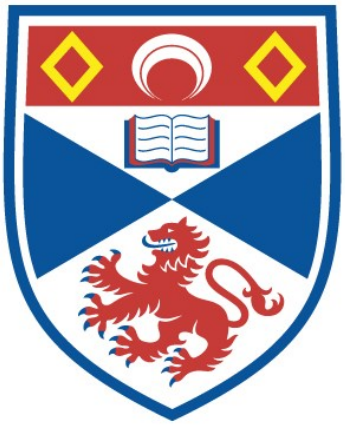
SELF-DECEPTION AND ITS INTERACTION WITH INTROSPECTION

Guglielmo Guarrasi

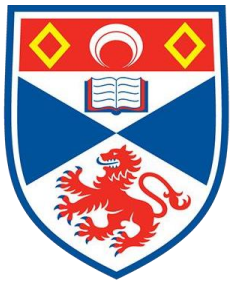
A Thesis Submitted for the Degree of MPhil
at the
University of St Andrews



2022

Self-Deception and Its Interaction with Introspection

Guglielmo Guarrasi



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Master of Philosophy (MPhil)

at the University of St Andrews

May 2022

Acknowledgements

I want to thank the University of St. Andrews for having allowed me to pursue my passion in philosophy at a master level. My biggest thank goes to the Simon Prosser, my supervisor. His help and supervision have being invaluable for the realisation of this dissertation. This last year and a half have been very tough for me. It is thanks to his patience and understanding that I have been able to comfortably finish my project. Additionally, I wish to thank the Lucy O'Brien, a professor from my undergraduate studies, for making me discover this topic.

Abstract:

Self-deception is a tricky phenomenon to define, especially once we realise the further complications its interaction with introspection might lead to. In this dissertation I am to analyse these two phenomena to show their compatibility. In fact, I am going to argue that, although self-deception is an instance where introspection fails whereas the latter can cause the end of a self-deceive state, no problematic interference happens between them. To reach this aim I am going to rely on the studies conducted by Nisbett and Wilson on instances where subjects fail to introspect certain mental states or processes. I will use this to argue that it is then possible for a subject to fail to introspect certain aspects of their mental life and, thusly, self-deceive. In turn, I am going to show how having areas where introspecting gives us the wrong result is not a major problem as it is a characteristic it shares with sense perception, which is something we are used to rely on.

Candidate's declaration

I, Guglielmo Guarrasi, do hereby certify that this thesis, submitted for the degree of MPhil, which is approximately 33,487 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2018.

I confirm that no funding was received for this work.

Date 15/05/2022 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of MPhil in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain

only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 15/05/2022 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Guglielmo Guarrasi, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 15/05/2022 Signature of candidate

Date 15/05/2022 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Guglielmo Guarrasi, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 15/05/2022 Signature of candidate

Self-Deception and Its Interaction with Introspection

1. Introduction

The topic of self-deception originates several problems within the field of philosophy. There is large disagreement over its definitions and over which cases might count as proper cases of self-deception. At a basic level, self-deception is considered the process through which a subject S acquires a belief against their best evidence (Deweese-Boyd, 2016, introduction). However, this is just the skeleton, some early definition most people could agree on, of the phenomenon that I am going to discuss in this paper. It should be noted although I am here describing self-deception as a process, it can also be a state, namely the state a subject enters after the process of deceiving themselves. In fact, generally, the belief obtained through self-deception is then retained through time. It is in these instances that we talk of a self-deceived subject. In this dissertation I will usually refer to self-deception as a process since that is its most discussed aspect by philosophers. It is more crucial and harder to define. However, I am not going to skip over discussing how a subject maintains their self-deceived state when this becomes relevant.

Philosophical theories of self-deception can be classified under two branches: intentionalists and motivationalists (Deweese-Boyd, 2016, sections 2-

3). The former tries to model self-deception in a way that matches regular deception. Motivationalists place no such connection between the two types of deception, but instead frame self-deception as happening because there is something the subject would want to be true but believes to be false. In this sense, S has some 'motivation' to enter the self-deceived state (Deweese-Boyd, 2016, section 3). On the other hand, the name 'intentionalists' derives from the fact that self-deception, according to them, should be characterised as a process involving intentions. In this respect, self-deception mimics interpersonal deception, which is usually considered an intentional action (Mahon, 2015, section 3). What separates the two is that the target and the subject of self-deception are the same person. This intentional character is the most noteworthy difference between intentionalists and motivationalists. In fact, for motivationalists intentions do not play a role in the self-deceiving process. A question one might have after hearing this is whether it would be possible for a subject to enter a self-deceived state intentionally in a motivationalist account. Such a possibility is going to be discussed in section 3.3 of this dissertation as many topics need to be discussed to be in a position to answer it.

However, something that is already clear from this is that the distinction between the two branches is not absolute. There are accounts, such as my own, that encompass characteristics of both. However, it is true that all accounts will fall closer to either depending on which characteristics are chosen to describe the core of the phenomenon. For example, my account works within a

motivationalist framework. Even though it shares some traits usually attributed to intentionalists, the fact that I do not consider intentions responsible for self-deception, and blame desires for it, puts me on the motivationalist side. In fact, my characterisation of self-deception draws from Mele (1997, 1999) and Hubbs (2018). More precisely, it incorporates some central claims from both and applies some crucial modifications drawing from the intentionalist framework and analysing its interaction with the process of introspection. The reason for these choices will be presented and defended as I proceed. My main aim in this dissertation is to present a satisfying account for the process of self-deception and describe its relation to introspection.

As I will discuss in chapter 3, the coexistence of the two gives rise to a lot of scepticism. In fact, on the one hand, it seems that the existence of introspection could make us doubt that of self-deception. Depending on how the former is characterised, it might be hard to imagine how we could possibly be able to deceive ourselves. Humans are generally thought to be able to access their mental states and processes in some sort of way. So, how can we simultaneously be the perpetrator and receiver of deception for the same belief? If we can perceive our deception as it is happening, we should be able to halt it so that we would not fall for it. In turns, the possibility for a process like that of self-deception can make us question the reliability of introspection. Should we reconsider the importance we attribute to it if we can be so wrong about our own mental life? Although rejecting the existence or effectiveness of either self-

deception or introspection might appear the easiest way out of this dilemma, this is not the path I am going to suggest we take. On the contrary, I will argue that we are both capable of introspecting and self-deceiving with some effectiveness, and explain how this is possible without the two phenomena interfering with one another.

Even though these are going to be my two main aims in this dissertation, there is a number of secondary objectives I have in writing this. The best way to present them will be to provide an overview of the rest of this piece... In the next chapter I am going to introduce the phenomenon at the centre of this dissertation: self-deception. In discussing it, I am going to show why it is such a hard concept to define and give my best attempt at explaining what it is without relying on one specific model for it. In doing so, I am going to show that self-deception should not be reduced to other neighbouring mental phenomena (i.e. wishful thinking and akrasia), but that it is worth investigating as its own thing. Then, in 2.1, I will use Mele's account to give an overview of the two branches in models for self-deception and explain why one should be preferred over the other. This will lead to the next section, 2.2, where I introduce Hubbs's thoughts on self-deception to show how I incorporate them in my own account. In 2.3 I am going to analyse a particularly difficult sub-set of instances of self-deception, the twisted cases. In this way I will discuss how my account can be defended against this problematic category of examples. I will, then, account for it in a way similar to how I deal with straightforward

cases of self-deception. Finally, I will conclude the second chapter with 2.4 by arguing for the benefit of including parts of Mele and Hubbs's accounts in my model.

This will give me the opportunity to move to the third chapter and focus on the topic of introspection. In addition to characterising what this phenomenon is, I am also going to defend it from some doubts that have been raised against it by Schwitzgebel. Then, in 3.1, I will be in a position to present the model of introspection I believe is the most accurate, Lycan's, together with the model it generated from, Armstrong's. In 3.2 I am going to answer the main question of this dissertation: can someone who is capable of introspection self-deceive? In this way I will argue how the two phenomena are compatible showing the results of some studies analysed by Nisbett and Wilson (1977). Finally, in 3.3, I am going to go back to a question that will have been forced to be on hold for the majority of the dissertation. Namely, I am going to wonder whether there is a way the two branches for models of self-deception can be compatible given what I will have said about introspection. There I am going to show that, although it is possible, it requires a serious revision of one of the core characteristics of the intentionalist branch. That will be where I am going to end my dissertation.

2. Self-Deception

I should immediately start by addressing the biggest concern with self-deception I mentioned in the introduction: this is the question of whether there is such a thing as self-deception in the first place. What does it mean that we might be capable of deceiving ourselves? Such a claim, without further clarification on the phenomenon, sounds puzzling. Broadly speaking, many things could be considered self-deception. For example, I could misjudge something and start acting in accordance to that belief, or start imagining fictional scenarios as true (e.g. what if I had superpowers?). Cases like these are ones where a subject acts in accordance to a belief that is either false or it is something they believe to be false. So, in a way, these are cases where someone is deceiving themselves either intentionally or not. However, these do not seem particularly interesting as instances of irrational behaviour like true cases of self-deception seem to be: there are interesting conversations that can be had about them individually, but they do not seem cases worth discussing together as their own category. They are too varied and they do not really resemble what we usually think of when we think of self-deception. There is a further issue here, there is two other phenomena that seem relatively similar to what is usually described as self-deception: wishful thinking and akrasia. Both are cases of irrational behaviour and, as I am going to discuss, share similar features with the general thought of believing something in the spite of evidence.

For all these reasons, self-deception stays on unstable grounds. There does not seem to be a compelling reason to consider it philosophically. Especially since even those tempted to accept it as worth discussing might try to reduce it to the other two phenomena to explain all putative self-deception scenarios. My aim in this section is to show that there is a defined set of cases that should be considered self-deception and might spark an interesting philosophical conversation. Also, I am going to show that these cannot be reduced to other examples of irrational behaviour regarding the formation of beliefs like the two aforementioned phenomena. In order to do this I will start by looking at what we can say self-deception is without yet committing to a specific view.

Here, I need to provide something more to show we should consider self-deception its own, interesting, phenomenon. To do this, I need to take a step back to look at what we have available to grasp self-deception. There is not much that it is available for such an inquiry, but some place to start is needed. The most straightforward way to try defining it might be to start from its name. This means trying to think of scenarios that can be described as "self-deception" to see what this can say about the phenomenon. This does not bring us far since I have already talked about the fact that such cases do not seem interesting as one unified phenomenon. A second attempt might be separating it into "self" + "deception" and looking at what the definitions might have to say. This is not of much help in giving a clear set of cases as what counts as an act of deception

is not obvious either. The dictionary would usually define it as the act of hiding the truth (often to gain an advantage). So, we would have it that self-deception is the act of hiding the truth from oneself. This, once again, is exactly how we end up with a broad variety of cases that do not seem to have much in common except this very definition.

Another point to start from might be considering what I have already presented as the least controversial claim about self-deception, namely that it is the acquisition of a belief against one's own best evidence. This narrows the set of phenomena that can be included, but it is not yet enough. This claim is still contentious since there are many phenomena that involve the acquisition of a belief against one's best evidence. Wishful thinking and akrasia are themselves an example of this. But there is more that can be said about self-deception when we think of our most basic notions for it. Namely, when we picture a scenario were a subject self-deceives, we have in mind some situation where the subject comes to believe something that contradicts the evidence available to them. This is what qualifies it as an irrational thing to do: the subject is not adjusting their belief to the evidence in front of them. This happens because the subject has some additional reason that interfere with the evaluation of what is true or false. Usually, this is thought to come about because the subject wants something to be true even though that does not appear to be the case. This is something I think the great majority of proponents of self-deception would agree on. It is not much to work with for an account yet, but it is all I can say

without taking a standpoint on the matter. Somehow, though, it is already too specific to encompass all cases of "self" + "deception".

For a case where we would say that a person is hiding the truth from themselves but is not self-deceiving, imagine a scenario where I see something that I mistakenly take to be a snake among the grass. This makes me form the, quite reasonable, belief that I should not go into the field since it is dangerous. My mistake here is hiding the truth from me (i.e. that the field is perfectly safe to venture in). So, such a case could be listed as an instance of self-deception if that were all that would be needed to qualify for it. However, this does not have anything of what I said in the previous paragraph. I am not forming a belief in spite of the evidence available to me. Also, there seems to be no apparent reason to call the belief I form in this scenario irrational as I have many good reasons to avoid the snake that I believe to be hidden in the grass. There are many other instances where something similar would happen. Is this the end of the enquiry as no set of cases can be called self-deception and form an interesting phenomenon?

It might be clear by now that the mistake is trying to identify self-deception with any instance of "self" + "deception". Namely, the common tendency we have to believe that whatever name we have given something dictates what that something is. A name can serve a purpose in understanding what the named thing is, but it does not have a normative power over it. Especially in certain cases, the name attributed to something might not reveal

anything about its essence. One famous case, as remarked by Voltaire (1773, p.338), is that of the Holy Roman Empire, which was none of the three things composing its name. I suggest that something along these lines goes for the phenomenon usually referred to as "self-deception". It certainly has the traits defining the set of instances of "self" + "deception", but it is a subset of this rather than the whole set. When philosophers think of self-deception and try to define it, they are not thinking of a name. Rather, they are thinking of the certain phenomenon we refer to when we have self-deception in mind. What this phenomenon should be called is a minor issue and not one that will be addressed in this paper. So, here, I am going to continue referring to this phenomenon as "self-deception". However, if the name will not help us understand what it is, how can I proceed to show that there is such a thing as cases of self-deception?

Considering the tough situation it has already put us in, the next best option is to look at some cases that are generally considered paradigms of self-deception and match everything I have been saying so far about the characteristics scenarios of self-deception should have. My aim here is to show that there is something going on with them. If I can demonstrate that these have some relevant differences from other cases that fall in the category of deceiving oneself, they will be worth considering as a category on their own. I am going to present two very similar examples. One where the belief acquired is something the subject wants to be true, and one where it is something the

subject would not want to be true. Having both will be useful for some discussions later...

Unfaithful wife: Frank and Kate have been a couple for a long time. Recently, though, he has been having the suspicion she might be cheating on him. He has found many clues pointing towards Kate having an affair with someone else and she has been acting differently lately. Having to break up with Kate would cause a lot of pain to Frank though. For this reason, in spite of the evidence collected, Frank believes that Kate is faithful to him.

Faithful wife: Henry and Lauren have been a couple for a long time. Henry has no reason to think Lauren is cheating on him. In fact, she has not been anything but a good spouse to him. However, Henry cannot shake the feeling that she might nevertheless be cheating on him. Having this thought makes him more careful to notice the potential signs Lauren might be unfaithful to him. We can imagine that, because of previous experiences, Henry is terribly afraid of not realising his partner is cheating on him. For this reason, despite the evidence available to him, Henry believes Lauren is unfaithful to him.

Now, there is one thing that is off with both of these cases. The two subjects had evidence in favour of one belief, but they decided to ignore such evidence to hold the opposite belief. Thus, we would consider Frank and Henry irrational in their decisions for how to deal with their respective situations as they fail to adjust their beliefs to the evidence they posses. There is nothing too out of the ordinary about these two scenarios: both are situations that might

happen in everyday life. They are not highly specific or absurd ad-hoc examples created to prove a point. These are instances of ordinary irrationality that match the characteristics of the self-deception I want to defend.

What strikes us about these is that, although the subjects had reasons for their respective behaviours, both of them formed a belief opposite to the one they should have. Unlike the case with the fake snake, here the subjects are not making a mistake in their evaluation of relevant facts for the belief they are forming. Quite the contrary, actually, in Faithful and Unfaithful Wife the subjects are well aware about the facts and have good reasons not to believe what they end up believing. The reason they ignore or misevaluate such evidence is because they have some motivations to form the belief they end up forming. For Frank it is his desire not to go through a break up whereas for Henry it is his insecurity caused by previous experiences. Such cases are peculiar since we normally know to differentiate between what seems likely and what is desirable to us. It is unlike cases of simple mistakes. On this ground, I would argue that these are cases that are worth looking into as they cause unusual behaviour. It might be too early to make the jump and claim this is self-deception, but it is useful enough to have two examples to rely on for the rest of this dissertation.

So, how do we know whether these are actually cases of self-deception? They certainly possess the traits necessary according to what said so far: the subjects are hiding the truth from themselves and this is an irrational action of

theirs. However, we might be tempted to explain this by appealing to other mental phenomena that recur in our everyday life. We would definitely not consider Frank or Henry crazy for their beliefs and actions, but simply irrational. Self-deception, though, is not the only instance of everyday irrationality. There are other instances where we do not properly adjust our beliefs to the evidence available to us. I have already mentioned them: wishful thinking and akrasia. These are two other phenomena that are used to explain similar situations. In order to avoid redundancy, one might be tempted to show that the two examples I provided, like any other potential scenario of self-deception, could be led back to them. For this reason, I need to show that neither phenomena can be employed to explain what happens to Frank and Henry. I am going to start with wishful thinking.

As Szabados (1973, p.203) points out, when we think of wishful thinking, we are more thinking of wishful *believing*. This can be considered another case of an imprecise name just like "self-deception". In fact, if we were to take is as wishful *thinking* it would be very easy to demarcate how the two differ. Wishful thinking as the name would intend could be understood as a situation where the subject is merely entertaining the possibility of something they wish for to be true. Just like when I imagine a world where the temperature is always optimal, I am thinking about something I would like were the case, but I am in no way giving it any credibility. I am far from believing it. If defined this way, wishful thinking is completely different from self-deception. However, it would

also be different from what we normally think when we picture wishful thinking. Usually, we imagine someone trying to convince themselves of something they would like were the case just like someone telling themselves "everything is going to be alright".

It is when we formulate it in this way that the difference with self-deception gets narrower. For both phenomena we have the case of a subject that forms the belief of something they wish were true in spite of the available evidence. So, what disqualifies Faithful and Unfaithful Wife as cases of wishful thinking? I think two issues can be raised here. The most obvious one is limited to Faithful wife and cases similar to it. Namely, wishful thinking struggles to explain those cases when the belief acquired by the subject is something they would not want to be true. What Henry accomplishes by believing Lauren is cheating on him is that he is going to be more observant of her behaviour. This will also allow him to be more protected emotionally since he will know in advance what might be coming his way. This is convenient to him since the payoff of not being open to an unwanted surprise is in favour of this decision. Still, this does not make "Lauren is cheating on me" something Henry wishes to be true. The whole reasons he starts thinking it is because he would not want that to be the case. So, it makes little sense to say that this is an instance of wishful thinking. On the other hand, self-deception can handle cases such as

Faithful Wife without creating contradictions of this sort: the subject is forming a belief that goes against their best evidence.[1]

This is already enough to show that not all cases of self-deception can be reduced to cases of wishful thinking. However, it is possible to show that even cases similar to Unfaithful Wife cannot be conflated with it. What marks them as self-deception is the strength of the belief the subject has acquired. Namely, how much the subject has considered possible defeaters for it and, consequently, how convinced they are of it. Take Frank's scenario. Now, imagine a situation where he has not considered as much evidence, but it is the simple desire that Kate is not cheating on him to make him believe she is not in fact cheating on him. I think a description like this better matches the idea we have of wishful thinking. A scenario, like the one described in the original example, where he has evaluated multiple pieces of evidence to formulate a stronger, although biased, belief is more in line with the phenomenon of self-deception.

Szabados (1973, pp.203-204) makes the point that Frank would be more inclined to accept evidence against the belief that Kate is not cheating on him in the wishful thinking case. This might not be enough to move him towards the opposite belief, but he would at least accept it as evidence counting against the one he is currently holding (*ibidem*). On the other hand, in a case matching the

[1] Cases like this one are a problematic set of instances of self-deception called "twisted cases". I will discuss them at length further into the dissertation to show that, albeit with some difficulties, many accounts of self-deception can include them in their view.

characteristics of self-deception, Frank is more likely to devalue the evidence or to outright dismiss it as counting against his belief (*ibidem*). This aspect of self-deception is crucial to explain how a belief caused by self-deception might be so resilient. In other words, it clarifies why a subject is capable of maintaining a belief for an extended period of time in spite of what their previous and any future evidence says they should believe. In fact, any account of self-deception, in order to make sense, must explain how the subject can resist the evidence unlike wishful thinking that has a more transient nature. What exactly I believe these processes that prevent the subject from acknowledging evidence against their belief as such is something I will reserve for when I will delve into specific accounts of self-deception. What I can say here is that there is a clear difference in the behaviour of a subject experiencing wishful thinking and one self-deceiving. This, together with everything else I have said, should be enough to clarify that self-deception cannot be reduced to wishful thinking.

I can now move to akrasia. Akrasia, also called weakness of will, can be defined as "failing to do what one knows it to be best to do; what one does instead is something that one desires to do, but believes it to be best not to do" (Gardner, 1993, p. 34). So, it is the case of a subject that does something they know is not what they should be doing to do something they desire to do. Immediately, it is easy to notice some terminology is in common between akrasia and self-deception: there is the common theme of opting for something that is desired although it would be more appropriate. However, it is possible

to notice that there is something that might disqualify cases like Frank and Henry as potential akratic subjects. In its basic definition, akrasia talks about actions whereas self-deception is about beliefs. Clearly, actions and beliefs are two separate things. The bullet is not quite dodged yet though. One might raise the point that forming a belief could technically count as an action. In that case there would still be room to reduce self-deception to something akratic subjects do. If that were the case, philosophy should focus on akrasia since that would be the source of irrationality.

Forming a belief is certainly something we could say "we do", but that hardly seems enough to count it as an action. After all, we tend to think of forming beliefs more of a reaction rather than an action (Chignell, 2018, section 3.4). When we are questioned about some matter or we wonder about it ourselves, we come out with a belief regarding it. In this sense, the process of belief formation is more like the knee-jerk reflex where the subject does it because of an external stimulus rather than as an intentional action. This makes it questionable whether something like this could be considered akratic behaviour. Can something that happened because of an involuntary reaction still be considered weakness of will? In any case, there is room to argue that forming a belief is something that is indirectly under our control (Mele, 1987, pp.55, 110), which seems to suggest it as an action. Mele's idea is that we put ourselves in a situation where we gather evidence and information regarding a topic, which leads us to form a belief. In this sense we have indirect control over

it. So, then, even if the belief formation itself is not akratic, the process that led to it is. Even with this in mind, I do not think this is enough to qualify believing as an action. I can jump from a tree and that is an action, but it seems farfetched to say that my falling to the ground is an action too. I have no control over it nor can I stop it. This is the case even though I might have been the one to put myself in the condition of falling. For this reason I doubt being part of a chain of events that started from actions of mine is sufficient to be considered an action.

To delve deeper into this matter, I should go further into the field of philosophy of action. I aimed at showing the difficulties in trying to argue that simple akrasia can explain putative self-deception cases. I do not consider it worth doing more than I already have. The main reason for this is that Mele shows that another form of akrasia, which is pertinent to belief, already exists (1987, p.109). This would parallel the version relating to action in being defined as the formation of the belief that *p* when the subject holds a conscious judgment against *p* (Mele, 1987, p.112). Additionally, the judgment has to be the subject's best judgment, namely something that defeats other judgments they might have regarding what belief to form.[2] This formulation is not so easy to disqualify as the previous one. All we need to do to make it compatible with the paradigm examples of self-deception I have provided is rephrasing the events

[2] There is more to the definition Mele provides, but it is not going to be useful to understand the difference with self-deception. An in-depth discussion of akrasia is beyond the scope of this paper.

as Frank and Susan not being able to bring themselves to believe what they know/think to be true to opt for what they want to be true.

Mele himself does not think this means self-deception could be reduced to akrasia regarding beliefs, or incontinent beliefs.[3] Rather, he considers the two phenomena distinct. His reasons, though, are mostly tied with his particular view of self-deception. This means I cannot use them since at this stage I am trying to show that self-deception is a real phenomenon. It would be unfair to defend a specific branch when I am not yet discussing which model should be preferred. This is especially true since I believe there is no need to adopt a specific view to show that self-deception is not akrasia about beliefs either. I am going to borrow Mele's terminology, though. In fact, I am going to be distinguishing between incontinent beliefs and irresponsible beliefs, where the latter is the kind relevant to explain self-deception. When we speak of akrasia we refer to a subject's best judgment. This encompasses all the judgments available to the subject in a situation where the one that emerges as their best one is that with the strongest influence (Mele, 1987, pp.110-111). This is not necessarily tied to the epistemic state of the subject. There might be some usefulness to believing something the evidence presents as wrong even though we might usually be inclined to adjust our beliefs to what the evidence points to. But the reason for this inclination of our is that, usually, believing what the evidence indicates is our best judgment.

[3] These are two names he uses interchangeably since incontinent is a translation of "akratic" sometimes found in translations of Plato and Aristotle.

Take Faithful Wife, Henry's not following his judgment to form a belief based on the available evidence allows him to do something he considers worth the trade (i.e. make it easier to figure out whether Lauren is cheating on him). This qualifies Henry's belief as irresponsible because it fails to match the evidence available to him (*ibidem*). By irresponsible Mele means that the belief formation happens "in the teeth of evidence", so, those cases when the subject is not justified in believing what they believe (1987, p.111). It is irresponsible since it is not the way our belief-formation faculty should be used. This is why we would be inclined to consider Faithful Wife an instance of self-deception. However, Henry's belief is still in accordance with his best judgment. What follows is that his belief is not incontinent. What Mele (*ibidem*) means by this is that it is not something the subject could not help but believe although it contradicts their best judgment. It is such a belief that would make Henry subject to the type of akrasia pertinent to beliefs. Even though his belief is likely wrong and Henry is aware of this, it can be something it is in his best judgment to believe. In this case, his best judgment could be summarised with "you can hold an irresponsible belief if the pay-off makes it worth it". So, Henry is still acting in accordance to his best judgment even whilst holding an irresponsible belief. Such a rule is not something Henry consciously thinks of... Rather, it is part of a series of mechanisms that opt for the belief that is more convenient to him (e.g. because it gives him a feeling of safety) rather than the one he considers more likely based on the evidence available to him. Analysing how these mechanisms can influence our belief formation is exactly what my

dissertation sets out to do. For now, what is important is that this is a case where the subject's best judgment trumps their other judgments. So, this seems to disqualify akrasia as a possible analysis for Henry's situation.

More can be said about how to distinguish akrasia about beliefs from self-deception. I do not think the phenomenology matches. When we think of the akratic person, we talk about someone that has formed a judgment regarding what they should believe. This describes a person that is more aware of the situation. The person might not realise their akratic behaviour, but they seem to know that what they are doing is wrong. What characterises akrasia is, in fact, the inability of the subject to act accordingly to what they are well aware they should do. I do not think the same can be said of self-deceivers. These are most often unaware of their behaviour. Subject of self-deception are often dismissive of the phenomenon and, as discussed when talking about wishful thinking, tend to ignore potential evidence to the falseness of their belief. This is why self-deception is a type of deception whereas akrasia is weakness of the will: self-deceivers hide something from themselves whereas akratic subjects are in a situation where they are incapable of doing what is right.

Finally, this ties up with a distinction offered by Gardner. He describes akrasia as a phenomenon where the subject is passive, unlike self-deception where they have an active role (1993, p.34). Namely, it is some contingent situation (e.g. an addiction) that weakens the will of the subject making them unable to believe what they judge to be best to believe. They do not have a real

agency over this. The same cannot be said about self-deceivers that are not subject to some external condition that weakens their will. Whether we consider self-deception to be intentional or not, the series of biases and strategies the subject deploys to form the belief against the evidence provided to them is a product of their mind. It is something the subjects do to themselves. An interesting case to consider to better separate the phenomena is that of a person whose akrasia makes them self-deceive. Think of a smoker in the 1960s... At the time there was reason to believe smoking was harmful to the smoker's health but this had not yet been proven. Here, the smoker, is clearly affected by a condition that weakens their will, namely addiction to nicotine. This makes them less likely to want to stop smoking. At the same time, though, they are someone who cares for their health and would not want to harm it with cigarettes. However, their addiction makes it so that they really do not want to quit smoking to the point that they might underestimate the studies pinning smoking as harmful. For this reason, they might ignore relevant evidence because their norm of believing what seems more likely is trumped by their incapability to quit smoking. This means that their belief that smoking is not harmful to them is irresponsible. However, it is not incontinent since it is a belief they have in accordance to what they consider their best judgment (i.e. I should continue smoking) So, the smoker ends up being self-deceived regarding whether smoking is harmful to their health.

What the smoker's example show is that even in instances where both akrasia and self-deception take place, it is possible to distinguish the two. Although the smoker's akrasia is part of the reasons they start deceiving themselves, it is a different phenomenon entirely. Their disregard for important evidence is motivated by their akratic behaviour, but is not part of it. So, for what shown here and the previous reasons, I do not think self-deception can be reduced to akrasia even if we accept that akrasia could extend to beliefs. With this I have achieved my aim for this introductory part as I have, hopefully, provided reasons to believe self-deception to be a real phenomenon independent from other cases of irrationality. In turn this might have brought some clarity to some of the characteristics that define this peculiar phenomenon. This means it is now the time to discuss which model of self-deception should be preferred.

2.1. Mele's Self-Deception and the Intentionalists

Since my account departs from Mele's and shares a few similarities with it, I think my best course of action will be to start by presenting his view first. This will also allow me to clarify the distinction between intentionalists and motivationalists, and why I prefer the latter kind of account. From here onward I will use Unfaithful wife as reference for examples when the situation does not

specifically call for something different since that one is the simplest and most basic example. If we take S to be the subject of self-deception (e.g. Frank), ¬B the content of the belief caused by self-deception (e.g. "My wife is cheating on me") and B its opposite, Mele characterises the process of acquiring ¬B as self-deception thusly:

> "M1. The belief that ¬B which S acquires is false.
>
> M2. S treats relevant, or at least seemingly relevant data
>
> to the truth value of ¬B in a motivationally biased way.
>
> M3. This biased treatment is a nondeviant cause of S's
>
> acquiring the belief that ¬B.
>
> M4. The body of data possessed by S at the time provides
>
> greater warrant for B than for ¬B." (Mele, 1997, p.95).[4]

As I will discuss in a moment, Mele considers a desire D of S's (e.g. Frank's would be the straightforward 'I wish for my wife not to be cheating on me') to be responsible for their self-deceived state (*ibidem*). For this reason, Mele falls on the motivationalist side of the spectrum. Now, I should point out that the conditions presented above are considered jointly-sufficient conditions by Mele

[4] Here I am quoting Mele's ideas and phrasing, but I am using my notation. Instead of using 'p' to talk of the belief, I prefer using 'B'. This is for consistency with the rest of the dissertation. Also, I marked the points of the list with an M to make it clearer in future references that this is Mele's account I am talking about. Finally, I have reversed B and ¬B since the content of Frank's self-deceived belief is in the negative form.

himself (*ibidem*), so none of them individually is to be taken as necessary or sufficient even though there is room to consider M1 a mandatory part of his account. In fact, Mele treats it as a lexical point that the belief ¬B must be false for S to be able to be self-deceived into it (*ibidem*). This is Mele's way of making sure that self-deception is an instance of deception as defined by the dictionary. For him, the way someone hides the truth from someone else is by convincing them of a false statement (*ibidem*). For now I will not challenge this interpretation. After all, this first condition is the biggest source of disagreement between me and Mele so it will come up later, in 2.2, when I present my account. But something to point out for now is that, for Mele, Frank can be self-deceived only if Kate is actually cheating on him so to make his belief that ¬B false.

I will continue with the other conditions. M2 is the trait I have referred to when talking about wishful thinking and its differences with self-deception. Namely, this is the condition referring to those processes that explain how self-deception can lead a person away from the evidence they have to the point where they form a belief opposite to the one that they should hold.[5] In Unfaithful wife, how can self-deception make Frank overlook obvious pieces of evidence in front of him? Frank is supposed to be perfectly capable of understanding certain behaviours as suspicious and common to unfaithful people. Imagine Kate to occasionally come home smelling like the perfume

[5] When I say here or henceforth "should believe" I mean it in an epistemic sense as it is the belief the subject has reasons to believe.

worn by a friend of hers, Frank is not going to miss this. It is true that self-deception is regarded as an instance of irrationality (Deweese-Boyd, 2016, section 1), but this does not mean Frank is not in a position to connect the dots and realise that the perfume is evidence Kate might be cheating on him. Rather, he is supposed to be a subject of the ordinary irrationality anybody could experience. After all, self-deception is considered a fairly common phenomenon. So, there has to be something that makes Frank able to dismiss evidence like this as counting against his belief regarding Kate's faithfulness in a way a wishful thinker would be incapable to do. Mele (1997, p.94) provides psychological evidence to show how sometimes what a subject desires can influence what they believe. And it is because of these phenomena that will be presented in the next paragraph that Mele points to the desire D as responsible for S's self-deception. This is what makes him a motivationalist.

Mele talks about four different phenomena that henceforth will be referred to as 'Mele's biases': negative and positive misinterpretation, selective focusing/attending, and selective evidence-gathering (*ibidem*). The first two consist in the subject recognising incorrectly whether a certain information should play a role in their belief formation process. Their desire D causes S to evaluate some evidence as undermining or supporting ¬B when this would actually be neutral to it (*ibidem*). For example, it could be the case that Frank completely avoids considering the perfume as evidence towards B when this is a common trope in situations when a partner is cheating. Or we can imagine

that Frank takes the fact that Kate has not changed the way she speaks to him as evidence that she might still be faithful to him when this information alone does not mean much. The other two biases, on the other hand, pertain to the evidence gathering process. Selective focusing/attending describes S's behaviour in focusing more on the evidence that favours ¬B rather than B. For example, Frank could give more weight to Kate's generally sweet behaviour than to the flirty way she talks to her friend, where the latter is clearly a more impactful information than the former. Finally, selective evidence-gathering describes something more radical, namely how our desires can make us look for more obscure evidence and overlook the one that is in front of us simply because only the former support ¬B (*ibidem*). In my example we could imagine that Frank takes the effort to look through Kate's love messages to himself whilst ignoring the fairly obvious display of a romance between her and her friend happening in front of him.

All these phenomena are the sorts of biases point M2 refers to. The next condition is just to make sure that they are the rightful cause of S's acquisition of the belief that ¬B. Nondeviant clauses are now very common within philosophy to make sure that the putative causes of a process work as intended. As such, this point does not require much discussion. I will limit myself to providing an example where point M3 is not respected to show why it is needed. In Frank's scenario imagine that, after finding some clues of the fact that Kate is cheating on him and spending an extraordinary amount of time

collecting those that she isn't, he goes to get drunk in a pub. While he is there he sends a message to his friend Henry about all the evidence he has found. Because of the alcohol, Frank forgets everything he has done the previous day, but he finds Henry's message telling him about his faithful wife, Lauren. However, Frank misreads it and thinks Henry is giving him his opinion on Kate. Trusting a more objective, external, point of view, Frank forms the belief that Kate is not cheating on him. Because of this mistake, Frank ends up having good reason to believe that Kate is not cheating on him. Mele's bias plays a minimal role in the formation of ¬B. Additionally, it does not have a straightforward cause-effect relation to the formation of the belief that ¬B. It is due to mere chance that Frank misread the message sent by Henry. This would disqualify Frank as a self-deceiver because such a scenario lacks what makes self-deception interesting: it lacks the irrational character.

I will move to Mele's final condition, M4. This point is characteristic of self-deception as it is one of its most central aspects and, possibly, the one that makes it interesting to us. It is one I have been referring to as central to the basic idea of this phenomenon. In normal circumstances, if the data available to S were to point to their desired belief ¬B, it would be odd to define their belief that ¬B as some form of deception. It would merely be an instance of someone's adjusting their belief to match what most likely seems correct. Take Frank as an example. If he formed the belief that Kate is cheating on him from the data available to him, we would not find anything particularly interesting about his

situation. This holds regardless of what beliefs he might have previously had. Imagine he at first assumed her to be faithful, but changed his belief upon noticing the perfume. This seems what we ordinarily do when we act rationally: adjust our beliefs to what we think is most likely to be the case.

I will take some space here to talk about how intentionalists would characterise Frank's scenario. My interest in doing this is to extensively discuss why I think a motivationalist framework is preferable to an intentionalist one, and take this opportunity to flesh out more of what is characteristic of the motivationalist view. As I mentioned in the introduction, intentionalists point to intentions instead of desires to be the cause of the deception as they model self-deception on the basis of interpersonal deception (Deweese-Boyd, 2016, section 2). So, the reason Frank deceives himself is because that is what he intended to do, it is part of his plan to start believing something he does not believe to be true. For an intentionalist, self-deception is close to a case of deception where deceiver and deceived are one person. This is why intentions are necessary for it. In fact, we usually do not say that X can deceive Y if X did not intend to do so (Mahon, 2015, section 3). X might make a mistake and condition Y, but X is not properly deceiving Y unless they have this intention. Another difference is that intentionalists require S to both hold the beliefs that B and that ¬B. This is used instead of point M1 as an interpretation of the deceptive nature of self-deception. This is how the subject, for them, is hiding the truth from themselves. In fact, we tend to call someone a deceiver when

they are aware of the deception, and this entails that they do have an opinion about what the truth is. When we think of interpersonal deception, it does not matter much whether what X is trying to convince Y of is true or not (*ibidem*).[6] As long as X believes it to be false, we would consider them a deceiver.

So, an intentionalist would say that, somewhat aware of his inability to handle a break up, Frank forms the intention to believe that Kate is not cheating on him. Deep down he remains aware that this is not the case but, nonetheless, he follows his intention and forms his belief regarding Kate's faithfulness. So, what is relevant here is that it does not matter the exact reason Frank forms his intention. The important part is that he has it and that that is the reason he ends up forming his self-deceived belief. Also, the intentionalists does not need to know whether Kate is actually cheating on Frank; this is not relevant for determining whether Frank is self-deceiving. What matters is that Frank holds both beliefs at the same time. Once again, this is because that is what they consider characteristic of self-deception. Since Frank is performing the act of deceiving someone (i.e. himself), he needs to be aware of the deception and, thusly, believe it to be false. On the other hand, differently from cases of interpersonal deception, in a self-deception scenario Frank is also the deceived. Thus, he is necessarily someone that ends up believing the false statement, or he would have avoided the deception. This means that Frank will form the belief

[6] It should be noted that the definition of interpersonal deception is itself under much discussion. Some philosophers would disagree with the claims made here about what classifies as X deceiving Y. However, the trend intentionalists tend to follow is the one I am referring to here. According to these, unintentional cases of what would otherwise be deception is more a case of 'misleading' (Mahon, 2015, section 3).

that the deception points to. So, according to the intentionalists, he needs to be holding both beliefs so that he can qualify for both the roles a self-deceiver takes in the deception.

Since this is a branching point in the making of a model of self-deception, I should address here the issue of which of the two branches should be followed. This will come in handy later when I present my full account and talk about how it differs from Mele's. Generally speaking, I am in favour of the motivationalist one. There are many reasons why I do not consider intentionalist models as compelling. For brevity, though, I am going to present the main objection against them raised by Mele. The reason for this is that I think it is something that intentionalist accounts cannot really manage to respond to. Additionally, I am going to present another of Mele's criticisms against the intentionalists. This is one that I do not consider successful. I will still present it since it helps better explain the intentionalist view and it will be helpful later since it is one of the areas where I agree with the intentionalist model. I will start with this latter one.

Mele calls it the static paradox. With this he raises worries against the possibility of a subject holding two opposite beliefs at the same time, which he considers impossible (1997, p.93, 95-96). It seems in fact problematic that a subject might be holding contradictory beliefs. This is a necessary aspect of any intentionalist model, though, since they believe self-deception to be like interpersonal deception: they need the subject to be a deceiver. However, self-

deceivers are people perfectly capable of reasoning and of employing the same mental faculties someone not subject to self-deception is capable to. So, they would realise that they are holding opposite beliefs. Furthermore, Mele takes a scenario like this to be impossible in the first place as it contradicts logic to postulate that someone might believe and not believe something. Nevertheless, if we take Unfaithful Wife, an intentionalist would have it that there is at least one moment where Frank is holding both beliefs at the same time.[7] Although what the conditions for deceptions are is itself a controversial topic, the general consensus is that the deceiver must be untruthful in their deception. Namely, they must believe that what they are trying to convince their victim of is false. Thus, a subject of self-deception, at least in the initial phase, should believe that the content of their self-deception is false. So, if this were to lead to a paradox as Mele indicates, no intentionalist view would be able to work.

A reply to the static paradox can be found in Gardner (1993). There he postulates the Principle of the Possibility of Contrary Beliefs (Gardner, 1993, p.23). Said principle (shortened into PPCB) claims that there is an important difference between believing both that B and that ¬B, and believing and not believing that B. Only the latter is a problematic contradiction since it affirms the truthfulness of two inconsistent matters of facts. But this is in no way implied by the former: believing something and its opposite does not entail

[7] Some views would say that Frank holds both beliefs for the entire period, and even afterwards whilst others would just have a brief moment for their coexistence in the subject when the second belief is formed. This does not matter, though, Mele's static paradox applies to all cases since they all require the synchronic holding of opposite beliefs for at least a brief period.

believing and not believing either of the two. On the other hand, the former formulation is not as problematic: it is not a logical contradiction. We can imagine someone having opposite beliefs unlike imagining someone that believes and does not believe something. Another distinction I would like to add is that between holding a belief with the content "B and ¬B" and have two different beliefs one that B and the other that ¬B. In self-deception scenarios we are imagining the subject to be in the latter situation as the beliefs come from different sources (i.e. one has been there from before the self-deceiving process started whilst the other is the result of it). So, the subject has not yet put the two beliefs together, which contributes to why they have not realised their condition. Still, even though the threat of a paradox in logic has been avoided, this does not mean the entire problem is solved. It is still questionable how the same person could synchronically hold opposite beliefs. Clearly, any person that is not the subject of full-blown irrationality would amend this situation after recognising it by discarding one of the two beliefs. As stated earlier, subjects of self-deception are thought of as merely under ordinary irrationality so they would not continue holding both beliefs. I will appeal to the nature of beliefs to explain how they can end up in such an irrational situation without being this irrational themselves.

A fairly widespread view is that beliefs need not be always occurrent. Rather, they have a dispositional character (Schwitzgebel, 2019, section 2.1). Take the property of being frangible of a piece of glass as an example: it is

not manifest at all times. It only manifests when the glass is struck and breaks. However, we do not want to say that the glass is only frangible when it breaks; we consider it to always be frangible. A similar reasoning can be applied to beliefs. They can be occurrent, as in present in the forefront of our mind, but need not be in order to be ascribable to a subject. The belief that dogs are animals is something a great variety of humans hold. People will usually not think twice in answering positively to a question on the matter. Such a process does not require anybody to ponder and successively form a belief. For this reason, we say that this belief is dispositional. Namely, it is available to use for when the subject needs it. On the other hand, I take a belief to be occurrent whenever this is consciously present in the subject's mind. It is something the subject is aware of whereas dispositional belief might need to be retrieved to become conscious.

This notion is what I consider necessary to answer what I had left unanswered of the static paradox. What I offer is that the subject S is never having both beliefs occurring to them simultaneously. The undesired belief is at the forefront of the subject's mind when the process of self-deception begins. However, as the process takes place and S ends up believing the desired one, the opposing belief gets buried.[8] At least one of them is always in the back of their mind. This is what avoids the clash that would inevitably lead to them reflecting on their contradictory beliefs from happening. With a process similar

[8] This can be applied to intentionalist accounts, but also to any model, like mine, that imagines the subject holding opposing beliefs.

to that of the biases described by Mele, S will tend to dodge the topic entirely since it causes discomfort with the increasing realisation of holding contradictory beliefs. Avoidance behaviour is typical of self-deceivers (Funkhouser, 2005, p.300). Gardner (1993, pp-26-27) observes that, upon noticing their condition, the subject will be almost forced to discard either belief (which one is irrelevant and it mostly depends on the situation). In fact, he points out that realisation is the most common way out of self-deception (*ibidem*). The subject may, autonomously or not, notice a discordance in their behaviour or beliefs. This will trigger an internal confrontation that ends with the winning belief.

Something fishy might seem to be going on here. How is it that S does not recall both of the opposite beliefs when thinking about whether B or ¬B? When thinking of topics that are relevant for it, it would be logical for both beliefs to resurface to the forefront of S's mind. Here, we enter the field of introspection. So, before I can answer this question, I need to discuss this other phenomenon. I reserve this for later as introspection will be the focus of the next chapter. For now, I will focus on a final worry that might be raised against this topic. Namely, can a dispositional belief cause behaviour? I have argued that S never holds both B and ¬B at the same time in the forefront of their mind. However, I have also described how the subject displays behaviour of both beliefs. This is a point that will be important in favour of my account, especially against Mele's. In fact, a tension in the subject's behaviour

and between this and their claims is an aspect characteristic of self-deception scenarios (Fernandez, 2011, pp.381-382). Namely, the subject will occasionally act like someone who believes ¬B would, but is occasionally going to inadvertently act in accordance to the belief that B. Clearly, though, it is not the case that S constantly switches occurrent beliefs when the situation demands it. Phenomenally, this does not seem to describe what happens. But, more importantly, especially in the case of behaviour pointing towards the truthfulness of B, I have claimed that this happens even while S is not attentively focusing on whether B or ¬B. This shows that the belief cannot be occurrent in those instances. However, I do not consider this an issue since I deem dispositional beliefs perfectly capable of causing behaviour. Imagine going to your usual supermarket. Once there, if I asked you to walk home, you would not need to actively think of which route to take. You clearly have beliefs on how to go back home, but need not make them occurrent in order to reach your destination. Still, what causes you to make it back home is the dispositional beliefs you hold in the back of your mind at all time.

I can now move on to the next objection raised by Mele: the dynamic paradox. The dynamic paradox is more challenging as it attacks the core of the view. This asks how a subject could be able to *intentionally* deceive themselves. It is, in fact, generally thought that awareness of a deception is the biggest counter to it. Think about cases of interpersonal deception. Would we say that Y is likely to be tricked when aware of X's intention to deceive them? Fortunately

no, or we would be bound to be deceived far more often than we already are. Similarly, Mele (1997, pp.98-99) says, when you are both the deceiver and the deceived, you cannot fall for your own intentional deception since you are aware of it.

As one can imagine, intentionalists have tried coming up with responses to this puzzle. However, I do not find them convincing. These answers revolve around positing a fragmented mind. Namely, they describe the subject's mind as divided so to have one part being the deceiver and one the deceived. This would bring self-deception even closer to interpersonal deception as here we are talking of two separate entities even though they are both part of the same subject. What metaphysical status we are supposed to give these entities varies between theories and is often unclear. However, they all share the same issue: it is a case where something is theorised ad-hoc to solve the problem at hand. There does not seem to be any other reason for the presence of these entities to explain the process of self-deception. Additionally, I am not convinced this part of the theory finds any underpinning in the phenomenology of the subjects and those around them. Take Gardner's explanation. He describes these entities as Proust's selves (1993, p.28). These, lacking the requirements for personal identity, do not divide the subject into multiple people. However, they are "sub-divisions on the temporal axis of the person" (*ibidem*). If I understand this and what comes after correctly, it means that each of these selves represents

different perspectives of the subjects, which Gardner calls "phenomenological sets" (*ibidem*).

I will show how this would work with Unfaithful Wife. Frank comes to believe that Kate is cheating on him because of the evidence provided to him. So, one of Frank's selves has this perspective regarding Kate's unfaithfulness and related matters (e.g. it might contain the opinion that he should break up with her). From this and the knowledge of how painful going through a break up would be for him, he decides to believe that Kate is not cheating on him. This is achieved because Frank's intention to form this belief would influence the way he looks at Kate.[9] This is his way of hiding the truth from himself. The end of this process is the formation of another self whose perspective is that Kate does not cheat on him. According to Gardner, we are to find confirmation of this when we notice that Frank displays different behaviours when he acts according to the belief that Kate is cheating on him and the one that Kate is not (1993, pp.27-28). Namely, when we notice the conflict in his behaviour that is characteristic of self-deceivers. Finally, the existence of the two selves prevents Frank from realising the existence of this self-deceived state of his since he fails to hold both perspectives at the same time. Doing so would be his way out.

I found this argument unconvincing. In addition to being less parsimonious than accounts like the motivationalists that do not need to posit

[9] I am not going to go too much into details on this part of the account since it is not relevant to understand the selves. But, basically, Gardner does not think an intention can directly affect a belief since believing is not an action to him. Thus, the intention would influence his active thinking and ways of describing it , which in turn would influence his beliefs.

additional entities within the subject, it just does not seem to me like a description capable of explaining the phenomenon. If these selves are not to be considered different persons with different minds, it seems hard to conceive of how Frank is not going to notice this state of his. I do not think the dispositional quality of beliefs is of any help here. These selves are not just holding a belief, but different perspectives. These are not exactly things that can be buried in one's mind... Also, how does Frank switch from one to the other? There does not seem to be a third self in charge of deciding which perspective is adopted. Would Frank not notice that the ways he looks and acts around Kate changes occasionally? From the way it is described, it does not seem these selves are subconscious objects of his mind like non-occurrent beliefs. To be able to hold intentions and similar mental faculties, the subject should be aware of them. So, I do not think arguments along the lines of the one I have given to answer the static paradox can be used in this instance. Therefore, this is not enough to dismiss the dynamic one.

Finally, there is something here that can be used to better understand one of the defining traits characterising self-deception. Views like Gardner's that try explaining the dynamic paradox with a fragmented mind are pushing the subject's contrasting behaviour too far. It is true that a self-deceiver is someone that will display incoherent behaviour depending on the circumstances: acting in accordance to the belief the evidence points to when distracted, but in accordance with the belief they want to be true when thinking

of the bigger picture for their actions. But is it not too much to imagine a subject effectively torn into two that switches back and forth from the two selves? Nothing as convoluted as this is usually observed in a self-deceiver. When caught in error (i.e. the subject is observed acting as if B were true when they have been claiming to believe the opposite), the subject will defend their view, but might give the impression to not truly be convinced by what they are saying. That is because deep down they hold the other belief. If we imagine a person fragmented into two selves composed by different sets of perspectives, we could expect them to be more headstrong about them. In fact, the self-deception might be lost once the two selves are separated... The self composed by the set of perspectives pointing towards the more desirable outcome, which should compose the core of the self-deception process, seems to provide good reasons for believing the more desirable belief. After all, this self is an entire set of perspectives according to which that belief seems to follow. So, when the subject is adopting this perspective, are they not going to be justified in holding the belief they are supposedly self-deceived about?

There are two final points raised by Fernandez (2011, p.387) in favour of motivationalism that deserve to be mentioned in a debate between the two trends. The first is that it cannot be accused of being created ad hoc. There is independent evidence showing how the mechanisms used by motivationalism to explain self-deception are truly part of how our mind works (*ibidem*). He is here referring to what I have presented as Mele's biases. These phenomena are

already known to operate in us by psychologists. Motivationalists simply appeal to them without making them up. The same cannot be said by intentionalists, who need to show that we do actually form an intention of deceiving ourselves before entering a self-deceived status on top of having to show that whatever model they construct to answer the dynamic paradox actually exists. The other advantage is its parsimony (*ibidem*). In fact the only view motivationalism requires us to accept (except itself) is that, whenever we have some desire connected to what we are currently forming a belief about, this desire might influence the way we collect and weigh the evidence in favour of it (*ibidem*).

So, now that I have given reason to move in a motivationalist framework, it is time to discuss one point of uncertainty I have left unanswered earlier: Mele's first condition for self-deception (i.e. that what the subject ends up believing because of self-deception is false). This falls into the debate that wonders whether the belief resulting from self-deception should be false or untruthful. In other words, when Frank starts believing that Kate is faithful to him, what is the necessary condition for self-deception? The first option is to say that this belief has to be false so to imagine that Kate is actually cheating on him. However, on another interpretation, the requirement is for Frank to have held, at some point, the opposite belief like an intentionalist would do. This is what would make the belief untruthful since Frank would be forming a belief he himself does not consider true regardless of whether it actually is. In M1,

Mele goes with the former interpretation as he considers it a lexical point from the concept of deception (1997, p.95). I am not so certain about this interpretation. Firstly, as hinted before, the literature on deception has not expressed a final verdict on whether untruthfulness or falsehood are required for it (Mahon, 2015). But, more importantly, I have already discussed how I do not think we should let the name of this phenomenon guide us in explaining it. Mele did not consider an untruthful self-deceived belief because of the static paradox, but I have already explained with the Principle of Contradictory Beliefs and dispositional beliefs how it is not a problem whether the subject holds both beliefs at the same time. So, is there something that can make us lean towards one interpretation or the other?

I think so... Earlier I have said that we do seem to have a general idea of some characteristics an account of self-deception should have. One of these is the inconsistency in behaviour that the subject shows. This is what I explained earlier when presenting the Gardner's use of the selves: the subject will display behaviour in accordance with one belief under certain circumstances, but act completely differently in others. This is what Fernandez calls the tension of self-deception (2011, pp.381-382), and it is generally considered crucial in marking something as this phenomenon. Going back to Frank's scenario, when confronted with direct questions on whether he believes that Kate is faithful to him, he will reply that he does and act accordingly. However, Frank will be often caught displaying behaviour pointing to his belief that Kate is cheating on

him when inattentive. He might get annoyed at her more easily or immediately look away if he sees her phone receiving a message. These inconsistencies between what he claims to believe and how he acts together with his inconstant behaviour seem to point out that the subject is holding both beliefs at the same time. This is because each belief is causing behaviour in accordance to itself. I find it harder to imagine this sort of tension in a subject that is only holding the more desirable belief.

In fact, with Mele's account, it is unclear why the subject should occasionally act accordingly to the belief they have evidence to think is true. After all that is not a belief they hold. This seems a big flaw since it does not seem to capture what is a marking trait of self-deception. Finally, something that could be considered odd about an account like Mele's is that whether a certain situation constitutes a case of self-deception is determined by a factor external to the subject (Fernandez, 2011, pp.388-389). After all, if it turns out that Kate was not actually cheating on Frank, Frank is disqualified as a self-deceiver because it turned out he was correct. According to Mele, his situation would not be one where someone is hiding the truth. Since self-deception is supposed to describe the irrational behaviour of certain subjects, it seems bizarre to me that the final result is determined by something external to the mind. The reason Frank behaviour was interesting is because he starts believing something that he originally did not consider true. Whether he had been correct

in his original assessment should not matter in analysing the rationality of this choice.

2.2. Hubbs's Confusion Account and My Model for Self-Deception

Before I can start presenting my own account of self-deception, I need to talk about an idea that was recently advanced by Hubbs (2018). In his paper, he does not provide a full account of what it means to be self-deceived. However, he gives a very convincing account of how a belief that ¬B is formed. According to him, this is the result of a confusion S has about two forms of satisfaction that believing ¬B might generate. He distinguishes between what he calls a thumotic and an epistemic satisfaction (2018, p.28). The latter is straightforward in that it is the satisfaction S gains from finding out what they believe is a truth. It is the pleasure associated with discoveries and understanding that derives from believing what is warranted by the evidence (*ibidem*). What he calls 'thumotic satisfaction', instead, derives from the arousal in the subject of positive emotions. Hubbs (2018, pp.35-36) specifies how these often derive from how we picture ourselves in relation to the judgment of others. For example, after some achievement we would consider us deserving some praise by an observer whereas we would feel object to blame when failing. It is important to note that no actual observer needs to be present

nor any actual praise or criticism needs to be issued. The feelings Hubbs believe generate thumotic satisfaction can stem even from how the subject imagines themselves being judged by others if they were present. They can come from situations like when we do something embarrassing and think "What if somebody had seen me?". These actions that we imagine would give ourselves positive or negative remarks are those generating thumotic satisfaction.[10]

This thumotic satisfaction is Hubbs's invention after having done research in the neurobiology of emotions. Such research shows how all emotions cause some level of arousal in the individual (Hubbs, 2018, p. 35). Said arousal is independent of whether the emotion is positive or negative, but it depends on the specific emotion and the circumstance that caused its arousal (e.g. it increments if something we are scared of gets progressively closer). Although we are usually capable of distinguishing the two, we sometimes confuse them. This particular form of confusion is what results in self-deception (Hubbs, 2018, p.37). Hubbs dispels the doubts regarding whether the two satisfactions could be confused by showing that if two things look similar enough, it might be possible that a subject can usually distinguish them, but still occasionally make a mistake (*ibidem*). A typical example for this might be that of two identical twins: although there are ways of distinguishing between the two that someone might be aware of, it does not entail that they will never

10 Hubbs admits that this account of thumotic satisfaction is incomplete without a more specific characterisation. However, he considers it sufficient as long as it makes thumotic satisfaction clearly distinguishable from epistemic satisfaction (Hubbs, 2018, p.36).

be mistaken about who is who. Hubbs advances the additional example of aluminium and molybdenum (i.e. a chemical element very similar to aluminium) to show that it is possible to confuse two things even without being aware of the existence of one. So, it does not seem unlikely that people might confuse between the two satisfactions.[11]

This confusion between thumotic and epistemic satisfaction is clearly insufficient to characterise self-deception since it needs not result in the acquisition of a belief, however it is very valuable in the making of a fully-fledged theory. Before showing the advantages that it brings, I will present how it ties up in my proposed view with the modifications I have made to Mele's account. I believe a process of acquiring a belief that ¬B is self-deception if and only if:

P1. S has had the belief that B.

P2. S maintains the belief that B even after acquiring the belief that ¬B.

P3. It involves some confusion for S between what they think is likely to be true and what they would like to be true (*ibidem*).

P4. S's currently available data provides greater warrant to B rather than to ¬B (Mele, 1997, p. 95).

[11] I am going to provide a more detailed account of this process later. It is not crucial for the present discussion, and might stir confusion. However, it will be important in section 2.4 where I present the combined account of Mele and Hubbs's theories.

P5. S has a bias in treating evidence while pondering whether B or ¬B (*ibidem*).

P6. Such treatment is a nondeviant cause of the retention of the desired belief (*ibidem*).

I have presented these six conditions as necessary and sufficient for self-deception. I do not deny, though, that there might be scenarios that contrast with some of these and we might still want to call these self-deception. To be certain about it, a review of every single scenario would need to be made. Since I have not yet found a scenario I would consider self-deception where one of these conditions did not apply, I have yet to find a reason to make one or more of these not necessary.

The points P4, P5, and P6 need no explanation since they are the same from Mele's account. The reasons for their presence here are the same I discussed previously. A part which slightly differs from Mele's is the difference in my P6 from his M3. I need to address this. Both are put as the nondeviant clause to make sure that Mele's biases play the role they are thought to play in regular scenarios. However, whereas Mele presents them as a condition relevant for the acquisition of a self-deceived belief, I only deem them necessary for its retention. The reason for this is that I do not consider Mele's biases relevant for *entering* a state of self-deception. I consider them important, but only for the subject to remain self-deceived. Going back to Frank, if we take Mele's biases to have a causal role in his formation of ¬B, we will have to say

that Frank's biased treatment of the evidence available to him starts immediately. Even though this happens without any intention of his, the way he influences his own evidence gathering and weighing is quite complicated. If Mele is right, though, Frank does not yet have the belief that Kate is not cheating on him. What is making him treat the evidence available to him in such a biased way then? The behaviour Frank displays in this early stage of self-deception seems unjustified. As I characterised it above in 2.1, Mele thinks this biased treatment is caused by the desire D. However, that seems to give too much explanatory power to a desire over someone's behaviour. I do not deny the presence of the mechanisms recognised by psychology, but merely that these are caused by a desire.

On the other hand, I think Frank is already self-deceived when displaying this biased behaviour. His desire has already caused the formation of ¬B, which is certainly enough to explain why he is treating information in such a biased way. This is where I distance myself from Mele and turn to Hubbs. In fact, I think the biased treatment of the evidence comes into play only at a later point. In his entering a self-deceived state, it is the confusion I put in condition P3 that is to blame. This is the confusion I characterised earlier as the one Hubbs argues happen between thumotic and epistemic satisfaction. So, in order, Frank has a desire D for Kate not to be cheating on him. Because of D, Frank gains thumotic satisfaction from the thought that Kate is faithful to him. For example, he is already imagining how he would feel good about himself

while in a happy relationship with the woman he loves. However, Frank mistakes the positive emotions caused by considering ¬B (i.e. "Kate is not cheating on me") for the type of satisfaction he would feel if he thought he was forming a warranted belief. Because of this confusion regarding the satisfaction he is feeling, Frank thinks he has 'discovered' that Kate is faithful and forms ¬B. Now, this newly acquired belief is what causes Frank to treat any evidence towards the truth or falsity of ¬B according to his bias as described by Mele.

The confusion between the two types of satisfaction does not need a reason to happen, unlike the biased treatment of evidence Mele describes. According to Hubbs, the subject is simply making a mistake, they are not doing anything that needs to be explained with some mental state. Additionally, appealing to Hubbs's account of entering a self-deceived state has another advantage. This allows the proposed view to easily avoid an objection like the one raised by Bermudez to Mele. This is the selectivity problem (Bermudez, 1997, p.108), which indicates how we constantly form desires for something we wished were true. However, a large quantity of these instances remain closer to the wishful thinking I discussed in the beginning of this chapter, and do not cause us to enter a state of self-deception even though they may involve a bias in the treatment of relevant data. So, Bermudez (*ibidem*) asks to show how we can differentiate between instances when we have cases of self-deception and those when we do not. Hubbs's account allows us to make a clear-cut differentiation. Self-deception is characterised by the confusion already

described between thumotic and epistemic satisfaction. This will be the criterion for marking cases of self-deception. If such confusion is present, it will be an instance of self-deception, otherwise it will not.

From how I have described it, the bias aspect of the account may seem unnecessary. It is something related to self-deception as it is its consequence, but it could appear out of place in a characterisation of it. This should not make us believe that my account could explain self-deception without it though. I do not think the retention of the belief that ¬B could be explained without Mele's biases. The confusion account introduced by Hubbs works in explaining how the subject comes to be self-deceived, but does not do much in justifying how they remain in that state. It would be unreasonable to postulate that S continues being self-deceived because they keep making the same mistake regarding the two confusions or because they never come to question the belief they have obtained again. At the same time, though, once S has acquired the belief that ¬B, it would be unusual for them to just discard it. Something like this would be closer to the situation of pondering whether ¬B is the case, followed by the realisation that B is actually true. This does not really describe self-deception, or at least the majority of its cases. The subjects usually retain their belief through extended periods of time. This includes occasions when they face evidence that B is the case. For example, Frank could often be present while Kate is obviously flirting with her friend, or observe her not being as affectionate to him as she used to be. Accumulating these would surely lead

him to doubt and eventually discard his belief that ¬B. However, instances of self-deception are usually more enduring than that, which is why shaking someone out of this state my prove very hard. This part of the phenomenon is what is explained through Mele's biases. Considering how S treats data relevant for ¬B, it is not hard to see how they can avoid the dismantling of their belief. There is more to say on the retention of the belief that ¬B, since not all the evidence against it is external to S. It seems likely that upon occasional reflection on whether B or ¬B, S might find themselves holding both beliefs. To discuss this, however, I need to talk about introspection so I will once again leave this for the next chapter. Before I do this, though, there is one final thing to discuss. In fact, my account must pass a particularly complicated challenge for all motivationalist views.

2.3. Twisted Cases of Self-Deception

If there is one thing all motivationalist accounts share, it is giving a desire the original causal role that generates self-deception in the subject S. However, it is perfectly possible to consider scenarios where S is self-deceived into believing something negative or that they would not want to be true. These are usually considered a sub-category of instances of self-deception that Mele calls twisted cases (1999, p.117). Opponents of motivationalist accounts wonder, in the cases

where the belief resulting from self-deception is worse for the believer than their original belief, how a desire could be responsible for self-deception. Namely, why would S form a desire for something they do not wish were true? I have already presented an example of a twisted case with Faithful wife. In this scenario Henry cannot shake the fear that Lauren might be cheating on him despite her being completely faithful to him. We can imagine her noticing this and even trying her hardest to prove herself to be faithful, but Henry is going to be biased in his treatment of the available evidence and will keep holding his belief. So, Henry ends up being self-deceived into thinking that she is cheating on him. This happens because Henry, based on his fear that ¬B (i.e. "my wife is cheating on me"), misinterprets the signals of Lauren's faithfulness as ways of preparing him for the break up.

Although the evidence currently available to Henry clearly points to his wife being faithful to him, he retains the opposite belief thanks to the biased treatment he is having for such evidence. Assuming he is still holding the dispositional belief B that Lauren is not actually cheating on him, we seem in a position to describe Henry's situation as that of a self-deceiver. The two only components missing for this scenario to qualify for my account are the presence of the desire itself, and the confusion between thumotic and epistemic satisfaction. However, I am not in a position here to rule these cases out merely because of these two conditions. In fact, that would beg the questions against an intentionalist that is capable of construing a model of self-deception capable of

accounting for the twisted cases. For similar reasons, we cannot simply ignore twisted cases of self-deception by saying that they are not true self-deception or some "inverted self-deception". To do that a motivationalist would have to say that what determines whether something is true self-deception is whether the desire at the beginning of the process is positive for the subject. Since the reliance on desires is one of the core differences from intentionalist accounts, such a reply would make the argument for motivationalist views circular. An answer to the twisted cases of self-deception must be found for an account to be successful in explaining the whole phenomenon. This involves finding an explanation to what led Henry to desire for his wife to be cheating on him.

Intentionalists themselves are unscathed by the twisted cases objection since they do not rely on desires but on intentions. It is much simpler to explain how a subject might have an intention to believe something negative. There might be some further motives to explain why, in these rarer instances, the subject might wish to believe something they consider negative (Nelkin, 2002, p.393). For example, in Henry's case, what makes it desirable to him is the same fear that he has of being cheated on that can make him want to believe that that is the case. Believing this can have a number of advantages for him. For instance, he would be able to more easily spot unfaithful behaviour and prepare for the break up. Since in an intentionalist framework the subject is directly responsible for their self-deception, it is easy to imagine Henry realising these advantages and making himself believe Lauren is cheating on him. A similar

discourse about further motives can be said about other cases of twisted self-deception.

Some motivationalists tend to mimic the intentionalists' response to answer this objection. Nelkin does it in the most straightforward way: she moves the desire D a level higher. Instead of saying that S desires ¬B to be true, she characterises self-deception as caused by a desire to believe that ¬B is true (Nelkin, 2002, pp.393-395). For the most part this is going to work out similarly to the intentionalist interpretation... Henry knows that if he is afraid he might be cheated on, he will be more careful so that he is very unlikely to be cheated on without noticing. Since Henry absolutely does not want to be cheated on behind his back, he will have a desire to believe he might be cheated on in order to lower his chance of such an unwanted event to actually happen to him. This desire would cause him to actually end up with the belief that Lauren is cheating on him, which will make him enter a self-deceived state. Additionally, a point from evolutionary biology can be drawn upon here. In some circumstances, the drawback of assuming the worst scenario is balanced out and surpassed by the possible gain deriving from it. In his life, if Henry is more careful, he will be less likely not to spot his partner cheating on him. If Henry was right in thinking them unfaithful, he might actually avoid a lot of pain. On the other hand, if no such risk was present, his only loss is that he is going to be more paranoid of his partners. Arguably, in Henry's scenario, the reassurance that he is not going to be cheated on makes his extra worries worth it. Views

like the one proposed by Nelkin are generally called second-order motivationalism (Fernandez, 2011, p.386). The name derives from the fact that they present as the cause for self-deception a desire about a belief, thus, a second-order desire.

However, similarly to Mele's, my account as described is an example of first-order motivationalism. We characterise D as a simpler desire for something to be the case: a first-order desire. Thus, a reasoning along the lines of Nelkin's is not so straightforward to apply. There is something Mele (1999, p.120) discusses for his account that can be applied to mine too as it does not involve any of the areas where our accounts differ. Mele introduces Friedrich's PEDMIN (i.e. *p*rimary *e*rror *d*etection and *min*imization) (*ibidem*). This explains that when presented with some options, we have a tendency for minimising crucial errors, where what is crucial for the subject depends on their desires (Mele, 1999, pp.120-122). Namely, the subject will consider what the consequences of each possible outcome will be (e.g. continuing a relationship with someone that might be cheating on them or breaking up with them) to determine on which basis it is more convenient to act. The subject will certainly do everything in their capabilities to avoid a crucial error (i.e. acting against an important desire of theirs) (*ibidem*). Consider Henry, who absolutely wants to avoid a situation where he discover his partner has being cheating on him. This makes such a scenario a crucial error for him. As such he will try avoiding choices that can lead to such circumstances. Since the belief that Lauren is

cheating on him is the one that will make him less likely not to notice the signals, Henry will want to believe it (Mele, 1999, pp.123). In this sense, Henry has a belief D that causes him to believe something he did not wish were true.

For clarity, I should explain here that this reasoning and evaluation of errors and possible outcomes is not a conscious thought in the subject's mind. Henry is not sitting down to carefully examine what consequences his beliefs are going to have on his life. Rather, this is something his brain is doing in calculating what the best action would be. I already referred to evolutionary biology and instances of assuming the worst scenarios... Something similar can be said about these scenarios. Think of the most common fight or flight decision animals, including humans, sometimes have to take. A dog that hears a faint rustling in the bushes has virtually no time to pick whether to try fighting whatever might be hiding or running away. It is instinct that is going to dictate what the dog will do. Similarly, these decisions that bring a subject to desire for something negative so that they might avoid an error happen without a conscious evaluation by the subject themselves. I will go into more details on the topic of how the brain constantly makes models and predictions for how the subject should act in a moment, but this should be enough to clarify the fact that these sort of decisions can happen without the subject being directly aware of them. It is only after the belief has formed that the subject is going to be more in control of what happens within their mind. However, at that point, the confusion between the two satisfaction and their biased treatment of evidence

will make them self-deceive. The fact that there is proof of behaviour along the lines of what Friedrich named PEDMIN in most animals, including humans, shows that Mele's answer to twisted cases is not created ad-hoc to save his account from this objection.

2.4. Mele and Hubbs Together

So far, I have been borrowing somewhat freely from Mele and Hubbs to incorporate parts of their theories with my views and form a complete account of self-deception. I have given explanations for these choices, but some doubts might still arise. Will the final result be too convoluted? Would it not be more parsimonious to modify one and work from there? Before moving to the topic of introspection, I think it would be helpful to take a moment to wonder on these issues. What I believe is that each account has its merit, but does not quite capture the entirety of the phenomenon. In this way I am going to be in a position where I can describe step by step the process of self-deception. I am going to start with Hubbs's. He draws from Barrett's (2017) discoveries on affects and how these influence our behaviour.

To explain affects, it is useful to start from a capacity we have been shown to have: interoception. This is responsible for conveying to the brain all the sensations from our body (e.g. things like the hormones in the blood,

sensations from our internal organs, the feeling of clothes on our skin...). These are felt by us as simple pleasant, unpleasant, or neutral feelings (Barrett, 2017, p.56). They do not quite manage to cause an emotion, they can go unfelt when they are not relevant for the working of our brain. In fact, Barrett explains that the brain constantly works to predict what scenario the person is going to be in and how they should act in said scenario. Simply reacting to external stimuli would be inefficient and impractical (Barrett, 2017, pp.59-60).[12] The way the brain predicts the scenarios we might end up in is by simulating them from the data available. Part of this data are the person's affects the brain consider relevant and, thusly, puts to the forefront of their mind (e.g. the neutral feeling of the weight of your tongue is not something you generally have, but it has now been brought to your attention). An example could be that of a gymnast that is about to compete. As the moment draws closer, the brain is well aware that this act will need to draw a substantial quantity of energy and has to prepare to focus and execute precise movements in a certain way. For this reason the brain will give the input to draw from the energy reserve to be ready for the effort and to release the hormones necessary for these actions.[13]

So, how does this work with self-deception? Take Unfaithful Wife... The thought that Kate might be unfaithful is unpleasant to Frank whereas the one

[12] I am not going to go into full details on this account since I am not competent enough on the matter and it is not going to be needed for discussing self-deception. I am going to explain the parts that I consider relevant as they come up.

[13] This is the sort of instinctual analysis and simulation I was alluding to when presenting PEDMIN. All these actions performed by the brain are not usually conscious unless the person is making an effort to focus on them specifically.

that she is still faithful is at least neutral. Hubbs notes that we, as living beings, have a tendency to avoid what is painful or dissatisfying and to do what is satisfying (Hubbs, 2018, p.35). For this reason, Frank's brain will try to avoid thinking about the possible unfaithfulness of his wife to think of the opposite scenario. I have already presented the two types of satisfaction: thumotic and epistemic. To put things together, in Unfaithful Wife the thumotic satisfaction is the positive affects Frank would have when thinking of Kate as faithful to him. There is no epistemic satisfaction generated by this situation since Frank is not discovering anything here. According to Hubbs, what happens in scenarios like these is that the subjects confuse the two types of satisfaction and mistakenly believes to be experiencing the epistemic kind. So, when Frank asks himself whether his wife might be cheating on him, his brain predicts the two scenarios that can generate from answering this question by drawing on all the current affects of Frank's. The imagined scenario where Kate is faithful will be immediately recognised as more pleasant because of the thumotic satisfaction it generates. However, he will mistakenly believe this satisfaction to be epistemic. So, he will believe he has discovered something new (or confirmed something he already knew) and maintain the relevant belief. This would be how the subject enters the self-deceived state.

Now, one question that could be asked here is whether this is possible. Is it possible for a subject to be so wrong regarding their affects and for this to influence their behaviour so radically? According to the evidence brought by

Barrett, this can happen quite commonly. In fact, she points out that when someone experiences affect without knowing the object or event that caused it, they are more likely to treat it as information about the world instead of realising that it came from their experience of it (2017, pp.74-75). This is called affective realism. An example provided by her is that of judges who were observed being more likely to deny parole to a prisoner if this was requested right before lunch. The explanation advanced is that they were unable to recognise that the negative affect they were experiencing was caused by the feeling of hunger coming from their stomach. Instead, they projected this negative sensation onto the prisoners and interpreted it as a gut feeling that they were not trustworthy. So, because of this, they were more likely to have a bad feeling about conceding parole (*ibidem*).

The one presented is the most significant example, but there are more everyday cases that Barrett reports where affective realism interferes with the subject's decision making. These are the result of the experiments of the psychologist Gerald L. Clore... One of these shows that people report feeling happier on sunny days. However, this was the case only when they had not been asked about the weather (Barrett, 2017, p.75). What this means is that the subjects would be mistaken in their ascription of happiness if they did not know the source of it. They would project the positive affect they were feeling from the sunny day onto something else, namely their general mood. For this reason, Barrett recommends to try having interviews on sunny days since

interviewees are, on average, rated more negatively on rainy days (*ibidem*). In this case the interviewer is projecting the negative affect derived from the weather onto the applicant and value them worse than they otherwise would.

For this reason Hubbs's confusion account can work. Frank not recognising the affect he feels and thusly mistaking one type of satisfaction for the other is no different from the judge misinterpreting their hunger for a gut feeling about the prisoners. The fact that this sort of mistake in evaluating the origin of someone's affect is a common phenomenon gives room to argue that it is something that can be extended to all potential cases of self-deception. At this point, since the belief is an epistemic discovery for the subject, Hubbs says it is now going to be something they will defend from possible challenges (2018, p.38). If, for example, we imagine some of Frank's friends trying to highlight the evidence he had collected of his wife's unfaithfulness, we can imagine him dismissing it on the grounds of the perceived epistemic security he feels. This gives at least some reason to imagine why the subject is going to uphold the self-deceived belief over time. It is a partial answer to the challenge I raised earlier regarding whether Hubbs's account was able to explain the retention of the self-deceived state in addition to give an argument for its beginning. I will come back to this in a moment when I discuss the flaws I see in Hubbs's model.

The final part of his account I need to talk about is how it would deal with the infamous twisted cases. It is, like I mentioned earlier, often more complicated for a motivationalist account to explain how or why a subject

might end up in this scenario. For this, we can take the example Faithful Wife where Henry ends up forming the belief that his wife Lauren is cheating on him because of his fear for the possibility of such a thing happening. Similarly to other motivationalist accounts, here it is immediately clear that what has been said about the regular cases of self-deception cannot work in such a straightforward way for this scenario. In this case the subject forms the belief that has the more dissatisfying feeling. So, similarly to first-order motivationalists that might struggle to explain why the subject would wish for something negative, Hubbs has to justify why in these cases S opts for the belief that generates less thumotic satisfaction.

To answer this problem, Hubbs considers what the husband might say when asked about his belief (pp.40-41). According to Hubbs, there are still ways this might be satisfying to him. Here, Hubbs is reflecting on the etymology of the word 'satisfaction'. All it means is doing enough... However, nothing in the word itself ties it with a feeling of pleasure. The association of satisfaction with positive feelings is a contingent connotation given in more recent times. We can satisfy all sorts of things. This could be applied to thumotic satisfaction as well. The example Hubbs proposes is that holding the belief that Lauren is cheating on him would allow Henry to angrily hold a code of honour (2018, p.41). Then, the fact that he can be the one correct brings him satisfaction. Or we might want to say that he would at least like to know how things are. Similarly to a situation where a friend is angry at us and we want to know why, this

information is not going to help necessarily but it will give us that feeling of closure because make things have been made clear. To gain a thumotic satisfaction we might need to discover something that causes us pain. This would still be a case where the subject has felt a thumotic satisfaction and confused it with an epistemic one (finding out an unpleasant truth here) (*ibidem*). Similar scenarios can be imagined for other instances of twisted self-deception.

Now that I have presented Hubbs's account of self-deception, I am going to talk about what I consider its flaws. There are two areas where I find it unsatisfactory... I will start with the one that involves the twisted cases I just presented. My issue with it is that it seems to me an ad-hoc modification of the core view. Although it is true that there is no reason to postulate that both types of self-deception have to be explained in the same way, they have to be compatible in one view. I found puzzling how pleasant affect, which was the core of the model for regular cases, is dismissed in talking about the twisted ones. Especially since I am not sure the view can work without it. Something like the positive relief one might feel by being able to hold a code of honour would be outweighed by the negative affect caused by the feeling of anger, and betrayal from a loved one. But, even if concede that somehow the affect experienced by Henry is a positive one, there is a major issue. Hubbs assumes the subject to be able to confuse the thumotic satisfaction with the epistemic one. So, we have to imagine these two to be somewhat similar in terms of affect.

But, when we take the epistemic satisfaction of finding out an unpleasant truth, it is clear that we have a negative affect. Therefore, it is puzzling that a positive affect generates a thumotic satisfaction that can be confused with an epistemic one characterised by a negative affect. It seems unlikely that he would confuse two things on opposite sides of a spectrum. So, I do not think Hubbs account can manage to explain twisted cases whether satisfactions can be considered something that can be caused by negative affect.

The other area where I find this view lacking is how it explains how the subject keeps holding the self-deceived belief. I think the reason provided to be insufficient. As imagined by Hubbs himself, the husband might find himself in situations where other people challenge the belief he formed. To be an account of self-deception we have to imagine that, in most cases, subjects in this situation will brush off the challenges either by considering them irrelevant or by using the same points in support of the acquired belief. According to Hubbs, the reason the subject does not yield to the challenges is that they are convinced they have discovered something truthful. I do not think this can work as a justification since, usually, people are willing to reconsider their opinions even when they have a high degree of confidence in them. To postulate that a subject would be unwilling/incapable to do so without any further mechanism being in place does not seem enough. For this reason I had incorporated Mele's biases here to explain the retention of the belief.

I will now move to the flaws I found in Mele's account. His model of self-deception is more delineated so here I will be presenting the parts that are relevant for what I am going to say. I have already introduced the set of phenomena I have been calling Mele's biases (1999, pp.218-219). These show that people tend to interpret information in a way that suits what they would like to be true better by misinterpreting evidence both in favour or against their desired outcome. Plus, people are more willing to go out of their way to find evidence in favour of their view whilst they might devalue how much some other information goes against their view. This, together with the fact that more vivid data tends to be given more weight to and the unfortunately common confirmation bias (i.e. we have a tendency to look for evidence in favour of our hypothesis) make an unbiased belief formation impossible. To explain what process chooses which belief is more preferable, Mele present multiple views offered by philosophers. The one I have already introduced is PEDMIN. Although this is originally used to talk about the twisted cases, Mele argues that there is no reason to believe it cannot encompass the regular cases as well (p.124).

To discuss this I am going to talk about the features of PEDMIN I have not already mentioned. The most important aspect of it that makes it work for these scenarios is that PEDMIN is not aimed at truth (Mele, 1999, p.121). Its objective is minimising the errors that would be crucial for the person. Whereas this often coincides with getting things right, this needs not be the case. This

explains why in certain situations a certain person might self-deceive whereas others may not: this mechanism functions in a subjective rather than objective way. It is in the cases where PEDMIN does not aim to get things right that self-deception occurs. Simply put, when the subject has some motivation for something to be the case, this is taken into account in the process just described and gets put above being correct in the testing of hypotheses by the brain. In Unfaithful Wife, it is Frank's desire to live a calm relationship that makes it so that hypotheses which allow him this will win. So, he ends up forming the belief that his wife is faithful to him. It is clear that the same process can be applied to a twisted case like Faithful Wife. We can imagine Henry to be someone who would consider being unaware of their partner cheating on them a primary error. Thus, hypotheses that will make it more likely to avoid just that, like one that will make him extremely cautious, will be preferred. At this point the biases described above will grant that the subject forms and keeps holding onto that belief since any treatment of evidence will be inevitably modified to better conform to the preferred hypothesis.

With this I do not need to add anymore to what I have already said regarding Mele's account; this is the part relevant for this discussion. As I already mentioned earlier, I only have one objection to this part: it lacks explanatory power. What I mean is that even though it does an excellent job at clarifying how the process works, it does not explain why this happens. Why does the subject end up in a situation where all these phenomena trigger in a

way that leads them to self-deception? The model remains mostly vague about this, we do not know what becomes characterised as a primary error. Whereas Hubbs has affect to explain this part of the problem, Mele's view does not have much. Since PEDMIN is not something that happens consciously in the subject's mind, it seems a stretch to say that it incorporate the subjects opinions in the hypothesis it tests. Additionally, it is unclear what motivates the subject to act in a biased way in the evidence collecting and evaluation. I do not think the desire that eventually causes self-deception can be offered as the cause of these.

It will probably be clear, at this point, that what I think is lacking in Mele's account is something like Hubbs's confusion. Whereas that model might be lacking in other areas, I think it explains this part very well. It is also grounded in the studies conducted by Barrett's team. In turn, I think that model can benefit from the processes described by Mele to fix the parts I deemed insufficient. The only question left at this point is: are the two models compatible? I think they are. Both revolve around mental processes (i.e. Mele's bias and PEDMIN or affect) that the subject is not directly conscious of and influence their belief formation process. In fact, I think PEDMIN can easily be linked to the way Barrett says the brain formulates models to predict how a certain action is going to go and regulates the body accordingly. This is exactly what is done when our brain is not consciously evaluating different hypothesis to consider which one will minimise the amount of primary errors. And, as

mentioned in the previous paragraph, the addition of affect provides something the PEDMIN model can base its prediction on for hypothesis testing.

The best way to properly show how the two models would work combined will be to show it. The first part will be from Barrett's and Hubbs's. Namely, the process will start with subject influenced by affect as the brain evaluates which simulation is more pleasant than the other. By combining all the relevant information available, Friedrich's PEDMIN is going to evaluate the consequences of the various scenarios that may take place. In this way, the brain can take in consideration what the things the subject wants to avoid are, and what the most agreeable result is. This makes such a process subjective from person to person, and subject to an external influence like a desire for something to be true. This should allow the view to account for both regular and twisted cases of self-deception since the fact that something that is usually considered negative might have positive consequences is taken in consideration. At this point the subject confuses the satisfaction felt in the overall more desirable scenario with the satisfaction we feel when we discover something. Thusly, the subject starts deceiving themselves. From this point onward, the subject will be biased in all their evaluation of evidence pointing towards either belief, which will allow them to remain self-deceived.

3. Introspection

I can finally move to the second main topic of this paper: introspection. I have already mentioned that it is reasonable to think this other phenomenon might interfere with self-deception. In fact, it might even make us wonder whether self-deception is even feasible in the first place. Considering that humans are generally thought to be creatures able to introspect their mental states and processes, there is reason for perplexity here. In my account I have argued that self-deception involves the subject holding contradictory beliefs at all time. This is where the issue lies. Even though the two beliefs are never occurrent at the same time, the subject should eventually be in a position where they would introspect their contradictory opinions and would have to discard one. This is what would lead to that realisation Gardner (1993, pp.26-27) notices is characteristic of the coming out of self-deception. In fact, we have assumed the subject in the paradigm examples to be generally rational people. By 'generally rational' I refer to any person that is not afflicted by some condition that makes them act completely irrationally to the point of having no issue holding two opposite beliefs. However, some irrationality must be attributed to them, after all, they are self-deceived subjects of (i.e. somebody that forms a belief in spite of the evidence available to them). Nonetheless, we have to imagine the subject to be someone that, upon realising they are holding both the belief that B and

that that ¬B, would evaluate them and discard one. So how is it possible for the self-deceivers to hold both for an extended period of time?

Before I can jump into explaining why introspection does not cause a threat to my view, though, I need to clarify what I take introspection to be. Although not as controversial as self-deception, this is still a highly discussed topic in philosophy. To explain it, many different views have been advanced. So many that it is possible to distinguish different branches and sub-categories within these branches, many more than for self-deception. For this reason, I cannot just talk of introspection as if there was only one way of understanding this concept. Rather, I am going to present one account and use that to analyse how it will interact with self-deception. The one I have decided to follow is Lycan's as it is the one I believe better captures the phenomenon of introspection. Additionally, I will rely on Nisbett and Wilson's discoveries regarding the fallibility of introspection to show that there is a space for self-deception to occur. I should note, though, that I am not going to make a full discussion and argument over introspection as a faculty. This dissertation focuses on self-deception. What I need here is an account of introspection that can be considered convincing in explaining this capacity of ours. This is merely to show that introspection does not invalidate the existence of self-deception.

Ultimately, my choice to follow Lycan's model is due to how well I believe it fits with the other aspects of our mental life... Its focus on attention makes it very relevant for the more recent discussion over the role of attention

in discussing consciousness (Watzl, 2011, pp.726-727). Additionally, I believe a model that considers introspection an active sense we have fits better with its phenomenology as I am going to discuss in 3.1. I will provide some reasons to believe that it is not because of the precise structure of Lycan's account that the two phenomena are compatible though. Given the space, I would have liked to explain how self-deception works in relation to introspection, not Lycan's introspection. However such work would need a lengthy discussion over all the numerous models for introspection. So, I will show that this specific model does not strand too far in character from the main trend in models of introspection and, thusly, provide reasons to believe other models might work similarly with self-deception. In this way, any model that does not outright reject self-deception should be in a position where they can easily adapt their accounts to encompass this other phenomenon. On the other hand, those models of introspection that leave no space for self-deception will be in a though position. I believe the burden of proof falls on them. In the second chapter I have offered many reasons to believe self-deception is a real phenomenon that happens. So, any view that cannot account for it has to explain why this is not a problem for them.

Now, for the reason just mentioned, in presenting Lycan's introspection, I have decided to keep in mind the general guidelines provided by Schwitzgebel in the *Stanford Encyclopaedia of Philosophy* (2014, section 1.1). There, he delineates the general characteristics a certain process has to respect in order

to be called 'introspection' in contemporary philosophy of mind. Thus, if I manage to show the compatibility of Lycan's model with self-deception, the other accounts are likely to be similarly compatible. In the rest of this paragraph I report Schwitzgebel's conditions from the Stanford Encyclopaedia of Philosophy (*ibidem*); in the next paragraphs I will proceed to explain them. Presented in short, these require the introspective process to be: (A) aimed at mental states and processes; (B) targeted only at those mental states and processes of one's own mind; (C) restricted to those mental states and processes that are current or in the immediate past of our mental life. Additionally, he explains that it is fairly common for accounts of introspection to include the following characteristics: (D) the relation of the introspective process linking the introspected mental state or process and the resulting one is direct or immediate; (E) this same relation must detect the introspected mental state or process and be separated from it ontologically; (F) introspection requires effort so that is not constant, effortless, or automatic.[14]

Condition A is a fairly straightforward requirement that simply makes sure that introspection is only concerned about the mental. Knowing the position of one's arm is not introspection although knowing the feeling about having one's arm in a certain position might be (*ibidem*). With condition B we want to make sure that not all means of finding out about someone's mental

[14] In the following two paragraphs I am going to give a brief characterisation of each of these. The original source on the Stanford Encyclopaedia (Schwitzgebel, 2014, section 1.1) goes more in detail with their explanation. I would recommend reading that for a more in depth analysis.

conditions are considered introspection. There are plenty of ways to find out about some mental state or process belonging another person (e.g. a simple conversation) that we would consider far from introspection. With this condition we have clarified that introspection is limited to one's own mental states. C rules out all those processes that inform us of non-recent mental states or processes of ours. For example, this exists to make sure that something like memory is not itself introspection. The way we can introspect distal mental states or processes is by firstly recalling them with memory, and then introspecting the result of the first process.

The second half of the conditions is a bit trickier to define since these were introduced with less obvious intents. D is a way of making sure that only processes simply consisting of 'looking inward' to our mind and examining some state or process count as introspection. So, something like deductive reasoning or reading a diary about oneself do not count as introspection since they are an indirect way of finding out about ourselves. Condition E clarifies that there must be some interaction between the introspective process and the introspected mental state or process (e.g. introspection detects a desire and forms a belief about it). For instance, something like me thinking 'I'm thinking about my thinking' is ruled out by this condition (*ibidem*) since there is not a real causal interaction. The connection between the two is that one would not exist without the other. With condition F introspection is taken as an active process we have control over like listening rather than something more passive

like hearing.[15] In this sense, introspection is thought as something that can be turned on and off, and requires some effort for the one doing the introspection.

Before I can move forward to Lycan's model of introspection and see its interaction with self-deception, though, there is something I should take a moment to discuss. Schwitzgebel raises several sceptical doubts against introspection. In fact, he is worried that the majority of philosophers seems to uphold an optimistic stance regarding the knowledge of one's own mental states and processes (2008, pp.245-246). Schwitzgebel agrees with the common idea that introspection is a very important faculty, but he considers it unreliable both in the sense that it does not always work and that it often makes mistakes (2008, pp.246, 265). Specifically, he thinks that there are certain instances in particular where introspection often bears the wrong results, but that, even in what are generally called 'favourable circumstances', it often leads to error. Most people would concede that we are bad at introspecting when we are drunk or in a situation of distress, those are the 'unfavourable circumstances' after all. It does not come as too much of a surprise that we would be wrong when introspecting in those states. Schwitzgebel thinks that we are not much better in our everyday life. His only concessions are those cases where we experience a simple quality (e.g. redness) or a pain since these could hardly go wrong (2012, p.2). Cases like these are considered incorrigible. When we

[15] Here I am thinking of hearing as the reception of sounds. So, as something the subject has only a very minor control of. They can cover their ears or try to focus on something else, but they cannot completely block their hearing faculty.

experience a simple quality, it is true that we are experiencing regardless of whether this is actually present in the world around us. So, these instances of introspection are always correct.

Although I agree with Schwitzgebel that introspection is not an infallible tool of our minds, I want to show that his judgment is too harsh against it. There certainly are cases where we think we would be correct with our introspection and we are not. However, I do not think these should be considered such a threat to the extent advanced by Schwitzgebel. We might often err in 'unfavourable circumstances', but we are not that commonly wrong in our everyday life. Also, if we are aware of what these unfavourable circumstances are, we can know not to rely on introspection when they take place. We seem to be perfectly comfortable relying on our sense of balance even though when we are drunk we know we are not as reliable. Is there a reason why introspection should have such high standards of reliability? As someone, following Lycan, who believes in an inner-sense version of introspection, I model it in a way that resembles sense-perception. Namely, I consider introspection a sort of sense of our brains to scan themselves and retrieve a certain mental state or process.[16] Since introspection is modelled around the other senses, I do not expect such faculty to be perfectly reliable. Similarly to how each of our senses is well known for being subject to error, introspection has its flaws that make it an imperfect tool for accessing our minds. However, I

[16] I will come back to what this entails when discussing Lycan's model.

deem it dependable enough to be something we can compare to the senses of sense perception. Just like we still rely on these in our everyday life in spite of their mistakes, I do not think minor issues, especially if common only in certain scenarios, should be enough to dismiss introspection.[17]

I am going to proceed in order with the different problems raised by him starting with the simplest case: having a visual image. Schwitzgebel asks us readers to think of something familiar to us, like one's home (2012, p.3). Then, he compares this to the actual perception of a real object in the outside world. Unlike the latter, our visual image lacks many components. For example, before we give them one, objects might lack a colour, or we might not be able to keep picturing some parts of the image whilst we focus on others (*ibidem*). More importantly, he points out that, until asked these substantial questions about the image, we have most likely never thought about them before. We would have said that the image in our head was complete: it is only on further inspection that we introspect how imperfect our imagination is. On the other hand, we do not have the same struggles with sense-perception: we do not need much further thought to answer substantial questions on how we see something (2012, p.4). Finally, he points out how we are yet to find any sort of

[17] What I am doing here might seem in contrast with when I claimed I aimed at defending introspection as a whole from Schwitzgebel's doubts. The fact that I am drawing a comparison with our mainstream senses to account for it makes this part of my dissertation very specific to the inner-sense views that Lycan's is part of. However, there would be not much of a point in defending each view since I am only going to talk about inner-sense views. If the others are not capable of resisting the difficulties raised by Schwitzgebel, there is reason to doubt they are accurate models of introspection in the first place. So, they would not be something I should defend. What I need to do here is make sure that at least one model of introspection can find answers to Schwitzgebel's criticism.

correlation between the different degrees in richness of imagery people report to have and their capability in performing tasks where this is helpful (*ibidem*). Schwitzgebel takes this to show that we do not know much about our mental imagery. This would be a problem for introspection since it means we are not using it correctly when trying to understand how it works, which makes it questionable to claim that it is something we can rely on.

There are several points to raise in response to this worry. To begin with, I think it is important to keep in mind that the concept of mental imagery is a bit vague... As Thomas (2018) notices, with it we might refer to different capacities, and not all of these involve a vision-like image. For example, the way I usually imagine something is more conceptual than visual: I have very vague pictures in my head whilst everything else is described similarly to how one would note down information about something. At least, this is how my act of imagination appears to me when I think of it. I am not very artistic compared to what often people report experiencing when imagining something. So, I believe, demanding for something to be imagined as a picture, like in the example provided, might lead some people outside of their usual act of imagining. As such, they would definitely need more thought to answer questions about it since it is an action they are unfamiliar with. This is not yet my answer to Schwitzgebel, but a worry I had when I encountered this example.

Regarding introspection, though, I do not think this issue should be of much concern. I think it is acceptable that we might be mistaken regarding how

certain processes of our mind work. When we employ something like our imagination, we do not usually monitor how we are doing so. This is because we have no need to as we are perfectly capable of performing whatever task we needed it for without having to actually control it. In the model of introspection I am defending, introspection is considered an active faculty. As such, we are not constantly introspecting all the inner workings of our active mind. Rather, we need to specifically focus on something to introspect it. We seem to go in a sort of 'auto-pilot' mode whenever something does not require our active attention.[18] So, outside of philosophy we have not had much reason to worry about how imagination works. Thus, it is not something the majority of people will have thought for long about and introspect to analyse its functioning. This would explain why we are still at a state where we do not know much about it.

At the same time, I do not think our incapability to answer substantial questions about our faculties is limited to introspection. Our senses are not immune from gaps of a similar kind. The comparison with vision is a bit unfair as this is the sense that receives the most attention and might be the most used in our everyday life. For this reason, even non-scientist like me can answer more substantial questions about it since so many facts about it end up becoming common knowledge. However, think of taste... Most people would claim that our flavour reception is entirely determined by it. However, this is

[18] The concept of this auto-pilot mode is something that will be clarified better firstly when I present the different types of consciousness Armstrong distinguishes and, then, when I talk about attention.

not true as smells have been observed to be a relevant component to it. Part of the taste we feel is influenced by what we are smelling: eating whilst closing your nose limits how much flavour you feel. Our ignorance regarding certain aspects of our senses applies to circumstances when there is a clear error in our perception too. To stick with taste, think of the protein known as "miraculin" from the miracle fruit... This has the property of affecting how we perceive sour tastes making them feel sweet in our mouth because of how it reacts with our taste receptors (Zeece, 2020, p.225). This is a clear and quite noticeable circumstance where our perception of taste is incorrect (i.e. imagine biting into a lemon and feeling only a sweet taste). However, we have been unaware of the reason for this change until the last part of the previous century. So, I am not sure why Schwitzgebel seems to think this is a problem for introspection in particular. The fact that there might be more things we do not know about it could indicate that it is a harder phenomenon to describe.

Although I have been using taste to show that introspection is not the only sense that we are not fully knowledgeable about, even vision is not entire safe from this sort of errors... Someone that does not know how movies are made, would say to be witnessing motion in the scenes on the screen. However, it is a known fact that movies are just a series of quickly reproduced still images to give the illusions of a moving one. This means that our vision does not give us the correct representation of what is in front of us. So, people unaware of this fact mischaracterise their own perception since they are unaware that the

motion they experience from movies is actually a perceptual illusion. These sort of mistakes happen because there is no tell within the perception itself that shows some of our assumptions might be wrong. Discoveries like these require us to know certain facts about how are senses work to be able to explain how these mistakes take place. So, it goes without saying that the more research is performed on one, the more we will be able to account for.

For these reasons, I think it is not worrisome if introspection is not capable of easily answering some questions on the workings of our mental processes. I have explained this is something it shares with sense-perception and why this can come to be depending on how much we know of a certain phenomenon. Another issue pointed out by Brons is with the tests that are employed to figure out certain aspects of our faculties. The example he takes to answer Schwitzgebel is specifically one of those used to figure out how much a person's vividness in mental imagery correlates to perform certain tasks. What he argues for is that even people that seem unable to form mental images (i.e. people with aphantasia) can compensate with other faculties (2019, pp.16-17). So instead of performing the task as intended by the test, namely by using imagination, they do it some other way. Take the mental rotation test... This is a test where the subject is asked to figure out which of a series of shapes is the same as the one given but seen at a different angle. The problem with this test is that the images are usually all available to the subject, who can simply compare back and forth between them to figure out the correct answer without even

forming a mental image (*ibidem*). So, even people with aphantasia might be able to perform them and pass the test. Thus, this shows that one reason we might not know much about certain aspects of our mental faculties is because the way we find such information is flawed.

The next topic regards the introspection of emotions. Once again, Schwitzgebel questions how good we are at introspecting them (2012, pp.5-6). He claims that in his experience, he is not able to characterise these as well as he would be with the characteristics of an external object (i.e. a desktop). The phenomenology is complicated to define and it is easy to end up misjudging what emotion one might be experiencing (*ibidem*). For Schwitzgebel, this is not a matter concerning language. It is not that we lack the words to properly characterise a clear phenomenology or to use this to differentiate among them (2008, p.250). Rather, it is our mental states that are not fully clear to us like an external object would be. The characters of the joy one might feel during a party are not as clear as the characters of a tomato in front of you. So, this is another case where our introspection seems a process not as effective as sense-perception.

Once again, I find contentious the use of vision as a term of reference. I think that things already become a bit foggier when we think of hearing. Possibly, some characters of a melody playing are going to be quite clear to those hearing it. But what about timbre, pitch and other aspects of music you are not quite used to pay attention to? Do these characters appear as clearly? In

any case, I disagree with Schwitzgebel on the fact that the issue is not linguistic although I agree that it is not *merely* linguistic. I will better explain this with an example that actually happened to me... Some time ago I was talking with a friend of mine about a character from a videogame (i.e. Byleth from Fire Emblem Three Houses) and we ended up debating whether their hair was a shade of blue or green. There was no doubt that we were looking at the same picture and that there was no abnormality about lighting or our vision. This was nothing like the infamous black and blue dress. However, neither of us really knew how to call that shade of colour; it definitely is something in-between the green and blue colours, but not of a shade you would often see. Now, what I am getting at here is not what Schwitzgebel discards... I do not want to say that the problem for us was that we were lacking the correct word to call the colour (2012, p.6). Our issue was a bit deeper: in the less unknown area composed by the shades of the green and blue, we had different standards for how to distinguish which colour a particular shade belonged to.

I think we have a similar problem with emotions. I am going to talk about joy, but this goes for all of them... Think about what you know about how feeling joy feels like. What is it like to feel joy? This area is a bit problematic since it depends on what we take mental states to be (i.e. is it the neuronal activity or something else?). Regardless of this, we do have some guidelines on how to tell when we are feeling joy, but these are vague. Surely, we know we must feel a rush of energy and be feeling good when we are joyous. But so is the

case when we are excited for something or when we are satisfied. Is it always obvious which one we are feeling? Schwitzgebel suggests this is because we are bad at introspecting. I think it is quite the reverse. When we grow up we are told which shape is which and we are quite easily able to distinguish among them early in our lives. The same goes for other properties relevant to the senses... However, we are given vague, if any, descriptions on what it really feels like to have a certain emotion. So, we end up being bad at introspecting our emotions because the way we have learnt how to qualify and differentiate among them is flawed in the first place. This is not too dissimilar from the scenario I presented with that videogame's character. It is because we lack the relevant information to make a fine-grained differentiation that we are incapable of attributing the shade we see to the correct colour. It is true that such cases of uncertainty are, once again, more common with introspection than with the other senses. However, I think the reason for this is quite simple.

Objects are external and available for everybody to experience. By induction, it is easy to figure out of what property people are referring to even if they do not know the word. Show me a bunch of different shapes of the colour green and I will figure out that "green" is a type of colour and not shape. On the other hand, emotions are personal. People may feel different about certain situations. I am not advocating for a relativism of emotions... I think everybody probably feels each emotion approximately in the same way. However, not the same event will have the same effect on everyone. This will

cause problems with identifying which emotions is which. For me, a new beginning might bring joy, for someone else that is a cause for excitement... Imagine trying to explain a colour or any other physical property to someone without being able to have them experience it. This is where the similarity with the shade of colour of that videogame character's hair continues. To prove who was right, my friend and I started searching on the internet images of particular shades of blue or green. However, we quickly notice how hard of a task this was. Without a proper expert, it is hard to find proper names for specific shades. Try searching for any shade of colour in particular. You will quickly realise that even among the first images you find, you are going to find different variations of that shade further proving the point that the categories we use for them are not well defined at all.

It might seem that I am arguing against my aim here. Because even though language is the cause of the problem, it might seem like introspection is still faulty. I do not think this is the case. When we introspect joy, we are experiencing all the relevant features that characterise what it feels like to feel joy. There is no mistake there. The problem comes later when we give a name to these symptoms. However, this is a mistake in linguistic usage, not in introspection. It is comparable to somebody seeing a powder blue and calling it a light blue. There is nothing faulty with this person's vision, it is just that they do not really know what 'powder blue' corresponds to and cannot determine whether it is an instance of that or some other shade of light blue.

There is one final instance where Schwitzgebel thinks we are bad at introspecting. This is regarding our own traits and behaviours. He makes multiple examples; I am going to present one for each. To start with traits, we would not be very surprised if some degree of error were to be present in our own evaluations since it is quite challenging to attribute certain traits to oneself. However, Schwitzgebel reports of studies where the comparing of self-attributions of certain traits (e.g. creativeness, laziness) and peer-attributions for the same person correlated negatively (2012, pp.12-13). Namely, people who reported themselves to be relatively creative or lazy were on average not as likely to be described as such by their peers. His example regarding behaviours is that of a sexist professor. This is the example of a professor that considers himself an advocate of feminism and equality between the genders. However, every time the occasion presents, he displays sexist behaviours: he might get more surprised when a female student says something correct, or he might be more critical of women than men during class (2012, pp.10-11). Clearly, these displays of behaviour show a sexist attitude of his. However, he is sincerely convinced that he is not sexist.

Although these are two different cases, I think the same explanation might be given to both. In fact, there is one crucial aspect these two have in common: it is easy to imagine a reason as to why someone would want to think of themselves the way they describe themselves. This applies to all the cases presented by Schwitzgebel in his papers (2008, 2012), and those examples I have

thought of myself. Simply, as we live through our lives, each person forms an idea of themselves that might be more or less close to reality. We all have certain beliefs about our traits, behaviour, and so on... These are the traits that contribute to make us figure out who we are. Depending on the person, though, these belief may be influenced by who we would like to be, by who we fear we might be, or a mixture of the two. As such, we have a reason to form a desire to actually be that way. It should be clear by now what direction I am taking here: I consider these instances simple cases of self-deception, both of the straightforward of twisted kind. In fact, these are situations where someone wants something to be true, starts believing it is, and acts accordingly to it. It is this process that makes us unreliable in reporting our own attitudes or traits. This may sometimes be noticed by their peers who might form a very different opinion regarding the person's traits or behaviour, which makes it so that self-attribution of them can even correspond negatively to peer-attribution.

So, is this a case where introspection is at fault? For once, yes. The interference of self-deception makes us introspect the wrong trait or attitude. For this reason, we end up with the wrong belief about ourselves. This is a phenomenon that I will have to discuss in further details later as the interaction between the two phenomena is the topic of this entire dissertation. For now I am left with one question: how bad is this for the view that introspection is an overall reliable tool for us comparable to our senses? I do not think it is particularly bad. Self-deception happens in determinate circumstances and can

be prevented with the right attitude. When the subject thinks rationally, they should not be deceived into thinking something wrong. Self-evaluation is known for being hard, and we are aware we should be taking people's opinion of ourselves with a grain of salt. So, this makes self-deception unreliable just in the sense that it might occasionally cause the wrong result and only in specific circumstances.

Once again, though, I do not think this is particularly different from what can happen with the senses. Think about priming... Imagine a wall with some small dark spots in a corner. It is common for people that have specific phobias to see these as the object of their fear. Those scared of spiders will see them as spider, those scared of cockroaches will see them as cockroaches... This is a case where something else (a fear in this case) interferes with the perception of the subject. Many other examples might be made also with the other senses. Take some fruit flavoured sweets, are you really capable of telling them apart or even guess what flavour they are without seeing what you are putting in your mouth? Do they really taste like the actual fruit if you do not already know what they are supposed to taste like?

I take this to be enough to answer the doubts expressed by Schwitzgebel. Although, we agree that introspection is not the infallible tool some philosophers think it is, it is not so unreliable as Schwitzgebel makes it to be. What I have shown is that regardless of some flaws that might be more present in unfavourable circumstances, it is a process that can be trusted to do its job

like we do with the mainstream senses we rely on every day of our life. In fact, it is not unusual or exclusive to introspection to find areas where it might not be as reliable as we would like it to be. What this means is that we should not take them for guaranteed, but we should have enough reasons to still rely on it. This is a good point to start discussing the model of introspection I will be using for the rest of the paper. In fact, the fallibility of introspection itself is one of the core features Lycan presents, and one that will help explain how we can self-deceive. However, before I introduce his model, I should present Armstrong's. In fact, Lycan's departs from his. Thus, a brief overview of how Armstrong formulated his model will be helpful to understand Lycan's.

3.1. Armstrong and Lycan's Models for Introspection

To begin with, Armstrong too proposes an inner-sense view regarding introspection. In fact, he models introspection on the basis of sense perception. To be more specific, he uses proprioception as a term of comparison to explain how introspection works (Armstrong, 1980, pp.61-62). Briefly, proprioception is the faculty to perceive one's own body. For instance, a person needs not rely on any other sense to locate their own limb. We can feel where our body parts are located without having to look for them or touch around to find them. Similarly, introspection would be a sense that provides information to its user

about their own mental states and processes. The closest thing to a centre of this sense, similarly to vision's being the eyes, would be the mind itself. However, similarly to proprioception, introspection is better understood to lack a proper organ (Armstrong, 1980, p.62). This should not be too problematic. It is true that we tend to think of a specific body part when we think of our senses, but things are not as clear when we think of senses different from the common five ones (e.g. proprioception itself, balance...). In this way, introspection seems to be similar to what is usually called 'consciousness' in non-philosophical contexts. Namely, introspecting would be comparable to being aware of one's own mental states and processes. Armstrong accepts this although he does not want to say that there is a perfect overlap between this faculty we nowadays call 'introspection' and the more broad consciousness.[19]

In fact, Armstrong classifies three kinds of consciousness: the minimal sort, the perceptual one, and the introspective consciousness I have just presented simply as 'introspection'. Minimal consciousness is the sort of thing that might be thought to be absent when we are in a dreamless sleep or under total anaesthetic (Armstrong, 1980, p.55). Thus, we can imagine the lack of this as characterising those very rare situations in which the subject does not have occurrent mental states or processes. It is, thusly, a form of consciousness we can almost always be said to possess. Perceptual consciousness is the sense that

[19] A small point to note here is that Armstrong does not refer to what we usually call 'introspection' with this name. As I am going to explain in the next paragraph, he refers to it as 'introspective consciousness'.

makes us conscious of what is around ourselves. This refers to those instances when we are using our senses (Armstrong, 1980, p.58-59). This is presented as a form of consciousness because we do seem to understand a sense in which a person that is dreaming and, thus, disconnected from reality, is not conscious (*ibidem*). Finally, introspective consciousness is what I have presented in the previous section as Armstrong's introspection.

A famous example presented by him to pinpoint the phenomenon is that of the lorry-driver (Armstrong, 1980, pp.59-60). Imagine someone driving a lorry for a long period of time; it would be fairly common for them to catch themselves distracted. Due to the repetitiveness and relative simplicity of the task at hand, after the journey, the driver is likely to notice not being able to recall every single moment of their drive. However, it is clear that they were enjoying the other two kinds of consciousness distinguished from introspection. It is otherwise mysterious how the lorry-driver could have been able to drive successfully and without causing an accident if lacking perceptual consciousness (Armstrong, 1980, p.60). They must have had some sort of awareness to be able to successfully navigate their surroundings. It goes without saying that they were also minimally conscious since they were capable of having mental states. So, why is the lorry-driver not able to recall some bits of their journey?

What they are missing is the third kind of consciousness Armstrong delineated. Even though the driver had a mental life and was aware of the

world surrounding them, they were not monitoring it or other aspects of their mental life. Their mind was sort of off just focusing on the task at hand. This is the state I described as similar to an 'auto-pilot' mode when replying to Schwitzgebel. That sort of state we enter when doing something that does not require our attention so that we can focus on nothing in particular. Armstrong takes the lorry-driver as a good reason to think that there is a third kind of consciousness that cannot be reduced to the first two. This is his reason to believe there is a need for introspective awareness (*ibidem*): the sort of awareness that looks inward to make us aware of our own mental states and processes. Namely, a function that serves the same role introspection usually is given, solidifying that Armstrong is simply using another name for it.

The process described here follows all of Schwitzgebel's conditions but the last one: the introspective process is characterised by Armstrong as automatic and effortless. A subject does not need to do anything in particular to enter a specific disposition to introspect their mental states or processes. Although this is a faculty that is sometimes turned off, like when our mind is distracted like for the lorry-driver, whenever it is working we do not need to do anything more than to be aware of our mental life. This is similar to how we do not need to do anything in particular to feel our body, we can just feel it. Both phenomena can be linked back to the way Barrett (2017, pp.57-60) says the brain examines the affect coming from our body to make a prediction and formulate what it is best to do. Similarly to the process of interoception Barrett describes,

proprioception and Armstrong's introspective consciousness too continue monitoring our body and mind respectively. However, their results are brought in evidence only when they become relevant for what we need at the particular moment. Surely, some effort is needed to focus on a specific part of it, but that is only when we want to focus on a specific mental state or process: the general phenomenon requires no effort on our behalf, we are simply aware of our mental life.

Regarding the other conditions of Schwitzgebel's... Introspective consciousness is clearly something concerned with the mental (A), and limited to the subject's own mental states and processes (B). Its scope respects the temporal proximity condition since, without relying on memory, we would not be able to be aware of our mental life of yesterday (C) just like I cannot use proprioception to recall where my leg was yesterday. It is not a coincidence that Armstrong (1968, p.324) compares it to a self-scanning device: only present mental states and events can be introspected by a subject. The remaining two conditions are also adopted since Armstrong describes this sort of awareness as a simple flow of information about tokens of our mental activity (D and E) (1968, pp.326-327). So, the only missing one is condition F. This is the main difference with Lycan's own characterisation of introspection. In fact, for him introspection is not automatic, we do need to do something in order to introspect a mental state or process. I am going to show why after presenting his account.

Lycan compares introspection to a scanner too (1987, p.72) and talks about it as a perception-like faculty (1996, p.14). To be more precise, he talks of it as a second-order representing of our psychological states and processes (*ibidem*). Namely, it is a faculty that individuates a lower-order mental state or process in our mind and transforms it into a higher-order one. I shall make an example here, imagine I am wondering whether I desire becoming famous or not... To find an answer to this question introspection would need to locate my first-order desire to be famous, or lack thereof, by focusing on it and formulate the second-order belief 'I (do not) have the desire to be famous'. As Lycan draws from Armstrong, he also believes the aim of introspection would be to refine one's mental processes in order to accomplish more complicated actions (Armstrong, 1980, p.65 in Lycan, 1996, p.14). Imagine the process of looking for your phone in your room. You do not just randomly go searching from place to place hoping to simply bump into an object you would recognise as your phone. No, first and foremost you have the mental image of your phone clear in your mind so that you do not need to waste time observing those objects that do not match it in shape, size or colour. Also, there is definitely going to be a method in your search: you will be thinking of those places you are more likely to have it in and start from those. So, it is clear how such an action requires a constant afflux of information from different sources. This requires the subject to be aware of their surroundings as well as to look inside and figure out their beliefs and opinions (e.g. about possible places the phone could be in, other options for finding it...). Without something to actively coordinate your

activities in organising what is necessary at the moment and retrieve the correct information, finding your phone would be a much slower and more challenging task.

For this reason, unlike Armstrong, Lycan does not qualify introspection as constant, effortless, or automatic. For him introspection is not a type of awareness we most often have on without any action required on our part, rather his introspection is more similar to what we generally call 'attention'. Lycan's introspection consists in the activation of attention mechanisms in our mind to focus on a particular state or process and create a higher-order one as described in the previous paragraph.[20] So, to be introspecting something, a subject needs an attention mechanism in action that 'scans' it and reports it to the subject's mind (1987, p.72). Similar to Armstrong's lorry-driver example, Lycan also provides an example of these attention mechanisms in action (1996, pp.16-17). Look at whatever is in front of you right now. Notice how you are capable of shifting the focus of your attention without doing anything with your eyes (*ibidem*). Even while looking at the monitor of your computer, you are able to draw your attention to the keyboard and start forming beliefs and all sort of other mental states about it. For the cause of this shift of focus we should rule out vision. After all, you did not do anything with your eyes, its sense

[20] In one of the two texts by Lycan I am referencing, he uses the word 'consciousness' more than 'introspection'. However, he has previously clarified in his discussion of the concepts of consciousness that there is multiple ways to interpret what philosophers mean with the former word. The instance of it he intends to discuss is mental/process consciousness (1996, pp. 3, 13), namely the awareness of our own mental states and processes, which is what we would simply identify with introspection.

organ. The change happens because your mind is focusing on something different than what it was focusing on before. For this reason Lycan wants to say that this is a case where we can see introspection in action (*ibidem*): what part of our perceptual experience is being observed is dictated by what our mind is focusing on.

In this example of Lycan's you would be described to have all three types of Armstrong's awareness. You have mental states and are aware of your surroundings by stipulation. So, the conditions for the two most common types of consciousness, the minimal sort and the perceptual one respectively, are satisfied. In addition, we would say that you are aware of these aspects of your mental life since Lycan describes you focusing on different parts of it. Then, even the conditions for introspective awareness are satisfied. But Lycan's view needs something more, namely it needs for you to focus on those specific mental states to say that you are introspecting them. It would not be enough to have it that you are generally monitoring the inner workings of your mind. For you to be aware of a mental state in this model, your attentions needs to be focused on it. This means that Lycan's model is more restrictive in terms of how many items of our mental life we can introspect at once. Whilst Armstrong would say that any element currently present in our mind is being introspected since they satisfy introspective awareness, we can only have our attention on a limited subset of those.

I prefer this second model to Armstrong's. The main reason for this is that his seems to better match a general idea of introspection. Phenomenally speaking, it seems to me that whenever we require some information about some mental event or process of ours, we do need a moment of thought to 'access it'. The process is extremely fast, but it is not instantaneous like awareness is usually taken to be. This seems more in line with an active model of introspection than with a passive one like Armstrong's. Additionally, I find Lycan's introspection more credible because of its more restricting quality. Imagine I asked your opinion on flat earth theories, the importance of fashion, and whether euthanasia is ethical or not. Arguably, you would need to ponder on each of these separately rather than being able to quickly provide an answer to them in succession. However, Armstrong would say they are already introspectively conscious to you if we assume they are things you have thought about before. You might need to articulate your thoughts in three different parts, but they should quickly become available to you. Rather, I would say you need to focus on these individual ones to be able to form/retrieve your opinion on them and answer my question. The reason for this is that you need to recall through memory the relevant beliefs, and combine them into an articulated opinion.[21] Armstrong's model for introspection would have you able to report your opinion on them all together, and not in quick succession.

[21] Here, I am not in a position to explicate when exactly introspection comes into play since that depends on the account. For Armstrong the recalling and opinion forming would both be the result of introspection. On the other hand, Lycan would argue that the recalling itself would not

Lycan's introspection does not only meet condition F, but the others too. From stipulation, it is described as starting from some lower-order mental state or process (1996, p.14). Thus, its target is only the mental as prescribed by condition A. At the same time, Lycan characterises introspection as a *self*-scanning device of the mind (*ibidem*), which makes it qualify for condition B. Similarly to Schwitzgebel's condition C (2014, section 1.1), the attention mechanism he takes to be introspection is limited to ongoing mental states and phenomena so that we would need to rely on memory to firstly recall something about the past. Introspection is also described not to involve any convoluted process involving external factors; it simply focuses on some lower-order mental state or process to elaborate and produce a higher-order one just like in Armstrong's version. This makes it meet condition D. Finally, condition E is met too since the result of a scanning process is causally connected to its input, but not ontologically dependent on it. The two are distinct entities. So, for the reasons of the previous paragraph, and the fact that Lycan meets all of Schwitzgebel's conditions, I think his account should be preferred. Thus, this will be the one considered for the rest of the paper when talking of introspection.

Now that I have clarified what I mean by 'introspection' and provided reasons to believe this is a phenomenon we can rely on, I can move on to talking about its interaction with self-deception. As a reminder, the issue is

---

be introspective. For him introspection would be limited to the combination of the beliefs and articulation of an opinion.

answering how we can deceive ourselves if we are able to introspect our mental states like our beliefs. Lycan's account might seem easy to work with. In fact, with a more active model where introspection requires our attention, it seems more likely that S's suppressed belief that B might slip unnoticed by introspection. However, it is hard to accept that whenever S is pondering about something where whether B or ¬B is relevant, the belief that B will not be introspected. Surely, this might be the case sometimes, but saying that this never happens to them seems a hard position to commit to. I have talked about how Lycan's believes introspection to be aimed at refining our mental life in order to accomplish more complicated actions (Lycan, 1996, p.14). This means that in deciding how to act in a determinate scenario, it should eventually evaluate all the mental states and process relevant for it. Saying that it regularly fails to scan the belief that B is too convenient.

Although it is true that avoidance behaviour works really well with something attention-based like Lycan's introspection, it is still mysterious how a self-deceiver could be so capable at avoiding a single belief in particular even if we accept that this is one that brings displeasure to the subject. If our Frank from Unfaithful Wife is planning a party to hold at his house, he might avoid inviting the person he believes Kate to be cheating on him with. He might dismiss this initial refusal of an invitation as a mistake and convince himself there would be no problem with him joining them. But, how is it possible that Frank does not once introspect his belief that Kate is cheating on him with this

man whilst pondering whether to invite him? As a relevant belief to the action in course, it seems a perfect candidate to introspect. If there is nothing in place in Frank to prevent him from recalling this belief, B should be a readily available information to him. Definitely, introspection is not a process we have full control of. This means that Frank is not able to introspect what he wants to avoid a specific belief.

3.2. Can Someone Capable of Introspection Self-Deceive?

I have already talked about the putative infallibility of introspection. In fact, the issue just presented would truly be problematic for someone who imagines introspection to always be successful in its efforts. The problem still stands though... Here we are looking at a systematic and recurring failure to introspect the relevant belief. This is not something that can be justified by merely accepting introspection to be a phenomenon that can be faulty on occasion. So, something more will be needed in order to claim that it is possible for self-deception to occur even in subjects perfectly able to introspect. Although not as pessimistic regarding introspection as Schwitzgebel, Lycan argues for its fallibility. Since he is characterising it as a (inner) sense, this allows that one "[...] might mistake and misdescribe the contents of one's own experience" (Lycan 1996, p.17). Namely, Lycan concedes that introspection can fail as a

phenomenon. This is not restricted to those instances where the subject is simply mistaken when reporting about the result of their introspection. Sometimes I might lack the language to express my own experience (e.g. try describing colours to someone who has been blind since birth), but this is not a failure to introspect. Here I am failing as a reporter, but introspection itself is successful.

On the contrary, Lycan agrees that introspection is fallible regardless of the effectiveness of reports on it. A clarification must be made here, though. When he speaks of the fallibility of introspection, he does not imply that it is possible for a subject to properly miss some mental state or phenomenon in their mind. Namely, he speaks of false positives (1996, pp.19-21), but not of false negatives. What I mean by false negative is a scenario where introspection can fail at locating its targeted lower-order mental state or process. Namely, it is an ideal type of error to allow self-deception to take place. On the other hand, a false positive is the situation where some introspection happens without anything being introspected. An infamous example in philosophy for this would be that of the phantom limb feeling some amputees have in the part of their body that is missing. Additionally, the mistakes Lycan seems to allow are those that happen when the higher-order state or process is formulated.

Still, I do not think the theoretical possibility of false negatives in a view of introspection is particularly problematic. Once again, false negatives are something that happens in cases of sense perception too. When you fail to see

the person dressed as a monkey walking among some basketball players, nobody would say the light reflecting from this person did not bounce to your eyes correctly. You did see it in that sense. What you did not do is see it as in noticing their presence. Without getting into a conversation on the role of mental faculties in sense perception, I think I can safely say the error was in the part of vision coming after the eyes. Which show that there is instances of false negatives in senses different from introspection too.

However, showing that this theoretical possibility would not be problematic for Lycan's view does not do much to show that false negatives actually happen. When we tend to think of cases where me misuse introspection, we usually do not seem to think that introspection might fail to individuate some mental state or process in the subject's mind. We think the mistake must be located later when the subject has to characterise its content. There S can make a mistake and formulate an erroneous higher-order mental state or phenomenon. For instance, I might mistake my feeling of anger towards crooked pictures that appeared in a dream for something that really happened to me yesterday. Therefore, I would be wrong in claiming to have felt anger towards crooked pictures yesterday because I am characterising my experience incorrectly. The introspection yielded the result I was looking for; no mistake was made there. For this reason, though, self-deception still has not been saved from introspection yet. There is no way around it: a false negative is exactly what deception needs. The possibility of false positives makes space for

situations where S can fail to introspect B. However, from the scenario here described, it seems introspection would eventually report both B and ¬B. It is not the case that S merely mischaracterises their belief that B to self-deceive into believing that ¬B. More work needs to be done to explain this phenomenon.

I believe the key to this can be found in Nisbett and Wilson's work (1977). There they show a multitude of examples where someone introspecting is unaware of relevant and major aspects of their own introspective processes. I will take one of the experiments they report to show how these can be useful. The one I find most relevant for the discussion at hand is an experiment conducted by Nisbett and Schachter. In this, subjects were exposed to an increasingly intense series of electric shocks (Nisbett and Wilson, 1977, p.237). Some of the subjects were given a placebo pill beforehand that would supposedly produce those effects commonly associated to electrical shocks (e.g. heart palpitations, breathing irregularities, etc...) (*ibidem*). The results showed the subject that took the pill were able to tolerate four times the amperage of shocks. The putative reason for this would be that they attributed the initial symptoms to the pill they had taken, which would lead them to tolerate them more. However, this did not show in the following interviews (*ibidem*). In fact, when asked why their tolerance was higher than others among other questions, the vast majority of the participants of the experiment that had taken the pill denied having thought about the pill or attributing it the fault for some

of the symptoms. Rather, they came up with some excuses or were unable to answer why they had been able to resist for longer than others (*ibidem*).

This shows that the participants were not aware of the reasoning they had undergone during the experiment. There is virtually no reason to doubt that the belief that some of the effects they were experiencing were caused by the pill and not the electrical shocks played a role in their resistance to said shocks. After all, it seems too much of a coincidence that specifically these people lasted longer in the experiment (it should be noted that this is not the only experiment of this sort with this kind of results). So, since we know the pill was a placebo, it is clear that it was the belief they held regarding its effectiveness that played a causal role in making them more tolerant to the shocks. However, this is something the subjects were completely unaware of since they denied thinking about the pill. Nisbett and Wilson do not discuss this, but the belief attributing the fault of the stimuli to the pill was most definitely not an occurrent one. In fact, as some of the patients even claimed, they were "[...] too busy worrying about the shock" (*ibidem*). However, I have already explained how a non-occurrent belief may cause some behaviour of the subject. The part that is of interest for our discussion comes when the subjects were asked to introspect about their own experience. There is no reason to believe the subject able to resist the shocks for longer that did not claim to have thought about the pill while being electrocuted might be under self-deception. The subjects did not form the belief that the pill had any effect on their

resistance to the experiment, they had simply not thought about it. So, none of the typical phenomena related to self-deception could have taken place. However, this is still a case where some relevant and important fact is overlooked by introspection. This experiment provides us with a case of a false negative. As Nisbett and Wilson say, the subjects were aware of the result of their thinking, but were not able to reconstruct its process (1977, p.232). The subjects were perfectly capable of understanding their level of pain and whether it was too much to endure. However, they could not provide reasons for why they were able to endure it, even in a more relaxed moment of self-reflection after the experiment.

As a disclaimer I should note that Nisbett and Wilson in their paper (1977) sustain a model of introspection closer to Armstrong's since they talk of introspection as awareness. However, in presenting it, I have already adapted the language they use to fit introspection as Lycan describes it. For example, I talked of attention rather than awareness. I do not believe this changes in any way the result of the experiment, since the two models are still fairly similar, and whether introspection is taken to be more active or passive does not play any role here. In fact, the subjects were interviewed in a moment of calm when they were able to think of the content of their minds during the experiment and the reasons for their behaviour (Nisbett and Wilson, 1977, p.232). Thus, it is not the case that the shocks were capable of distracting them from thinking about the pill and act as a more straightforward obstacle for introspection. So, it is not

the case that modifying the model of introspection used changes significant portions of the results of the experiment or its implication. There is something else that is shown by noticing that the subjects failed to properly introspect even when they were in a condition of relative calm (i.e. during the interview with no electric shock going through their body). Namely, even in a case where there is nothing thwarting their attempts to introspect, they fail to locate a certain mental process (i.e. the reasoning they made during the experiment). So, this shows that a false negative has been found in the framework of the adopted view of introspection.

This can be used in defence of self-deception since it makes space for the non-occurrence of B to S even when they are wondering whether B or ¬B is the case. This is equivalent to the subjects not being able to attribute their higher endurance to the pill. Additionally, it should be noted that, whereas the reason for the subject's resistance to a higher amperage of the shock experiments is unclear, in self-deception cases there is a desire motivating them to believe that ¬B. At a subconscious level this might come into play when determining whether the subject believes that B or ¬B. If Frank is considering whether he should ask Kate why her shirt smells like another person's perfume, memory would most likely provide him with both information B and ¬B. However, his desire D would prevent him from actually introspecting B and forming the

belief that this is also something he believes.[22] Together with the avoidance behaviour and Mele's bias as explained in the previous chapter, I have given reason for the robust endurance self-deception typically displays. Such combination prevents S from realising their self-deceived state. However, it also shows the fragility of self-deception the moment S is able to identify D as the cause for their belief that ¬B. Once that is clear to S, the desire is not really able to interfere with S's introspective process, which will lead them to realise they are holding both beliefs that B and that ¬B. With good conditions, S will be able to recognise the bias they have been adopting in evaluating the data available to them and discard ¬B as unwarranted by the available data.

Before moving to a final worry that might surge from what has been said so far, I think this is the moment to put together everything I have said so far to clarify the whole process. In fact, I have noticed that, usually, to present a scenario involving a self-deceived subject as objectively as possible, this is described starting in medias res. For this reason, the sequence of events that lead the subject S to be self-deceived about the belief that ¬B might be hard to grasp. My aim here is to illustrate and clarify the whole picture to see its interactions with the other processes that have been discussed so far and show how these tie together. For consistency, I am going to use the example of Unfaithful Wife where Frank maintains the belief that Kate is faithful to him in

[22] The details of how introspection fails in this specific case of self-deception is just a speculation of mine. I am not a psychologist and cannot make a definite case-study off of a fictional scenario. Here I only want to make a hypothesis explaining how I think the desire D could come into play in Henry's failure to introspect.

spite of the evidence he has had the chance to collect that point to her having an affair because of his desire of not going through a divorce.

Now, it is important to clarify that when I say that Frank must already have some evidence regarding whether B or ¬B, I am not thinking of a fully-fledged argument or even a list of points. Anything that could be considered pointing towards B rather than ¬B would work for this point of discussion. Often, even when we have never properly thought about something we do not need to consider it before giving what is usually called a "gut reaction". If, out of the blue, you asked me whether I would want to go shopping tomorrow I would immediately reply "Absolutely no" because of my distaste for it, but upon consideration I might realise I do need to buy some clothes and change my mind. Similarly Frank, when asked whether it could be the case their partner is cheating on them might quickly reply "Maybe..." because he has possibly noticed some behaviour of their partner hinting that B might be the case and has not yet thought of the implications. What I want to say with this example is that the early stage of self-deception needs not be a stage where S has already been deep in thought about whether B is the case as often seems in self-deception scenarios.

So, what matters here is that Frank has a belief, regardless of his confidence in it, that B is the case. S does not even need to have realised they hold the belief that B, as long as they have it. There is a reason why I believe that this step is needed. Allegedly, if the scenario had S form the belief that ¬B

without having a prior belief that B, we could still say that S is forming a belief *in spite* of the evidence available to them. We could even say that they are motivated by a desire in doing so and have all the other relevant conditions for self-deception be met. However, I would argue that this would not qualify such a scenario as one of self-deception. It could be called "motivationally jumping to the wrong conclusions" or "irrational belief formation". What it is lacking is the deception part of self-deception. These cases are definitely worth looking into for a reflection on irrational action, but what sets self-deception apart from similar phenomena is that S is convincing themselves of something that deep down they know is not true. So, we need S to believe that B is the case.

It is at this point that the other mental state necessary for self-deception is formed in Frank: the desire D that ¬B were the case. This moment is probably going to be when he gives some thought about the consequence each scenario would have and realises ¬B to be the more desirable one for them. This needs not be the case as in a scenario where wishful thinking turns into self-deception as described at the beginning of the second chapter. However, until this point, the ordering of each of these steps is not as important as it will be later. All we need is for S to have a belief that B as per the evidence available to S and a desire that ¬B. This is the point where the confusion between the two different types of satisfactions described by Hubbs come into play. Whereas the satisfaction S is feeling contemplating a scenario where ¬B is the case is a thumotic satisfaction, S mistakenly believes it to be an epistemic satisfaction.

Here, S mistakes the satisfaction from imagining something pleasurable with the satisfaction of finding out a truth.

Something to point out here is that it is possible that a subject already had the belief that ¬B before they started gathering evidence that B. Frank's case is a good example of such a case. We can imagine that Kate has not always been unfaithful to him. So, there was a time when Frank's belief that she was not cheating on him was both rational and correct. For obvious reasons, this is not yet a case for self-deception. The problems start when the evidence favouring the belief that B started being more than the one in favour of his precedent belief that ¬B. In fact, this is the moment when the latter belief becomes irrational as it does not conform to the available evidence. Even then, though, we would only call Frank a self-deceiver if he did not adjust his beliefs. There might be a time when Frank might not have yet thought about whether Kate is faithful to him or not even after receiving evidence on this topic, he is not self-deceived then. It is only in the moment that Frank decides to keep believing that she is not cheating on him that he becomes self-deceived.

So, to go back to the more general discourse, since Frank is now convinced they have figured out something new, they form the belief that ¬B. Here it is important that the subject is not merely revaluating the evidence to come to a different conclusion. They are not changing their mind. He, in fact, still believes B even when forming the belief that ¬B. Not only we want the subject to hold both beliefs at the same time, but we also want this confusion

between the two kinds of satisfaction to be a non-deviant cause for his self-deception. Both of these are required for the phenomena to be recognised as self-deception. If all of this applies, we can consider S a subject of self-deception. However, for a full description of the process, I cannot stop here. Self-deception can be a long lasting phenomenon. Then, I need to explain how both beliefs can be retained to clarify why S remains self-deceived.

At this stage the belief that B is merely dormant in a dispositional sense. So it might become occurrent in certain occasion but will not usually be active in Frank's mind. On the other hand, the belief that ¬B will be the occurrent one every time it is relevant. Here is where the work on the unreliability of introspection helps us make sense of this scenario. It is generally agreed that we are not constantly aware of all of our beliefs... If we were, S would figure out they hold two opposite beliefs and discard either of them, exiting the self-deceived state. But this sort of realisation does not occur so easily.

Mele explains why this is the case. In fact, this is no ordinary situation where someone rarely thinks about a certain topic (e.g. that the freezing temperature of water is 0 °C) and, thus, rarely entertains their own belief about it. Rather, Frank might often ponder about Kate's faithfulness and yet fail to realise his self-deceived condition. This is because of the phenomena I have called Mele's bias: negative and positive misinterpretation, selecting focusing/attending, and selective evidence-gathering. These provide an explanation of why Frank's belief that B will hardly ever be occurrent. He is, in

fact, not going to be objective in his analysis of further events relating to the self-deceived belief and, thus, make the possibility for a defeater to the belief that ¬B unlikely to arise. At the same time he is going to display the typical avoidance behaviour towards this topic that is common in self-deceived subjects. Since, deep down, he does have the belief that B, he is going to avoid as much as possible considering that scenario since it is less pleasant to him.

Her, it is important to notice how introspection works. As I have previously explained following Lycan's model, this can be imagined as attention mechanisms that focus on specific mental states or processes to bring them to the forefront of the subject's mind and, thus, making them aware of these. This is what happens with the belief that ¬B. Whenever the question arises, S scans their mind and focuses their attention on this specific belief to have the thought that ¬B is the case. Now, there is more senses we can be said to be conscious of our mental life. Following Armstrong's terminology, here I am talking of what is closest to introspective consciousness. This is a more active way of using the attention mechanisms. In the field of attention this could be considered a top-down approach where the subject, from "above" scans for a specific item in their mind and brings it to their attention. However, it is generally accepted that we do experience another kind of consciousness of our mental life, Armstrong's perceptual consciousness. Namely, the type that the lorry-driver enjoys while driving without paying too much attention to the task at hand. I believe this to be more of a bottom-up approach, where it is the

different stimuli that attract the subject's attention. This makes them more passive to their surrounding, which explains why most of it is not as easily retained. This could be so that we are free to concentrate on something else. After all, according to Lycan, introspection is aimed at allowing us to perform more complicated and refined tasks. Otherwise, we go in this sort of auto-pilot mode when there is no need for us to focus on the task at hand since it requires no further developing.

I believe this differentiation between types of attention can be used to explain S's behaviour after self-deceiving themselves into believing that ¬B. Mele's biases show why it is possible that the subject does not come to realise their self-deceived state, but this does not yet explain a characteristic trait of self-deceived subjects: their conflicting behaviours. For example, whereas S will mostly be observed acting in accordance with the belief that ¬B (e.g. not objecting to their partner going on a trip with a friend), they will be occasionally caught displaying the opposite type of behaviour (e.g. obsessively checking social media to monitor said trip). So, when S is aware of the pertinence of their action to the question of whether B or ¬B, they can be described to follow a top-down approach to the introspection of their beliefs and, consequently, the actions these cause. On the other hand, when distracted, the subject can be described to follow a bottom-up approach that may display beliefs that are usually suppressed by S and cause a type of behaviour opposite to their usual.

3.3. Intentions in a Motivationalist Framework

Now that I have presented the role of introspection within self-deception, I can answer another question I have been putting on hold since the introduction. Is it possible, with the motivationalist account I have been arguing for, to have a case where someone intentionally deceives themselves? Take Claude for example... He is about to compete in a match and decides to psych himself up before its start. Nothing uncommon here, it is just the usual "You can do this. You are going to win." Imagine that all the conditions for self-deception are met. So, he actually started with the belief that he was not going to make it, but after psyching himself up, he formed the belief that he can make it. We postulate that Claude has been really tired from sleep deprivation. This means that his performance will be far from stellar, and he is aware of this. Thus, Claude will have more reasons to believe that he is not going to make it rather than that he will. However, he treats this data in a nondeviant biased way by convincing himself that it is not a particularly bad scenario, and that it will barely have an effect on the match. Finally, we imagine that Claude's reason to do this is that he has heard that this procedure often has actual effect on people in his situations. Namely, believing that you are going to make it does actually have a series of psychophysical effects on you that do increase your chance of making it. Thus, this creates a confusion in Claude between the thumotic pleasure of thinking he is going to make it and the epistemic pleasure of having

found out that ¬B (i.e. 'I am going to make it') is a true statement. All in all, this seems a case of straightforward self-deception: the subject starts believing something they would like to be true in spite of what the evidence says they should believe.

The difference is that Claude is intentionally deceiving himself since he has made a conscious decision to start this psyching-up process. As a consequence, this means he will be aware of what he is doing while psyching himself up. To characterise what is happening here, I have to differentiate two scenarios. In the first one, Claude retains his intention during the whole process. I would argue that in this scenario, Claude's attempt is going to fail. In favour of this belief of mine I can bring Mele's dynamic paradox, the one I had no issue with for my account as it did not involve intentions. Here it is relevant since these are being brought back into the picture. As Mele would point out (1997, pp.98-99), in this thought-experiment, Claude is intentionally trying to deceive himself. Thus, he is in a position where he is aware of the deception taking place, which should make it impossible for him to fall for it. How can he fall for his own deception if he is aware of it? He cannot trick himself into forming the belief that he will win whilst he is consciously thinking of his scheme. Unless the intentionalists find an answer for the dynamic paradox that cannot be disputed, scenarios of deception like this one should not be pursued.

This leads us to the second scenario, here Claude at some point forgets of his initial intention to deceive himself. We can imagine him so drawn into the

process of psyching himself up that he does not bother remembering why he had started it. This means that his original intention is not present in the forefront of his mind anymore. As shown in defence of my argument with the experiments analysed by Nisbett and Wilson, it maybe be possible that a subject fails to introspect some parts of their mental states and processes even when these are relevant for the situation they are in. To grant this, we clearly need to grant that an intention can cause behaviour in a subject even when it is not occurrent just like I have argued for dispositional beliefs. Otherwise, Claude's self-deception would stop since the cause that made him enter his self-deceived state is not present anymore. I do not see how Claude could intentionally deceive himself after having lost the intention to do so since this was its cause. If self-deception should be understood as an intentional action like intentionalists want us to believe, it seems that it would be needed for the intention to remain the cause of the act of deceiving himself. If not, we could argue that he interrupted his self-deception to do something an intentionalist would not recognise as self-deception with their model.

Therefore, the intentionalist must bite the bullet here and argue that Claude retains his self-deceiving intention. Simply, this is not present at the forefront of his mind because it is dispositional just like his suppressed belief that he will not make it. However, if we make such a move, I think we end up blurring the lines between desires and intentions too much. Usually, the difference highlighted between these two mental states is that, unlike desires,

intentions are more committing towards their end goal (Setiya, 2018, section 1). Namely, they usually involve some planning and serious consideration of the subject to reach the intended result. However, desires here have been presented as capable of directly causing behaviour. Such behaviour is efficient and overall consistent in making the subject act in a certain way, which is why self-deceived subjects have been described as generally successful in displaying behaviour in accordance to their belief that ¬B. If we concede that intentions can be dispositional and still be called intentions while never emerging as occurrent to the subject, we seem to end up with something very close to a desire.

At this point, I should say that whether my account classifies as 'motivationalist' or 'intentionalist' does not matter much. I have described it as the former since I consider it more similar to this kind of view. However, if the main distinction between the two accounts (i.e. whether desires or intentions should be blamed for self-deception) is so blurred, I do not see a reason to hold this separation. Thus, I am willing to include these modified intentions among the possible cause for a state of self-deception if the intentionalists decide to commit to the concept that intentions can be dispositional in a way that S is mostly unaware of their ongoing self-deception. This can happen with my account since I have already moved closer to intentionalist views by holding that we can ascribe to the self-deceived subject both beliefs that B and that ¬B at all time. The desire/intention dispute was the most noticeable demarcation between the views. But, if desires can be robust and intentions can be

suppressed, I do not see further reasons to say this weaker version of the intentionalist account could work. I would still say my view is to be considered motivationalist since it is closer to that branch of models for self-deception rather than the intentionalist one for how the two were originally constructed.

4. Conclusion

It is now time to finish this dissertation. I have, in fact, reached my original aims. In the second chapter, I have introduced and explained the concept of self-deception. Starting from Mele's account, I have characterised the most interesting peculiarities of this phenomenon in application to the scenarios described in Unfaithful Wife and Faithful Wife. I have argued that these are necessarily instances of self-deception as other phenomena fail to characterise or include them correctly. Using Mele's characterisation of self-deception as the archetype of a motivationalist account, I have explained how an intentionalist account would interpret these paradigm cases of self-deception. In this way I have explicated the two main differences between the two kinds of account: the intention/desire distinction, and how intentionalist require the subject to hold opposite beliefs whereas motivationalist need the desired belief to be false. I have discussed, then, which account I think should be preferred by introducing

the paradoxes Mele raises against intentionalists (1997, pp.93, 98-99), and some further motives to prefer a motivationalist account.

Subsequently, I have introduced Hubbs's proposition for a characterisation of self-deception and used it to present my full account. I have addressed the parts where my opinion differed from Mele's and explained why I believe my view does not fall for his static paradox although it moves closer to an intentionalist account. In response, I have presented reasons in favour of my account against Mele's. Finally, I concluded the first section by presenting some final worries one could have with self-deception as characterised in this paper. Most notably, I have explained how my view can appeal to an answer similar to the intentionalists' described by Mele to solve the problem with twisted cases of self-deception.

In the third chapter, I moved to the topic of introspection. Using Schwitzgebel's guidelines for what are generally considered the traits a theory of introspection should follow, I have presented Armstrong's and Lycan's accounts. Although sharing many similarities, I have clarified why I think Lycan's should be preferred considering how it, phenomenally, better corresponds to our experience of introspection and its following all six of Schwitzgebel's conditions. Then, I have used this theory of introspection as a model to see whether self-deception is truly possible for creatures that can introspect like us. Thanks to Nisbett and Wilson's (1977) discussion of several psychological studies, I was able to create a breach in the infallibility of

introspection big enough to justify the existence of self-deception. However, I have defended it from the accusation of being too unreliable to be useful. My proposition is that subjects might sometimes fail to introspect certain mental states or processes of theirs even when this is relevant for the situation they are in. This, tied together with the generally acknowledged view that beliefs can be dispositional, grounded the possibility for self-deception.

In the final section, I debated whether in a motivationalist framework, intentions can be the cause for self-deception like an intentionalist would argue. To this I have generally answered negatively with the exception of a scenario where what is usually considered characteristic of intentions (i.e. the recurrent awareness the subject has of them and their usually committing nature to the end goal) is lost. In this particular scenario I question what is left to distinguish an intention from a desire. As such, if the premises for this scenario are met, I have concluded that these intentions can be the cause of self-deception. More interesting conversations can spark in discussing how these two topics interact. However, this will have to wait for another time. For now, I have achieved all I had set out to do in this dissertation.

Bibliography

- Armstrong, D. M. (1968). *A Materialist Theory of the Mind*. London: Routledge.
- Armstrong, D. M. (1980). 'What is Consciousness' in *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press, pp.55-67.
- Barrett, L. F. (2017). 'The Origin of Feeling' in *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.
- Bermudez, J. L. (1997). 'Defending Intentionalist Accounts of Self-Deception'. *Behavioural and Brain Sciences*, vol. 20 (1), pp.107-108.
- Brons, L. J. (2019). 'Aphantasia, SDAM, and Episodic Memory'. *Annals of the Japan Association for Philosophy of Science*, vol.28, pp.9-32.
- Chignell, A. (2018). 'The Ethics of Belief'. Last accessed 18th of March 2021. Available at: https://plato.stanford.edu/entries/ethics-belief/
- Deweese-Boyd, I. (2016). 'Self-Deception'. Last accessed on the 22nd of August 2019. Available at: https://plato.stanford.edu/entries/self-deception/
- Fernandez, J. (2011). 'Self-Deception and Self-Knowledge'. *Philosophical Studies*, vol. 162, pp.379-400.
- Funkhouser, E. (2005). 'Do the Self-Deceived Get What They Want?'. *Philosophical Quarterly*, vol. 86, pp.295-312.

- Gardner, S. (1993). 'Ordinary Irrationality' in *Irrationality and the Philosophy of Psychoanalysis*. Cambridge: Cambridge University Press, pp.15-39.
- Hubbs, G. (2018). 'Self-Deceptive Resistance to Self-Knowledge'. *Les ateliers de l'éthique*, vol. 13 (2), pp.25-47.
- Lazar, A. (1999). 'Deceiving Oneself Or Self-Deceived? On the Formation of Beliefs "Under the Influence"' in *Mind*, vol.108 (430), pp,265-290.
- Lycan, W. G. (1987). *Consciousness*. Cambridge, MA: MIT.
- Lycan, W. G. (1996). *Consciousness and Experience*. Cambridge, MA: MIT.
- Mahon, J. E. (2015). 'The Definition of Lying and Deception'. Last Accessed on the 22nd of August 2019. Available at: https://plato.stanford.edu/entries/lying-definition/
- Mele, A. R. (1987). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford University Press
- Mele, A. R. (1997). 'Real Self-Deception'. *Behavioural and Brain Sciences*, vol. 20 (1), pp.91-102)
- Mele, A. R. (1999). 'Twisted Self-Deception'. *Philosophical Psychology*, vol. 12 (2), pp.117-137.
- Nelkin, D. (2002). 'Self-Deception, Motivation, and the Desire to Believe'. *Pacific Philosophical Quarterly*, vol. 83, pp.384-406.

- Nisbet, R. E. and Wilson, T. D. (1977). 'Telling More than We Can Know: Verbal Reports on Mental Processes'. *Psychological Review*, vol. 84, pp.231-259.
- Schwitzgebel, E. (2008). 'The Unreliability of Naive Introspection'. *The Philosophical Review*, vol. 117 (2), pp.245-273.
- Schwitzgebel, E. (2012). 'Self-Ignorance'. *Consciousness and the Self*. Cambridge University Press.
- Schwitzgebel, E. (2014). 'Introspection'. Last accessed on the 22nd of August 2019. Available at: https://plato.stanford.edu/entries/introspection/
- Schwitzgebel, E. (2019). 'Belief'. Last accessed on the 22nd of August 2019. Available at: https://plato.stanford.edu/entries/belief/
- Setiya, K. (2018). 'Intention'. Last accessed on the 22nd of August 2019. Available at: https://plato.stanford.edu/entries/intention/
- Szabados, B. (1973). 'Wishful Thinking and Self-Deception' in *Analysis*. Vol. 33 (6), pp.201-205.
- Thomas, N. J. T. (2014). 'Mental Imagery'. Last accessed on the 17th of April 2021. Available at: https://plato.stanford.edu/entries/mental-imagery/
- Voltaire, F. M. A. (1773). *Essais sur les moeurs et l'esprit des nations*. Electronic copy last accessed on the 22nd of June 2021. Available at: https://archive.org/details/essaissurlesmoeu03volt/page/n7/mode/2up
- Waltz, S. (2011). 'The Philosophical Significance of Attention' in *Philosophy Compass*. Vol. 6 (10), pp.722-733.

- Zeece, M. (2020). 'Flavors' in *Introduction to the Chemistry of Food*. Elsevier.