can drive and opts to nevertheless drive home with their car. If the driver is merely pretending that they can drive while being exhausted but does not actually believe it, then it is not clear why they would decide to hit the road. The stakes are certainly high here as driving while being exhausted can result in serious accidents, yet the self-deceived person does not abandon their belief.[28]

3.2 Belief-Based Accounts: Minimalism

As I have already noted, belief-based deflationary accounts agree that the agent's doxastic stance toward the friendly proposition is one of belief, not pretense (Mele 2001, Van Leeuwen 2008). As I have also already noted, in virtue of being deflationist, such views also reject the intentionality and contradictory beliefs requirement. But this skeleton can be elaborated upon in various ways. In this section I examine the prospects for belief-based deflationism via consideration of the most influential way of developing the theory, which I call *minimalism.* The *minimalist* account argues that self-deception is a phenomenon where the individual non-intentionally self-deceives by endorsing a false belief that results from an examination of evidence unduly influenced by their own interests and

[28] This example is inspired by Porcher 2014: 323–324.

motivations (Mele 2001). According to minimalism, the following four conditions are jointly *sufficient* for someone (S) to be self-deceived:

2. The belief that *p* which *S* acquires is false.
2. *S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.
2. This biased treatment is a nondeviant cause of *S*'s acquiring the belief that *p*.
2. The body of data possessed by *S* at the time provides greater warrant for ~*p* than for *p*. (Mele 2001: 52-53)

These four conditions are all that is *sufficient* for an individual to be considered self-deceived. In this sense, the account is rather *minimal* for it does not presuppose any additional feature. According to the minimalist account the belief that Sally has cancer is simply a *hypothesis* that Sally entertains in the form of the question: "Do I have cancer?". Sally fears that this hypothesis might be true, as well as wishes it was not. Call Sally's fear and wish the *motivational elements*. When testing whether the hypothesis that she has cancer is true, Sally's motivational elements bias her treatment of evidence in a way that she is not aware of. The motivationally biased evidence then causes Sally to endorse the convenient belief that she does not have cancer. Importantly, Sally can be aware of her motivational elements, but what she is not aware of is that these elements are causally influencing her examination of evidence in an epistemically problematic way.

According to the minimalist account, a case of self-deception can be construed as follows:

a) Sally fears that the hypothesis that she may have cancer might be true, and wishes that the opposite, more welcome, hypothesis that she does not was true.
b) Sally tests the hypothesis that she may not have cancer by gathering evidence in order to assess its truth value.
c) Sally's motivational elements bias her treatment of evidence in an epistemically problematic way.

d) The biased treatment of evidence causes Sally to endorse the false belief that she does not have cancer.[29]

Let me underline that this is a deflationist view. The difference here as compared to intentionalist accounts, broadly construed, is clear from step a) and it consists in *how* the self-deceived individual comes to believe the friendly proposition. While for intentionalists the belief that Sally has cancer is acquired intentionally, according to the minimalist account the belief results from motivationally-biased hypothesis testing that occurs non-intentionally. Let me clarify this last point.

Suppose Sally is a gifted oncologist who leads an active lifestyle and never had serious health problems. During a checkup she is informed of a mass resembling a tumor. As a result, Sally starts suspecting that she may have cancer. She is not sure that that her suspicion is warranted. The test results are merely initial and might turn out differently later on. However, the mere entertaining of the possibility that she may have cancer generates anxiety: if she really had cancer, she would not be able to see her kids grow up, and she would also subject her family to months of agony. Naturally, the thought that she may have cancer generates severe discomfort in Sally, while the thought that she may not instead produces instant relief. She consciously wishes that the proposition that she does not have cancer were true.

Thus, Sally starts testing whether the hypothesis that she may have cancer is sufficiently justified. However, Sally's motivational elements (i.e. her fear that she may have cancer and desire that she may not) direct her attention to the more comfortable proposition that she does not, and, as a result, she initiates testing for whether the latter is true. Evaluating whether she does not have cancer,

[29] Tough Mele claims that the self-deceptive belief must be false because deception by definition involves a false belief, I do not take this as being a necessary requirement. One can deceive themselves about a belief that accidentally turns out to be true, and still count as self-deceived. This is because even if their belief is true, the *etiology* of how they acquire it is epistemically problematic.

instead of whether she does, shapes Sally's epistemic investigation in particularly problematic ways. Sally may start collecting evidence that confirms that she does not have cancer, while discounting evidence that shows that the opposite is true. This is one way in which Sally's epistemic investigation might be biased.

To clarify, suppose one asks you whether you think of yourself as being a good friend. In order to answer the question, you will test whether the proposition 'I am a good friend' is true. In doing so, you will contemplate all those instances where you listened to your friends' struggles, you successfully kept promises and supported your friends' projects. In other words, you may picture only those times where you have been in fact a good friend. On the contrary, if the proposition in question was 'I am not a good friend,' you would recall instances where you have failed at being a good friend, thus confirming that the earlier proposition is true. The hypothesis testing can be influenced depending on which proposition one is focusing on. This is why Sally's motivational aim has what is called a "directional influence" on her beliefs, where directional influence involves examining the evidence to "favor a particular, predetermined outcome" (Avnur & Scott-Kakures 2015: 11). She initiates what she believes to be an objective examination of evidence with the aim of verifying whether it is true that she has cancer. However, she also has the additional aim of coming to believe that she does not have cancer because she thinks she is merely seeking reasons that will allow her to conclude her investigation. Sally does not only have an interest in believing what is true, but also has an interest in satisfying her own desires.

Sally then, aims at assessing whether it is true that she does not have cancer, but she also wishes that she does not have cancer. In this sense, directional influence is an "especially pernicious kind of influence that calls into question the belief's justification" (Ibid.). Directional influence can manifest in several ways. Sally might gather and pay attention to only a selective part of evidence. She might remember that a friend of hers had the same test result that was later found to be mistaken and focus her attention to that kind of evidence, instead of the test result. She might also mistakenly interpret the evidence. For example, she may credulously believe confirming data (a friend's experience), while scrupulously testing data that do not support her favorable hypothesis (test results). She might not realize that the test results constitute better evidence than her friend's unusual experience. From this it follows that Sally does not intentionally bring herself to believe that she does not have cancer, rather she finds herself in holding such belief as the consequence of an unconscious bias that shaped her epistemic investigation.[30]

The minimalist account is further supported by empirical evidence that shows how people examine hypotheses differently depending on whether the hypothesis in question is a particularly welcome one to accept. According to Yaacov Trope and Akiva Liberman, hypotheses that are more comfortable to embrace are subject to less scrutiny, while uncomfortable hypotheses, due to particular interests and desires that influence their testing, are more rigorously analyzed (Trope & Liberman 1996).

4 Which Is the Best Account? Intentionalism vs Deflationism

Recall that the dynamic paradox involves questioning the successful outcome of intentional self-deception by stressing that the self-deceived person would see through their own intentions, while the static paradox highlights the implausibility of the self-deceived person holding two contradictory beliefs at the same time without them undoing each other. Intentionalism proposes to solve the two paradoxes by compartmentalizing the mind, but this solution is particularly problematic because it posits a metaphysical structure of the mind that is psychologically implausible. Yet it seems impossible for intentionalism to avoid the paradoxes without committing to the compartmentalization of the

[30] Recall how Talbott's account also suggests something similar, thus making it virtually undistinguishable from belief-based deflationary accounts.

mind. Intentionalism then faces the difficult choice of either succumbing to the two paradoxes or committing to a metaphor of the mind which raises further issues. This dilemma, which seems to have no solution, constitutes a reason to prefer deflationary accounts. Deflationary accounts solve the dynamic paradox by eradicating the intentional element in self-deception, and they avoid the static paradox by deflating one of the two beliefs to a weaker doxastic attitude. As I explain below, deflationism's solution to the two paradoxes is not free of drawbacks, but unlike intentionalism, these drawbacks can be solved (as I explain in Chapter 3). Thus, deflationary accounts are more plausible than intentionalist accounts because they advance a solution to the paradoxes that is not as problematic.

4.1 For and Against Belief-Based Accounts

Among deflationary accounts, I think there are good reasons to prefer belief-based accounts over pretense accounts. Briefly, belief-based accounts i) better capture the self-deceived person's behavior, and ii) give a more substantive characterization of the irrationality of self-deception. Let me explain i) and ii) in more detail.

With respect to i), it is often the case that self-deceived individuals display both an avoidance strategy and self-destructive behavior (Funkhouser 2005). The former consists in the self-deceived subject avoiding coming into contact with evidence that may run contrary to their favored proposition. Eric Funkhouser provides the example of Nicole, who is self-deceived that her husband is not having an affair with her friend Rachel. In the evenings when her husband claims to be out with friends, she avoids driving by Rachel's house for fear of spotting her husband's car (Ibid., 302). The latter involves the self-deceived person behaving consistently with their professed proposition, even when doing so is potentially detrimental to them. An example of this is the case of the driver who is self-deceived that they can drive while being exhausted, and does nothing to prevent the worst (adapted from Porcher 2014: 323–324).

While a belief-based account is able to capture both self-destructive behavior and the avoidance strategy, pretense accounts can only properly explain the latter. A belief-based account can explain the avoidance strategy by arguing that despite the self-deceived person believing their preferred proposition *p,* they lack full confidence about its truth. They believe *p* yet *suspect* that *p* may not be entirely true, and fear of finding out that *p* may be false explains why they so carefully avoid confronting contrary evidence. Recall the example of Mitchell who is self-deceived that he is not going bald but poses at a certain angle when being photographed and does not allow his wife to tussle his hair. If belief-based deflationism is correct in arguing that the self-deceived person genuinely holds the friendly belief while merely suspects the unfriendly one, then Mitchell's behavior is to be explained by positing that he does *believe* he is not going bald, yet he in the back of his mind he *suspects* he is and that is why he poses at a certain angle and does not allow his wife to tousle his hair. Doing so prevents him from being in the situation where he would need to confront evidence in favour of this suspicion.[31] The self-deceived subject's belief that *p* also explains why the self-deceived person engages in self-destructive behavior. If the self-deceived person believes that *p*, then it makes sense that this belief would guide their actions, even when this leads to potentially detrimental consequences.

Pretense accounts explain the avoidance strategy by appealing to the idea that the self-deceived person merely pretends that *p* while believing that not-*p*. It is precisely because the self-deceived person believes not-*p*, pretense accounts claim, that they avoid confronting counterevidence. After all, if Nicole genuinely believed that her husband was not having an affair, what would be the problem in driving by Rachel's place? It is because she *does not* genuinely believe that he is faithful, and merely pretends that he is, that she avoids driving by Rachel's home. However, if the self-deceived person merely *pretends* that *p* while believing that not-*p,* explaining self-destructive behavior is more

[31] In addition, others have suggested that the reason why Mitchell's inconsistent behavior can also be explained by appealing to the idea that the self-deceived person is not fully confident in the truth-values of their belief, but merely possesses a high degree of conviction that their belief is true. See Lynch 2012.

challenging. As the example of the exhausted driver shows, if the driver is merely pretending that they can drive while being exhausted but does not actually believe it, then it is not clear why they would opt to drive given the danger of the situation. According to belief-based deflationism, the driver's choice to hit the road makes sense because it is simply the by-product of the driver believing that they can drive. If one is self-deceived that they can drive while in a deep state of exhaustion, and genuinely believes that, one will drive even if driving reveals itself to be self-destructive.

Advocates of pretense accounts would reply that, according to their characterization, the pretense plays the same role that a belief does with respect to its vivacity and motivation to act, which is what may explain the driver's behavior. This may be true, yet a belief is a superior candidate for producing actions. Imagination and desires may share some traits with beliefs when it comes to the way they generate actions, but as some have pointed out, it is counterintuitive to think that they share the same motivational role (Porcher 2014: 309, Van Leeuwen 2009: 232). Imagining that *p* may yield behavior similar to believing that *p*, yet it would be implausible to think that imagining that *p* shares all the "characteristic effects" that believing that *p* has on behavior (Van Leeuwen 2009: 232). To say that the two share *some* characteristic effects would certainly be more plausible, but it would not provide adequate explanatory grounds to account for the self-deceived individual's self-destructive behavior.

With respect to second problem of irrationality, belief-based deflationism provides a more robust characterization of why self-deception is irrational. According to pretense accounts, neither the process nor the state of self-deception is irrational. There is nothing irrational in the process because there is nothing irrational in merely pretending that *p* is the case, and there is nothing irrational in the state of self-deception because the agent still believes the proposition with greater warrant. The irrationality of self-deception lies in the fact that a projective attitude (the pretense) plays the role that is usually played by a receptive attitude (the belief).

But if this exhausts what goes awry in self-deception, then it is a rather thin characterization because the irrationality seems to occupy a marginal role in the structure of the phenomenon. Compare the characterization with belief-based deflationism. Here self-deception qualifies as irrational because of the etiology of the belief. The *process* of how the self-deceptive belief is formed is perniciously influenced by non-epistemic factors. Thus, the irrationality is internal to the phenomenon. This is a more substantive characterization because the irrationality is grounded *within* the structure of self-deception, contrary to pretense accounts where it is somewhat tangential to it.

Despite these positive aspects, belief-based deflationist accounts and especially their most familiar form, minimalist accounts, also face some serious challenges and objections. As I will explain, minimalist deflationist views have been criticized for failing to account for some aspects that any theory of self-deception should accommodate, such as explaining the characteristic phenomenology associated with self-deception and why the phenomenon is subjectively irrational (the *phenomenological* challenge and *irrationality* challenge, respectively). The minimalist version of belief-based deflationism has also been said to face two objections that I call here a *passivity* objection, and a *uniqueness* objection.[32]

Belief-based deflationism, and especially minimalism, does not seem well placed to capture important phenomenological aspects, namely the mental tension typical of self-deception (Audi 1997). This difficulty stems from the theory's commitment to self-deception as belief. If belief-based deflationism is correct in claiming that the product of self-deception is a belief, then this seems to portray the self-deceived person as a genuine believer which, as a result, makes it unclear why they should experience any form of internal conflict.

Belief-based deflationist accounts do mention that the self-deceived person *suspects* that their unfriendly belief might be true, and this may indeed help in explaining the experiential tension. The

[32] The labels are mine, but the content of the objections and challenges is not: they are objections and challenges that are well-known in the literature on self-deception. In this chapter I am simply reconstructing them to then show how the doxastic violation account is able to accommodate them.

self-deceived individual believes *p* but suspects that not-*p* might be true, which may capture the self-deceived person's unstable condition. However, belief-based deflationist accounts also posit that self-deception occurs without any awareness on the part of the individual, which makes explaining the tension more challenging. If the self-deceived person genuinely believes *p* and thinks that the belief is simply the result of a regular epistemic investigation, then it is hard to see how one could account for the internal conflict. Some have pointed out that the self-deceived person's "knowing insincerity" stems precisely from the fact that they are aware that they believe a proposition they ought not to (Scott-Kakures 1996: 49). But if belief-based deflationist accounts eradicate any awareness from the self-deception, this point cannot be explained.

As for what I called the *irrationality challenge*, this derives from the general agreement that an essential aspect of self-deception is that it is *internally irrational*. That is, the self-deceived individual does not merely depart from general canons of rationality, but from canons that are relative to the agent (Davidson 1985, Scott-Kakures 1996, Van Leeuwen 2008). Some go as far as to argue that it is "a brute fact that our vernacular conception of self-deception is such as to make internal irrationality a fundamental aspect of the phenomenon" (Scott-Kakures 1996: 32). Yet there are reasons that justify why self-deception's irrationality is specifically subjective.

Robust intentionalism motivates the notion of internal or subjective irrationality by appealing to several claims. First, the theory claims that self-deception is the intrapersonal version of ordinary deception. If the deceiver and deceived are situated in the same mind, then it is clear that what goes awry in self-deception occurs internally to the subject. In this respect, intentionalism is particularly well-equipped to capture the idea that the self-deception is not merely a *deception*, but a deception that occurs within the *self*. A second claim that helps explain the subjective irrationality is that the self-deceived person violates the principle of total evidence that they *themselves* endorse. This claim validates the idea that the departure of rationality that occurs in self-deception is relativized to the subject.

Naturally, deflationist accounts face a more difficult challenge in accounting for internal irrationality because such theories reject intentionalism's picture of the mind. Yet the motivation to capture subjective irrationality makes sense for deflationism because properly characterizing the subjective irrationality also provides an argument for distinguishing self-deception from other instances of irrationality that may not be subjectively irrational. However, by locating the irrationality of self-deception in its etiology, minimalist belief-based deflationist accounts cannot explain why self-deception's irrationality is specifically subjective.

Granted, it may not be rational to hold a belief purely on non-epistemic reasons, but it is not clear according to minimalist accounts what standards the self-deceiver is going against when deceiving themselves. Alfred Mele seems to hint at the idea that the self-deceived person is violating somewhat internal standards when he mentions that they select and examine evidence in a way that they "normally" would not (Mele 2001: 106). This suggests that the self-deceived person possesses a habitual way of investigating evidence and when deceiving themselves they are departing from it. However, Mele does not add anything more illuminating on this point, and instead proposes what he calls the "impartial observer test" (Ibid.). The impartial observer test is meant to identify cases of self-deception. He says:

> if S is self-deceived in believing that *p*, and D is the collection of relevant data readily available to S, then if D were made readily available to S's impartial cognitive peers (including merely hypothetical people), those who conclude that *p* is false would significantly outnumber those who conclude that *p* is true. Call this 'the impartial observer test' (Ibid.).

This remark seems in tension with Mele's previous point about the individual's habits, and instead seems to locate the irrationality if the phenomenon *outside* the self-deceived person, implying that they are violating external standards, standards that other non-biased peers would be ready to follow. I conclude that minimalist forms of belief-based deflationism will have particular trouble meeting the irrationality challenge.

The passivity objection, which is an objection specifically to minimalism, has two components: the first component involves minimalist accounts' failure to confer a special or important role on the *self* in self-deception, and the second component involves minimalist accounts' characterization of self-deception as a phenomenon that simply *happens* to an individual, in which their agency is not involved. With respect to the first component, minimalist accounts seem to erase the role of the *self* because they portray self-deception as process of motivationally-biased hypothesis testing where the agent is not particularly involved. If according to minimalist accounts self-deception is non-intentional and the self-deceived person slips into their own deception, then it is unclear what the role of the *self* is. With respect to the second component, minimalist accounts seem to portray self-deception as a passive phenomenon that merely *happens* to the self-deceived person instead of one that the agent actively brings about. It seems that the agent themselves does not *do* anything when self-deceiving and this passive view of self-deception fails to preserve thinking of the phenomenon as ordinary 'deception', thus compromising its intuitive appeal.[33]

Related to this is the uniqueness objection to minimalism, which argues that if self-deception's irrationality is exhausted merely by its etiology, as minimalism argues, then it is hard to see how the phenomenon could be different from other irrational phenomena that also involve a problematic etiology, but would not qualify as self-deception. For example, wishful thinking also involves the biased examination of evidence in favour of a particular belief, but one would not consider it an instance of self-deception. Recall that robust intentionalism is able to distinguish them by pointing out that self-deception is inherently intentional while wishful thinking is not. Some have attempted to

[33] See also Bach 1997: 105 who seems to agree that self-deception is not a passive phenomenon because it involves counteracting the truth that is "dangerously close at hand and must be repeatedly suppressed".

draw a similar distinction when suggesting that while in wishful thinking there is no confrontation with contrary evidence, in self-deception the individual recognizes and resists (through rationalization) the unfriendly evidence (Szabados 1973: 204). However, this distinction does not help. If the self-deceived agent recognizes the unfriendly evidence (whatever doxastic attitude that involves) then it is hard to see how such a claim can be reconciled with the claim that self-deception is non-intentional. How can the self-deceived person slip into her own deception without any awareness while at the same time recognizing contrary evidence? Minimalist accounts seem incapable of attributing a unique status to self-deception lest it downgrade the phenomenon to an ordinary bias.

*

I began this chapter by dividing theories of self-deception into three families: robust intentionalist, weak intentionalist, and deflationist. I argued that deflationist approaches were the most plausible of the three. I then subdivided deflationism into pretense theories and belief-based theories and argued that the latter approach is more promising. However, I have ended the chapter by raising four challenges and objections to the most familiar way in which belief-based deflationism has been developed, which I called minimalism. What is the solution? As I will argue in coming chapters, the solution is to hold that belief-based deflationism should be developed in a different way, which I call the doxastic violation account.

CHAPTER 2

The Doxastic Violation Account

The aim of this chapter is to present my preferred account of self-deception which I have called the *doxastic violation account*. Why, you might wonder, advance a new view of self-deception, considering the many that are already available in the literature? There are two reasons why presenting my view of self-deception is worthwhile: the first reason pertains to the analysis of self-undeception, and the second reason involves the relevance of the doxastic violation account in the philosophical panorama.

To start, what motivates my discussion of the doxastic violation account is the methodological assumption that a characterization of the phenomenon of self-undeception, the central topic of this dissertation, is dependent on what one takes self-deception to consist in. Thus, in order to fix the reference for the discussion on self-undeception, it makes sense to endorse and defend a particular theory of self-deception.

But in addition to the motivations that pertain to self-undeception, the doxastic violation account is also independently relevant. The account overcomes the objections, outlined in Chapter 1, that minimalism faces. And given that minimalism emerges in Chapter 1 as the most plausible rival of the doxastic violation account, showing the superiority of the account provides an important contribution to the debate on self-deception.

Furthermore, my analysis of the doxastic violation account also offers a fresh look at the notion of epistemic norms as attitudes that can possess a non-doxastic valence to agents who endorse them. The relevance of my novel approach not only emerges with respect to showing how the doxastic violation account captures self-undeception, as I will show in Chapter 4, but also with respect to phenomena that go beyond self-undeception. As I will suggest in the concluding chapter, my improved notion of epistemic norms can be employed in the analysis of conspiratorial beliefs to explain why they persist over time, and in particular why they so tenaciously resist revision.

In this chapter I present the doxastic violation account, which I then defend in Chapter 3. I then draw upon the conceptual resources laid out in both this Chapter 2 and Chapter 3 to advance my account of self-undeception in Chapter 4.

1. Situating the Doxastic Violation Account

The doxastic violation account is a type of belief-based deflationist view. Like all belief-based deflationist accounts, the doxastic violation account claims that self-deception occurs non-intentionally and it results in a belief. However, in contrast to minimalism, the doxastic violation account claims that motivationally biased hypothesis testing is not sufficient to declare an individual self-deceived. Recall that minimalist accounts claim that the following four conditions are jointly *sufficient* for someone (S) to be self-deceived:

> 1. The belief that *p* which *S* acquires is false.
> 2. *S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.
> 3. This biased treatment is a nondeviant 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*. (Mele 2001: 52-53)

The doxastic violation account departs from minimalist accounts by objecting that the four conditions above, while they sufficiently capture general instances of motivated reasoning, they are *not* sufficient for considering an individual self-deceived because self-deception must involve the violation of the agent's own epistemic norms.

We find the antecedent of the theory in Neil Van Leeuwen's formulation who, to my knowledge, is the only one who explicitly emphasizes the role of the agent's epistemic norms in self-deception. In his paper "Finite Rational Self-Deceivers", Van Leeuwen presents his view, yet does not fully articulate it given that the paper's focus is a different one. Because his presentation is rather schematic, my aim here is to flesh out and expand on Van Leeuwen's account so as to make the theory serviceable to my project.

According to the doxastic violation account, an individual is in a state of self-deception if and only if:

> (i) she holds a belief,
> (ii) that belief is contrary to what her epistemic norms in conjunction with what evidence she has would usually dictate, and
> (iii) a desire, with content appropriately related to the belief formed, causally makes the difference to what belief is held in an epistemically illegitimate fashion. (Van Leeuwen 2008: 195).

Here self-deception is characterized as consisting of both a *state* and a *process*. The state of self-deception captures what self-deception *is,* metaphysically, while the process of self-deception explains *how* self-deception occurs, psychologically. That is, self-deception is a state that involves in the self-deceived person believing a comfortable proposition, and it is a process that results in a desire which causally influences the formation of the self-deceptive belief (Van Leeuwen 2001: 112).

In this respect, both minimalist accounts and the doxastic violation account describe a similar process. The two theories identify a desire as the epistemic culprit, and are compatible with Trope and Liberman's idea that individuals have an easier time accepting evidence in favor of a particularly welcome belief. What differentiates the doxastic violation account is the explicit addition of the condition that the agent violates *their own epistemic norms,* which minimalist characterizations do not mention.

2. Epistemic Norms

The emphasis on the agent's epistemic norms raises the question of what exactly these are. According to the doxastic violation account, an agent's epistemic norms are "patterns of belief formation" that the agent "usually follows and that are rationally justifiable" (Van Leeuwen 2008: 195). "Rationally justifiable" implies that the agent who holds an epistemic norm is able, on reflection, to give reasons for why they hold their norm. This is of course different from requiring agents to provide these reasons *at the time* that they acquire the pattern of inference, as individuals often come to embrace epistemic norms without knowing why. It is however perfectly reasonable that individuals might, if asked or pressed, be able to point to relevant reasons for why they hold a certain norm upon reflection. It is not required that these be *good* reasons to the eyes of an external rational observer, but merely that the agent is capable of providing them.

The epistemic norms that the agent holds need not necessarily have a connection to truth. Individuals may hold all sorts of epistemic norms, from, say, distrusting experts to believing grandma's medical advice. If a conspiracy theorist oddly believes official theories about climate change, or if one resists their grandma's suggestion to spread olive oil on a mosquito bite to lessen the itching, then both cases qualify as violation of one's own epistemic norms, regardless of whether those norms track truth in the first place. In this sense, the doxastic violation account does not focus on which epistemic norms the individual *should* hold based on their connection with truth, but rather it focuses on how norms govern individuals' behavior on the basis of whether the agent themselves alone believes them to track truth.[34]

[34] This remark raises interesting cases. Suppose an individual agent is self-deceived, for she violates her own epistemic norm. For example, S holds an epistemic norm that establishes that she ought not to believe a certain group of people for they are rationally inferior. S's epistemic norm is not only mistaken (i.e. it fails to track truth) but it is also perniciously morally loaded. On my view, if S satisfied all the requirements of self-deception, including violating her own epistemic norm, she would still qualify as self-deceived.

According to the doxastic violation account an agent's epistemic norms are patterns of belief formation that, in virtue of them being their own, agents believe they have at least a *pro tanto* reason to follow. In this sense these norms *guide* agents' epistemic behavior with respect to how they think they *ought* to form beliefs.

This last remark shows that my conception of agents' epistemic norms departs from Van Leeuwen who claims instead that they are exclusively patterns of belief formation or epistemic habits. To start, that agents believe they have at least a *pro tanto* reason to follow their own epistemic norms. With respect to epistemic habits, instead, it is not obvious that agents believe that they have a reason to follow their habits. In fact, habits may be arbitrarily followed without agents believing that they have reasons that explain why they hold a particular habit. Agents' epistemic norms instead can be supported by reasons that defend, from the agents' own perspective, why agents' hold them in the first place.

With this I do not intend to say that agents' epistemic norms are not patterns of beliefs, but rather that they are not *exclusively* patterns of beliefs. Consider Catholicism, for example. If one is Catholic, then one may abstain from eating meat on Friday. But this is not merely a pattern that one follows. It is not held arbitrarily, for one can provide reasons in its support. The same reasoning can be applied to epistemic norms. One can have a *habit* of believing experts because one holds the epistemic *norm* that one ought to defer to experts.

But what is it meant exactly when one says that an agent holds an epistemic norm and how does one tell whether an agent holds an epistemic norm?

2.1 How Epistemic Norms Operate in Individuals' Psychology

The working definition of epistemic norms I provided above seems to imply that the only way to detect whether an individual holds a specific epistemic norm is *behavioral* – that is, by observing their epistemic practice. A behavioral analysis is certainly *one* way to infer an individual's norms. If, for example, Mia reliably trusts experts then one can infer that she holds the norm that one ought to defer to experts. A behavioral approach is thus helpful in uncovering the kind of norms an individual may hold, but it cannot be our exclusive means. The reason why is that it is perfectly possible for individuals to exhibit behavior of a certain kind without this necessarily being the result of an epistemic norm that the individual holds (Rorty 1983). Individuals may, for example, repeatedly manifest behavior that conforms to standards of beauty, and yet not necessarily hold the epistemic norm that one ought to conform to standards of beauty.

Consider this passage from Roxane Gay's recent book Hunger:

> I'm a feminist and I believe in doing away with the rigid beauty standards that force women to conform to unrealistic ideals. […] I believe it is so important for women to feel comfortable in their bodies […] I know, having grown up in a culture that is generally toxic to women and constantly trying to discipline women's bodies, that it is important to resist unreasonable standards for how my body or any body should look. […] What I know and what I feel are two very different things. […] I am not comfortable in my body (Gay 2016: 17-18).

In this passage Gay expresses her discomfort in her body that would lead one to infer that she holds the norm that one ought to conform to certain beauty standards. Her discomfort seems to be caused by her awareness that her body does not meet traditional standards. Yet, on reflection, Gay does not hold that norm. This mismatch between behavior and norms is what justifies the fact that a behavioral analysis cannot be the only means to extrapolate epistemic norms. One must not only observe agents' epistemic practice, but one must also listen to the reasons that individuals, upon reflection, are able to provide for why and whether they hold a specific norm.

That epistemic norms can be accessed upon reflection relates to the idea that epistemic norms can be held *implicitly*. That is, agents may hold a variety of epistemic norms, some of which may be held unknowingly. Consider the example of an epistemic norm of the sort "one ought to defer to experts". Suppose Mia is known among her friends for being someone who always follows her doctor's advice. It would be legitimate for her friends to infer from Mia's behavior that she holds the epistemic principle of deferring to experts. That is, it would be legitimate for her friends to infer that it is *because* Mia holds the norm that she ought to defer to experts that she typically follows her doctor's advice.

However, it would not at all be unreasonable for her friends to wonder if Mia is aware that she holds such an epistemic norm, for it is not obvious that she is. Mia's norm that one ought to defer to experts could be held implicitly and accessed only upon reflection. One could imagine Mia's friends interrogating her about the reasons for her actions (e.g. following her doctor's advice) and Mia discovering, through introspection, that it is *because* she believes one ought to believe experts that she acts in a certain way. Given this, it would be overly demanding to expect individuals to be aware of their epistemic norms at all times. Any reasonable account that aims at providing a psychologically plausible characterization of epistemic norms must make room for the possibility of these norms being held implicitly.

2.2 Bland vs Loaded Epistemic Norms

If I am correct in claiming that epistemic norms can be held implicitly, then presumably it is fair to assume that not all epistemic norms are held the same way. Agents hold a variety of epistemic norms, some of which may be held more *strongly* than other because of a particular investment on behalf of the agent, others instead may be held more neutrally. Let me now explain this point in more detail.

Quine famously claimed that individuals' beliefs can be conceived as a web where some beliefs are held in the center, and others are located at the periphery. Those beliefs that stand at the periphery are beliefs that the individual may invest with little importance, beliefs that the agent may not especially *care* about. Beliefs that stand at the center may be instead beliefs that the individual is particularly attached to because, for example, they are considered more reliable. Quine says: "[w]e all hold, for example, that those [beliefs] gained from respected encyclopedias and almanacs are more to be relied on than those gained from television commercials" (Quine 1978: 9).

Quine uses the metaphor of the web of beliefs to clarify that individuals, in addition to holding beliefs, also hold "higher-order beliefs about beliefs" (Ibid.). A rational agent may hold the belief that *p*, and the higher-order belief that locates *p* at the periphery because, say, *p* was acquired from a television commercial, and thus less reliable. On the contrary, one may hold an arithmetical belief acquired from an encyclopedia coupled with the higher-order belief that it should be held at the centre because of its reliable source. The belief's location also "guides" the way individuals assess evidence (Ibid.). When individuals detect a "conflict" among beliefs, they may collect and assess evidence in order to revise one or both beliefs or simply reject the more peripheral one (Ibid.). If one considers perceptual evidence to be particularly reliable, for example, then a belief obtained through observation may resist revision. As Quine argues, "we agree that what we think we see is usually there" (Ibid.). Seeing does not directly imply believing, yet "it goes a long way" (Ibid.).

Quine's terminology can be applied to the case of epistemic norms, with appropriate differences. While Quine advances his metaphor to highlight a distinction that pertains to the domain of beliefs, I wish to remain neutral with respect to the kind of doxastic attitude epistemic norms are. Yet, I aim at utilizing Quine's view as an analogy to capture a distinction between what I call *loaded epistemic* norms—that is, norms that stand at the center of the individual's psychic economy and that the agent is particularly invested in—and *bland epistemic norms,* that is, norms that stand at the periphery and which the agent is not invested in. The distinction that I draw by employing the central and peripheral predicates is however different from the one that Quine proposes. Quine appeals to doxastic features, such as reliability, to differentiate between central and peripheral beliefs; my distinction instead involves both doxastic and non-doxastic elements.

Roughly, a bland epistemic norm is a norm that serves as a tool to navigate the environment and does not contribute to shaping the agent's life in any other non-doxastic way. A loaded epistemic norm is instead a norm that in addition to allowing the agent to navigate the environment also possesses non-doxastic features such as being *emotionally charged*, *identity-defining*, or *community-building*.[35]

Consider for example the norm that I ought not to believe the clock at the post office because I know it to be broken. This norm serves as a tool to navigate the environment in the sense that it is operative in my decision making and it shapes my actions. If, for example, I am at the post office and I need to plan when to go to buy groceries, I will not form my beliefs by relying on the clock. The norm also functions as a premise for both practical and theoretical reasoning where the former involves reasoning with respect to "what to intend" and the former amounts to reasoning with respect to "what to believe" (Harman 1997: 434). An example of practical reasoning is, for instance, someone intending to quit smoking on the premise that smoking is an unhealthy habit. Similarly, one could rely on the very same belief that smoking is unhealthy with respect to theoretical reasoning and conclude that one's own favourite restaurant, which just made the decision to designate some areas smoke-areas, made the right decision.

The norm according to which I ought not to believe the post office clock can be utilized as a premise for practical reasoning. I might, for example, avoid making decisions based on the time that the clock displays. But it can also be used as a premise for theoretical reasoning in the sense that I may form beliefs on the basis of the premise that I ought not to believe the clock. Instead of forming the belief that it is 3 o'clock, for example, I might form the belief that I do not know what time it is, because I know the clock to be broken.

[35] There might be other features that render a norm loaded. It is not my aim here to cover all features, but only the ones that I take to be salient to fix the reference for discussion.

What I said above should clarify what I mean when I say that bland epistemic norms allow individuals to navigate the environment. Now consider a different epistemic norm: I ought to trust my father. This norm, just like the one I discussed above, allows me to navigate the environment and can be employed as premise for both practical and theoretical reasoning. When, say, I intend to trust my father with something or, when deliberating what to do, then I am employing the norm as premise for practical reasoning. Similarly, if I am quick to believe my father's words when he asserts that, say, the household finances are in good shape, then this is an instance of theoretical reasoning where I form a belief by relying on the norm that I ought to trust my father. From what I said so far it may seem that the two norms I discussed—that I ought to distrust the clock at the post office and that I ought to trust my father—are of the same kind. Yet the two show one radical difference. The difference is that the former norm seems to be one whose whole function is to be employed as an instrument to navigate the environment, while the latter norm is *not* merely an instrumental tool, but it plays additional roles.

For one, the norm that one ought to trust their father speaks to the person's identity. If, say, one exhibits behavior indicative that they indeed distrust their father, then not only might this cause the agent to experience guilt directed at the action itself, but it might also cause the agent to experience shame as they wonder what kind of child would distrust their parent. In this sense, the norm plays an informative role as to what kind of person one is. The norm that one ought to trust their father also modulates access to a particular community, specifically the community of family members who share the same norm and who are bound together by the same conviction that their father is an honest, good person.[36] Participating in this community may carry a sense of belonging that triggers positive emotions in the individual. Because of these emotions involved, maintaining the norm that one's father

[36] The notion of community will have to be carefully qualified. If the notion of "community" is too weak, then the risk is that every norm will be loaded.

ought to be trusted is also an emotionally charged norm. Thus, acting in ways that run contrary to this norm, as well as revising or abandoning it, results in a process whose emotional cost is particularly high.[37] From this it should be clear that appropriate ways to detect whether a norm is loaded involve examining their i) revision-process and ii) the emotional cost of behavior that is inconsistent with the norm. Let me now pause to explain i) in more detail. (I will explain ii in Chapter 3).

In general, the process of revising doxastic attitudes involves updating attitudes when new information calls into question the attitudes' truth-values. Peter Gärdenfors 1988 provides the following example to clarify how the revision-process takes place with respect to beliefs. Suppose one purchases a ring that they believe is made of gold on the basis that the seller assured them it was made of 24 carat gold. While repairing their boat, they notice that the "sulfuric acid" that they are using stains the ring (Gärdenfors 1988: 1). They remember from their chemistry classes that the only acid that "affects" gold is aqua regia (Ibid.). Now, they believe that their ring is made of gold, but they also know that the only thing that can stain gold is aqua regia. The fact that the ring is stained implies that it is not made of gold. The proposition that the ring is made of gold is justified by the jeweler's guarantee while the proposition that it is not made of gold is justified by their knowledge of chemistry. Since their knowledge of chemistry provides a better reason to abandon their previous belief that the ring is made of gold, their now believe that my ring is not made of gold. [38]

[37] This however raises the question of whether a violation of a loaded norm can be non-emotionally laden or more generally, result in a lack of investment on behalf of the agent. Consider Bill, for example. Bill has very strong commitments to religious values and presumably, some epistemic norms associated with those commitments are loaded. Suppose he violates those norms, yet he has very little occurrent feeling because Bill is simply an emotionally "cool" person. This is a case of a violation of a loaded norm that is not emotionally laden. However, I doubt a violation of a norm that is loaded in the way I describe it is not followed by any emotional tension. It is implausible that a norm that is either emotionally charged, community-building, or identity-defining if violated does not generate any reaction on behalf of the agent. If that is the case, as it seems to be for Bill, then the norm he holds might not be loaded at all, and might simply amount to some other kind of attitude.

[38] This example is an adapted and simplified version of the one that Gärdenfors provides. See Gärdenfors 1988: 1 for a more detailed exposition.

The example shows how the agent revised their belief that their ring is made of gold when faced with information that calls that belief into question. They hold a belief that their ring is made of gold; they receive a new epistemic input (the ring is stained) that, if true, would lead to the belief that the ring is not made of gold. In order to preserve consistency, the agent revises their belief set by abandoning one of the beliefs; in particular, the one that they take to be less reliable (the one based on the jeweler's assertion). The belief that is less reliable stands at the periphery of their web of beliefs while the other more reliable one stands in the center.

Epistemic norms that more tenaciously resist revision are likely to be loaded ones, while norms that are more easily sacrificed during revision are likely to be bland norms. By "resisting revision" I mean that the revision process is not automatic. That is, because the norm is loaded the agent has an investment in the norm and thus an interest in not revising it.[39] Due to the agent having an interest in maintaining the norm unaltered, the revision process is experienced as painful for the agent. Thus, in order to revise or abandon the norms the agent may need a more substantial amount of evidence than they would need if the norm to be revised was bland. If, for example, one holds the norm that one ought to trust one's own father and is later confronted with evidence that the beliefs formed in accordance with that norm are false, then this calls into question the norm itself. The reason why is that the falsity of the beliefs formed by following that norm indicates that one ought *not* to trust their father. But if the norm is loaded and speaks to the agent's identity and is also emotionally charged then its revision would entail dramatic, unpleasant consequences. As a result, because its revision is so costly, the agent may demand an increased amount of evidence in order to proceed to revise the norm than they would require if the norm was instead bland.

[39] I am not saying that *all* norms that resist revision are loaded norms. There might be other reasons that explain why some norms are not easily revised. My point is that when identifying the reasons that explain why certain norms might be difficult to revise, one should also contemplate that those norms *might* be loaded. That is why I say that norms that resist revision are *likely* to be loaded norms.

To summarize what has been said so far, bland and loaded epistemic norms can be defined in the following way

> Given an agent S who holds the epistemic norm *n*, *n* is a *bland epistemic norm* iff *n*'s main and only function is to serve as a tool for S to navigate the environment.[40]

> Given an agent S who holds the epistemic norm *n*, *n* is a *loaded epistemic norm* iff *n*'s function is to serve as a tool for S to navigate the environment (call this the *doxastic function*), and *n* plays at least one of the following additional roles (call these *non-doxastic functions*):
>
> a) *n* speaks to S's identity in important ways;
> b) *n* builds and modulates access to a specific community (or specific communities);
> c) *n* is emotionally charged.[41]

The relationship between the doxastic and non-doxastic function is conjunctive. If an epistemic norm has a doxastic function, but lacks any non-doxastic function then that norm is not loaded, but qualifies only as a bland norm. If instead a loaded norm does not play any doxastic function but satisfies at least one among a), b), and c) then that norm cannot be said to qualify as a proper epistemic norm. This is because I take it to be constitutive of a norm that it plays *at least* a doxastic function. Thus, if a norm does not play any doxastic function it cannot be said to be a norm. In this sense, an attitude that is only emotionally charged, speaks to one's identity in important ways, and modulates access to communities—in other words, an attitude that plays *exclusively* non-doxastic functions without serving as a tool to navigate the environment—does not amount to an epistemic norm, but rather to another attitude altogether (a desire, for example).

[40] By "environment" I mean both *physical* and *social* environment. If, for example, someone holds the norm that they ought to wear jeans at the synagogue, then this norm is a loaded one in the sense that it plays a doxastic function (it helps the agent navigating their *social* environment) but it also plays a non-doxastic function because it allows them to be part of a community (i.e. the community that gathers in the synagogue).

[41] Again, a), b), and c) are examples of non-doxastic functions. I am open to adding more conditions if needed.

If instead a norm plays a doxastic function and, in addition, also plays a non-doxastic function, but satisfies *only one* among a), b), and c), then that norm can still qualify as loaded. It is sufficient that it plays some sort of non-doxastic role to attribute the label of loaded norm. This implies that there may be norms that satisfy all three conditions, and some that only satisfy one. The norm that I ought to carefully scrutinize evidence before embracing a belief may a) speak to my identity (if I am a scientist, for example) and b) allow me to access the scientific community; while, say, the belief that I ought to trust my best friend may only satisfy the condition of being emotionally charged (it is not clear that it would allow me to access particular communities or contribute to my identity in crucial ways). In this last case, one may question whether the norm even deserves the label of loaded, considering that is only satisfies one requirement. I argue that satisfying only one requirement is indeed sufficient, and this will prove to be useful in Chapter 3 to explain why some self-deceived individuals experience a lesser *degree* of the mental tension characteristically associated with the phenomenon, when it comes to violating epistemic norms that are loaded.

2.3 How Loaded Epistemic Norms Differ from Motivated Beliefs and Besires

Even though I remain neutral with respect to the kind of doxastic attitude that loaded norms amount to, a legitimate worry that stems directly from the definition I provided is whether they are different from instances of *motivated* attitudes (and motivated beliefs in particular). I think there are good reasons for keeping loaded norms and motivated beliefs distinct. Usually, the notion of ‘motivated’ is predicated on the etiology of the belief involved. When assessing whether a belief is motivated, its motivated aspect is meant to indicate that something went awry in the formation of that belief. Alfred Mele, for example, argues that motivation is able to “bias” the way we form beliefs (Mele 1993: 25). In this sense, a motivated belief is one whose formation has been unduly influenced by non-epistemic factors. With respect to motivated beliefs, there does not seem to be anything particularly notable about the beliefs themselves. They are regular beliefs acquired in an epistemically problematic way, and that is what makes them motivated.

Now, loaded norms can certainly be formed in epistemically problematic ways but while the distinctive feature of motivated beliefs is their flawed etiology, this is not the case for loaded norms. Loaded norms are not merely norms whose formation may be unduly influenced by non-doxastic factors (although they might be), they are norms of a certain *kind*. They are identity-defining, emotionally-charged, and community-building. Loaded epistemic norms are norms that *can do* many things, so to speak, for their holders. While with respect to motivated beliefs the notion of 'motivated' refers to *how* beliefs are formed, the notion of 'being loaded' is attributed in relation to the kind of norms they *are*. This is in fact what my discussion aims at capturing: it aims at fleshing out not *how* loaded norms are formed but *what* they are. Thus, the difference between motivated beliefs and loaded norms rests on the idea that while a flawed etiology exhausts what motivated beliefs are, it does not exhaust what loaded norms are.

Similarly to the objection about motivated beliefs, one might also ask how loaded norms differ from *besires*. According to Humean accounts, emotional and non-doxastic features are not constitutive of beliefs, which are instead motivationally inert. Desires, on the other hand, motivate actions. A besire is a belief that *p* accompanied by a desire that *p* (Smith 1994). Now, if the norm "I ought to trust my father" motivates its holder to behave accordingly, then it seems that my definition of loaded epistemic norms is dangerously similar to the one of besires. That is, in order for a norm to be motivationally operative it must amount to a belief coupled with a desire. Earlier, I have defined loaded epistemic norms as norms that both motivate the agent and are emotionally charged. Yet, I take besires and loaded norms to be different.

When defining loaded norms as norms that can be emotionally charged, I do not mean that loaded norms *themselves* possess non-doxastic features, rather I refer to what loaded norms can *do* for those who hold them. In particular, loaded norms are able to give back to those who hold them what I have called *non-doxastic returns*. It follows that by emotionally charged norms I do not mean that the norm *itself* is emotionally charged, but that there is an exchange that occurs between the norms and its holder. The trade-off is that *if* an agent S holds a norm *n*, the norm *n* is emotionally charged if holding *n* gives back to the agent non-doxastic returns in the form of positive emotions. If I hold the norm that I ought to trust my father, I will get positive feelings in return. That is, forming beliefs in accordance with the norm that I ought to trust my father will, in return, make me feel good about myself. It will preserve my self-identity in allowing me to think of myself as a good daughter, it will modulate access to my family, and generate positive emotions. Yet while besires motivate action, loaded norms do not.

To clarify, consider pain relievers. When one says that the common medicine known as Advil is a painkiller, one does not mean that Advil *itself* is a painkiller, rather one more properly points to the interaction that occurs when one takes the medicine. *If* one takes Advil, *then* Advil may relieve pain in return. Now, loaded norms are different from besires because while it is constitutive of besires to have non-doxastic properties, the non-doxastic returns involved in loaded norms are not constitutive of loaded norms, rather they are the outcome of the exchange.

A further question is how to explain the etiology of loaded and bland norms. It may appear, as I have said earlier, that defining norms such as "I ought not to trust the post office clock" and "I ought to trust my father" as bland and loaded ones is an arbitrary move. I want to now briefly explain how norms may come to be loaded or bland. To start, there are no objective criteria according to which norms qualify as loaded or bland; what constitutes bland or loaded norms is *agent-relative.* To use an example, it is not uncommon for individuals to attribute emotional meaning to mundane objects. A blanket may come to acquire affective meaning because, say, it is inherited from a deceased loved one even though to the eyes of an external observer it might be *just* a blanket. Norms are no different. Whether a norm is bland or loaded does not depend on the norm's content, but on the agent's attitude towards the norm. Some epistemic norms are bland in the sense that the agent does not have any particular attitude towards them, others are instead loaded; they may define the person in important ways or carry some sort of emotional attachment. What helps identify whether a norm is loaded or bland is *in part,* but not only, is to observe the kind of reaction the agent shows when the norm is violated. If this reaction is strong, so to speak, then this norm can be said to be loaded. If instead, the agent shows a relatively unremarkable reaction then this is indicative that the norm involved is bland. An example will clarify this point.[42]

Suppose I hold the epistemic norm that I ought not to form any factual belief from fiction. Yet, while reading *The Legend of Sleepy Hollow* I might form the belief that Sleepy Hollow is an actual town. A friend points out to me that the town is simply fictional and does not actually exist. In discovering that my friend is right, I might revise my belief and realize that in believing Sleepy Hollow was real I formed a belief that is at odds with my epistemic norm. I may quickly revise it and abandon it without feeling any particular emotion for having violated my norm. In fact, I may even laugh in thinking that I have been so gullible as to believe that Sleepy Hollow really existed. My reaction is indicative of the fact that the norm I hold is not a loaded one. It is a norm that I do hold, but that I am not attached to in any particular way. It does not define who I am as a person, nor do I feel an emotional attachment to it. It simply serves as a tool to navigate the environment. Thus, violating it does not constitute a significant transgression for me.

Now consider a different situation. Suppose I hold the epistemic norm that I ought to be rational. Suppose also that I am a scientist and thus the norm that I ought to be rational is a loaded one for me. It defines who I am as a person and it is one that I care about. Suppose also that I am living through a particularly stressful time of my life when nothing seems to be going right. Talking to

[42] However, the agent's reaction may not be observable to someone else.

my grandmother, she suggests putting an amulet on my key ring for good luck. In a moment of weakness, I buy the amulet and form the belief that things will get better. Later, a friend points out that my belief is irrational for there is no evidence that an amulet will improve difficult situations and that it is unlike me to give in to superstitious beliefs.

In acknowledging that my friend is right, I may not only abandon the belief that an amulet will improve my situation, but I may also feel ashamed for having violated my epistemic norm that I ought to be rational. This reaction is obviously sparked by the fact that the norm in question is a *loaded norm* that is crucially important for me.

3. How Epistemic Norms Operate in Self-Deception

In this final section I explain how epistemic norms factor into self-deception. So far, I have argued that agents' epistemic norms are patterns of belief formation that in addition also *guide* agents' epistemic behavior with respect to how they think they *ought* to form beliefs. Thus, epistemic norms may govern agents' behavior without agents having to form judgments about whether a situation requires following a specific norm. Agents' epistemic norms are norms that can be held implicitly and are closely tied to motivation and action. They are also regulative, operative in their epistemic deliberation. They are norms of the sort of 'I ought to believe experts'.

I have said that agents' epistemic norms guide their beliefs. If Mia holds the norm 'I ought to believe experts', then this norm motivates her to act in accordance with it. Thus, if she is presented with a situation where an expert affirms that smoking causes cancer, Mia infers that smoking causes cancer. Note that, as Doug Kremm also argues, this does not guarantee that Mia will reliably believe experts; it only ensures that *if all goes well*, she is always going to believe experts. In particular, it shows that if Mia's epistemic norm has not changed and she fails to believe experts, then she is committed to seeing her having violated that norm as a "malfunction" (Kremm (MS): 19).[43]

3.1 Normal vs Abnormal Situations

Let me now introduce a distinction between what I call *normal* and *abnormal* situations concerning agents' epistemic norms. Situations where all goes well are situations that I call *normal* situations. I use the expression "all goes well" to capture those situations where the agent faces no *interference* or *impediment* that prevents them from believing in accordance with their epistemic norm.[44] Forming a belief in accordance with the agent's own epistemic norm is the outcome of a process that consists in several stages. The first stage involves the agent holding a specific epistemic norm *n*, say, the norm that involves believing experts. The second stage involves the agent confronting a situation where the deployment of that norm is called for. A situation of this sort could be Sally confronting a scientist declaring that smoking causes cancer. The third stage involves the epistemic norm kicking in, and the fourth stage results in Sally coming to believe that smoking causes cancer. A normal situation is one where this process unfolds *smoothly*.

To use a different scenario, consider a young lawyer who just joined a law firm. Suppose she holds the norm "I ought to trust my skills". A normal situation would be one where she proceeds to trust her skills, without any particular interference that prevents her from doing so. But suppose that, during the opening of her first trial, she is suddenly assailed by imposter syndrome thoughts.[45] This is an abnormal situation in the sense that the following of her norm is rendered more difficult by interfering elements. In normal situations agents *always* follow their epistemic norms. Abnormal situations instead, are situations where the process described in terms of stages that I explained earlier

[43] I borrow this point from Doug Kremm who makes a similar remark with respect to practical commitments and *akrasia*. I believe that, with some tailoring, his view can also be applied to instances of self-deception. See Kremm (MS).

[44] I am again, indebted to Doug Kremm, who makes a similar argument about practical akrasia. I borrow his argument for the case to be made for epistemic norms, with some tailoring due. See Kremm (MS): 5-7 for a more detailed description of his view.

[45] A similar version of this example can be found in Sarah Stroud's and my paper "Self-Control in Action and Belief" (*Philosophical Explorations*, forthcoming).

fails to proceed smoothly. That is, abnormal situations are situations characterized by interfering elements that may potentially decrease the likelihood that the agent comes to form a belief in accordance with their epistemic norm.

Self-deception is an *abnormal* situation where the interfering element is constituted by the agent's interests and desires.[46] For example, suppose that Sally holds the norm that one should believe experts. She is later confronted by a situation where the application of her norm is called for. A situation of this sort is Sally witnessing a scientist declaring that smoking causes cancer. The epistemic norm kicks in, but Sally is also a heavy smoker and has an interest in believing that smoking does not cause cancer. Together with the epistemic norm kicking in, her desire that smoking not cause cancer *also* kicks in. The desire represents precisely the motivational element that renders the process of Sally coming to believe that smoking causes cancer more laborious. Sally's epistemic norm would lead her to believe that smoking causes cancer, but her motivational state would lead her to believe otherwise.

In this sense, Sally's desire is an element that interferes with the process of her coming to follow her epistemic norm. Of course, Sally does not necessarily have to conclude that smoking is safe. She can perfectly come to believe in accordance with her epistemic norm, no matter how powerful her desire is. Yet, from her perspective it is tempting to believe that smoking causes no harm, and it is with this in mind that I suggest that it is likely that Sally will give into the temptation of distrusting experts, unless she successfully exercises *self-control*.

3.2 Self-Deception and Self-Control

I take self-control to involve both *awareness* and *intervention*.[47] With respect to awareness, I take the agent's awareness of her own motivational states, and the influence that these may exert on the interpretation of evidence as a background condition for exercising self-control. That is, exercising

[46] See Kremm (MS): 19 where he makes a similar remark with respect to carrying out "practical commitments".

[47] These two aspects of self-control are somewhat similar to what Kremm (MS) calls "practical" or "operative awareness". See Kremm (MS): 20 for a more detailed discussion.

self-control becomes a live option only if the agent is aware that their desires, wishes, interests, and so on may unduly shape their evaluation of evidence with respect to a specific proposition. Let me briefly explain this last claim.

I take self-control to be called for only when framed against the background of a temptation, otherwise any mundane action such as eating, walking, and taking our parents to the doctor would count, mistakenly, as a self-controlled action. A factor constitutes a temptation when it influences an agent "whose influence, or degree of influence, they do not endorse". An example is one being tempted to indulge in an extra slice of chocolate cake that exerts an influence on them they do not approve of.[48]

With respect to self-deception, the proposition according to which smoking is safe constitutes a temptation for Sally. Being a heavy smoker, coming to believe that smoking does not cause cancer is an alluring prospect, one that would be very tempting to believe was true. In this sense, this alluring perspective can be said to influence Sally's examination of evidence, similarly to how the way a slice of chocolate cake exerts its influence. Sally, however, does not endorse this influence for her epistemic norm involving trust in experts pulls her, so to speak, into the opposite direction.

Temptation as well as awareness of one's own motivational elements are the background conditions for self-control to prevent self-deception from occurring. Now one might point out that my suggestion that one can avoid self-deception by exercising self-control is problematic because while one is typically aware of exerting self-control, self-deception can come about without awareness. Self-deception does happen without awareness, but the notion of self-control that I propose does not require awareness of self-deception. Rather it posits that the self-deceived person be aware of the motivational valence that is attached to the comfortable proposition in question. And this is not an

[48] I borrow this definition from the examination of self-control and temptation Sarah Stroud's and my paper "Self-Control in Action and Belief" (*Philosophical Explorations*, forthcoming).

unreasonable request. The self-deceived individual can be perfectly aware of their own wishes and interests, and they in fact often are. Awareness of one's own motivational elements is thus a necessary condition for avoiding self-deception. If one must exert self-control over one's own motivational elements, then this certainly requires the agent to be aware of those. And if one is aware of the valence of the comfortable proposition, one should be vigilant. One may not be aware of being in the process of self-deceiving, but one should be aware that it may be a possibility.

However, awareness of one's own motivational elements is not a sufficient background condition for exercising self-control, for two reasons. First, self-deceived individuals are not aware of *how* those motivational elements *bias* their evaluation of evidence in favor of a particular proposition, and second, being aware of one's own motivational states alone is not a sufficient background condition for exercising self-control because it is not the case that every time an agent wishes that *p* does this wish directly influence their coming to believe *p*.

Thus, a notion of self-control that successfully facilitates resistance to self-deception will have to tackle these last two aspects and prevent motivational states from biasing the examination of evidence. How to control motivational elements from biasing one's own epistemic investigation, considering that we cannot *directly* control our biases? I suggest that the notion of self-control we should adopt is one that does not control the interference of motivational elements *at the moment* in which the interference is occurring, but rather the notion of self-control that we should adopt is one that controls the bias *indirectly* by shaping the environment of the individual in such a way to prevent the epistemic investigation from being biased.

A similar idea is used when preventing instances of implicit bias. Individuals are often unaware that they may have implicit bias and for this reason controlling their biases at the very time they are occurring is challenging. But this does not mean that implicit bias cannot be controlled at all or that there cannot be any awareness that can facilitate such control. One can be aware that, because implicit bias is so widespread, they may have implicit bias towards certain groups of people and when encountering a member of that group, one may employ resources that shape their rationality before the bias is triggered. One may, for example, slow down their thoughts, carefully focusing on the details of the situation. In this way, one is not controlling the bias at the moment that it is occurring; rather one is controlling the environment so that the bias is prevented from being triggered. I suggest that a similar reasoning can be applied to self-deception. Let me now explain my suggestion in more detail.

Self-deception is not problematic because agents hold a variety of desires, hopes, and wishes. Rather, self-deception is problematic because, and when, agents *let* those desires, hopes, and wishes unduly guide their coming to believe a certain proposition. In light of this, the background condition for self-control to be employed is the agent's awareness that in those circumstances where the agent has an interest at stake, their motivational states may shape their belief-formation in epistemically pernicious ways.[49]

In light of this, employing self-control involves some sort of intervention aimed at resisting giving into a temptation. In the domain of action, this intervention usually takes the form either of an act of will or environmental control. The former involves a resolution to, say, not reach out to take the extra slice of cake. This form of self-control is a synchronic resource that directly influences the course of action. The latter instead involves strategies the agent may deploy in order to prevent themselves from facing a potential temptation in the first place. An example is someone who knows they have a weakness for carbs refusing to have bread in their house. The temptation to eat bread is less likely to arise if there is no bread to be found. This form of environmental self-control is not a sheer act of will, but a resource aimed at neutralizing the influence of a tempting factor indirectly.

In order to avoid falling into the grip of self-deception, the individual must exercise self-control in order to resist the temptation of believing a comfortable proposition. In our familiar case

[49] I advance ways in which the self-deceived agent can gain this awareness in Chapter 4.

of Sally, who is confronted with the reality that smoking causes cancer, exercising self-control involves resisting the temptation of distrusting experts and instead sticking to her epistemic norm of believing experts, thus forming the belief that smoking poses a risk to her health. Before explaining what self-control with respect to one's own norms involves in the domain of beliefs, I wish to give an example of how agents exercise self-control with respect to norms that pertain to the domain of action as it will be a helpful reference for the epistemic counterpart.

Suppose I hold the norm that 'I ought to never raise my voice'. Suppose that I always follow that norm, but while having an argument with a friend I find myself raising my voice without even realizing it. This situation—having an argument—is a situation that is abnormal in the way I have characterized above. It is a situation where there are some elements at play, in this case anger, that interfere and make it more difficult for me to follow my norm. Elements that can interfere and eventually undermine our carrying out of our norms are varied: desire, anger, hate, hope, and so on.

The point that I wish to make here is that much the same can be said about the case to be made for self-deception. Recall the example of Mia who holds the epistemic norm of believing experts. Suppose she is also a heavy smoker. When she hears experts declaring on TV that smoking causes cancer, Mia has an interest in smoking not causing cancer, and here lies the trouble: following her norm, she would get a result that she cannot stomach. Mia is thus experiencing a temptation to believe the more comfortable proposition that smoking does not cause cancer, and if she does—if she believes that smoking does not cause cancer—then she is violating *her own* epistemic norm and can be considered self-deceived.

Recall that an abnormal situation involves motivational elements that interfere and make it more difficult for the agent to follow their epistemic norm. A normal situation is a situation where such elements are not present. Self-deception is an abnormal situation where there is a motivational element that decreases the likelihood that the agent will stick to their own epistemic norms. The problem in self-deception then, is the agent's failure to exercise self-control successfully with respect to following their own epistemic norms and this failure occurs without the agent realizing it.[50] As Van Leeuwen says, we "slide" into self-deception, and similarly, I say, we *slide* into violating our epistemic norms (Van Leeuwen 2008: 196).

This does not mean that we do not *do* anything when we self-deceive. Self-deception is not a *mistake* that merely *happens* to us. It is, however, something that self-deceived individuals *let happen*. Recall the example of raising one's voice during an argument. It often happens that the individual does not realize that they are raising their voice *at the moment*, but only after their interlocutor points that out. This means that presumably they do not know at the moment that that is what they are doing; nevertheless, they let it happen. Had they succeeded in exercising self-control prior to that, they would not have raised their voice. But they did not succeed. They *let* their anger interfere, failed to stick to their norm and raised their voice.

The same occurs with self-deception. If Mia had successfully exerted self-control, she would not have *let* her interest cloud her epistemic practice to the point of violating her norm of believing experts. The agency involved in self-deception is thus a failure to exercise self-control.[51] In this sense, we do not intentionally decide to violate our norms, just like we do not intentionally decide to raise our voice, but we *let it happen*. Of course, we do not need to exercise self-control in order to follow our norm of believing the experts when the experts are giving us truths we appreciate; we need it when they give us truths we cannot stomach.

[50] Whether the agent *attempts* to exercise self-control or indeed does not even try is irrelevant from my argument for they both constitute a kind of incontinence. Aristotle in *NE 7.7* calls the former "impetuous" and the latter "uncontrolled".

[51] Note, however, that in self-deception the agent *has* the ability to exercise self-control, at least in the sense of what Mele calls "simple ability" whose sufficient condition involves an agent A-ing at a time, where A-ing is the action of exercising self-control (Mele 2003: 448). This is important because where we could have exercised self-control and failed to, is a situation where we can be deemed responsible for the failure. Hence, self-deception does not rule out moral responsibility.

This line of reasoning raises a worry of inconsistency. If a failure of self-control is what leads to the violation of one's own norms, then this seems to imply that, conversely, one needs to exercise self-control in order to follow one's own epistemic norms. The problem is that exercising self-control requires conscious effort, while earlier I said that we follow our own epistemic norms without any need to even think about it suggesting a contradiction.

The worry of inconsistency though, relies on a misunderstanding. It is true that no conscious effort is required in order to follow our norms, but this applies only to normal situations. In normal situations, we follow epistemic norms automatically, like habits. And this makes sense: when everything goes well, we do not need to consult a textbook in order to know what to do.[52] We already know how to proceed because we have internalized our norms. It is in abnormal situations, where things are atypical, that we do not know what to do and we go back to the textbook for guidelines. Going back to the textbook is a metaphorical expression that encapsulates what I mean by exercising self-control. Self-control is not called for when everything goes well; it is called for when things *do not* go well. We do not need to exercise self-control in order to stick to the norm that prescribes that we not raise our voice when we are having a pleasant conversation with our friend; we do need it when we are having an argument with our child and she is getting on our nerves.

The problem in self-deception is not so much that the self-deceived subject does not hold their epistemic norm, for the self-deceived person *does* hold them; rather they fail to stick to them.

[52] Pollock 1987: 66 makes a similar remark when arguing that "norms can govern your behavior without your having to think about them".

CHAPTER 3

Beyond Minimalism

Now that I have explained in more detail my preferred theory of self-deception, the doxastic violation account, in this chapter I want to defend its superiority by arguing that only by adopting this specific type of belief-based deflationism, can one meet the challenges and respond to the objections levelled against the theory that I discussed in Chapter 1. Before turning to my argument, I want to briefly recap the challenges and objections that minimalism faces. Recall that minimalist accounts face a *phenomenological* and an *irrationality* challenge, and are vulnerable to the *passivity* and *uniqueness* objection.

Belief-based deflationism, and minimalism in particular, faces a *phenomenological challenge* because it is not obvious how it can capture the mental tension typical of self-deception. This is challenging because of the theory's commitments that self-deception results in a belief and it occurs without the self-deceived person being aware of it. If belief-based deflationism is correct in claiming that the product of self-deception is a belief, then this portrays the self-deceived person as someone who genuinely and stably holds their comfortable belief. Yet if the self-deceived person has made up their mind, then it is unclear why they would experience any internal conflict. Minimalist accounts specify that the self-deceived person *suspects* that their unfriendly belief might be true, but this does not suffice to explain the experiential tension. A suspicion may not be a sufficiently strong enough doxastic attitude to capture the conflict that nags the self-deceived person, for suspicions can be easily rationalized away. Furthermore, minimalism (as well as all belief-based deflationist accounts) posits that self-deception occurs without any awareness on the part of the individual, which pushes back the question of how to account for the experiential tension. If the self-deceived person believes not-*p* and thinks that this belief is simply the result of a regular epistemic investigation, then it is hard to see how one could account for the internal conflict.

Due to its characterization of self-deception as a motivationally biased hypothesis testing minimalism also faces an irrationality challenge. One of the few uncontroversial claims within the self-deception literature is that it is a phenomenon of internal or subjective irrationality[53] where the self-deceived person does not merely violate "canons of good reasoning" but their own "standards of epistemic rationality" (Scott-Kakures 1996: 34). That is, the self-deceived person believes "what [they] also believe to be ruled out epistemically by what [they] believe" (Ibid.). Capturing the subjective irrationality of self-deception is a challenge that *any* theory of self-deception must meet, and minimalism has difficulty accommodating it.

By arguing that what goes awry is the way the self-deceptive belief is formed, minimalist accounts locate the irrationality of self-deception in its *etiology*. In doing so, minimalism renders unclear in what sense self-deception is a phenomenon of *subjective* irrationality. While it may not be rational to form a belief under the influence of desires and interests, a motivational bias is not *per se* a form of *subjective* irrationality. Mere motivational bias is too thin a notion here because it does not necessarily involve a violation of the self-deceived person's own epistemic standards.

The passivity objection argues that because belief-based deflationism, and minimalism in particular, commits to the claim that the individual unknowingly slides into their self-deceit, the resulting picture of self-deception is of a *passive* phenomenon that does not involve any agency on behalf of the individual. The drawback is that conceiving of self-deception in these passive terms compromises a rather intuitive way to think of the phenomenon as mirroring ordinary deception, which is by definition intentional.

[53] I use the terms "subjective" and "internal" interchangeably.

Ordinary deception involves an agent A, who believes *p,* deliberately deceiving an agent B into believing not-*p*. in light of this, self-deception is simply the intrapersonal version of ordinary deception, where A and B are the same person (Davidson 1985, Audi 1997). Characterizing self-deception as intentional is intuitive because it does justice to the idea that the agent is involved in self-deception in a way that is similar to A deceiving B. However, intentional self-deception also raises many issues, which has in part motivated theories like deflationism to *deflate* the phenomenon's intentional element. Yet, if self-deception is non-intentional, then it is unclear what the agent's role is if the individual does not actively cause their self-deceit. In this sense, deflationism, and minimalism in particular, seems to erase the *self* from self-deception rendering the phenomenon a condition that simply *happens* to the individual – hence, the passivity objection.[54]

The idea behind the *uniqueness objection* is that if self-deception is simply a non-intentional process that results from a motivationally biased hypothesis testing, as minimalism argues, then it is unclear how it is any different from other phenomena that also involve an epistemically problematic etiology, but do not qualify as instances of self-deception. For example, *confirmation bias* involves a biased examination of evidence in favour of a preferred belief, yet it is importantly different from self-deception. Wishful thinking involves an agent *believing* what they instead merely *desire*, but this does not warrant the conclusion that wishful thinking and self-deception are the same (Szabados 1974, Bach 1981).

In the next section I show that the doxastic violation account meets the challenges and objections faced by minimalist accounts.

[54] This objection also applies to pretense accounts but given that I have already rejected pretense accounts on the basis of other reasons, my target here is belief-based deflationism, and minimalism in particular.

1. Meeting the Phenomenological Challenge

As I mentioned in Chapter 1, self-deception is associated with a certain experiential tension.[55] The self-deceived person experiences internal conflict, discomfort, and is nagged by persistent doubts about whether the self-deceptive belief they profess to endorse is true.[56] The presence of such phenomenology is further supported by empirical evidence that shows that tendencies to self-enhancement (i.e. the tendency to take credit for one's own accomplishments while discounting the role played by external factors), which could be construed as a kind of self-deception, may generate "a deep […] sense of uneasiness" (Randall et al. 1995: 1161). Belief-based deflationism, and minimalism especially, has a hard time making sense of this tension because the view argues that self-deception results in an outright belief. That is, if the self-deceived person has made up their minds about what they believe, it is unclear why they would experience any tension.

The doxastic violation account is well-equipped to explain the mental tension. In particular, I suggest that the persistent doubts that plague the self-deceived person can be explained by appealing to the idea that the self-deceived person holds a belief that is at odds with their own epistemic norms. An individual's epistemic norms are norms that the individual holds, believes to have at least a *pro tanto* reason to follow, and obeys without needing to exert any conscious effort. Because the agent is naturally disposed to believe in accordance with their epistemic norms, violating them generates a similar result to when one breaks a *habit*. Consider for example the common habit of shaking hands when meeting someone. Suppose you decide to unlearn the habit of shaking hands due to health measures imposed by the current Covid-19 pandemic. You know that breaking the habit of shaking hands is a justified one, yet knowing that does not prevent you from nevertheless being *disposed* to

[55] Self-deception is also associated with a behavioral tension. See Funkhouser 2005.

[56] The mental tension is not considered a conceptually necessary component of the phenomenon, but it is nevertheless considered characteristic of self-deception. Meaning, most cases of self-deception are accompanied by it. See Losonsky 1997, Graham 1986, Lynch 2012.

shake hands with the next person you meet. This is because one is naturally disposed to follow that habit nonetheless. Furthermore, interrupting the habit of shaking hands generates a certain unease. You might feel as if you are being rude and disrespectful in refusing to shake hands, even though you know it is the right behavior to adopt. This unease is not caused by the fact that you are breaking a habit, but by the fact that you are breaking *your own* habit. You probably would not feel as if you are acting rude if you did not have the habit of shaking hands in the first place.

The same goes for epistemic norms: violating one's own epistemic norms generates a certain phenomenology, which manifests through doubts. These doubts are challenging to dismiss not because the agent violates *an* epistemic code but because the agent violates their *own* epistemic code. Michael Losonsky hints at a similar idea when he says that "it is as if the cognitive mechanism cannot help but respond to the force of the possessed evidence, although the motivational structure is able to override it" (Losonsky 1997: 122). In this sense, the self-deceived person feels plagued by doubts not because unfriendly evidence is "dangerously close at hand", as has been famously argued, but rather because the individual's epistemic norms are their own and when they are violated, they fight back, so to speak (Bach 1997: 105).

It is the agent's epistemic norms then that cause the individual to be sensitive to the force of good evidence because the evidence is deemed good in virtue of the norms. So the norms identify the evidence that one ought to pay attention to and since the force of good evidence goes against the agent's friendly belief, the individual is not entirely certain about whether their self-deceptive belief is true or false. All things being equal they would be certain, but their motivation interferes with the following of their norms. The fact that the self-deceived individual is violating their own norms also explains the *phenomenology* of the mental tension. The anxiety and worry that the self-deceived person experiences are caused by the fact that, as Van Leeuwen says, the agent is violating *their own standard*, rules that *they themselves* hold. Just as there is a phenomenology associated with instances where we violate our moral code and we act wrongly according to the norms we already have, so there is a phenomenology associated with its epistemic counterpart. This point becomes especially clear if we also think that epistemic norms, just like moral norms, can be *loaded*.

A loaded epistemic norm is a norm whose content consists in a loaded doxastic attitude. As I argued in Chapter 2, loaded attitudes are those that can be emotionally charged, contribute to our personal identity in important ways, and can modulate access to particular kinds of community. Epistemic norms gain loaded features, that is, non-doxastic functions, over time. Factors that may contribute to norms acquiring non-doxastic functions may involve, for example, the kind of job one has. The epistemic norm "I ought to trust scientific evidence" may become particularly loaded if, say, it is a scientist who holds the norm. Believing that one ought to defer to scientific evidence may be a loaded attitude insofar as, among other things, it allows a scientist to be part of a community of fellow scientists. In fact, believing scientific evidence is a shared norm among scientists. Endorsing that norm may be the qualifying factor that maintains the scientist's access to the community. Abandoning or violating the norm would indeed result in being excluded from the community. Along similar lines, if the scientist has a passion for their job, then it may very well be that this norm also carries certain emotional features. Finally, the norm may also contribute to the scientist's personal identity in crucial ways.

The property of being loaded is *agent-relative*, meaning that it is dependent on the individual. The same epistemic norm that may be loaded for an individual, may not be loaded for another agent. The predicate of being loaded also comes in *degrees*, meaning that one epistemic norm may be more or less loaded than another. For a scientist, the epistemic norm "I ought to trust scientific evidence" may be more loaded than the norm "I ought to trust my family members". I myself hold the norm that "I ought to seriously engage with an argument when assessing it", that "I ought to never trust my partner about directions" and that "I ought to be suspicious when my cats act as if they have not been fed".

Yet as a philosopher, the norm that “I ought to seriously engage with an argument when assessing it” may be *more loaded* than the norm that I should not trust my cats. Similarly, with respect to moral norms the norm “I ought not to murder” may be more loaded than the norm “I ought not to lie”. Murdering an individual may result is feelings of alienation with consequent disruption of identity in a way that lying may not.

As I said, part of the reason why some norms are more loaded than others, is that some define who we are as individuals in important ways. One may think of themselves not only as a husband and father, but also, for example, as the kind of person who does not lie. I may think of myself as a philosopher and also as somebody who seriously engages with an argument when assessing it. Holding a loaded norm that defines the agent’s identity consequently makes it the case that the agent has some sort of investment in the following of that norm. Some norms contribute to our identity so crucially that one may have more than a mere epistemic interest in acting in accordance with them. Those emotional features that epistemic norms can carry make it the case that we may *care* to act in accordance with them because, given that they define who we *are*, failure to do so may result in our becoming an individual who acts in a manner we disapprove of.

1.1 Why the Doxastic Violation Account’s Explanation of the Mental Tension is Superior

The view that I have defended, according to which the tension in self-deception is explained by the self-deceived agent violating their own loaded epistemic norms, is superior to minimalist accounts and even those proposals aimed at improving the theory in order to accommodate the tension. The most prominent proposal comes from Kevin Lynch (2012). As a reaction to the criticism that minimalism struggles to explain the mental tension, Lynch advances an improved version of the theory aimed at capturing the tension.

Lynch endorses minimalism, but suggests that the claim according to which the self-deceived person *believes* their comfortable proposition should be replaced by the claim that the self-deceived person possesses a high degree of (unwarranted) *confidence* about the proposition in question. Degrees of confidence track how convinced an individual is of a proposition *p* and allows us to distinguish instances where one has a high degree of confidence that *p* without necessarily believing that *p*.[57] Abandoning talk of belief provides a more promising explanation of the tension because it supports Lynch's claim that the self-deceived person "may not manage to fully convince themselves of what they want to be true" (Lynch 2012: 440). In fact, those who fall victim to self-deception are "normal people", according to Lynch, "intellectually able and rational" thus they are "generally sensitive to the force of good evidence" (Ibid, 442). For this reason, Lynch argues that despite their effort to rationalize, the self-deceived person fails to completely counteract the "unwelcome evidence" (Ibid., 440). Of course, uncertainty alone is unlikely to generate "nagging" doubts, but Lynch argues that the self-deceived person *cares* that a certain proposition is true. Thus, a combination of *uncertainty* and a *stake* in the issue is sufficient to account for the self-deceived person's persistent doubts as well as tension and discomfort (Ibid, 440).

I wish to remain neutral on whether I side with Lynch on the issue regarding degree of confidence *versus* belief, as my position on that specific debate does not impact my argument. Yet my account of the tension departs from Lynch with respect to what *causes* the self-deceived person's appreciation of counterevidence, and it does so by advancing a competing explanation. My account of the mental tension suggests that the self-deceived individual is sensitive to evidence at odds with their self-deception not because they are normal, intellectually able and rational people, but because (and only if) the evidence they encounter conforms to the epistemic norms they already have. Those persistent doubts that plague the self-deceived individual result from them holding a belief that violates their own epistemic norms and not, as Lynch argues, from a combination of uncertainty and a stake.

[57] Lynch admits that the notion of degrees of confidence leaves unexplained how beliefs and degrees of confidence relate to each other and whether beliefs simply are propositions that one is highly confident in. He sets aside these questions as they are not strictly relevant to his project (Lynch 2012: 439).

According to Lynch, the cause of the uncertainty is the sensitivity of the self-deceived person to the force of good evidence, which, in turn, is caused by the fact that the self-deceived individual is a normal person, intellectually able and rational. Lynch's claim seems to imply that if one is a normal person, intellectually able and rational, then they are likely to be sensitive to the force of good evidence. This claim may seem like a truism, yet a closer examination shows that it is psychologically implausible. The reason why, I suggest, is that one can perfectly be a normal, intellectually able and rational person and yet not be sensitive to the force of good evidence. (Lynch does not explicitly say whether by "rational" he has in mind objective or subjective rationality. I show later that his argument is problematic in either case.)

Consider the case of Mia. Mia usually defers to experts. While reading an academic paper, she learns that evidence shows a causal link between smoking and cancer. Taking that piece of evidence seriously, as being good evidence, implies *de facto* deferring to experts. Mia already holds the principle that one should defer to experts; thus judging that evidence to be good aligns with her prefixed belief. As a result, she is sensitive to the force of that evidence. Mia's sensitivity to evidence then depends on whether *she* judges it to be good. Yet whether she judges it to be good—and this is the crucial point—depends on whether her judgment aligns with her norms governing how she typically acquires beliefs. Mia's case shows that her sensitivity to evidence is not a function of being a normal, intellectually able, and rational person, rather it is a function of the encountered evidence matching with her epistemic principles.

Consider now a different case. Frank is usually skeptical towards experts. He believes that experts are pretentious, overly-educated liberals with their own political agenda to pursue. While he is watching TV, a journalist quotes an academic paper that shows clear evidence that smoking causes cancer. Frank, however, is not sensitive to the force of that evidence and fails to form the belief that smoking causes cancer. The reason why is that judging that evidence as good runs contrary to his epistemic norm that experts should not be trusted.

Frank is an example of someone who is a normal, intellectually able, and rational person, and yet not sensitive to the force of good evidence. In fact, Frank *is* being rational, albeit *subjectively* rational, because he is believing in accordance with the epistemic principles that he actually has that experts are not to be trusted. One may reply that Frank is being *objectively* irrational for he is violating abstract canons of rationality that presumably would dictate to defer to experts. This may very well be the case, but self-deception is a phenomenon of subjective irrationality, so appeals to objective irrationality are irrelevant here.

To recap, if Lynch has in mind subjective rationality, then Frank constitutes a counterexample to Lynch's claim because Frank is someone who is subjectively rational but is not sensitive to the force of good evidence. If instead Lynch has in mind objective rationality, then Lynch's claim would appear to be vindicated by Frank's case given that Frank is being insensitive to the force of good evidence and he is not being objectively rational. But, as I said above, claims that pertain to objective rationality have no relevance when it comes to self-deception. This is because self-deception is a phenomenon of subjective irrationality, where the agent does not violate abstract, objective standards of rationality (that they do not hold) but rather departs from their own canons of rationality. Appeals to objective irrationality would also pose a problem with respect to explaining the nagging doubts that haunt the self-deceived individual. In fact, I suggest that violating abstract canons of rationality that one does not hold is unlikely to generate the persistent doubts typical of self-deception, but only regular doubts (if any at all).[58] Here is why.

[58] I borrow this line of reasoning from Lynch himself who, however differently from what I suggest, argues that if the tension was only caused by uncertainty (as opposed to a combination of uncertainty and a stake in the issue) it would only generate regular doubts (Lynch 2012: 440).

Suppose for the sake of argument that self-deception involves a violation of abstract canons of rationality. Suppose also that it is a public epistemic norm that experts' opinions in their area of expertise are more reliable than those of non-experts. Further suppose that the majority of scientists claim that smoking causes cancer, on the basis of reliable evidence. Frank, a heavy smoker, has an interest in thinking that smoking does not cause cancer. While listening to Alex Jones declaring that medical experts are wrong, Frank self-deceives by forming the belief that smoking does not cause cancer. In being self-deceived, Frank is violating public epistemic norms. In fact, he believes Alex Jones who is not an expert in medicine, over scientists who are indeed experts. However, suppose that Frank does not hold these norms; he does not usually take seriously what experts say because he thinks they are liberal snobs with their own political agenda to pursue. The question is: does violating public norms that Frank does not hold generate nagging doubts?

I argue that it does not. I take it that a persistent doubt is a doubt that constantly haunts its subject, so to speak. But the doubt is persistent because, and *only if*, one takes the unfriendly evidence close at hand to be *good* evidence. (Presumably, it is easier to dismiss evidence one judges to be poor.) Thus, the doubts are persistent because, and only if, the unfriendly evidence is good according to one's own epistemic norms. I am not suggesting that Frank cannot have doubts. He may very well have doubts about whether smoking causes cancer, but even in the eventuality that doubts may arise, they are unlikely to be persistent.[59] They would be persistent only if he deems the evidence to be good which, in turn, is a function of Frank's epistemic norms.

But this last claim shows that the element of *stake* that Lynch posits is explanatorily redundant. The idea that some epistemic norms involved in self-deception are our own and may even be loaded insofar as they, say, shape our identity, is enough to explain why violating them generates

[59] We can imagine, for example, Frank in a social situation surrounded by epistemic peers who believe smoking causes cancer who challenge his thoughts on the matter. It may be the case that due to that epistemic pressure a doubt might arise.

uncomfortable feelings of anxiety and unease. Similarly, just as we may feel guilty when violating our moral code, so we may feel anxious and worried when violating our epistemic code. The element of *stake* need not be posited as an external element that contributes to explaining the mental tension because the role that the stake is supposed to play in Lynch's theory is already incorporated in the definition of an agent's epistemic norms.

Certainly, as I mentioned at the beginning of this chapter, it is not the case that every instance of self-deception is accompanied by mental tension. The mental tension typical of self-deception is not conceptually necessary, but only characteristic of the phenomenon. That is, most core cases of self-deception are accompanied by it, while some may be accompanied by it to a lesser degree, and others may not manifest any tension at all. This may appear to be a potential problem for the doxastic violation account because it is not obvious how the theory could explain this variation in degree, particularly with respect to those instances of self-deception that display no tension at all.

However—and this is the second advantage of my proposed view—the doxastic violation theory, and in particular the idea that the predicate of being loaded can come in degrees, helps explain why some cases of self-deception may be marked by acute tension while some others may not. My suggestion is that the degree of mental tension present in self-deception is a function of how loaded the violated epistemic norms are. The more loaded a norm is, the greater *likelihood* that its violation will generate tension. Consider again the epistemic norm "I ought to trust scientific evidence". If a scientist holds this norm, and then indulges in the comfortable but mistaken belief that they are healthy despite overwhelming scientific evidence that points to the contrary, then they are violating the norm that they ought to trust scientific evidence, and may be considered self-deceived. In this case, I suggest that the tension generated by violating the epistemic norm in question may be substantial because the epistemic norm the scientist endorses is particularly loaded.

On the contrary, if the scientist violates an epistemic norm that is not loaded, the tension that follows from the violation may not be as intense. An example of this is the scientist holding the epistemic norm according to which they ought to never trust their cats when they act as if they have been fed. This norm is not particularly loaded: it is not emotionally charged, does not modulate access to any community, nor speaks to the scientist's identity in important ways. Yet it is a norm that they follow and that they formulated on the basis of observing their cats' behavior. If the scientist, on a generous impulse, trusts their cats and feeds them, then the violation of this norm may generate less tension compared to the tension generated by the violation of the norm that one ought to believe scientific evidence. Thus, the mental tension generated in self-deception is *proportional* to the degree to which the violated epistemic norm is loaded. When an individual violates a norm that they do not endorse, no tension may follow. If an individual violates a loaded norm that they hold, then the highest degree of tension may manifest. The individual may experience persistent doubts coupled with an uncomfortable phenomenology. If instead the individual violates a bland norm that they hold, they may experience persistent doubts, but with no associated phenomenology. What causes the mental tension is thus *overdetermined*. The tension is generated by i) whether the norm is endorsed and ii) whether the endorsed norm is loaded. In particular, i) is responsible for generating the mental tension and ii) accounts for the *degree* of tension.

In sum, on the one side stands Lynch's view which provides an unsatisfying answer to the causal question and posits a combination of uncertainty and stake to account for the mental tension. On the other side stands the doxastic violation account that I believe provides both a more plausible answer to what causes the phenomenology of self-deception and is more parsimonious because it explains the mental tension only by relying on the concept of an agent's epistemic norms.[60] I suggest

[60] Van Leeuwen also argues that his view of self-deception as a violation of the agent's own epistemic norms is more parsimonious, but he claims that with respect to intentionalist views of self-deception that posit that the agent holds

that the latter is preferable to the former, not only because it is more plausible and parsimonious, but also because it has the advantage of meeting the remaining irrationality challenge as well as respond to the passivity and uniqueness objections.

2. Meeting the Irrationality Challenge

What I have just said with respect to the phenomenological challenge shows how the doxastic violation account is well-placed to explain the subjective irrationality of self-deception. In fact, the subjective feature of the irrationality of self-deception emerges when the type of irrationality that is perpetrated is one where there is a departure from rules or canons that are *internal* to the agent, as opposed to rules or canons that belong to outside observers but that the agent does not hold. Given that this is what subjective irrationality involves and that in order to account for the subjective irrationality a theory must be able to explain how the agent is violating internal canons, it follows that the doxastic violation account is naturally well-equipped to explain this aspect. This is because integral to the theory is the idea that the self-deceived individual violates their own epistemic norms, which is another way of claiming that they departs from their internal canons.

Compared to minimalism, the doxastic violation account provides a superior characterization of self-deception's subjective irrationality. Recall that the reason why minimalist accounts have difficulty explaining self-deception's subjective irrationality is that they characterize self-deception as a motivationally-biased form of hypothesis testing and thus it locates the irrationality of self-deception exclusively in the etiology of the self-deceptive belief. The subjective aspect of self-deception's irrationality comes to life when the agent is believing against their own standards in a way that appeal to motivational bias does not achieve.

two contradictory beliefs at the same time (Van Leeuwen 2008: 195). I have indeed argued that his view is more parsimonious even with respect to deflationist theories like Lynch's.

At times, Mele does seem to hint at the idea that the self-deceived individual is violating some sort of internal standards. For example, he notes that the self-deceived subject selects and examines evidence in a way that they "normally" would not (Mele 2001: 106). This suggests that the self-deceived person possesses a typical or habitual way of assessing evidence that they then depart from when gripped by self-deception. Unfortunately, Mele does not develop that idea, and instead proposes what he calls the "impartial observer test" discussed in Chapter 1:

> if S is self-deceived in believing that *p*, and D is the collection of relevant data readily available to S, then if D were made readily available to S's impartial cognitive peers (including merely hypothetical people), those who conclude that *p* is false would significantly outnumber those who conclude that *p* is true. (Ibid.).

But as we can see, this remark seems is in tension with Mele's claim about the individual's habits, and instead seems to locate the irrationality of the phenomenon in the agent's violation of *external* standards, that is, standards that other non-biased peers would instead follow. The doxastic violation account is thus better-equipped to capture self-deception's subjective irrationality because postulating that the self-deceived individual violates their own epistemic norms brings to light the subjective feature of self-deception's irrationality.

2.1 A Potential Objection

Despite the doxastic violation account being able to explain self-deception's subjective irrationality, some have argued that the theory does not fully succeed because positing the notion of violated epistemic norms is not by itself sufficient to appropriately capture the idea of subjective irrationality (Scott-Kakures 1996). What is crucial is that the self-deceived person *knowingly* violates their own epistemic norms and the doxastic violation account cannot account for this awareness because, by definition, the theory argues that the self-deceived person *unknowingly* violates their epistemic norms (recall that Van Leeuwen claims that the self-deceived subject *slides* into their self-deceit).

According to this objection, unknowingly violating one's own epistemic norms cannot explain the mental tension, and in particular that "knowing insincerity" that is typical of self-deception (Ibid., 49). Scott-Kakures for example suggests that when avowing their professed belief, the self-deceived person betrays a knowing insincerity which involves the "dividedness of mind", the conflict that arises from partly recognizing the truth, and yet refusing to recognize it (Ibid., 49). Because the self-deceived person "is aware of something from which he wants to flee, and yet this flight is also something that he condemns", this knowing insincerity demands the involvement of self-reflection, which implies that the self-deceived person must somehow know that in believing *p* they are violating their own epistemic standards (Ibid., 49).

The objection is right in claiming that the doxastic violation account presupposes that the self-deceived person *unknowingly* violates their epistemic norms, but it is wrong in claiming that this prevents the doxastic violation account from successfully explaining self-deception's irrationality. The objection seems to suggest that the knowing insincerity experienced by the self-deceived person can only be accounted for if one posits that the self-deceived person is somewhat aware that they are violating their own epistemic norms. But, maintaining that the self-deceived person is not aware that they are violating their own epistemic norms I suggest does not imply that one cannot capture the conflict characterized as "knowing insincerity".

The self-awareness argument claims that the knowing insincerity experienced by the self-deceived person is a consequence of their critically reflecting on their own mental states, which in turn implies that they must know that they should not believe *p*. From this it follows that conceding that the agent is knowingly violating their own epistemic norms is the only way to solve the irrationality challenge. But this would be too quick a conclusion. I suggest that there is an alternative way to explain the insincerity that does not force us to postulate that the self-deceived person *knows* that they ought to believe otherwise.

I have argued that the self-deceived individual violates norms that they themselves hold, that they usually follow similarly to habits of thought, and that can guide their actions without them needing to explicitly contemplate them. This means that the agent's epistemic norms are able to operate *below the surface* of awareness. When agents follow their epistemic norms, they do not necessarily *know* that they hold a certain epistemic norm nor that what they are doing at the moment is believing in accordance with a norm. Agents can certainly *come to know* their epistemic norms upon reflection, but one need not posit critical reflection in order to account for how the norms exert their guiding force.

There are two reasons for why I think agents are not often aware they hold certain epistemic norms. One reason is psychological. It is simply a fact of psychology that we absorb epistemic norms, most often during our upbringing, in ways that are implicit. We might, for example, come to hold the epistemic norm that experts are to be trusted by observing our parents reliably believing experts. This is what I mean when I say that epistemic norms operate *below the surface* of awareness. The second reason is pragmatic. Agents cannot always consult an epistemic textbook when deciding what to believe as this would be in the long run both time-consuming and labor-intensive. It is in the interest of agents that their norms guide their behavior automatically, and this habitual aspect is a *positive* advantage: it saves time and energy.

In the light of this, I suggest that the knowing insincerity experienced by the self-deceived person can be explained by appealing to the idea that the agent holds and deploy their epistemic norms without being aware of them, which is another way of saying that the norms are *internalized*. That is, they are internal to the agent's mental economy and guide their actions similarly to habits of thought. When the norms are violated, *even if unknowingly*, they may generate unease, which I propose can play the role Scott-Kakures assigns to "insincerity". This knowing insincerity then does *not* indicate that the self-deceived person has employed critical reflection; rather it indicates that the self-deceived subject has violated epistemic norms that they have internalized. The self-deceived individual experiences discomfort not because they know that they should believe *p* but because they are violating a rule that they do hold. In this sense, the feeling of discomfort can arise without the agent being aware that the cause is the doxastic violation. This seems to make sense if one thinks that we often experience feelings and emotions before knowing *why* we experience them. Consider the following case.

> Phoebe and Miriam are happily married. One day Miriam expresses an interest in leaving the big city to move into the countryside. She is tired of the frantic rhythm and long hours and wants to retire somewhere quieter. Phoebe accepts and they start making plans. When in town, they decide to go see a house with the intention of buying it. While walking around the different rooms, Phoebe is suddenly struck by an anxiety attack. She cannot breathe. As a result, she is forced to get out of the house to get some air and slowly starts feeling better. After having returned to the city, Phoebe begins interrogating herself about the reason why she had that anxiety attack and comes to the conclusion that the attack is a reaction to her not being comfortable with the idea of moving to the countryside.

This example shows how often we are able to directly access our desires and preferences only upon reflection which in turn may be triggered by a simple bodily reaction. Before exercising critical reflection, Phoebe does not know the *why* of her anxiety attack. Only after having exercised critical reflection does she reach the conclusion that her anxiety attack has been caused by her unwillingness to move to the countryside. Yet, her unwillingness is perfectly able to still *exercise its force* through bodily reactions without Phoebe knowing that that is what is going on. Similarly, I suggest, the self-deceived person's discomfort is caused by the internalization of their own epistemic norms, which can be accessed upon reflection.

3. Responding to the Passivity Objection

The passivity challenge has two components: the first component involves conferring a special role on the *self* in self-deception, and the second component involves rescuing the individual's *agency* in self-deception. The passivity challenge arises because minimalism seems ill-placed to address both of these components. With respect to the first component, minimalism has difficulty accommodating the role of the agent because it portrays self-deception as a process of motivationally-biased hypothesis testing where the agent is not particularly involved. The agent is not aware of their self-deceit nor do they go against epistemic canons that are their own (as we have seen in the irrationality challenge). With respect to the second component, minimalism seems to portray self-deception as a passive one that merely happens to the self-deceived person, thus eradicating the individual's agency.

As I said in Chapter 1, there is a strong intuitive pull to the idea that self-deception is not merely a bias that occurs to some unfortunate individual, but rather a phenomenon where the agent is crucially involved. Robust intentionalism does justice to this idea but as a result, it faces the dynamic and the static paradox. As a reaction to the two paradoxes, minimalism deflates the agent's intentions and characterizes self-deception as a form of motivated bias where no intention to deceive is involved. This strategy successfully avoids the paradoxes, but it has the unforeseen consequence of almost completely erasing any agency.

Now one might argue that erasing the *self* is not in principle a drawback. It makes sense to eradicate the *self* from self-deception if keeping it commits us to a characterization of self-deception that is problematic (i.e. one that portraits the self-deceived person as aware of their self-deception). Yet I suggest here that the doxastic violation account is able to preserve the role of the *self* in self-deception in such a way that it does not force us to embrace an incoherent characterization of self-deception.

The doxastic violation account preserves a role for the self by positing that the self-deceived person violates not merely abstract epistemic norms, but *their own* epistemic norms. The self that is implicated is the one that has internalized and committed to the relevant norms. The doxastic violation account also rescues the agency in self-deception. By arguing that the self-deceived person actively *violates* their own norms, the account also does justice to a characterization of self-deception where the individual *participates in bringing about* their self-deceit, thus departing from minimalism's passive view of self-deception. Self-deception is not a mistake that merely *occurs* to the individual; it is not merely an automatic response to a painful truth that is difficult to embrace. The individual *allows* the self-deceit to occur by letting a motivational element interfere in the etiology of their belief. One might note that the agency involved in *letting the self-deception happen* may not be an especially thick notion of agency because it does not evoke an active contribution on behalf of the agent. Yet the notion is sufficient to justify the claim that self-deception is not merely a mistake, which is the issue in question. Furthermore, arguing that the mental tension characteristic of self-deception is caused by the agent violating their own epistemic norms, validates an additional kind of experiential tension, an *agent-centered* tension, that is not available to minimalism. Let me explain this last claim in more detail.

According to Lynch's view, the experiential tension in self-deception is primarily directed at the self-deceptive proposition, and it involves nagging doubts about whether the friendly proposition the self-deceived person professes to endorse is true. It follows that the uncertainty in question is *directed towards the belief.* The self-deceived person's anxiety revolves around the suspicion that the belief they profess to endorse may be false. Lynch's view of the mental tension is what I call a *belief-centered* view where nothing, or very little, picks out or identifies the 'agent'. Yet, self-deception includes the word 'self' for a reason: the phenomenon does not consist simply of a deception about a proposition. It is a deception that *also* involves the agent in some crucial respects.

The doxastic violation account identifies the cause of the mental tension as the violation of the agent's own epistemic norms. The agent's epistemic norms would lead the individual to conclude that *p*, yet they believe not-*p*. Because the agent is going against their own epistemic standards, a certain degree of conflict and uncertainty is to be expected. In this sense, the doxastic violation account captures the *belief-centered* tension Lynch discusses, but it also validates what I call an *agent-centered* tension, that is, a tension that is directed at the agent themselves *qua* perpetrator of the epistemic violation.

Forming a belief that is at odds with one's *own* epistemic norms must make a difference with respect to the kind of tension that is generated compared to the belief-centered tension Lynch discusses. In addition to the anxiety about the belief, the self-deceived person may also experience a sense of *unease* towards themselves because it is *they* who have failed to live up to their own epistemic norms. Not only may the agent be haunted by questions concerning the belief ("Is not-*p* the case?"), but the doxastic violation may also give rise to a phenomenology directed at the agent that manifests through underlying feelings of being *out of sorts* ("Am *I* right in believing not-*p*?"). The feeling of being out of sorts is caused by the violation of the epistemic code that the agent holds and usually follows. Since the epistemic code is part of the agent's mental economy, the feeling of being out of sorts can be described as an experience of *alienation* where the agent does *not quite feel like themselves* because they hold a belief that departs from the epistemic rules that they themselves set up and usually follow. In this sense, the phenomenology associated with the violation of one's own epistemic norms is specifically *about* the agent. Paul Noordhof hints at a similar idea when he claims that "the problem with self-deception is that [self-deceived individuals] seem to avoid accepting a certain proposition and have anxiety over, or lack of confidence in what *they are up to*" (Noordhof 2009: 45, emphasis added). Richard Holton, albeit for different reasons, also suggests that self-deception is in part "about the self" (Holton 2001: 53).

The agent-centered tension may make more sense if we think about a parallel with the moral domain. When violating their own moral code, agents may not only experience negative emotions associated with the morally wrong action just performed, but they may also experience negative emotions associated with themselves as the *person* who performed a morally reprehensible action. When one violates one's own moral norms, they may not only feel guilty about the action, but also feel shame directed at the agent themselves *qua* the subject that performed the reprehensible action.[61] As a result, agents may wonder whether having performed such a reprehensible action betrays who they are, or who they have become, as *persons*. They may wonder if having performed such a morally wrong action indicates a lack of integrity. My suggestion is that the same can be said about self-deception. Similar to the moral domain where violating one's own moral code may be accompanied by an action-centered as well as an agent-centered phenomenology, so in the epistemic domain violating one's own epistemic norms may be accompanied by a belief-centered as well as an agent-centered tension. The violation theory thus provides us with the conceptual resources to solve the passivity challenge by doing justice to the intuition that there is a *self* in self-deception. By relativizing the violation of epistemic norms to the agent, the account validates the agency of the self-deceived person in bringing about their self-deceit.

4. Responding to the Uniqueness Objection

Recall that by claiming that self-deception is a form of motivationally-biased hypothesis testing, minimalism risks portraying the phenomenon as being indistinguishable from other similar

[61] This is of course assuming that one does not *want* to violate one's own moral code, but is forced to do so in having to choose between two options with neither of them ideal. Similarly, there might be good reasons to violate one's own epistemic norms. If one is a partialist, for example, then one might not apply the same evidential standards to everyone but interpret evidence more charitably if this is particularly embarrassing and involves a friend (see Stroud 2006, Crawford 2019). However, these are rare cases that I avoid here and focus on paradigmatic instances.

biases, such as confirmation bias, other similar irrational instances, like wishful thinking, and mere instances of bad reasoning.

Some deflationists attempt to resolve the uniqueness problem by suggesting that wishful thinking and self-deception are distinct phenomena insofar as in the former there is no confrontation with contrary evidence, while in the latter the individual appreciates disconfirming evidence (Bach 1981; Johnston 1988).[62] However, this distinction does not fully address the challenge. If it is correct that the self-deceived person *recognizes* the unfriendly evidence (whatever doxastic attitude that involves), then it is hard to see how such a claim can be reconciled with the claim that self-deception is non-intentional. Thus, the *uniqueness* challenge consists in conferring on self-deception a *unique* status in order to avoid downgrading the phenomenon to an ordinary bias.[63]

Showing how the doxastic violation account avoids the uniqueness challenge involves answering two questions that need to be handled separately. The first question is whether *conceptually* self-deception can, by definition, be distinguished from faulty forms of reasoning (like wishful thinking, confirmation bias, bad reasoning); the second question is whether *practically* many instances of such forms of reasoning overlap with instances of self-deception.

With respect to the first question, framing self-deception as a deviation from one's own epistemic norms distinguishes it both from wishful thinking and confirmation bias because *definitionally* neither of those phenomena involve a violation of the agent's own epistemic norms, in the sense that the notion of epistemic norms does not belong to the definition of wishful thinking. Definitionally, wishful thinking consists in the agent believing what they desire, without any necessary departure from subjective canons of rationality (Davidson 1985: 205). This last one is a crucial feature of self-

[62] Al Mele rejects this distinction on the grounds that it reduces self-deception to a characterization that is "overly narrow" (Mele 1983: 375).

[63] Other attempts to distinguish self-deception from wishful thinking have involved, for example, arguing that self-deception constitutes a failure of self-knowledge (see Scott-Kakures 2002).

deception, instantiated by the agent's violation of their epistemic norms. The reason for this claim is that self-deception is *subjectively* irrational while wishful thinking, albeit still irrational, is not necessarily a phenomenon of *subjective* irrationality because there may not be a departure from one's own epistemic standards. Since violating one's own canons of rationality is the defining feature of subjective irrationality, it follows that what distinguishes the two must lie in the violation of subjective epistemic norms. Yet *practically*, it seems reasonable to infer that many instances of self-deception might overlap with cases of wishful thinking and confirmation bias. This is because, *typically*, one violates their own epistemic norms when engaging in wishful thinking or confirmation bias because individuals generally hold epistemic norms that involve believing in accordance with the evidence and not in accordance with one's wishes. In claiming that practically some instances of wishful thinking can overlap with instances of self-deception I am not contradicting the definitional claim. I am not saying that *all* instances of wishful thinking are instances of self-deception, for this would be equal to saying that the two phenomena are definitionally the same (thus contradicting the definitional claim). I am simply conceding that despite the definition of wishful thinking need not involve the violation of epistemic norms, practically some instances might be more complex to distinguish.

Does the doxastic violation account differentiate, definitionally and practically, self-deception from mere instances of poor reasoning? Answering this question requires a more involved discussion because it raises the additional question of what is meant by 'poor reasoning', specifically when this is framed against the background of epistemic norms. Presumably there are different kinds of poor reasoning. In what follows I examine different types of poor reasoning and show how they are distinct from instances of self-deception. I will argue that while not all instances of bad reasoning are self-deception, self-deception does instantiate bad reasoning because one certainly cannot be said to have reasoned well when violating their own epistemic norms. Thus, if an individual engages in poor reasoning it is not necessarily the case that they are also self-deceived, but if they are self-deceived then they are also manifesting poor reasoning.

To start, one example of a type of poor reasoner is someone who holds *objectively bad norms*. Someone who usually distrusts experts or only reads sources that they expect to agree with, cannot be said to hold good epistemic norms, yet someone who holds bad epistemic norms is different from someone who is self-deceived. In fact, the idea of holding bad norms assumes that there is such a thing as objectively good norms one can depart from. Self-deception instead does not warrant any judgment that pertains to grounds of objectivity because it is a phenomenon of *subjective* irrationality where what goes awry is relativized to the agent.

Another type of poor reasoner is also someone who fails to form beliefs in accordance with what their epistemic norms would imply. For example, someone who usually does not believe anything psychics say might fail to recognize that such epistemic norm also implies that they should not believe their friend Stacy while she is "reading their cards." Surely this type of poor reasoning might appear similar to an instance of self-deception the way the doxastic violation account describes it. One might be tempted to think that a failure to form beliefs in accordance with one's epistemic norms (which I refer to as poor reasoning) and a violation of those epistemic norms (which the doxastic violation account refers to as self-deception) are simply the same. Yet a *failure* is different from a *violation*. While a failure evokes the concept of error, a violation seems to be more purposeful. As I explained in Chapter 2, the violation of the self-deceived person's epistemic norms is caused by the fact that the agent *lets* their motivation interfere with their following their norms. In this sense the agent does not simply *fail* to conform to their epistemic norms due to, say, a temporary moment of gullibility; rather they take active part in violating their epistemic norms.

A further example of what bad reasoning might involve is someone who, unlike the self-deceived person who holds epistemic norms and violates them, *does not hold any* epistemic norm in the first place. Consider again the case of Frank. Suppose Frank believes NASA when they claim to have beaten the Russians to the moon, but he does not believe that global warming is occurring, despite NASA supporting the claim. Suppose further that this is not the only instance of inconsistency that Frank shows but that it occurs on a routine basis. When the CDC releases guidelines for how to protect oneself from the spread of the novel coronavirus, Frank follows them faithfully, yet he also believes that Covid-19 is "just a flu". Now suppose that experts declare that overwhelming evidence has revealed a causal link between smoking and cancer. A few days later, Frank is watching his favorite show where Alex Jones insults experts and claims that they are wrong: smoking does not cause cancer. Being a smoker, Frank has an interest in smoking not causing cancer. As a result, Frank believes Alex Jones over experts. Is Frank self-deceived? I argue that he is not. He is a merely a *poor reasoner*.

The reason why Frank qualifies as a bad reasoner, I suggest, is that contrary to self-deceived individuals who hold epistemic norms and then violate them, Frank lacks epistemic norms to begin with. That is, there is no epistemic guideline that Frank *reliably* follows. Frank does not seem to *usually* believe experts, but he does not *usually* distrust them either. One would then be tempted to argue that Frank's problem is that he is simply being incoherent, but it is precisely this incoherence that indicates that Frank does not hold any epistemic norms. There is no coherent set of epistemic principles that characterize most of Frank's choices.[64]

[64] The idea that an agent might not hold any epistemic norms is controversial, but it is not implausible. It seems that it would be challenging for someone to successfully navigate one's environment without the use of any epistemic norms at all. One could argue that even someone like Frank—other things being equal—might believe what his senses tell him – e.g., that it is raining, that it is a sunny day, that something exploded nearby. Now would this reliance on his senses amount to an epistemic norm? It certainly would so long as Frank *reliably* follows it. The definition of epistemic norms I adopt is behavioral. A necessary condition (but not sufficient) is that *n* constitutes an epistemic norm only if an agent S reliably follows it. If Frank reliably believes in accordance with his senses, then believing what his senses tell him qualifies as an epistemic norm in Frank's psychological economy. But if Frank often shows inconsistency in believing what his senses tell him, then one cannot say that relying on his senses constitutes an epistemic norm for Frank.

This is not to say that Frank is not also biased. His resistance to believing that smoking causes cancer is indeed motivationally shaped by interests and desires. My point is that Frank's case is an example of motivationally-biased belief-formation that does not amount to self-deception because no epistemic norms are involved. One should not conclude that Frank holds all his beliefs arbitrarily; rather Frank is *epistemically clumsy*, so to speak. *Epistemic clumsiness* is in fact what lack of epistemic norms amounts to: it is akin to navigating one's own environment without an *epistemic map*. In contrast, the self-deceived person does hold an epistemic map, but does not follow its directions.

Even though Frank is not violating any norm, however, would not repeated incoherence generate persistent doubts? After all, Frank is failing to hold even general epistemic norms. However, violating norms one does not hold is *not* sufficient to generate nagging doubts; at best it may generate *regular* doubts – that is, doubts that are not persistent.[65] The reason why is that if the agent does not hold epistemic norms, then these do not belong to the agent's mental economy, and thus they lack that persistent quality needed to capture the idea of experiential tension typically associated with self-deception.

It would be a different story if Frank *did* indeed hold the norm that he ought to believe experts. If Frank violated his own standards of rationality, then this would be sufficient to generate nagging doubts. Going against his own rules, which he usually follows, would cause a certain ambivalence towards the friendly proposition he professes to endorse. In this scenario, Frank would not be entirely confident that smoking does not cause cancer, and the reason why is that he would be violating norms that are part of his mental economy and thus are continually 'at play' in the background. The violation of one's own epistemic norms is thus what shows that not all instances of bad reasoning qualify as also instances of self-deception. That is, while self-deception is an example of poor reasoning, many

[65] This line of argument is based on Lynch 2012, who similarly, albeit differently, argues that if the mental tension characteristic of self-deception was only caused by uncertainty, the doubts that the self-deceived would experience would not be persistent. See Lynch 2012: 440 for more on this.

instances of bad reasoning are *not* instances of self-deception, and this is the sense in which the two are distinct.

A final, more general reflection follows from this distinction. Individuals do not deceive themselves because they do not show regard for what is true or because they fail to hold appropriate epistemic norms. This would be an oversimplification. Bad reasoners lack the norms, but self-deceived individuals do not. They do hold epistemic norms, yet in those circumstances when desires are involved, they do not successfully follow them.

*

In this chapter I have showed that only by adopting a specific type of belief-based deflationism, that is, only by adopting the doxastic violation account, can one respond to the objections levelled against minimalism and meet those challenges that any theory of self-deception should accommodate. It solves the phenomenological challenge by providing an account of self-deception's experiential tension that is more psychologically plausible and by validating an agent-centered tension that escapes the characterization of traditional deflationism. The agent-centered tension combined with the relativization to the norms of the self-deceived person also preserves a role of the self in self-deception, thus solving the passivity challenge.

The doxastic violation account better accounts for the unique status of self-deception by differentiating it from those allegedly similar phenomena such as wishful thinking, confirmation bias, and even instances of bad reasoning that instead do not involve any violation of the agent's own epistemic norms. Finally, the doxastic violation account better captures the subjective irrationality of self-deception because it explicitly emphasizes a sense in which the beliefs of the individual are at odds with their own canons of rationality, even if unknowingly.

CHAPTER 4

Self-Undeception: Coming to Our Epistemic Senses

We have now come to the final chapter of this dissertation where I discuss the core of my project: self-undeception. Though the previous three chapters are about self-deception, I consider them to provide essential preparatory ground for approaching self-undeception. To recap, I have so far argued that the doxastic violation account is superior to minimalism with respect to its characterization of self-deception. In Chapter 1 I argued that the dynamic and static paradoxes provide good reasons to prefer deflationist accounts (whether belief- or pretense-based) over intentionalist accounts. Recall that the dynamic paradox questions the successful outcome of intentional self-deception by arguing that the self-deceived person would see through their own intentions, while the static paradox highlights the implausibility of the self-deceived person simultaneously holding two contradictory beliefs about the same content without them undermining each other. Deflationist accounts solve the former paradox by removing any intentional element from self-deception, and avoid the latter by deflating one of the two beliefs into a weaker doxastic attitude.

Among deflationist accounts, which comprise both pretense accounts and belief-based accounts, I have given reasons for preferring the latter. I have then subdivided belief-based accounts into minimalist accounts and doxastic violation accounts and argued that doxastic violation accounts are superior. I argued that the doxastic violation account better explains the mental tension typical of self-deception, better accounts for self-deception as a unique phenomenon as opposed to regular instances of motivated reasoning, rescues the role of the self, and better captures the subjective irrationality of self-deception.

Now that I have laid out and provided reasons for why we should prefer the doxastic violation account with respect to self-deception, I argue that the doxastic violation account is also superior to competing theories with respect to its characterization of self-*un*deception because, unlike the other views, it provides an account of self-undeception that is *normatively* superior. That is, the doxastic violation account advances a path to self-undeception that is preferable to the one proposed by competing theories. The reason why is that, unlike minimalism and pretense-based accounts, the doxastic violation account provides a path to self-undeception that is motivated by self-reflection and that, in turn, leads to what the agent sees as a better epistemic future.

As I also mentioned at the beginning of the dissertation, my approach to self-undeception may seem somewhat unusual. Why frame the issue against the background of a particular account instead of offering a unified account of what self-undeception is? The reason why is that I take self-undeception to be a phenomenon that derives and unfolds, so to speak, from self-deception. As my metaphysical definition will make clear, to be self-undeceived is to *not* be self-deceived *anymore* with respect to a certain proposition. In light of this, it seems to make sense that one would propose an account of self-undeception that accords with one's preferred theory of self-deception.

Before turning to my analysis of self-undeception, I want to first begin by discussing a pre-theoretical example of an instance of self-undeception and stress why the phenomenon, despite having received little attention in the literature, warrants philosophical inquiry.

1. A Pre-Theoretical Example of Self-Undeception

In general, we can all agree that some people come to their senses. Consider for example Jane Austin's novel, *Sense and Sensibility*. There, one of the main characters, Marianne, has an unrequited love for John Willoughby who instead decides to marry another woman. Feeling rejected, Marianne impulsively takes a long walk in the rain, after which she develops a serious fever. Upon recovering, she suddenly comes to her senses, and realizes that Willoughby, who had been so insensitive to her feelings, would never have been able to make her happy. Marianne then decides to go on to marry Colonel Brandon, who has always been secretly in love with her.

Qualifications aside, self-undeception is not radically different from Marianne's experience: it involves self-deceived individuals coming to their *epistemic* senses. By 'coming to their epistemic senses', I mean that they come to abandon their self-deceptive proposition and embrace one that they take to be more warranted. Consider our familiar case of Sally.

> In the last few months Sally's life has changed dramatically. She has quit her job and made the decision to shave her hair. She goes to the hospital once a month to do her chemotherapy sessions, and her brother Charlie has temporarily moved in to assist her. While contemplating her new life, Sally occasionally finds herself thinking back to when she used to rationalize her test results and discount those suspicious headaches. Despite the ample evidence, Sally had for months insisted that she did not have cancer. But now that time seems far away, and Sally is painfully aware of the reality of her illness.

Like Marianne's case, Sally's is a pre-theoretical case of self-undeception. While in the grips of self-deception, Sally manages to somehow come to her epistemic senses and embrace reality. How and why her self-undeception occurred are issues I tackle in this chapter. Nevertheless, since theorizing about self-undeception is relatively uncharted territory, one legitimate question arises: why should one investigate the phenomenon in the first place?

To start, one *pro tanto* reason for why self-undeception warrants our philosophical attention is for completeness sake. In the literature on self-deception, there are two central questions worth answering. First, a *metaphysical* question: what is it for an individual S to be self-deceived about a proposition *p*? And second, a *psychological* question: how does an individual S come to be self-deceived about a proposition *p*? I presume that the same questions transfer to any case to be made for self-undeception. A complete theory of self-deception should be able not only to tell us what self-deception is and what are the underlying psychological mechanisms that bring it about, but it should also give us a story of how those very same mechanisms unravel, that is, how we self-*un*deceive, and what self-undeception amounts to.[66]

Completeness, however, is not the only reason why it is worth examining self-undeception. An additional reason is that, unlike its flip-side, self-undeception validates a third, *normative* question of what is a *preferable* or *desirable* way, in the sense of *modus operandi*, in which a self-deceived individual S should undeceive themselves. It is important to investigate this question because its exploration, as I will show, paves the way to issues that pertain to the agents' cultivation of what they see as a virtuous epistemic future.

While it is clear what the desideratum is with respect to the metaphysical and psychological question –that is, we look for an answer that is conceptually satisfying and explanatorily plausible, respectively –it is not clear what we should be looking for when answering the normative question. Meaning, it is not obvious what an account of self-undeception should look like in order to properly answer the normative question. In order to shed light on this unclarity, let me pause to explain the normative question in more detail.

As a way of motivating the normative question, let me start by explaining why I think self-undeception, unlike its flip-side phenomenon, validates a normative question. A helpful way to motivate the normative question is to explain why this question makes sense *only* when framed against the background of self-undeception and why it is not an appropriate one to raise with respect to self-deception. I have said that the normative question is a question about what constitutes a desirable way to self-undeceive. The reason why the normative question should not be applied to self-deception is that self-deception is a criticizable state because the self-deceived person holds a belief at odds with their canons of rationality. Thus, raising the normative question of what is a desirable way to self-

[66] Philosophers run the same line of argument also for other questions like well-being. Recent attention has been given to theories that are able to account for what constitutes *well*-being and *ill*-being. See for example Kagan 2015 and Tully 2017.

deceive seems out of place because one would not want to provide the agent with recommendations for how to achieve what is a criticizable state. Asking what the ideal way is to self-deceive then does not make sense, for the same reason it does not make sense to ask what is the best policy to be immoral or irrational.

Someone might reply that it makes sense to ask 'what the best policy is for achieving self-deception' if, consequentially speaking, self-deception brings about a positive outcome. After all, the self-deceived person's aim is to avoid a painful truth and in doing so, self-deception can at least aim at somewhat enhancing the self-deceived person's well-being in a way that facing reality would not. But though this may very well be the self-deceived person's *aim,* it is an aim that is not fully accomplished since, as I have explained in the previous chapters, self-deception is associated with an experiential tension characterized by feelings of anxiety and worry.

Yet even if self-deception was able to generate some comforting feelings by creating positive illusions, as some have attempted to argue, it still would not warrant the conclusion that self-deception is the *best* policy to achieve such feelings. (Taylor 1989, Taylor & Brown 1994, Taylor et al. 2000). That self-deception *can* trigger positive emotions is a descriptive claim which is irrelevant when we raise the "normative" question about whether self-deception is the best policy to trigger those emotions (Van Leeuwen 2009: 107). As Van Leeuwen articulates in his paper "Self-Deception Won't Make You Happy", if someone argues that it is bad to give sharp knives to children under the age of four, showing that in *one possible case* "a three-year-old child did something good with a knife (e.g. cut carrots) does not refute the point" (Ibid., 111, words rearranged).

I hope this clarifies why self-undeception validates a normative question while self-deception does not. Let me now explain what we should be looking for in a satisfying answer to the normative question. I said earlier that there are preferable and less preferable ways to self-undeceive. A desirable path to self-undeception is one that prepares the agent for what they take, by their own light, to constitute a *better epistemic future*. By "better epistemic future" I mean one that, through the employment of resources or tools acquired during the self-undeception process, can prevent future instances of self-deception.

What are the tools that are necessary to prevent self-deception? I take it that *self-reflection* is one good candidate. A characterization of self-undeception where the agent self-undeceives while exercising self-reflection, that is, by reflecting on their own condition of being self-deceived and coming to understand what went awry in their epistemic investigation, is a path that makes room for the agent's better epistemic future. Here is why.

First, given that self-deception is non-intentional and that the individual is not aware of their self-deceit, self-reflection allows the agent to gain awareness about their own condition. In this sense, self-reflection is a resource or a tool that provides the agent with the opportunity of recognizing what they see is a previous epistemic malfunction. A characterization of self-undeception where self-reflection is an integral feature not only allows for the agent to embrace reality, but it also allows them to *know why* embracing reality is warranted (it is warranted because the agent gains awareness of their epistemic shortcomings). In other words, the agent abandons their self-deceit, and also knows the *why* of such abandonment. Knowing why embracing reality is warranted is important because in providing the agent with an epistemic diagnosis of what went awry, the individual has the resources to learn from their epistemic shortcomings and prevent another future self-deceit from happening. Self-reflection and the agent's epistemic future then are tied together. Self-reflection provides the agent with the tools necessary to diagnose what they see as an epistemic malfunction and in the long run, it also provides them with the resources to prevent another self-deception. It is in this sense that a path to self-undeception that involves self-reflection sets up the newly undeceived agent for what they see as a potentially better epistemic future.

Surely self-reflection can contribute to such an epistemically improved future, but why is it so uniquely crucial? To answer this question and isolate why self-reflection is so important as a means to epistemic improvement, consider a path to self-undeception that does *not* involve self-reflection. Take for example someone who self-undeceives out of a lucky circumstance. Someone who is self-deceived that their son is not doing drugs suddenly hits his head and, upon recovery, now believes the warranted proposition that their son is a drug addict. Undeceiving in this way certainly does not involve any self-reflection. The agent, though successfully undeceived, does not know *why* their belief revision was warranted. In fact, the process of undeception merely *happens* to them, without making room for self-reflection as well as understanding their previous self-deceived condition.

Someone who undeceived out of luck is someone who is unlikely to learn from their epistemic shortcomings and, relatedly, is not a good position to prevent future instances of self-deception. The individual does not have the resources to improve their epistemic future nor cultivate themselves as a virtuous epistemic agent. For these reasons, a path to self-undeception that is motivated by self-reflection and that prepares the agent for future epistemic improvement is preferable to a path that involves neither of those aspects.

From what I said it should be clear in what sense the way we undeceive matters. However, I have not yet said anything about why the way we undeceive matters in a specific *normative* sense. What is it about a path to self-undeception that makes it preferable to another, *normatively* speaking? I suggest that the normative aspect of an account's superiority consists in the *recommendations* about how to self-undeceive that can be extracted from the account. Let me explain.

What kind of recommendations can an account that characterizes self-undeception as occurring out of luck give to self-undeceived agents? An account that describes self-undeception as a state resulting from lucky circumstances is not in the position of providing guidelines about how to self-undeceive because given that self-undeception simply occurs to the individual, a recommendation about 'being lucky' is not a useful suggestion. On the contrary, an account of self-undeception that involves self-reflection and that sets up the individual for a better epistemic future can provide suggestions about what steps the agent, and all agents in general, can take to foster conditions favorable to self-undeception. If self-reflection is triggered by a dialogue with a trusted friend, for example, a recommendation could be to cultivate personal relationships.

In what follows I argue that while both minimalism and pretense accounts provide a satisfying metaphysical and psychological characterization of self-undeception, neither provides a satisfying normative account of self-undeception. Both theories advance an account of self-undeception that is not motivated by self-reflection about the self-deceived person's condition, does not lead the self-deceived person to an understanding of their epistemic shortcomings, and does not provide them with the resources to improve their epistemic future.

On the contrary, the doxastic violation account advances an account that is superior to competing theories where self-undeception occurs by employing self-reflection and recognizing what the self-deceived person sees as a malfunction in departing from their epistemic norms. I will argue that this recognition can occur in different ways, including through interpersonal reasoning with an interlocutor or through the detection of an isomorphism with another individual's self-deceit, facilitated by empathy. As such, the doxastic violation account, unlike minimalist or pretense-based accounts, is in a normatively superior position in recommending that agents cultivate their interpersonal relationships as well as their empathetic abilities.

2. How Minimalism and Pretense Accounts Characterize Self-Undeception

The answer to the metaphysical question of what self-undeception amounts to is straightforward: for S to self-undeceive is for S not to be self-deceived *anymore* with respect to *p*. This answer is compatible with minimalism, pretense accounts, and the doxastic violation account.

However, answering the psychological question of how self-undeception occurs is more challenging because there does not seem to be one exclusive path to self-undeception. Presumably, each theory will yield a different account.

In this section I examine two psychological accounts of self-undeception that stem from minimalism and pretense-based accounts: self-undeception through *affective tipping* and self-undeception through *trumped incentives*. Self-undeception through *affective tipping* draws on psychological research and shows that motivated reasoners' rationalization collapses under the pressure of repeated disconfirming evidence. In "The Affective Tipping Point: Do Motivated Reasoners Ever 'Get It'?" Redlawsk et al. argue that when counterevidence is strikingly incongruent with their beliefs, motivated reasoners reach an *affective tipping point* where the reasoner's anxiety level increases to the point of forcing them to re-evaluate the evidence (Redlawsk et al. 2010). Redlawsk et al.'s account does not explicitly mention self-deception, but it can be utilized for self-undeception. The reason why is that since minimalism conceives self-deception as a species of motivated reasoning, Redlawsk et al.'s proposal can be transferred onto the theory. The idea is then to assess whether the mechanisms that unlock motivated reasoning can also unlock self-deception.

On the other hand, self-undeception through trumped incentives is instead a philosophical proposal advanced by pretense-based accounts and characterizes self-undeception in terms of costs adjustment (Gendler 2007). Pretense-based accounts suggest that when facing high-stake situations, the self-deceived person may abandon their self-deceit when the cost of maintaining it is *trumped* by the benefit of embracing reality.

2.1 How Minimalism Characterizes Self-Undeception: Self-Undeception Through Affective Tipping

A first, intuitive hypothesis for how self-undeception may occur is that, since the self-deceived person holds a belief at odds with the evidence, their self-deceit may collapse under the pressure of growing counterevidence. This is due to a negative emotional state that is caused by the evidence mismatch, which eventually leads the agent to revise their beliefs. This hypothesis is in fact one that is consistent with existing empirical research on motivated reasoning. Psychological studies show that motivated reasoners' rationalization skills are limited and reach a tipping point when confronted with growing disconfirming evidence.

A body of established and well-known psychological research shows that motivated reasoners do not update their beliefs in normatively correct ways (Kunda 1990, Betsch et al. 2010, Browne et al. 2015). When confronted with evidence that questions the justificatory grounds of their beliefs, rational agents are expected to adjust their doxastic attitudes by decreasing their level of confidence. Redlawsk et al. instead show that, due to the conative elements that influence their belief-formation, motivated reasoners "maintain and support their existing evaluations" (Redlawsk et al. 2010: 563).

In their experiments, Redlawsk et al. explore how voters' doxastic attitudes change upon receiving unexpectedly negative information about their favorite presidential candidate. Incongruent information, they argue, is not followed by "greater accuracy in evaluation", rather voters "mentally argue against" counterevidence and "bolster their existing evaluation" by recalling positive traits about their preferred candidate while discounting new negative information (Ibid., 567). This resistance to belief revision is explained by the mismatch between the voters' expectations about a candidate and the negative evidence encountered. The mismatch causes a negative emotional reaction that triggers resistance. Despite this resistance being tenacious, Redlawsk et al. show that it does not last under all circumstances. Their crucial finding is that when confrontation with disconfirming evidence is *continuous* and *especially threatening* (meaning, that it is particularly at odds with their existing beliefs), voters reach a tipping point after which they appear to "wise up" and initiate a correct updating of their evaluations (Ibid., 589). The experiment also shows that as voters reach a tipping point, their anxiety increases thereby indicating that anxiety may contribute to triggering a proper update of their beliefs (Ibid., 569).

Redlawsk et al.'s proposal does not mention self-deception, but only motivated reasoning. Yet their model can be applied to self-deception because minimalism argues that self-deception is a *species* of motivated reasoning. Their reference to anxiety can also be carried over to self-deception because, as I argued in Chapter 3, self-deception is commonly associated with an anxious phenomenology. If Redlawsk et al.'s model can be transferred to self-deception, as I am suggesting, then this means that their affective tipping point model can *also* map onto self-undeception. Let me now show how Redlawsk et al.'s proposal applies to minimalism.

Minimalism conceives of self-deception as a phenomenon where an examination of evidence unduly influenced by interests and motivations "non-deviantly" causes the individual to endorse a false belief (Mele 2001: 51). As we have seen, Sally may select or misinterpret the evidence in a way that just so happens to fit the more comfortable prospect that she does not have cancer. She may credulously believe confirming data, while scrupulously testing data that does not support her favorable hypothesis. According to minimalism, Sally may very well be aware of the interest she has in her not having a disease, yet she is not aware that such interest is shaping her epistemic investigation in epistemically problematic ways. In this sense, we can say that Sally *slides* into coming to form her self-deceptive belief whose flawed formation she is not aware of.

Despite the self-deceived person professing to endorse their belief, their self-deceptive state is associated with an uncomfortable phenomenology. Self-deception does not "feel good" for the self-deceived agent, who is nagged by *persistent* doubts and internal conflict that manifests through feelings of anxiety and worry. Due to a combination of uncertainty and a stake in the issue, the self-deceived individual is not totally confident in its truth-values and, as a result, their efforts to rationalize fail to fully counteract the "unwelcome evidence" (Lynch 2012: 440).

Now that I have laid out how Redlawsk et al.'s model maps onto self-deception, let me unpack how I think their proposal maps onto self-undeception. In particular, I show that Redlawsk et al.'s model validates what I call *self-undeception through affective tipping*. My conjecture is that, following Redlawsk et al., the tension generated by self-deception is what may lead the self-deceived person to a re-evaluation of evidence, and eventually to self-undeceive. Here is a case study of how self-undeception through affective tipping point may be construed.

Sally receives initial information from her doctor that indicates she may have cancer, yet she maintains her belief that she is healthy. Since her belief is false, we can presume that counterevidence may accumulate over time. Now, when it comes to small pieces of counterevidence, Sally easily rationalizes it away by mentally providing alternative explanations that can account for it. She may believe that receiving information from only one doctor is not sufficient to come to the conclusion that she has cancer or that diagnoses happen to be mistaken sometimes. She still professes that she is healthy, yet as she utters that, she betrays an anxious uncertainty that plagues her with recurring doubts. As counterevidence keeps mounting, we can imagine that Sally's health may begin deteriorating or that additional test results may uncontroversially show that she has cancer. At this stage, Sally's level of anxiety may increase and reach a tipping point where she simply cannot rationalize counterevidence any more due to its growing amount and striking incongruency with her existing belief. It is at this point that Sally may begin to update her beliefs according to the evidence.

2.2 The Limits of Self-Undeception Through Affective Tipping

Minimalism does not seem to provide a satisfying normative account of self-undeception because it advances a characterization that does not seem to constitute a desirable way to self-undeceive. To start, self-undeception through affective tipping relies on counterevidence continuously growing to the point of trumping the self-deceit because phenomenologically too costly to maintain. The model presents a portrait of self-undeception where it is not clear that the self-deceived person has employed any self-reflection on their condition. On the contrary, the agent abandons their self-deceptive belief because emotionally too laborious to rationalize, without initiating a process that leads to an understanding of what went wrong in their epistemic practice. In this sense, self-undeception seems to exclusively aim at lessening the agent's anxiety.

Because self-undeception through affective tipping does not involve any self-reflection on behalf of the individual who does not come to any understanding of their prior condition as self-deceived, it also limits the extent of which the agent's epistemic future can improve. If the self-deceived person does not need to acquire any realization that they were ever self-deceived, then they are unlikely to gain any helpful insight in how to improve their future epistemic practice. Granted, the affective tipping model does lead to belief revision, but this revision is somewhat *forced* on the self-deceived individual who does not show to have grasped their epistemic shortcomings.[67]

The affective tipping point model is also not in the position of being able to advance any recommendations for the improvement of the self-undeceived's future. The model is dependent on counterevidence becoming impossible to rationalize which may be a useful way to capture self-undeception psychologically, but not normatively. The reason why is that the only recommendation the model could advance is for the self-deceived person to wait, or hope, that counterevidence be *so*

[67] Someone might object that there can still be epistemic improvement even in the absence of a full understanding of one's epistemic shortcomings. After all, replacing one belief with another carries at least a minimal awareness that one has changed their mind and this may be still lead to some epistemic growth. I am not sure that this objection works. It has been suggested that when individuals replace their belief that *p* with a new belief that not-*p*, they do not concede that they have changed their mind and instead manifest a revisionist attitude according to which they profess to have believed not-*p* all along (Robin 2018). This reaction is, at least anecdotally, also shown by many individuals who are fortunate enough to self-undeceive. When pressed about their previous beliefs, they quickly respond that they "knew not-*p* was true all along". This makes me inclined to suspect that replacing beliefs does not obviously carry minimal awareness. Instead, replacing beliefs as a result of a self-reflection process seems better-suited to carry a more substantive level of awareness that sets up the agent for a future epistemic growth.

dramatically at odds with their existing beliefs in order for self-undeception to occur. This does not seem to be a suggestion that anyone who has the self-deceived person's best epistemic interests in mind would propose for it would require the self-deceived individual to maintain, at least initially, their unwarranted belief in the hope that the future may prove them wrong.

An account of self-undeception of this sort is akin to an account of how individuals overcome addiction by waiting to hit rock-bottom. This is a helpful account to have psychologically, but *ideally*, one would want to know what neutralizes addiction *before* addicts have to hit rock-bottom. Relatedly, the idea that one must hit rock-bottom in order to recover seems also unfalsifiable: if S has not recovered, then S must not have hit rock-bottom; and if S has recovered, then S must have indeed hit rock-bottom. The challenge then is to provide an account of self-undeception in those *moderate* circumstances that are *not* immune to rationalization and explain how, in *those* instances, belief revision may nevertheless occur. In light of these criticisms, self-undeception through affective tipping point does not seem to qualify as a plausible candidate for normative self-undeception.

2.3 How Pretense Accounts Characterize Self-Undeception: Self-Undeception Through Trumped Incentives

A second, promising account of self-undeception comes from pretense accounts, and most notably Tamar Gendler, which characterize self-undeception in terms of trumped incentives (Gendler 2007). Recall that pretense accounts conceive self-deception as a form of pretense where the self-deceived person does not believe the comfortable proposition they profess to endorse, but merely pretends to in the sense of make-believe, while often believing the more uncomfortable but warranted proposition.

Recall too that, according to pretense accounts, self-deception occurs in the following way. The self-deceived person begins imagining what it would be like if a particularly friendly proposition not-*p* was the case. In doing so, they merely engage in a wishful fantasy. Yet, entertaining the unfriendly belief *p* generates discomfort, so the individual minimizes interaction with evidence that supports it, instead directing their attention to those features of the world that nicely fit with not-*p* (Gendler 2007: 241). Over time, the pretense may intensify and shift from a *performative* kind of pretense, where the individual performs to their peers, to an *imaginative* kind of pretense*,* where the agent pretends to themselves. Both kinds of pretenses are behavioral; that is, the self-deceived subject merely *acts* as if not-*p* was true. Self-deception as properly conceived occurs when, as the pretense process unfolds, the individual's pretense comes to play the same role of beliefs in terms of vivacity and motivation to act.

In light of this, pretense accounts argue that self-undeception may occur when the subjective cost of maintaining the self-deception "trumps" the subjective cost of accepting reality (Ibid. 245). This may occur particularly in situations where the subject is forced to confront high-stakes choices. In these circumstances, the individual abandons their self-deceit because the subjective cost of sustaining the self-deceit is overridden by the subjective benefit of embracing reality (Ibid. 245). To see this more clearly, let us look at an example of trumped incentive in action.

Consider again the case of Sally who is self-deceived that she does not have cancer. We can imagine doctors presenting her with the following choice of starting a therapy that would give great health benefits to those who have cancer, and instead cause terrible pain to those without cancer, while not adopting the therapy would yield precisely the opposite effect.[68] If Sally genuinely believes that she does not have cancer, then she may very well decide *against* starting the therapy. But if she is merely pretending that she does not have cancer, as Gendler suggests, then she will instead allow her belief that she has cancer to guide her actions. In particular, when confronted with this high-stakes

[68] This example is based on the one that Gendler provides in Gendler 2007: 245.

forced choice, she will likely choose to start the therapy (Ibid., 245).

2.4 The Limits of Self-Undeception Through Trumped Incentives

Pretense accounts advance a characterization of self-undeception that on reflection does not seem to constitute a desirable way in which self-undeception should occur. According to pretense accounts, just as the self-deceived individual may slide into their state, so too may they slide out without necessarily gaining *awareness* that they were ever self-deceived. What seems to trigger self-undeception then does not seems to be the agent self-reflecting on their condition or on what might have possibly gone wrong in their epistemic practice. Rather, self-undeception through trumped incentives depicts the self-deceived agent undeceiving out of *convenience* by gravitating towards the option that at the moment is perceived as less costly. The agent *does* reconcile with truth, yet this reconciliation occurs because the warranted belief just so happens to be the most practically advantageous choice. This focus on advantageousness comes at the expense of an understanding of the agent's epistemic shortcomings. This is a drawback because the absence of self-reflection and understanding does not bode well for the self-deceived subject's epistemic future, and possible prevention of a future self-deceit.

Of course, those who are inclined to side with consequentialist views may be unsympathetic to this argument and object that, so long as self-undeception occurs, *how* it occurs seems to be irrelevant. Yet even from a consequentialist perspective self-undeception through trumped incentive falls short. The reason why is that the agent's epistemic life, so to speak, does not end at the time in which self-undeception occurs. If one is interested in what (to the agent) constitutes an epistemic improvement, promoting a *reflection-free* route out of self-deception is not ideal because it make more difficult the agent's access to those epistemic goods that would indeed contribute to their growth. In this sense, self-undeception is not just a valuable state to be acquired in the short term as a matter of circumscribed alignment with reality, but it can signify the beginning of a diachronic process where the agent has the opportunity to become more virtuous, epistemically speaking. The absence of self-reflection however renders such possibility less likely and as such, pretense accounts face a similar drawback to minimalism: their characterizations of self-undeception make no room for what the agent may see as future epistemic improvement.

3. Static Self-Undeception

Surely there are circumstances where we accept truths out of situational opportunism because it just so happens that they turn out to also be practically advantageous, but these circumstances do not exhaust all the ways in which we reconcile with truth. Presumably, there are other ways to self-undeceive that do not bear on such a convenience, but that are the result of a self-reflection process that leads us to an understanding of why a change is needed. This point makes more sense if we consider a parallel with the moral domain.

There certainly are cases when performing the *right* action coincides with performing the most *prudentially advantageous* action. Consider for example the desegregation of public housing that took place in Yonkers, New York in the late nineties.[69] A federal judge, witnessing the city's opposition, ordered Yonkers to proceed with the desegregation, where failure to do so would incur a one-hundred dollar per day fine. The mayor decided to comply. His decision was surely the right thing to do, morally speaking, but it was also the most advantageous action because it protected the city from potentially going bankrupt. Yet surely there are also instances when we do the right thing regardless of any prudentially beneficial consequence. I think the same can be said for the epistemic domain. There are instances where we embrace truth even when it is not convenient to do so. Such is the case for those

[69] Former New York Times writer Lisa Belkin analyzes Yonker's case in her book *Show Me a Hero* (1999).

instances of self-undeception that I call *static*, that is, instances where an individual S may self-undeceive even though the subjective cost of maintaining self-deception is still lower than the subjective cost of accepting reality.

Static self-undeception may seem puzzling, for it does not seem to involve any convenient incentive whatsoever. The (unconscious) aim of self-deception is to avoid a painful truth, and here the self-deceived person abandons their state precisely by accepting that painful truth that drove their self-deceit in the first place. The defenders of minimalism and pretense accounts might ask what there is to be gained for the agent in self-undeceiving in this way. The two accounts might then question if static self-undeception is even plausible. In what follows I argue that it *is* plausible; in fact, more so than it seems. A structurally similar case in the domain of action will illustrate why.

> *Affair*. Chris is a committed partner who dearly loves Tom. Chris's parents raised him to be a kind person and taught him that betraying people's trust is one of the worst things someone can do. Chris has always been faithful to Tom and to all of his past partners. However, despite his love and commitment, he cannot help but notice that lately Tom has been insufferable. He has been avoiding housework, spending almost every night on the couch watching hockey, leaving Chris the burden of having to take care of their one-year-old child. Chris does not want a divorce and wishes Tom was not such an immature, selfish partner. He is still very much in love with him; yet one night he finds himself spending the night with Paul, whom he met at the grocery store while purchasing milk for his child. Chris feels guilty to have betrayed Tom's trust. He decides not to tell him, yet finds himself drawn intensely to Paul to the point of starting an affair. Despite his temporary, yet deeply fulfilling, happiness with Paul, Chris feels terrible, both towards Tom and also towards himself. He is ashamed to have become the kind of person who betrays the trust of those close to him.

If you empathize with Chris, as I do, you probably wish for his affair to end. This could happen in a variety of ways.

> *Ending 1*. Chris's affair becomes too costly to maintain. We can imagine Paul falling in love with Chris and threatening to tell Tom about the affair.

> *Ending 2*. Chris's marriage becomes more bearable. We can imagine that Tom makes an effort to be a better partner and that Chris's motivation for cheating consequently dissolves.

Note how *Ending 1* and *Ending 2* structurally echo self-undeception through affective tipping and self-undeception through trumped incentives where either reality is forced on the self-deceived person or it becomes more bearable.

These are plausible endings, but normatively, they do not seem to occur in a way that is desirable for Chris. Chris does not end his affair upon self-reflecting and understanding why his actions are morally wrong. On the contrary, he ends it because at the moment it is convenient for him to do so. Had Paul not threatened to tell Tom about the affair, Chris would probably still be seeing him. Had Tom not made an effort, Chris would still be unfaithful. As with self-undeception through trumped incentives and through affective tipping, the different endings I presented do not occur in a desirable way.

In light of this, a third, better way in which the affair could end is the following:

> *Ending 3*. Chris ends the affair *despite* its still being pleasant and Tom still being a bad partner: he ends the affair because he realizes that his action is wrong according to Chris's own moral code.

Ending 3 certainly seems possible. Not every right action stems from convenience. Sometimes we perform the right action *because* it is the right thing to do, even if it hurts. We do so because upon reflection we come to the conclusion that we have a moral code that importantly contributes to our personal identity, and there is a psychological price to pay for violating it. This price is that we may turn into a morally reprehensible person. If one is inclined to take *Ending 3* seriously, then one should also be inclined to take seriously static self-undeception. Given their structural similarity, *Ending 3*'s plausibility carries over to static self-undeception, at least *prima facie*. In fact, both *Ending 3* and static self-undeception involve a reconciliation (with epistemic norms in static self-undeception and moral norms in *Ending 3*), the source of which does not stem from mere contingent convenience or self-protection, but stems from self-reflection and an aim to be a better epistemic agent. If what I say about the plausibility of static self-undeception is correct, then the challenge is to explain how the self-deceived individual could ever cease to be in their state even in absence of any convenience.

I suggest that both minimalism and pretense accounts are ill-equipped to explain static self-undeception. Self-undeception through trumped incentives, for one, relies on self-deception becoming too costly to maintain. Yet in static self-undeception the subjective cost of both reality and self-deception remain constant, so it seems that one cannot apply the same line of reasoning that worked for self-undeception through trumped incentives. Self-undeception through affective tipping point also cannot account for static self-undeception because Redlawsk et al.'s suggestion is that motivated reasoners reach a tipping point when growing counterevidence is too incongruent to rationalize. Thus, the affective tipping model works only for those instances of self-deception where counterevidence grows continuously and generates growing anxiety to the point that it makes the self-deceit too costly to maintain, phenomenologically speaking. But in static self-deception there is no counterevidence growing because nothing has changed in the situation that the self-deceived person faces, so it is unclear how the affective tipping model would explain it.

The reliance on counterevidence continuously growing also raises another issue: cases where counterevidence accumulates are cases where the self-deceptive belief in question is false, and so increasing disconfirming evidence is the result of the belief failing to capture reality. Now, some, if not most, instances of motivated reasoning do produce a false belief. But not all of them do. Suppose for example that I am heading a search committee and propose to hire an applicant that I profess to be the most qualified. If asked, I can provide several reasons in support of the applicant who *in fact* happens to be the most qualified. However, in addition to those reasons, the applicant and I also happen to be close friends. I believe that they are the most qualified, but my belief-formation has been unduly influenced by motivational elements –that is, their being a close friend make me positively biased in their favor. This is an example of motivated reasoning that results in a belief that happens to be true.

Redlawsk et al.'s affective tipping point model does not work in this case because, due to the belief being true, no disconfirming evidence is expected to accumulate since the belief accurately captures reality. The account of self-deception modeled after Redlawsk et al. is also vulnerable to this criticism because the condition that the self-deceptive belief acquired be false is not a necessary one, according to the doxastic violation account. Thus, there may be instances of self-deception where the individual is self-deceived, yet the proposition in question turns out to be true (although these cases might be rare).

To recap, there are no features of minimalism and pretense accounts that one can appeal to in order to explain static self-undeception. Minimalism's characterization of self-undeception relies on counterevidence growing to the point of being impossible to rationalize, while pretense accounts rely on self-deception becoming too costly to maintain. Static self-undeception however does not hinge on either of these features, and if minimalism and pretense accounts cannot explain static self-undeception, then it is clear that additional conceptual resources are needed in order to explain it.

Yet, how might one explain why a self-deceived individual would be motivated to abandon their self-deceit if it is still the easier path to embrace? An insight comes from the above example of Chris. Recall that in *Ending 3* he notices that, while engaging in an affair, he has been departing for his moral code and elects to break up with Paul because he intends to recommit to his moral norms. Similarly, I suggest, the self-deceived person may abandon their self-deceit because through self-reflection they may coming to the realization that they have departed from their epistemic norms. Following such realization (that may occur in several ways, as I explain in the next section), the self-deceived person may elect to recommit to their epistemic norms. This characterization of self-undeception constitutes a normatively proper account of self-undeception where the phenomenon involves a reflection and recognition of a violation of epistemic norms.

From this it is clear that the doxastic violation account is a good candidate to properly explain static self-undeception because the notion of epistemic norms, unlike minimalism and the pretense account, is integral to the view. However, according to the doxastic violation account the self-deceived person "slides" into their self-deceit unknowingly, without being aware that they are self-deceived. The doxastic violation account then faces the challenge of explaining, psychologically, how the self-deceived person could come to realize that they have departed from their epistemic norms. This is the task that I take on in the next section.

4. How the Doxastic Violation Account Characterizes Self-Undeception

In this section I show how the doxastic violation account can capture static self-undeception. The doxastic violation account validates two paths that can explain how the self-deceived person could come to realize that they have departed from their epistemic norms: *dialectical self-undeception* and *isomorphic self-undeception.*

Dialectical self-undeception involves reasoning the self-deceived subject out of their own deceit by means of a dialogue, which I call *self-undeceptive interpersonal reasoning*. Here an interlocutor aims at making the self-deceived individual more *open* to the uncomfortable belief in question in a psychologically non-threatening way. By relying on a stance of trust, the interlocutor invites the self-deceived person to consider evidential reasons in favour of such openness. My conjecture is that openness to evidential reasons is achieved by the interlocutor inviting the self-deceived person to reflect on their departure from their epistemic norms and by encouraging a motivation to recommit to them.[70]

This is not to say that self-undeception must necessarily occur through interpersonal means. It is not necessary to wait around, so to speak, for an interlocutor to take on the undeceiving task. In isomorphic self-undeception, I propose that the individual may self-undeceive *intrapersonally* by relying on their own epistemic resources. According to this proposal, a self-deceived agent S recognizes an isomorphism between her own situation and the situation that another self-deceived agent B experiences. Self-undeception occurs when S is able to transfer her own judgment of B to herself and her situation. Witnessing B's self-deceit may lead S to the realization that they have also been violating their epistemic standards which in turn may lead them to re-commit to their epistemic norms.

Successfully detecting an isomorphism between S and B is, however, unlikely to occur in purely rational ways. Though only anecdotal, self-deceived individuals are often able to identify self-deception in others without transferring the same judgment to themselves. Recognizing an isomorphism is also more challenging when the individual in question is someone with whom the self-deceived agent has little in common and with whom they do not easily identify. For these reasons, I suggest that *empathizing* with the self-deceived individual B may create an emotional dynamic able to bypass the obstacles that would be in place if the recognition of the isomorphism occurred in purely rational ways.

Note how both dialectical self-undeception and isomorphic self-undeception hinge on the self-deceived person undergoing self-undeception even when maintaining their self-deceit would still

[70] Note that the two paths I propose here are not meant to be our exclusive means to self-undeception. Though I am open to the possibility that there might be other promising models, my primary aim here is to advance the debate on self-undeception.

be the most convenient option (which is what static self-undeception is about). Yet why would the self-deceived person be motivated to recommit to their epistemic norms? Recall that in Chapter 2 I suggested that the epistemic norms that the self-deceived person violates are *loaded epistemic norms*. That is, they are norms that may contribute to an individual's identity, modulate access to particular communities, or foster positive emotions. The fact that the agent holds these norms and that these norms also come to possess these non-doxastic features is what I suggested causes the self-deceived person to experience mental tension once the norms are violated.

Now suppose that Sally holds the epistemic norm that she ought to trust her expert colleagues. That norm is not simply a *bland* norm, rather it is a *loaded* norm. It contributes to who Sally is because she is a specific kind of agent: she is a doctor. So that norm informs her identity in important ways. It is also a norm that allows her to be part of the medical community. Violating that norm implies a disruption of Sally's identity, and this disruption is revealed by the fact that Sally experiences mental tension. Not only is Sally doubtful about her self-deceptive belief that she does not have cancer, but she is also doubtful about the epistemic practice that she herself has conducted. The feeling of alienation, of being *out of sorts* is indicative of such agent-centered tension. Now in recognizing that she has violated her epistemic norm, and that this norm is not just a bland norm, but partially makes her who she is, she may realize that the violation contributes to make her a person that she is not, and that she may disapprove of. Thus, my suggestion is that Sally's motivation to recommit to her epistemic norm may derive precisely from the motivation of being true to herself, or restoring the kind of person she is: a person who trusts her expert colleagues. Coming to this realization involves, and requires, a process of self-reflection. Note how this is also not a path that stems from convenience of prudential reasons: often our motivation of committing to our norms (epistemic or moral) comes at a great price as it involves accepting realities that we cannot stomach. In Sally's case, her reconciliation with her epistemic norms comes at the price of believing that she has cancer. The challenge is then to explain how self-reflection could be triggered.

4.1 Dialectical Self-Undeception

One of the defining aspects of the doxastic violation account is that the self-deceived person lets their motivational bias cloud their judgment to the point of endorsing a belief that is at odds with their epistemic norms. It follows from this that an important desideratum for a plausible account of self-undeception is one that is able to cut through the filters of the self-deceived person's lack of awareness.

Satisfying this desideratum is no easy task. As Redlawsk et al. also point out, any small piece of counterevidence is likely to trigger the self-deceived person's rationalization. This, however, is not a reason to conclude that attempting to talk the self-deceived subject out of their own deception is a lost cause. In the next section I argue that it is indeed possible to help reason the self-deceived person out of their own condition. One simply needs to do it in the *right way*. The right way is, I suggest, to trigger self-undeception through self-undeceptive interpersonal reasoning.

In a nutshell, self-undeceptive interpersonal reasoning occurs between two interlocutors, say, Sally and Charlie, one of which, Sally, is self-deceived that she does not have cancer. The reasoning focuses on two levels: emotional and rational. It works on an *emotional* level by pressing on Sally's openness to the uncomfortable belief that she has cancer and by relying on a stance of trust, and it works on the *rational* level, by showing Sally reasons in favor of why she should adopt such openness. If the reasoning is successful, at the end of the conversation Sally may internalize Charlie's words. As a result of this internalization, Sally may realize that she has been violating her epistemic norms and may recommit to them. Dialectical self-undeception however does not hinge on reality being *factually* easier to embrace, rather it involves the self-deceived person learning (through interpersonal reasoning) to better cope with it. More formally, I define self-undeceptive interpersonal reasoning as a kind of interpersonal reasoning between two agents, S and C, such that:[71]

i. S and C are two interlocutors where
ii. S is self-deceived about not-*p* and C is not self-deceived about not-*p*
iii. S trusts C, and
iv. C stimulates S's openness about *p* by providing practical reasons in favor of such openness.

Given this, S ceases to be self-deceived with respect to a proposition *p* dialectically if and only if:

1. S is self-deceived about not-*p*,
2. S engages in self-undeceptive interpersonal reasoning, and
    1. Self-undeceptive interpersonal reasoning is a non-deviant cause for S coming to abandon the belief not-*p*.[72]

Let me now unpack how dialectical self-undeception may unfold. To start, there is a background condition that must be in place for self-undeceptive interpersonal reasoning to be effective. The background condition in question builds on the distinction that Peter Strawson draws between an "objective" and an "interpersonal" stance (Strawson 1962: 13).[73]

[71] I borrow the term interpersonal reasoning from Strawson 1962. See also Manne 2014 for a use of interpersonal reasoning that resembles mine in many ways.

[72] Despite the formal characterization I just provided, I do not intend to identify a set of necessary and sufficient conditions for self-undeception. My aim is more modest than that—to instead isolate the most salient features of the phenomenon.

[73] I am deeply indebted to Kate Manne's paper "Internalism: Sad but True?" that borrows the same distinction in order to support a form of reasoning where one of the two interlocutors come to see reasons for φ. My argument is similar to Manne's in many respects but utilizes Strawson's interpersonal stance as a background condition for dialectical self-undeception to be successful. See Manne 2014: 95 for more on how Manne employs Strawson's distinction.

Strawson famously argues that individuals have different ways to relate to each other: they may adopt an interpersonal stance, which involves reasoning with an interlocutor as equal, rational peer, or they may adopt an objective stance, where one treats their interlocutor as someone with whom there is no other "civilized attitude" than seeing them as someone to feel sorry for, "handled" or "cured" or simply to be "avoided" (Ibid., 10). For example, one may adopt an objective stance towards those who exhibit compulsive behavior, young children, or "incapacitated" people (Ibid., 13). An "incapacitated" individual, Strawson suggests, is an individual whose picture of reality is "pure fantasy", someone who does not live in "the real world at all" and who is "acting out of unconscious purposes" (Ibid., 13).

Though the self-deceived person does not qualify as neurotic or compulsive, one may still be tempted to view them as someone with whom to adopt an objective stance. After all, in tenaciously refusing to embrace reality, there is a sense in which the self-deceived individual is living in a fantasy of their own by failing to see what is indeed so transparent to others.[74] Yet I suggest that we resist the temptation of conceiving the self-deceived individual as a hopeless peer, and instead adopt an interpersonal stance where we treat the self-deceived person as a peer we *can* argue with. Since self-undeceptive interpersonal reasoning has the form of a dialogue, positing an interpersonal stance is a necessary condition for it to be effective. In adopting an interpersonal stance, we approach the self-deceived person as an equal peer, someone who is willing to take seriously others' perspectives as well as rationally capable of appreciating the force of reasons.

With that in mind, self-undeceptive interpersonal reasoning focuses on both a rational level and an emotional level. The reason for this double focus is twofold. First, a reasoning that aims at presenting counterevidence is bound to fail because, as also Redlawsk et al. point out, counterevidence

[74] Though the label of "incapacitated", or of someone to be "handled" or "cured" would be too strong and misrepresentative of the self-deceived.

would be dismissed by the self-deceived person's rationalization skills. This is especially true if we again think that self-deception is sustained by the desire of an uncomfortable proposition to be false and the fear that it might indeed be true. My conjecture then is that a successful dialectical self-undeception must take into account these conative elements that drive self-deception in the first place.

In our familiar case of Sally, this does not mean, however, that Charlie has to soften Sally's wish that she does not have cancer; rather Charlie must focus on stimulating Sally's openness to the uncomfortable belief that she may have cancer. In fact, the self-deceived person will probably always wish that their comfortable belief is true. A self-deceived mother will probably always wish that their son did not do drugs; a partner will always wish that their lover was not a liar. Similarly, Sally will probably always wish that she did not have cancer. Thus, I suggest that Charlie stimulates Sally's *openness* towards the hypothesis that she might have cancer. Still, this hypothesis is a rather dim one, so how could Charlie *realistically* facilitate Sally's openness to it?

My suggestion here is that this openness must occur in such a way that it is not perceived as psychologically threatening. Charlie may begin by sympathizing with Sally's anxiety, and by showing understanding for her situation. After all, who does not wish to be healthy? Charlie may then proceed by pointing to medicine's success in treating cancer as well as share research on pioneering therapies. He may point to Sally's strength and determination, and stress that for as frightening as it may seem, he is confident in her resourcefulness to fight the disease. With the right therapy and social support, Sally can win this battle. In appealing to Sally's personal resources, he may point out that it is unlike her to succumb to an unwelcome diagnosis. Charlie may remind her of all those moments where Sally, as a doctor, encouraged patients to come to terms with their illness. In stressing that she is not accepting her own diagnosis, Charlie may point out that she is going against her epistemic values that she herself endorses, and that greatly contribute to being who she is. Sally is not the kind of person who distrusts her colleagues, who disvalues evidence, and in general, hides away from reality. In appealing to who Sally is as a person, Charlie directly targets Sally's own epistemic norms and facilitates Sally's realization that, in not accepting her diagnosis, she has not been true to herself.

More specifically, Charlie may point out that Sally has violated a (loaded) epistemic norm that contributes to her identity, as well as allows her to be an integral part of the scientific community. Now, because the loaded norm possesses features that contribute to Sally's identity and to her belonging to a community that she values, Sally is motivated to preserve both aspects, and thus motivated to recommit to her epistemic norm. Her recommitment however should not be interpreted as stemming from self-preservation or convenience, because the self-deceptive belief is still the more comfortable belief to embrace. Sally self-undeceives *in spite* of her self-deceit still being the easier option.

Charlie then aims at easing Sally's openness to the prospect of her having cancer, and he does so by appealing to her epistemic norms. Yet, nothing so far would in theory prevent Sally from rationalizing away Charlie's words. What is needed for Sally to take seriously Charlie's suggestions, is a non-psychologically threatening element that provides the necessary condition for preventing the activation of Sally's rationalization. The element of *trust* fits nicely with this task. I want to suggest that Sally may be more likely to embrace Charlie's words if the two are in a *stance of trust* towards each other. I take trust to involve a "state of mind" where Sally is willing to take Charlie's words seriously, not in virtue of their content, but in virtue of the relationship in which they stand.

Trust does not imply that Sally may automatically believe what Charlie tells her nor does it imply that the trusted person must be an individual whom Sally already trusts. As Richard Holton points out, one can also *decide* to trust (Holton 1994: 63). This can be seen when one decides to trust one's own therapist, for example. Charlie's invitation to be believed is then an invitation to adopt a "stance of trust", more than to believe him *tout court* (Ibid., 75). It is not a statement that Sally "straightforwardly believes or disbelieves", but an invitation to take a stance of trust (Ibid.). It is only after having adopted such a stance that there is a sufficient background condition for allowing Sally to be more receptive and less inclined to rationalize Charlie's reasoning. If the reasoning is successful, Sally may internalize Charlie's words, where by "internalization" I mean that Sally *feels their force* and appreciates their epistemic import. As a result, the possibility of her having cancer may not be perceived as a reality that she has no power over but, on the contrary, as a challenge that she can handle.[75]

One might object that interpersonal reasoning still relies on altering the cost of beliefs. But this would be a misinterpretation of the model. Interpersonal reasoning does not rely on costs because, in self-undeceiving, Sally does not embrace an easy perspective. In embracing the warranted belief that she has cancer, she is *in fact* embracing a costly proposition.[76] Interpersonal reasoning facilitates her reconciliation with reality by providing her with the appropriate psychological resources to confront and accept the uncomfortable perspective that she has the disease. Thus, Sally cannot be said to gravitate towards the easiest path to accept. Quite the opposite. She might very well fail to overcome cancer, but thanks to interpersonal reasoning she now knows that she will not be alone, she will find in Charlie an *ally* who will support her through this difficult challenge.

4.2 Isomorphic Self-Undeception

Now that I have laid out a first path that might allow Sally to come to the realization that she has violated her own epistemic norms, let me turn to a second characterization of self-undeception which occurs *intrapersonally*, that is, without requiring outside intervention. The aim of presenting this

[75] Further evidence for interpersonal reasoning's success can be found in the psychological literature on cognitive behavioral therapy, which has been shown to have positive outcomes in treating delusional beliefs. See Kingdon & Turkington 1994.

[76] It might be objected that it could still be less costly, overall, to retain Sally's identity and the unwelcome belief that she has cancer rather than losing her identity and access to the community. From this observation it would follow that self-deceiving is in fact a more costly option for Sally. But if the objection was right, it would be unclear why Sally self-deceived in the first place. If the embracing the unwelcome belief had been easier than losing her identity, Sally would have naturally followed her epistemic norms. But Sally *did* deceive herself, and she did so precisely to avoid a truth that she cannot stomach.

second path is to show that self-undeception does not necessarily have to occur interpersonally—that is, with the self-deceived person biding their time until an external interlocutor reasons with them—but it can also occur in intrapersonal ways that draw upon the self-deceived person's own resources.

In general, identifying our own biases is a quite difficult task. It is often through conversations with other people who direct our attention to our behavior, that we become aware we might be biased (Poos et al. 2017). Given that self-deception is also a bias, and one which the self-deceived person is unaware of, dialectical self-undeception seems to be particularly well-suited to facilitating the process of undeception. But the plausibility of dialectical self-undeception should not lead one to conclude that it is our exclusive means to self-undeception. In fact, there are other viable paths, one of which I discuss in this final section. I advance a model of self-undeception that occurs intrapersonally, and in particular, by detecting an isomorphism between S's self-deceit and another individual B's self-deceit. Recognizing an isomorphism between the two instances of self-deception involves S seeing their own situation *in* B's situation and transferring their own judgment of B to themselves. Building on the suspicion that this last step is unlikely to be successful if occurring only in purely rational ways, I suggest that empathy can contribute in facilitating the self-deceived person recognition of the isomorphism. Detecting this isomorphism can in turn allow the self-deceived person to gain insight into their own condition by self-reflecting and coming to the realization that they have departed from their epistemic norms.

Here is a brief example of how isomorphic self-undeception can occur. Suppose Sally believes she is healthy and behaves consistently with her belief. Suppose also that, in ways that I will explain in a moment, she happens to interact with another individual, Beatrice, who experiences a situation similar to hers. Beatrice has compelling evidence that she has an autoimmune disease. She very much desires that not to be the case, and stubbornly rationalizes disconfirming evidence. Suppose also that, due in part to the striking similarity with her situation, Sally empathizes with Beatrice in such a way that she can feel the force of Beatrice's experience.

The mechanism of empathy can facilitate Sally's identification with Beatrice's perspective and, in particular, it may allow Sally to see her own situation in Beatrice's situation; that is, it may allow Sally to detect an isomorphism. This initiates a self-reflective process that may lead Sally to come to understand that, similarly to Beatrice, she has incorrectly interpreted the available evidence and, as a result, has not been true to her epistemic norms. My conjecture is that if Sally is able to transfer her own judgment of Beatrice to herself, this time with the realization that she has violated her epistemic norms, this epistemic process may lead her to abandon her self-deceptive belief.

Given this, S ceases to be self-deceived with respect to a proposition *p* through isomorphic self-undeception if and only if:

1. B and S are two individuals
2. S is self-deceived about not-*p* (S holds the belief that not-*p*), and
3. B is self-deceived about a proposition not-*p** (B holds the belief that not-*p**, where not-*p** has the same or similar content as not-*p*),
4. S empathizes with B.
5. S detects an isomorphism between S's self-deceptive experience and B's self-deceptive experience
    a. The detection of the isomorphism is a non-deviant cause for S coming to abandon the belief that not-*p*.

Let me now explain in more detail how an instance of isomorphic self-undeception can be construed. Let us consider again the case of Sally who is self-deceived about her not having cancer and, as a result, she holds the belief that she is healthy. Sally desires to be healthy and fears that she may not be. She can perfectly be aware of such conative components, but she is not conscious of the directional influence that these exert on her beliefs.

We can imagine Sally being an ordinary person leading an ordinary life. This includes having a wide range of *experiences*, from reading literature, enjoying theatre, and her favorite music, watching movies, conversing with friends, and so on. Suppose now that Sally is reading a book that portrays the story of a family over forty years. To her surprise, Sally shares several traits with Beatrice, the main character of the book. They are the same age, both with a degree in medicine, both parents of two. They also have strikingly similar personalities: stubborn and hopelessly optimistic. The book is captivating, and Sally often loses herself in the story. Due to her similarity with Beatrice, she naturally identifies with her.

In a rather moving scene, Beatrice's health begins deteriorating. She experiences sudden fevers, dramatic weight loss, blood losses, chronic fatigue, and so on, all of which force her to miss time from work. Rattled by this change of events, Beatrice knows those are clear symptoms that she would normally attribute to an autoimmune disease. Yet she finds herself justifying her weight-loss to a diet she has started the week before, and interprets her fever as simply the results of her stressful job. Appalled by her rationalization, Beatrice's colleagues avoid her, and Beatrice isolates herself even more. Her family is increasingly worried, and instead of accepting their attempt to help, Beatrice becomes aggressive. She does not trust her colleagues anymore, treats her patients coldly, and taking questionable advice she finds on the internet that her colleagues rightly question. Though Sally empathizes with Beatrice's feelings, she cannot help but notice how her character has changed. In fact, Sally can barely recognize the Beatrice that she got to know at the beginning of the book. She can almost feel Beatrice's changing and, immersed in her character, she can see her own situation *in* Beatrice's situation.[77]

Sally knows that it is unlike Beatrice to distrust her colleagues, and Sally also knows that it is unlike *herself* to hide from reality. Yet Beatrice has been drifting away from her identity. Is Sally drifting away from her identity the same way Beatrice is? And if so, why? Detecting an isomorphism between

[77] Empirical evidence suggests that there may be a causal connection between fiction and empathy development. For more on this see Kidd & Castano 2013. See also Panero et al. 2016 for a critique of their view.

the two situations can initiate a reflection of this sort that may eventually lead Sally to apply her own judgment of Beatrice to herself and her situation, thereby recognizing her own epistemic shortcomings. If Sally is able to maintain self-control and not let her fear cloud self-reflection, then her inquiry may lead to admitting that the reasons why she feels so out of sorts and not quite herself are due to a departure from her epistemic norms (which contribute to who she is as a person). Thus, the motivation to preserve her identity might motivate her to *recommit* to her epistemic norms, not because this is the easy path to endorse—admitting one has cancer is never easy, after all—but for the sake of being true to herself, and thus for the sake of realigning with her epistemic values.

In the scenario just described, empathy is revealed to be a crucial background condition that contributes to the detection of an isomorphism. The reason why empathy plays such a determining role, I suggest, is that identifying an isomorphism is presumably more challenging when the other individual involved is someone who the self-deceived person struggles to identify with or whose self-deceptive proposition is radically different. In fact, resemblance of manners, character, language may more easily trigger our empathy towards others (Hume 1739: 2.1.11), as well as with those whose emotions are directed towards the same "object" as ours (Maibom 2017: 24). In this sense, empathy creates an emotional dynamic that allows Sally to gain insight into her own condition and bypasses those obstacles that would be in place if Sally was to detect an isomorphism in purely rational ways.

Yet how does the detection of an isomorphism through empathy facilitate self-undeception? It does so, I suggest, in virtue of the fact that empathy often influences agents' future motivations to act (Batson 2011: 60). Sally may recall those instances when, as a doctor, she witnessed patients refusing to accept their health problems and her attempt to convince them otherwise. She may even feel embarrassed to have indulged in those very same behaviors that, as a doctor, she so effortfully counteracted, and almost feel as if in doing so she has betrayed her own values. On the contrary, detecting an isomorphism in rational ways may not generate a motivation to act, and thus may not represent a promising path to self-undeception.

It is the ability to recognize an isomorphism between similar self-deceits, and identify with the individuals who are experiencing them, that builds the conditions that allow Sally to gain insight into her own situation. Acknowledging her own irrationality can be the welcome outcome that stems from such process. Admitting her disease may be a painful truth to stomach, but it comes with the realization that in accepting it Sally is being truer to her identity and, indirectly, to her epistemic norms. She may realize that she is not necessarily more rational than her patients, and that no one is immune to the tricks of a rose-tinted lie, not even her.

4.3 Dialectical and Isomorphic Self-Undeception and the Superiority of the Doxastic Violation Account

I have now presented two ways in which the doxastic violation account could explain static self-undeception. I have argued that, for structural reasons, both minimalism and pretense accounts are ill-equipped to account for static self-undeception, since both of them rely on a dynamic, changing situation. In this section, I want to stress the connection between the two techniques I described, dialectical and isomorphic self-undeception, and the doxastic violation account. My aim is to defend the superiority of the doxastic violation account by showing why dialectical and isomorphic self-undeception are not in harmony with minimalism and pretense accounts. Further, I will show how the doxastic violation account does achieve the required harmony here.

To this purpose, two points are worth assessing. First, a *structural* point about whether the structure of minimalism and pretense accounts is even suited to employ dialectical and isomorphic paths to self-undeception; and second, a *practical* point about whether the two techniques can be predicted to successfully yield self-undeception (assuming that minimalism and pretense accounts can employ them). A negative answer to the structural point commits us to a negative answer to the practical question. If neither dialectical nor isomorphic self-undeception are structurally suited to be employed by minimalism and pretense accounts, then it does not make sense to ask whether the two techniques can be successful. However, below I suggest that there is nothing that structurally prevents minimalism and pretense accounts from employing a dialectical or isomorphic strategy, but the two techniques, if employed, are unlikely to be successful due to the internal features of minimalism and pretense accounts.

4.4 Dialectical Self-Undeception vs Pretense Accounts and Minimalism

To start, could Gendler's pretender have a trusted interlocutor encourage her to drop her pretense? I think she cannot. Dialectical self-undeception employs a form of interpersonal reasoning aimed at providing the self-deceived with reasons to abandon their self-deceptive state. As we have seen, for the doxastic violation account these reasons hinge on the valence that the (violated) *loaded* epistemic norms have for the individual. A trusted interlocutor can point to reasons that pertain to Sally's identity by suggesting that she has become the kind of person who does not trust her colleagues. Alternatively, the interlocutor can also point at Sally's ostracization from the scientific community and suggest that she is isolated from the colleagues she usually defers to. These reasons motivate Sally to act, so to speak, and revise her self-deceptive belief.

I am skeptical that pretense accounts can provide reasons that are sufficiently motivating to spark action in the same way Sally's revision process is triggered by reasons that appeal to her identity. Given that pretense accounts' characterization of self-undeception relies on incentives, and that static self-deception does not involve any situational change, pretense accounts must identify other reasons in order for interpersonal reasoning to even be a viable option. This is not a problem for the doxastic violation account because, as we have seen, the theory does not exclusively depend on a change of dynamic. Emphasis on epistemic norms makes available a wide range of reasons that the doxastic violation can appeal to when explaining static self-undeception. Yet pretense accounts do not make use of the notion of epistemic norms, so the reasons that interpersonal reasoning is said to provide must be found elsewhere.

One feature of pretense accounts that could be useful for this purpose is the individual believing the uncomfortable proposition (while pretending otherwise). Pretense accounts stress that although the self-deceived engages in a pretense that *p*, they often believe that not-*p*. If we then imagine what a conversation with a trusted interlocutor might look like for pretense accounts, the interlocutor could encourage Sally to abandon their pretense by appealing to the fact that, after all, she deep down believes that she has cancer. Already believing that she does is a reason to cease pretending otherwise. But would believing that *p* constitutes enough of a motivating reason for Sally to effectively abandon her pretense that not-*p*? I do not think so. I am skeptical that Sally, without a change in situation, could see sticking to her belief as an appealing incentive to abandon her pretense. If Sally considered her belief that *p* a good enough incentive to stop pretending, then this would render unclear why she self-deceived in the first place. It is precisely because she cannot stomach the belief that *p* that she pretends otherwise. It would be different if Sally's interlocutor could appeal to the fact that she should *care* to be the kind of *person* who is true to her beliefs for if Sally had such desire, this would constitute an incentive to let her belief guide her actions.

But invoking Sally's desire to be a certain kind of person is somewhat appealing to Sally's identity. And while this strategy might be suited for the doxastic violation account, it is not in harmony with pretense accounts since the theory places no emphasis on epistemic norms, and even less on the valence they have for those who hold them. Simply pointing out that the self-deceived already believes that *p*, without providing any additional motivation for why *p* should guide their actions, carries no motivation for the self-deceived individual to abandon their pretense. Pretense accounts *can* thus employ dialectical self-deception, but this technique is unlikely to successfully yield self-undeception.

A similar conclusion follows for minimalism, which stands in an even more disadvantaged position when showing how dialectical and isomorphic self-undeception can be successfully employed. As we have seen, pretense accounts can appeal to the self-deceived's belief that not-*p*. This appeal alone is, I suggested, unlikely to spark self-undeception, but it is at least a serviceable path. Minimalism, however, instead has virtually no reason that could be provided to the self-deceived to motivate them to abandon their self-deceptive belief. This is because according to minimalism, the self-deceived is conducting what they think is a regular epistemic investigation with no awareness that their conative elements (i.e. their desires and interests) are perniciously shaping the investigation's epistemic outcome. When confronted with a trusted interlocutor who encourages them to abandon their self-deceptive belief, the self-deceived person may then simply react by rationalizing the interlocutor's attempt. To the interlocutor's suggestion that Sally should admit having cancer because she can "win this battle", Sally may respond by agreeing that she could overcome cancer, if only she had the disease. In this sense, minimalism is particularly at a high risk of rationalization because the self-deceived's complete lack of awareness makes it the case that their mental states are fully secluded.

Contrary to the doxastic violation account (which can appeal to the agent's epistemic norms), and to pretense accounts (which can appeal to the self-deceived pre-existing beliefs), there is no element in the self-deceived's mental economy that minimalism can invoke in order to motivate the self-deceived to abandon their self-deceptive belief. A minimalist could reply that the interlocutor could appeal to the self-deceived's motivational bias by stressing that the self-deceived fails to realize that their interests are unduly influencing their epistemic investigation. But this is exactly the corrective approach that I earlier deemed unlikely to succeed, one that can be perceived as psychologically threatening and that the self-deceived can easily rationalize. Thus, even though there is nothing inherent in the structure of minimalism that makes dialectical self-undeception inapplicable, the agent's cloaked awareness nevertheless renders the technique unlikely to yield self-undeception.

4.5 Isomorphic Self-Undeception vs Pretense Accounts and Minimalism

So far so good for dialectical self-undeception. Let me now turn to isomorphic self-undeception and ask: can Gendler's pretender notice an isomorphism between herself and a fictional character? I think the answer here is less clear. While I think she could notice an isomorphism, I do not think this will be sufficient for self-undeception. In isomorphic self-undeception, the recognition of an isomorphism sparks a self-reflective process that leads the self-deceived individual to eventually re-evaluate the evidence. Now, what works in the doxastic violation account is that, at least in Sally's case, the motivation to re-examine the evidence is triggered by the recognition of the isomorphism followed by Sally's reflection on her identity. Similar to how the fictional character has been drifting away from the kind of person she was at the beginning of the story, Sally has also changed in ways that are unlike who she used to be. This recognition of Sally's identity-shift is inevitably tied to her epistemic norms because these are not detached from her identity. On the contrary, they contribute to making Sally the kind of person she is. Interest in preserving her identity, and being true to herself, are the motivating factors that spark a re-evaluation of evidence.

Given the absence of epistemic norms in both minimalism and pretense accounts, I am skeptical that isomorphic self-undeception could yield a successful self-undeception. This is not to say that minimalism's self-deceived or Gendler's pretender could not notice an isomorphism with a fictional character, for they could. For instance, we can imagine that Gendler's pretender recognizing an isomorphism with a fictional character who is also pretending that not-*p*, while still believing otherwise. But the recognition of the isomorphism is unlikely to carry any self-undeceptive import because identifying the isomorphism is only the first step to self-undeception. For self-undeception to occur, the recognition of the isomorphism must be accompanied by a motivating factor that leads to action, i.e. a re-examination of evidence. In the doxastic violation account, this motivating factor is the realization that Sally has changed and has isolated herself from the scientific community (all elements connected to the valence that Sally's epistemic norms carry). But without those motivating factors at play, the risk is that the isomorphism is motivationally inert, and moreover is easily rationalizable. After all, even anecdotally, individuals often see their character flaws reflected in others, without this always sparking action. For a change to occur, one must have a motivating factor that drives the change in the first place.

*

I started this chapter suggesting that the doxastic violation account's characterization of self-undeception is superior to its competing theories, specifically minimalism and pretense accounts, and I have proposed two paths to self-undeception that show how the doxastic violation account provides a characterization of the phenomenon that is normatively proper.

I have defined normatively proper self-undeception as a state that results from a self-reflective process (whether driven by reflection on one's own identity or access to communities) that leads the agent to recognize their own epistemic shortcomings. I have stressed the importance of this last aspect by suggesting that understanding one's epistemic pitfalls is helpful to prepare the agent for what they see as an improved epistemic future.

One way in which the now self-undeceived agent can benefit from understanding their previous epistemic shortcomings is to spark a motivation to be more self-controlled in the future. In particular, Sally may form what Peter Gollwitzer calls an "implementation intention" (Gollwitzer 1999: 494). Contrary to goal intentions, which simply identify an outcome to obtain, implementation intentions also specify goal-directed responses where the agent commits themselves to react to a certain situation in a specific manner. Instead of the simpler, goal-intention structure such as "I intend to reach *x*", implementation intentions have the structure of "When situation *x* arises, I will perform response *y*" (Ibid.).

Sally may form the implementation intention that when she is to assess a situation that comes with a desire for a specific outcome, she will commit to implementing strategies that ensure that she sticks to her epistemic norms. She may, for example, slow down her thoughts, remind herself of her past self-deceit, or whisper to herself that she is not the kind of person who "simply believes what she wants". (Implementation intentions can certainly take different forms depending on the specific object of self-deception.)

Yet what would motivate Sally to form implementation intentions? The motivation, I suggest, may derive from the conscious realization that violating her epistemic norms has led her to a compromised identity, and transformed her into a person that, on reflection, Sally does not feel comfortable with. In this sense, the motivation to preserve her identity and the motivation to follow her epistemic norms are intimately related.

If successfully employed, implementation intentions can certainly contribute to avoid future instances of self-deception and this, in turn, opens the path to an epistemic virtue theory of *preventing* self-deception (instead of undeceiving after the self-deceit has already occurred). As it should be clear by now, I do not think that preventing self-deception is a matter of coming to endorse new epistemic norms, such as valuing truth. Preventing self-deception can be easier than that: one simply needs to form implementation intentions with respect to those epistemic norms one already has. On similar lines, Van Leeuwen suggests that "self-deception can be better avoided by cultivating cognitive habits that neutralize the aspects of mind that give rise to self-deception. One can confront discomforting evidence and accept it for what it is" (Van Leeuwen 2008: 207). This of course is not to say that being self-controlled is an easy task; on the contrary, its difficulty may explain why many are poor at exercising it.

The epistemic benefits of undeception however do not end at the moment that the individual undeceives. The newly undeceived agent can benefit from the wisdom of his rational condition by committing to practical strategies that will prevent future instances of self-deception. Epistemic growth is a diachronic process, one that can only be achieved by grasping one's previous epistemic weaknesses and learning from them.

CONCLUSION

What lessons should we learn from this examination of self-undeception? By way of concluding, I want to bring together the insights that I think this dissertation has brought to the philosophical table, and I want to flag some remaining questions that I take to be worth pursuing in the future.

Philosophically, my examination of self-undeception provides new reasons for why we should adopt a particular theory of self-deception, that is, the doxastic violation account. Defending a particular view of self-deception, as I also mentioned in the introduction, is of course not a novel project, but defending a specific theory of self-deception by making claims that rely on an analysis of self-undeception is a non-standard approach because it indirectly offers an examination of the underexplored phenomenon.

Though I see the conceptual importance of self-undeception as an important contribution, the most insightful lesson that I take from this dissertation is one that pertains to emotional and non-doxastic factors. Examining self-undeception, and in particularly looking closely at the *dynamics* that can lead individuals to abandon their self-deceit, has encouraged me to think not only in terms of what is philosophically coherent, but also in terms of what is psychologically plausible. Let me explain.

If one conceives of self-deception in purely epistemological terms, then the phenomenon's culprit seems to be a rejection of evidence and an unwarranted belief. And this analysis would imply that if what goes wrong in self-deception is a false belief, then a correction of this belief should suffice to restore rationality. But reiterating evidence cannot be a psychologically plausible route to undeception. This is because an unwarranted belief is merely the end result of a more nuanced phenomenon, one fueled by motivations and interests. If one really takes those non-epistemic motivations and interests seriously, and moreover is interested in formulating a psychologically plausible theory of self-undeception, then the outcome, as we have seen in Chapter 4, is a self-undeceptive model that does not resemble what an epistemologist may be used to. It instead has a lot more to do with emotional and non-doxastic factors.

The more general lesson that we should take from this is to accept that we cannot counteract pernicious epistemic practices exclusively by correcting beliefs and reiterating evidence, because that strategy would only be targeting a *symptom* of a more insidious phenomenon. A more effective way of correcting flawed epistemic practices is to give priority to their *causes*. In doing so, we take seriously the suspicion that a hostile attitude towards evidence often has little to do with evidence and a lot to do with non-epistemic factors (such as motivations and interests). My coming to accept this last point has been the most illuminating outcome of this dissertation.

Where should we go from here? As the aim of this dissertation has been to open the debate on self-undeception, I think it is worth pursuing the project even further, particularly with respect to those remaining questions that could not be discussed here.

An issue that I think it is worth pursuing is one that I mentioned in the introduction, and it involves assessing whether the insights provided by an examination of self-undeception can be applied to similar phenomena such as conspiracy theories. The conspiracy theories that I have in mind are those that involve, for example, conspiratorial beliefs that the earth is flat or that vaccines are unsafe.[78]

One of the most pressing questions that both philosophers and other scholars have attempted to answer is what characteristics render these conspiratorial beliefs so resistant to revision. Research has shown that reiteration of evidence not only seems to be ineffective but, in some instances, it seems to even *backfire,* with individuals clinging to their conspiratorial beliefs more strongly (Horne, Powell, Hummel, and Holyoak 2015; Betsch and Sachse 2013). In light of this, the literature has moved to

[78] The claim according to which the earth is flat relies on the belief that science has conspired into tricking us into believe that the earth is round; and the claim that vaccines are unsafe relies on the argument that scientist and pharmaceutical companies hide the real risks of vaccines in exchange for profit. See Jolley and Douglas 2014.

explore alternative ways that can capture such recalcitrance (Tetlock 2002; Cichocka, Marchlewska, and Golec de Zavala 2016; Douglas, Sutton, and Cichocka 2017).

Though different phenomena, self-deception and conspiracy theories share similar aspects that can be fruitfully understood by employing similar insights. As I have said, self-deception is also resistant to evidence and correction, and crucially, it involves the violation of loaded epistemic norms. Now, conspiracy theories do not involve the violation of loaded epistemic norms, yet conspiratorial beliefs do seem to qualify as *loaded*. That is, conspiratorial beliefs seem to be beliefs that, over time, give back to those who hold them certain non-epistemic positive returns. For example, these beliefs (i) speak to agents' identity in important ways; (ii) build and modulate access to specific communities; and (iii) are emotionally-charged. These loaded, non-epistemic features are what may explain why conspiratorial beliefs are persistent over time. Developing meaningful relationships with other members of the flat-earth or anti-vaccine community can foster a sense of belonging and generate positive emotions. Gradually, these kinds of beliefs become integral to the individual's identity, and they may come to see themselves not merely as someone who holds these beliefs, but as a *flat-earther* or an *anti-vaxxer*.

If something like this is right, then this may help explain why these beliefs are so resistant to evidence. Flat-earthers or anti-vax individuals may have an interest or a motivation to belong to particular communities, or to preserve their identity. And if both of these are somewhat fueled by the conspiratorial beliefs, then individuals may have an interest in maintaining those beliefs even in the face of disconfirming evidence.

In light of that, I hope that this dissertation has contributed to our understanding of an undertheorized phenomenon in both conceptual and practical ways.

REFERENCES

Gärdenfors, P. (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press.

Aristotle, J. A.; Smith, D. P. & Chase, (1911). *The Nicomachean Ethics of Aristotle*. Dent.

Audi, R. (2008). "The ethics of belief: doxastic self-control and intellectual virtue". *Synthese* 161, 403-418.

Audi, R. (1997). "Self-Deception vs. Self-Caused Deception: A Comment on Professor Mele". *Behavioral and Brain Sciences* 20 (1):104-104.

Avnur, Y. & Scott-Kakures. D. (2015) "How Irrelevant Influences Bias Belief", *Philosophical Perspectives*, 29(1):7-39.

Bach, K. (1997). "Thinking and Believing in Self-Deception". *Behavioral and Brain Sciences* 20 (1):105-105.

Bach, K. (1981). An analysis of self-deception. *Philosophy and Phenomenological Research* 41:351-370.

Badwhar, N. (2014), *Well-Being: Happiness in a Worthwhile Life*, Oxford University Press.

Barnes, A. (1997). *Seeing Through Self-Deception*. New York: Cambridge University Press.

Batson, C. D. (2011). *Altruism in Humans*. Oxford University Press.

Baumeister, R. F.; Bratslavsky, E.; Muraven, M.; Tice, D. M. (1998). "Ego depletion: Is the active self a limited resource?". *Journal of Personality and Social Psychology*. 74 (5): 1252–1265.

Baumeister, R, Vohs, K. D., Tice, D. M. (2007). 'The strength-model of self-control'. *Current Directions in Psychological Science,* 16, 351-355.

Belkin, L. (1999) *Show Me a Hero: A Tale of Murder, Suicide, Race, and Redemption*. Boston: Little, Brown.

Bermúdez, J., (2000), "Self-Deception, Intentions, and Contradictory Beliefs," *Analysis* 60(4): 309–319.

Betsch, C. & Renkewitz, F. & Betsch, T. & Ulshöfer, C. (2010). "The Influence of Vaccine-critical Websites on Perceiving Vaccination Risks". *Journal of health psychology*. 15. 446-55.

Betsch, C., & Sachse, K. (2013). "Debunking vaccination myths: Strong risk negations can increase perceived vaccination risks". *Health Psychology,* 32(2), 146–155.

Bilgrami, A., (2006), *Self-knowledge and Resentment*, Harvard University Press.

Canfield, J. V. & Gustavson, D. F. (1962). "Self-deception". *Analysis* 23: 32-36.

Carnap, R., (1950). *Logical Foundations of Probability*, Chicago: The University of Chicago Press.

Cichocka, A., Marchlewska, M., Golec de Zavala, A., & Olechowski, M. (2016). "They will not control us": In-group positivity and belief in intergroup conspiracies. *British Journal of Psychology*, *107*, 556–576.

Crawford, L. (2019). "Believing the Best: On Doxastic Partiality in Friendship". *Synthese* 196 (4):1575-1593.

Davidson, D. (1985) "Deception and Division" in J. Elster (ed.) *The Multiple Self*, Cambridge University Press, 79-92.

Davidson, D., (1982), "Paradoxes of Irrationality," in *Philosophical Essays on Freud*, R. Wollheim and J. Hopkins (eds.), Cambridge: Cambridge University Press.

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current Directions in Psychological Science*, 26(6), 538–542.

Elster, J., (ed.), (1985), *The Multiple Self*, Cambridge: Cambridge University Press.

Fantl, J. (2018). *The Limitations of the Open Mind*. Oxford, UK: Oxford University Press.

Foley, R. (2009). "Beliefs, Degrees of Belief, and the Lockean Thesis". In F. Huber & C. Schmidt-

Petri (Eds.), Degrees of belief (pp. 37–47). London: Springer.

Friederich, J. (1993). "Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena." *Psychological Review,* 100: 298-319.

Friedman, J. (forthcoming). The Epistemic and the Zetetic. *Philosophical Review.*

Funkhouser, E., (2005) "Do the Self-Deceived Get What They Want?" in *Pacific Philosophical Quarterly*, 86:3, 295-312.

Gay, R. (2017). *Hunger: A Memoir of (My) Body,* HarperCollins.

Gendler, T. S. (2007). Self-Deception as Pretense. *Philosophical Perspectives* 21 (1):231 - 258.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493–503.

Harman, G. (1997). Practical reasoning. In Alfred R. Mele (ed.), *Review of Metaphysics.* Oxford University Press.

Hempel, C. (1965). *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science.* New York: The Free Press.

Henden, E. (2008). 'What Is Self-Control?'. *Philosophical Psychology,* 21, 69-90.

Heil, J. (1984). "Doxastic Incontinence", *Mind*, 93 (369): 56-70.

Heil, J. (1989). "Minds Divided", *Mind*, 98 (392): 56-70.

Hempel, C. (1965), *Aspects of Scientific Explanation*, New York: Free Press, 571-583.

Holton, Richard (2001). What is the role of the self in self-deception? *Proceedings of the Aristotelian Society* 101 (1):53-69.

Holton, R. (1994), "Deciding To Trust, Coming To believe", *Australasian Journal of Philosophy*, 72(1):63-76.

Holton, R. & Shute, S. (2007). Self-control in the modern provocation defence. *Oxford Journal of Legal Studies* 27 (1):49-73.

Horne Z., Powell D., Hummel J.E., Holyoak K.J., (2015) "Countering antivaccination attitudes". *Proceedings of the National Academy of Sciences of the United States of America.* s18;112(33):10321-4.

Hornsey, J. M., Harris, E. A., Fielding, K. S., (2018). "The Psychological Roots of Anti-Vaccination Attitudes: A 24-Nation Investigation". *Health Psychology* 37: 307-315.

Hume, D. (1739). *A Treatise of Human Nature.* D. F. Norton & M. J. Norton (eds.), Oxford University Press, 2000.

Hunter, D. (1996). On the relation between categorical and probabilistic belief. *Noûs*, 30, 75–98.

Jenkins, D. (2018). The Role of Judgment in Doxastic Agency. *Thought: A Journal of Philosophy* 7 (1):12-19.

Johnston, M., (1988). "Self-Deception and the Nature of Mind". In C. Macdonald (ed.), *Philosophy of Psychology: Debates on Psychological Explanation.* Cambridge: Blackwell 63--91.

Kagan, S. (2015). An Introduction to Ill-Being. *Oxford Studies in Normative Ethics* 4:261-88.

Kennett, J. (2013). 'Just Say No? Addiction and the Elements of Self-Control'. In N. Levy (Ed.), *Addiction and Self-Control.* Oxford: Oxford University Press.

Kidd, D., Castano, E. (2013). "Reading Literary Fiction Improves Theory of Mind". *Science.* 342. 10.1126

Kingdon, D. G., & Turkington, D. (1994) *Cognitive-Behavioral Therapy of Schizophrenia,* Guildford Press.

Kremm, D. (MS). "Expressivism, Agency, and Akrasia", draft.

Kriglanski, A. & Webster, D. (1996). "Motivated Closing of the Mind: 'Seizing' and 'Freezing'." *Psychological Review,* 103: 263-283.

Kunda, Z. (1990). The case for motivated political reasoning. Psychological Bulletin, 108(3), 480–498.

Lalich, J. & Singer, M. (1996), *Cults in Our Midst*, Jossey-Bass.

Lazar, A. (1999) "Deceiving Oneself or Self-deceived? On the formation of Belief "Under the Influence"" in *Mind*, 430:165-1-290.

Levy, N. (2004), "Self-Deception and moral Responsibility", *Ratio (new series)*, 17: 294–311.

Lewicka, M. (1992), "Pragmatic Reasoning Schemata with Differing Affective Value of a Consequent Logical Implication." *Polish Psychological Bulletin,* 23: 237-252.

Lord, C. G.; Ross, L.; Lepper, M. R., (1979) "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence", *Journal of Personality and Social Psychology*, 37(11):2098-2109.

Losonsky, M. (1997). Self-Deceivers' Intentions and Possessions. *Behavioral and Brain Sciences* 20 (1):121-122.

Lynch, K., (2012). On the "tension" inherent in self-deception. *Philosophical Psychology* 25 (3): 433-450.

Maibom, H., (2017). "Affective Empathy". In: H. Maibom (ed.) *Handbook of Philosophy of Empathy*. London: Routledge.

Manne, K. (2014). "Internalism about reasons: sad but true?". *Philosophical Studies* 167:89-117.

May, J. (2017). "Empathy and Intersubjectivity". In Maibom H. (ed.), *The Routledge Handbook of Philosophy of Empathy*. New York: Routledge 169-179.

McHugh, C. (2017). Attitudinal Control. *Synthese, 194*(8), 2745–2762.

Mele, R. A., (forthcoming). "Self-deception and selectivity". *Philosophical Studies*:1-15.

Mele, R. A. (2003). *Motivation and Agency*. Oxford University Press.

Mele, R. A., (2001). *Self-Deception Unmasked*. Princeton University Press.

Mele, R. A. (1993). "Motivated Belief". *Behavior and Philosophy* 21 (2):19 - 27.

Mele, R. A. (1987). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control.* Oxford: Oxford University Press.

Mele, R. A. (1983). "Self-Deception". *Philosophical Quarterly* 33 (October):366-377.

Nelkin, D. K. (2002). Self-deception, motivation, and the desire to believe. *Pacific Philosophical Quarterly* 83 (4):384-406.

Nyhan, B., & Reifler, J., "The Roles of Information Deficits and Identity Threat in the Prevalence of Misperceptions", forthcoming.

Noordhof, P., (2009). The Essential Instability of Self-Deception. *Social Theory and Practice* 35 (1):45-71.

Orlandi, M., Stroud, S. (forthcoming) "Self-Control in Action and Belief", *Philosophical Explorations.*

Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., Winner, E. (2016). "Does Reading a Single Passage of Literary Fiction Really Improve Theory of Mind? An Attempt at Replication". *Journal of Personality and Social Psychology* 11(5):46-54.

Paul, S. K. (2015*a*). 'The Courage of Conviction'. *Canadian Journal of Philosophy,* 45, 647-669.

Paul, S. K. (2015*b*). 'Doxastic Self-Control'. *American Philosophical Quarterly,* 52, 145-158.

Pears, D. F., & Pugmire, D. (1982). "Motivated Irrationality". *Proceedings of the Aristotelian Society*, Supplementary Volumes, 56:157-196.

Pedrini, P., (2014) *Che Cos'e' l'Autoinganno e Come Funziona*, Laterza, Bari.

Plant, Ashby & Peruche, Michelle & Butz, David. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology*. 41. 141-156.

Poos, J. M., Bosch van den, K., Janssen, C. P., (2017). "Battling Bias". *Computational Education* 111: 101-113.

Pollock, J. L. (1987). Epistemic norms. *Synthese* 71 (1):61 - 95.

Porcher, J. E. (2014). Is Self-Deception Pretense? *Manuscrito* 37 (2):291-332.

Randall C., Block, J., and Funder D. C., (1995) "Overly Positive Self-Evaluations and Personality: Negative Implications for Mental Health," *Journal of Personality and Social Psychology* 68: 1152-62.

Quine, W. V. O. (1978). *The Web of Belief*, McGraw-Hill.

Railton, P. (2014). "The Affective Dog and Its Rational Tale: Intuition and Attunement". *Ethics* 124: 813-859.

Redlawsk, D.P., Civettini, A., & Emmerson, K.M. (2010). The Affective Tipping Point: Do Motivated Reasoners Ever "Get It"? *Political Psychology, 31*, 563-593.

Robin, C. (2018). How Eerie and Unsettling it Can Be When People Change their Minds", *Crooked Timber.*

Rorty, A. O. (1988) "The Self-Deceptive Self: Liars, layers and Liars", in B. P. McLaughlin and A.O. Rorty (eds), *Perspectives on Self-Deception*, University of California Press, Berkeley, 11-28.

Rorty, A. (1983). "Akratic Believers". *American Philosophical Quarterly* 20 (2):175-183.

Rudman, Laurie & Ashmore, Richard & Gary, Melvin. (2001). "'Unlearning'" automatic biases: The malleability of implicit prejudice and stereotypes. Journal of personality and social psychology."81. 856-68.

Sanford, D.H. (1988). "Self-deception as rationalization", in *Perspectives on Self-Deception,* McLaughlin, B.P. & Rorty A. O. eds., University of California Press, Berkeley-Los Angeles, 157-169.

Scott-Kakures, D., (2002) "At 'Permanent Risk': Reasoning and Self-Knowledge in Self-Deception, in *Philosophical and Phenomenological Research*, 65 (3):576-603.

Scott-Kakures, D. (1996). Self-deception and internal irrationality. *Philosophy and Phenomenological Research* 56 (1):31-56.

Smith, A. (1853). *The Theory of Moral Sentiments*, New York: August M. Kelley Publishers, 1966.

Smith, M. (1994). *The Moral Problem.* Blackwell.

Stueber, K. (2006). *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*, Cambridge, MA: MIT Press.

Strawson, P.F. (1962), "Freedom and Resentment". *Proceedings of the British Academy*, 48, 125.

Stroud, S. (2006). "Epistemic partiality in friendship". *Ethics* 116 (3):498-524.

Sweller, J. (1988). "Cognitive load during problem solving: Effects on learning", *Cognitive Science*, 12 (2): 257–285.

Szabados, B. (1974). "Self-deception". *Canadian Journal of Philosophy* 4: 41-49.

Talbott, W. J. (1995), "Intentional Self-Deception in a Single Coherent Self," *Philosophy and Phenomenological Research*, 55: 27–74.

Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E., and Gruenewald, T. L. (2000) "Psychological Resources, Positive Illusions, and Health," *American Psychologist* 55: 99-109.

Taylor, S. E. and Brown, J. D. (1994) "Positive Illusions and Well-Being Revisited: Separating Fact from Fiction," *Psychological Bulletin* 116: 21-27.

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

Tetlock, P. E. (2002). Social-functionalist frameworks for judgment and choice: The intuitive politician, theologian, and prosecutor. *Psychological Review*, *109*, 451–472.

Thagard, P. (2004). What is Doubt and When is it Reasonable? *Canadian Journal of Philosophy* 34 (Supplement): 391-406.

Trope, Y. & Liberman, A. (1996) "Social Hypothesis Testing: Cognitive and Motivational Mechanisms", in E.T. Higgins and A.W. Kruglanski eds., *Social Psychology: Handbook of Basic Principles*, Guilford Press, New York, 239-270.

Tully, Ian (2017). Depression and the Problem of Absent Desires. *Journal of Ethics and Social Philosophy* 11 (2):1-16.

Van Leeuwen, N., (2009) "Self-Deception Won't Make You Happy", *Social Theory and Practice*, 35(1):107-132.

Van Leeuwen, N. (2008). "Finite rational Self-Deceivers". *Philosophical Studies* 139 (2):191 – 208.

Williams, B. (1973) "Deciding to Believe" in *Problems of the Self*. Cambridge University Press, 136-151.