Fixing Belief

Hiu Chuk Winnie Sung

UCL

PhD Philosophy

Declaration

I, Hiu Chuk Winnie Sung, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis is concerned with self-ascriptive belief. I argue that one's lower-order belief can be fixed from the reflective level. One reasons about whether *p* is the case and it is on the basis of one's endorsement of *p* that one comes to believe *p*. I argue that one's self-ascriptive belief can also be fixed from the reflective level. One reasons about whether *p* is the case and it is on the basis of one's endorsement of *p* that one comes to self-ascribe the belief *p*. I further suggest that it is possible for the reflective way of fixing lower-order belief to fail but the reflective way of fixing self-ascriptive belief to succeed. When this happens, one is in a state of believing that she believes *p* when in fact one does not believe *p*. This suggests that the state of believing that one believes *p* and the state of believing *p* are distinct states and that the state of believing that one believes *p* does not necessitate the state of believing *p*. It also raises a sceptical worry about whether one's self-ascriptive belief amounts to knowledge.

In Chapter 1, I situate my discussions in the existing literature, focusing on the constitutive view of self-ascriptive belief. In Chapter 2, I use an everyday case in which a subject self-ascribes the belief that *p* and is later surprised that *p* to motivate the possibility that there are different levels at which beliefs are fixed. In Chapter 3, I develop an account of ratiocination and argue that the conclusion of ratiocination is in the form of *I ought to believe p*. Hence, at the end of ratiocination, one is in a state of believing that *I ought to believe p.* In Chapter 3, I discuss how one's belief that *I ought to believe p* initiates a top-down fixation of the corresponding lower-order belief. I also discuss why it is possible for the top-down fixation process of a rational subject to terminate before it fixes the lower-order belief. In Chapter 4, I discuss the transparency account of self-knowledge. I first criticise the transparency account's claim that a rational subject's endorsing *p* necessarily leads to believing *p*. Someone who ratiocinates and concludes that *p* but does not believe *p* because the top-down fixation process terminates early is an example of how a rational subject can endorse *p* without believing *p*. I then draw on the transparency account to argue that from a rational subject's first-person perspective, if she self-ascribes a belief to herself and if she endorses that *p*, she will self-ascribe the belief that *p*. If this is right, then one can self-ascribe the belief that *p* because one endorses *p* but in fact does not believe *p* because one's endorsement fails to fix the lower-order belief. In Chapter 5, I return to the constitutive account, explaining why its central claim should be rejected. I also reject the incorrigibility thesis, which holds that a self-ascriptive belief that *p* entails the lower-order belief that *p*. Finally, I raise a number of puzzles concerning the epistemic status of self-ascriptive belief.

Impact statement

I argue in this thesis that reasoning can fix both higher-order beliefs and lower-order beliefs. I also argue that it is possible for reasoning to fail to fix lower-order belief while succeeding in fixing higher-order belief. It is possible to have a higher-order belief without the corresponding lower-order belief. If this is correct, this thesis rejects a popular account of self-knowledge—the constitutive account—which holds that the higher-order belief and the lower-order belief cannot exist without each other. This thesis also rejects the incorrigibility thesis that holds that, necessarily, if one believes that one believes *p*, then one believes *p*. My account thus challenges the dominant picture of self-knowledge.

My thesis raises a sceptical challenge to any account that defends the epistemic security of self-knowledge of belief. That the reflectively fixed higher-order belief can be empty—it does not have a corresponding lower-order belief—presses us to explain the epistemic status of higher-order belief that is formed on the basis of lower-order belief. In light of this possibility, the prospects of providing a unified account of the nature of self-ascriptive belief are dim. But the non-unified account—treating beliefs fixed at different levels differently—also faces serious problems. One of the problems is that, from the first-person perspective, a subject cannot tell the difference between a higher-order belief that is reflectively formed and a higher-order belief that is formed on the basis of lower-order beliefs. Nor can a subject tell the difference between having a higher-order belief that has a lower-order belief embedded in it and higher-order belief without the corresponding lower-order beliefs. Even if the subject's higher-order belief is formed on the basis of a lower-order belief, she could have been wrong about what she believes. How, then, can one's self-ascription of belief amount to self-knowledge?

Since I have to explain the nature of ratiocination and the way ratiocination fixes belief, a substantial part of this thesis is on the nature of ratiocination and the top-down fixation process. Accordingly, I defend a number of controversial claims concerning the nature of reasoning and belief-formation. For example, a common assumption of accounts of theoretical reasoning is that, when one reasons about *p*, one's conclusion is in the form of *p*. I argue that reasoning is self-consciously directed, the conclusion one draws is in the form of *I ought to believe p*. This is one of the main claims I appeal to in arguing for my thesis.

Contents

Chapter 1
Introduction

In this thesis, I aim to show that it is possible for a rational subject to believe that she believes *p* when in fact she does not believe that *p*. I argue that both a subject's higher-order belief and her lower-order belief can be fixed from the reflective level, and that the fixation can occur independently. It is possible, I argue, that the lower-order belief fails to be fixed while the higher-order belief is fixed.

The possibility of a rational subject's belief that she believes *p* without her believing *p* generates a broader worry about self-knowledge, that is, whether a subject can ever tell, from her first-person perspective in making a self-ascription, if she merely has a view on her mind. This is different from the worry that one could have a mistaken self-ascriptive belief in the sense that she believes that she believes *p* when in fact one believes not-*p*. In such a case, one makes a mistake about the way one considers the world to be. However, one is at least right in thinking that she has a view on the world. In a case where the subject believes that she believes that *p,* but in fact does not have a belief about *p*, she mistakenly takes her belief about her own mental state to be her belief about the way the world is when, in fact, all she has is a belief about her own mind. I argue that, perhaps surprisingly, it is reasoning that opens up this possibility of us being mistaken about our own minds.

In the first half of this introductory chapter, I set the scope for my discussion and clarify my assumptions about first-person self-ascription of belief. In the second half of the chapter, I situate my thesis by examining two prominent approaches. I explain why accounts of self-ascriptive belief that assume a constitutive link between higher-order belief and lower-order belief cannot accommodate the possibility that a subject believes that she believes *p* but does not in fact believe *p*. Finally, I offer a sketch of the general trajectory of my argument.

1.1 Preliminaries

This thesis is only concerned with present-tense self-ascription of belief. I will sometimes use 'higher-order belief' to refer to one's belief about her own belief and 'lower-order belief' to refer to one's belief about whether a certain proposition is true. This sets my use of 'higher-order belief' and 'lower-order belief' apart from some contemporary discussions that tend to use the two terms to track beliefs that results from two different systems or levels of cognitive functioning in humans. According to this two-system view, the lower-level is fast,

parallel, tacit, and automatic, therefore not subject to control; the higher-level is slow, serial, explicit, and subject to conscious judgment and assent.[1] On this two-system view, 'lower-order' and 'higher-order' beliefs would respectively refer to beliefs that are generated by lower-level cognitive functioning and those by higher-level cognitive functioning. Frankish, for example, uses 'level 1 belief' to refer to states that operate at the lower-order level and 'level 2 belief' to refer to states at the higher-order level.[2] For Frankish, level 1 and level 2 beliefs are of different kinds, with the former akin to opinion or a commitment to use *p* as a premise in reasoning, and the latter to behavioural dispositions.

Again, this is not the way I use 'higher-order belief' and 'lower-order belief' and I remain neutral on whether there are two kinds of beliefs that are generated by different systems of cognitive functioning. Even if these two kinds of beliefs mandate different ascriptive constraints (e.g., we may need different constraints for ascribing beliefs to creatures with language and creatures without language), this distinction only applies to cases where we ascribe a belief to others. Whenever I think about what I believe—that is, the way the world is from my perspective—I am already one level above my belief about the world, regardless of which system generates this belief. It is in this sense that I mean our beliefs about our own beliefs are at a higher-order level; in addition, it is in this strict hierarchical sense that I speak of higher-order and lower-order beliefs. For ease of exposition, I will sometimes use 'B*p*' to refer to a subject's belief that *p* and 'BB*p*' to refer to a subject's belief about her own belief that *p*. To keep the usage consistent, when philosophers use 'second-order' in their discussions, I will sometimes refer to it as 'higher-order'.

By 'present self-ascription of belief', I mean an assertion or thought of the form 'I believe that *p*'. I assume that the kind of self-ascription of belief that is relevant for this thesis meets two conditions: sincerity and identification. The sincerity condition stipulates that the subject is being sincere in her self-ascription, whether it is in assertion or in thought. This condition may be formulated as follows: if the subject believes that she believes that *p*, the subject believes that it is the case that she believes that *p*. This rules out cases where one supposes that she believes that *p* so that she can use her belief that *p* as a premise in her reasoning and decision-making. We need to be particularly careful when it comes to self-ascription in thought; a subject may have many thoughts constantly running in the background of her mind—some of them fainter than others. There might be a thought that a

[1] Cf. Dennett 1978; Frankish 2001, 2004; Kahneman 2011.
[2] Frankish 2004.

person she admires has done something that disappoints her but she is not willing to acknowledge the thought because it is too uncomfortable for her to acknowledge this thought. To keep our discussion focused, I will only be concerned with the kind of thoughts that are sincerely acknowledged by the subject, such as 'She is away from work' or 'Canberra is the capital of Australia'. These thoughts are like assertions, only not spoken aloud.

I draw a distinction between assertion about oneself and self-ascription of belief. When *p* is a proposition that concerns the subject herself, it is easy to confuse a mere assertion about oneself with a self-ascription of belief. This may occur when a subject affirms a proposition about herself in a way that parallels the way she affirms propositions about others, and yet she does not hold a corresponding lower-order belief about herself. Suppose there is a kind of mental disorder M, the nature of which is that the agent who has M does not believe that she has M. A subject can be convinced by compelling medical evidence that she has M. Based on this evidence, she asserts that 'I have mental disorder M'. Given the stipulations of the case, one way to interpret it is that the subject affirms a proposition about herself third-personally but does not believe in the affirmed proposition first-personally. This is not to say that the subject's assertion about herself structurally parallels her beliefs about other people. The subjective experience she has of affirming that she herself has mental disorder M is perhaps different from her experience of affirming that her friend has M. Moreover, suppose she sincerely believes that her friend has M, she is in a position to have both the lower-order and higher-order belief that her friend has M. But since the nature of M is such that the subject who has M will not believe that she has M, even if she sincerely believes the medical evidence, she is not in a position to form the belief that she has M. The affirmed proposition, 'I have disorder M' will not bring about a lower-order belief in the proposition. It is relatively uncontroversial to say that it is possible that one makes a sincere assertion about *p* without believing that *p*. The kind of assertions that are relevant here are those that amount to self-ascriptive belief, meaning that, when a subject asserts 'I believe that it is raining', she must sincerely believe that she believes that it is raining.

The identification condition stipulates that the subject identifies with the belief she ascribes to herself. It may be formulated as follows: if a subject believes that she believes *p*, she is conscious of the belief she ascribes to herself as her own belief. This condition rules out cases where one ascribes the belief *p* to herself but is alienated from the belief that *p*.

There are at least two possible ways in which identification fails. One possible way may be labelled as 'accessibility without ownership'. This occurs when the subject is conscious of a belief but is not conscious of the belief as her own. Some sufferers of

schizophrenia, for example, are conscious of certain thoughts but describe these thoughts as being 'inserted' into them by someone else. Details of schizophrenia aside, it could be said that a subject who suffers from a similar disorder is conscious of a lower-order belief, but does not take herself to be the originator of the lower-order belief. When alienated in this way, the subject will not ascribe the lower-order belief to herself. In our consideration of standard cases of self-ascriptive belief, we assume the subject identifies with the belief that *p* as her own.

A second way in which identification could fail may be labelled as 'ownership without accessibility'. This occurs when a subject identifies a lower-order belief as her own but the lower-order belief itself does not occur in the subject's consciousness. This usually occurs when subjects come to accept certain of their unconscious lower-order beliefs in a third-personal way, for example, through testimony. An often-mentioned example in the literature is how one can be convinced by her psychotherapist that she unconsciously believes that her sibling has betrayed her, and so comes to ascribe to herself the belief that her sibling has betrayed her. Suppose it is the case that the subject has the unconscious belief that her sibling has betrayed her. Given the nature of unconscious belief—the unconscious belief cannot occur in the subject's consciousness—the subject cannot be conscious of her own belief that she is betrayed by her sibling. Hence, as Moran points out, even if the subject comes to believe that she holds the unconscious belief, her belief about her lower-order unconscious belief is formed in a way that parallels the way she forms beliefs about others. Although the subject has a sense of ownership over the lower-order belief, she does not have access to it.[3]

More needs to be said about precisely how these two kinds of beliefs are different. For present purposes, it suffices to note that such alienated self-ascriptions are possible; however, these are not the focus of my study. Accordingly, I will assume that the lower-order belief ascribed by the subject to herself must figure in the subject's consciousness. It is not enough if the subject holds the belief that *p*, yet is not conscious of the belief that *p*. The higher-order belief must be formed based on one's own conviction, not on the basis of external evidence for her beliefs.

Stipulating this identification condition is crucial for my subsequent discussion because the possibility I put forward—that one can have higher-order belief without the corresponding lower-order belief—is not a case where the subject fails to identify with a

[3] Moran 2001, p. 85.

certain belief that she ascribes to herself, but a case where a subject sincerely identifies with a lower-order belief she ascribes to herself but in fact does not have the lower-order belief.

A final remark. My present concern is how a subject comes to believe that she has a certain lower-order belief. I shall simply assume that belief can be theorised as a binary state, regardless of whether or not it is analytically, metaphysically, or normatively reducible to a graded state.[4] In order to keep our discussion focused, whenever I discuss a subject having a belief, I assume that one has the all-out belief that *p*. My argument might also have implications for self-ascription of other mental states; these, however, will not be considered here. I will only be focusing on what the account has to say about one's beliefs about one's own beliefs.

1.2 Constitutive Account

If higher-order and lower-order beliefs are distinct states, one might worry that the relation between a subject's belief about her belief and her lower-order belief is a causal one. Such an account would have us think that forming beliefs about our beliefs parallels the way we form beliefs about the external world or other people.[5]

On the Cartesian model, when we self-ascribe a belief, there is a special cognitive faculty that operates over and above an inner mental item, which gives the subject some sort of privileged access to her own psychological states. The problem with positing something like an inner perceptual faculty is that it opens the possibility of the inner perceptual faculty breaking down and as a result of that, the subject becomes 'self-blind.' A subject might believe *p* without knowing that she believes *p*.[6]

There is another way to maintain that the higher-order and lower-order beliefs are distinct states, without appealing to an inner perceptual faculty. We could claim that the way we form beliefs about our own beliefs parallels the way we come to form beliefs about the beliefs of others. Ryle, for example, rejects the Cartesian model and turns the psychological outward; on the Rylean account, the accuracy of the statement 'I believe that *p*' is determined

[4] See Frankish 2004, Chapter 2 section 1.3, for detailed discussions of the distinction between binary (or flat-out belief) and partial belief. For Frankish, 'the term "belief"' can also track states that are "multi-track behavioural dispositions", which are non-conscious, passive, graded, and holistic' (2012, p.24).
[5] I consulted Coliva 2016 Chapter 7 and Parrott 2017 for their helpful summaries of constitutive views.
[6] See Shoemaker 1998 on why self-blindness is not possible.

not by the subject's having inner access to her mental items, but by observing her outward behaviour. We can determine the accuracy of a subject's self-descriptions by observing whether her behaviour indicates the presence of the mental state that she attributes to herself.[7] However, this account has limitations, as it cannot explain our intuition that there is something distinctively first-personal about self-knowledge.

A more popular approach posits that there is a constitutive link between a subject's higher-order and corresponding lower-order beliefs—I will call this approach the 'constitutive account'. This type of approach avoids the problems that plague the Cartesian model and the Rylean account. Since the relation between the higher-order belief and the lower-order belief is a constitutive one rather than a causal one, it avoids the possibility of self-blindness. Moreover, since a subject's belief about her own belief is necessarily true, it avoids the possibility that she knows her own beliefs in a third-personal way.

The core claim of the constitutive account may be stated as follows:

> Constitutive thesis: Given certain background conditions, one believes that *p* if and only if one believes that one believes that *p*.[8]

Different accounts might stipulate different background conditions. Broadly speaking, defenders of the constitutive account can be divided into two groups, disagreeing about the grounding relation between higher-order belief and lower-order belief. What I shall refer to as the 'factualist account' maintains that lower-order belief grounds higher-order belief. What I shall refer to as the 'non-factualist account' maintains that higher-order belief grounds lower-order belief.[9] Here, I will focus my discussion on two representative accounts from each group, one by Shoemaker and the other by Heal.

The factualist account holds that the lower-order belief that *p* is the more fundamental notion. It holds that a subject can only self-ascribe a lower-order belief when the lower-order belief is present. Such accounts usually appeal to subject's rationality and her possession of the concept of belief. Shoemaker, for example, maintains that a subject's higher-order belief that she believes *p* is necessarily constituted by her lower-order state of believing that *p,* because there is a rationality-based connection between lower-order beliefs and higher-order

[7] Ryle 1949.
[8] Coliva 2016, Chapter 7 provides a helpful summary of the constitutive account.
[9] I thank Mike Martin for helping me to formulate this.

belief. A subject knows what she believes as long as she is rational and possesses the relevant concept of belief. As Shoemaker writes:

> What I have asserted […] is a connection between self-knowledge and rationality; that given certain conceptual capacities, rationality necessarily goes with self-knowledge. It is entirely compatible with this that there are failures of rationality that manifest themselves in failures of self-knowledge. And such I assume we have in cases of unconscious belief.[10]

Shoemaker maintains that it is a requirement of full human rationality that one revises and updates one's belief system such that, when one encounters new evidence that contradicts *p*, one will adjust one's lower-order belief of not-*p*. Since being rational involves responding to evidence about one's beliefs, and since readjustments are rational only if they are made on the basis of one's awareness of the contents of one's own attitudes, being rational requires a subject to have higher-order attitudes in order to rationalise an adjustment of belief. Hence, it is necessary to postulate higher-order beliefs to explain how a rational subject can revise and update her belief system. Shoemaker in his later work posits a constitutive relationship between first-order beliefs and higher-order beliefs. He argues as follows:

> It is not the belief that *p* all by itself that accounts for the disposition to judge that one has it if the question arises; this requires in addition the possession of the concepts of belief and of oneself, and it requires a certain degree of rationality. Perhaps it is all of this, together with the belief that *p*, that constitutes the standing second-order belief that one believes that *p*. […] If a belief has the belief that *p* as an essential part, its possession cannot survive the loss of the belief that *p*.[11]

Shoemaker acknowledges that not all of a subject's lower-order beliefs are accompanied by higher-order beliefs. The lower-order beliefs that are constitutively 'self-intimating' are 'available' beliefs, such that the subject is 'poised to assent to their contents, to use them as premises in reasoning, and to be guided by them in their behaviour.'[12] On this picture, available lower-order beliefs partially constitute standing second-order beliefs; the available

[10] Shoemaker 1988: 208.
[11] Shoemaker 2009, p. 42.
[12] *Ibid*., p. 40.

beliefs are parts of the higher-order beliefs that self-ascribe them. Shoemaker emphasises that this is different from saying that available beliefs can cause higher-order beliefs. What the available lower-order belief can cause is the subject's affirming the belief that *p* when *p* is considered, but not the higher-order beliefs. Higher-order beliefs come into being with lower-order beliefs when the subject is aware of her lower-order beliefs. It is in this sense that available lower-order beliefs 'intimate' to higher-order beliefs and ground self-ascription of beliefs.

The non-factualist account holds that one's belief that one believes *p* is the more fundamental notion. For the non-factualists, it is only when a belief is entirely independent of judgement that it can be considered factual. But when a subject believes that she believes that *p*, her higher-order belief is already a judgement and hence cannot be considered factual. The proponents of non-factualist accounts tend to identify themselves as followers of the Wittgensteinian tradition. While it is unclear if Wittgenstein himself endorses a non-factualist account of self-ascription, he clearly questions the assumption that self-ascription is a report of the subject's lower-order belief. Wittgenstein writes:

> How did we ever come to use such an expression as 'I believe…'? Did we at some time become aware of a phenomenon (of belief)?
>
> Did we observe ourselves and other people and so discover belief?[13]

Wittgenstein seems to suggest that when one says, 'it is raining', or 'I believe it is raining', her point is not to have her audience infer something about her state of mind. In Wittgenstein's words, if my audience says, 'I see, this is how it seems to you now', the subject may reply, 'we're talking about the weather… not about me'.[14] The point here is that by prefixing 'I believe' or 'I do not believe' to '*p*', the subject has not changed the subject matter from *p* to her own state of mind.

Heal's account is an example of how we may understand self-ascription without taking the subject to be reporting her lower-order belief. Heal proposes a constitutive theory which does not require a prior first-order belief for self-ascriptions of belief. On this account, a higher-order belief constitutes the first-order state that it is about; a subject who sincerely utters 'I believe *p*' makes it true that she believes *p*. Heal draws a distinction between a

[13] *Philosophical Investigations* II x, section 86.
[14] Wittgenstein, *Remarks on the Philosophy of Psychology* I, 750.

mental state of belief and ordinary utterances.[15] According to Heal, it is only on the basis of a mental state of belief, rather than utterances, that we may know if the belief we ascribe to ourselves is true. As she writes, '[T]he existence of a second-level belief about a first-level psychological state is what makes it true that the first-level state exists'.[16] For Heal, a higher-order belief always brings with it a first-order mental state.

If there is a constitutive link between higher-order belief and lower-order belief, then our self-knowledge is epistemically secured. Since a subject's higher-order belief must contain the lower-order belief her belief that she believes *p* must amount to knowledge. But is there such a link? We can get a glimpse of the burden this assumption has created for both factualists and non-factualists from the way Shoemaker and Heal try to handle Peacocke's example of an administrator who ascribes to herself the belief that graduates from overseas universities are equally qualified as graduates from local universities, yet systematically favours local graduates when making hiring decisions.[17]

On Shoemaker's account, this is a case where the administrator mistakenly ascribes to herself the belief that not-*p*, while in fact she believes that *p*. The reason that it is possible for the subject to falsely believe that a belief with a certain content *p* is one's 'dominant belief', is that the subject fails to realise that she actually has a stronger belief not-*p* that contradicts *p*. Shoemaker thinks that, in this case of self-deception, 'the tendency of a belief to become available' is 'blocked'.[18] However, given Shoemaker's view that belief has a tendency to be available for a rational subject's access, it is unclear why the actual dominant belief has not made itself available in cases of self-deception.

Heal would likely argue that the administrator is not being sincere because 'the non-existence of appropriate behaviour is grounds for questioning the truth of a self-ascription of belief and at the same time grounds for questioning sincerity'.[19] Let us slightly modify Peacocke's example; suppose we are considering the same administrator, who works at a different university. She is very complacent and tends to make decisions that are in line with her direct superior's judgements. It so happens that her new superior thinks that the university should hire as many overseas graduates as local graduates. The superior's judgement factors into the administrator's decision-making process; as a result, the pattern of the

[15] Heal 2002.
[16] *Ibid*., p.4.
[17] Peacocke 2000, p.90.
[18] Shoemaker 2009, p.43.
[19] Heal 1994, p. 21

administrator's observable behaviour suggests that she does believe that overseas and local graduates are equally qualified. Now, on Heal's account, there would be little reason to think that the self-ascriptions of the administrator are not authoritative. She sincerely believes that she believes that local and overseas graduates are equally qualified and her pattern of behaviour is also consistent with the belief she ascribes to herself.

We may argue, from a third-person perspective, that in a counterfactual situation wherein her superior is indifferent, but she still hires significantly more local graduates, the administrator's pattern of behaviour would reveal her belief that local graduates should be favoured. However, from a first-person perspective, while she is in the actual context, she might sincerely believe that she believes that local and overseas graduates are equally qualified, and yet her actual pattern of behaviour is also consistent with the belief she ascribes to herself. It seems that we should be able to claim that the administrator is insincere, or that she is making a false self-ascription. However, on Heal's account, we seem unable to make either of these claims. This modified case suggests that the link between authority and infallibility, as Heal intends it, can only be established and remain at the level of BB*p*—that is, if the subject sincerely pronounces on her second-order belief that *p*, then it is the case that she believes that she believes that *p*. This may be a step away from stating that, since her self-ascribed belief BB*p* is authoritative and infallible, her self-ascribed B*p* is also authoritative and infallible. It seems that another plausible explanation of what is going on in the case involving the self-deceived administrator appeals to the thought that the administrator believes that she believes that local and overseas graduates are equally qualified, when in fact she lacks the belief that local and overseas graduates are equally qualified. She is neither self-deceived nor irrational. Such an explanation, however, is unavailable to the constitutive view.

We are now in a position to see what a successful counterexample to the constitutive account will look like. It needs to be structurally like the self-deceived administrator but without irrationality. The bulk of the coming argument can be understood as developing an example of this sort.

1.3 Outline

In this thesis, however, I will not argue directly against any version of the constitutive account. Again, my main claim is that conscious, self-directed reasoning opens up the possibility for rational subjects to believe that they believe *p* when in fact they do not. Still, if this claim is correct, then defenders of the constitutive account are in an unenviable position.

Suppose that it is possible for higher-order belief and lower-order belief to come apart. Accounts that model self-ascription on the constitutive account's assumption can accommodate the possibility of fallibility by specifying the normal conditions for self-ascriptions, such as the presence of rationality. They can say that the constitutive link between higher-order belief and lower-order belief only holds when these normal-making conditions are met. In this way, the constitutive account can accommodate fallible cases such as a husband believing that he believes that his wife is faithful, whereas his behaviour indicates that he believes that his wife is unfaithful. The constitutive account can explain such phenomenon in terms of irrationality and contradictory beliefs. They might say that the husband believes that his wife is unfaithful and believes that his wife is faithful. Due to irrationality, the husband only believes the belief that his wife is faithful; however, it still is the case that his higher-order belief and his lower-order belief are constitutively linked.

If we further suppose that the subject is rational, then the constitutive account cannot rely on these additional conditions to maintain the link between higher-order belief and lower order belief. Indeed, if it can be shown it is possible for a rational subject to believe that she believes *p*, when in fact she has no belief about *p*, then the constitutive account is cast in serious doubt. For the possibility of a rational subject having a higher-order belief without the corresponding lower-order belief breaks the constitutive link between a higher-order belief and a lower-order belief without triggering any plausible abnormal-making conditions. This is precisely what my thesis aims to show.

The argument spans five chapters. In Chapter 2, I use a case of surprise to motivate the possibility of a rational subject's having a higher-order belief that is not linked to any lower-order belief. This helps render my view intuitively plausible and motivates the idea that there are two ways of fixing beliefs; one is a top-down way that starts from one's reasoning, and one is a bottom-up way that starts from one's direct experience with the world. It is the former—fixation through reasoning—that gives rise to the possibility of higher-order belief without lower-order belief. The rest of the thesis will refine and defend this idea. In Chapter 3, I discuss a species of fixation through reasoning that is self-consciously directed, which I call 'ratiocination'. I argue that the conclusion of ratiocination is in the form of *I ought to believe p*. When one concludes ratiocination, one believes that one ought to believe *p*, and this belief in turn initiates a top-down belief fixation process. In Chapter 4, I argue that it is possible for this top-down fixation process to fail at the lower-order, leaving one without the belief that *p*. In Chapter 5, putting together the subsequent discussion, I present my account as to why it is possible to have a higher-order belief without

the corresponding lower-order belief. Since one does not necessarily form the belief that *p*, it is possible that, while one's mind rationally moves from believing that she ought to believe *p* to believing that she believes *p*, the lower-order belief that *p* is never fixed, hence the possibility of having only the self-ascriptive belief without the corresponding lower-order belief. The subject does not have to be irrational when she has a self-ascriptive belief without the corresponding lower-order belief, since from her first-person perspective, she is not able to tell the difference between forming a belief about her own mental state and forming a belief about *p*. In this chapter, I also draw on the transparency account of self-knowledge to show how it can be rational for one's mind to move from believing that she ought to believe *p* to believing that she believes *p* without having to rely on an inner sense mechanism or observing one's own behaviour. My account thus posits distinct states, yet it can avoid many of the commitments that come with the Cartesian model and Rylean account. In Chapter 6, I discuss the implications of my argument for self-knowledge.

Chapter 2
Surprise

2.1 Sam's surprise

Consider the following case: Sam has not seen a black swan before. One day, he learns that biological structures that contain melanin usually look black. He also learns that melanin is found in swans' feathers; that the colour of birds varies greatly depending on their environment; that some bird species in the northern hemisphere are white, while their southern hemisphere counterparts are black. He deliberates about what he learns and comes to conclude that there are black swans. He then sincerely believes that he believes that there are black swans. Sometime later, when Sam is on a visit to Australia and walking around a lake, he sees a flock of black swans. Sam has not forgotten what he learned nor the conclusion of his deliberation, yet he is surprised that there are black swans. How can we make sense of Sam's surprise? Before trying to answer this question, three points of clarification about the case require detailed discussion.

First, Sam's case is set up in an as neutral way as possible to minimise other factors from interfering in the process of higher-order belief formation. I assume that the subject matter of whether there are black swans is neutral and insubstantial for Sam. Whether he believes that there are black swans does not affect his life, his self-conception, moral standing, or anything that matters to him. By minimising these other factors that could have affected the process of self-ascription of belief, we can focus on the ways in which a subject's higher-order belief is normally formed, instead of being distracted by other concerns, such as the subject's having special motivation for making a mistake.[20]

It is possible that, for some subjects, some old beliefs are too "tenacious" to be updated, or, when the stakes are high, one might revert to one's old belief under pressure. In a case where the subject has some motivation to preserve the higher-order belief, it is conceivable that one might maintain the higher-order belief even though one's lower-order belief is dropped. A suspicious husband might, upon being reassured by his wife that she is not having an affair, come to believe that his wife is not having an affair and ascribe the

[20] There are other ways of telling Sam's story that we can say Sam is a motivationally biased believer, e.g., Pears (1984) and Mele (2003). I am only suggesting that it is possible that Sam is not influenced by motivation.

belief to himself. However, his old belief that his wife is having an affair can soon subvert the newly acquired belief. It could be the case that the husband is motivated to maintain the more comfortable and convenient higher-order belief of his wife not having an affair; for Sam, however, since we assume that the subject matter is neutral to him, the motivation for maintaining the mistaken higher-order belief is not clear. It is not obvious what stakes are involved in Sam's case that would motivate such reversion of belief. Let us thus rule out the possibility that Sam did believe that black swans exist when he concluded reasoning, but, motivated by practical considerations, he dropped this lower-order belief sometime between the time of deliberation and the time of seeing the swans. Even if he did acquire the belief that there are black swans and somehow revert to believing that there are no black swans, it is unclear why he would be motivated to maintain his false self-ascriptive belief that there are black swans, instead of also updating his higher-order belief to the belief that there are no black swans.

The lack of practical motivation also helps forestall a natural response to the case, namely, that something has simply gone wrong with the mechanism of self-ascription. Consider, for example, Peacocke's explanation of mistaken self-ascriptions:

> It is not the full explanation, and my exposition was peppered with occurrences of the qualifying phrase 'when all is working properly'. Someone can make a judgement, and for good reasons, but it may not have the effects that judgement normally do – in particular, it may not result in a stored belief which has the proper influence on other judgements and on action. A combination of prejudice and self-deception, amongst many other possibilities, can produce this state of affairs.[21]

Peacocke thinks that one can make mistaken self-ascriptions but this only happens when the process of self-ascription goes wrong (e.g. when prejudice or self-deception is involved). Although Peacocke's case also posits that the subject can ascribe to herself a belief but does not possess the corresponding belief, the theoretical work in Peacocke's case is different from Sam's case. Peacocke concedes that the methods of self-ascription can be fallible; the subject's 'prejudice and self-deception' could have prevented her from employing the method of self-ascription properly such that her judgement that *p* fails to produce a lower-order belief

[21] Peacocke 2000, p.90.

that *p*.[22] However, such fallibility does not threaten Peacocke's claim that a judgement that *p* will initiate the belief that *p* when 'all is working properly'.[23] For Peacocke, this case of mistaken self-ascription is a product of some failure of the normal mechanism of self-ascription. When things are working properly, one's judgement that *p* will initiate one's belief that *p* and, on the basis of the lower-order belief that *p*, one comes to believe that one believes that *p*. Once there is a positive account of self-ascription that assumes certain things about how a higher-order belief is formed, then all the unusual cases of mistaken self-ascriptions can be treated as abnormalities that do not threaten how one's higher-order belief is normally made. In short, Peacocke introduces his case to show that his account can handle cases of fallibility and provide a satisfactory explanation for mistaken self-ascriptions. When all is working properly, Peacocke maintains, judgement is the fundamental way to form first-order belief—judging that *p* gives one a reason for self-ascribing the belief that *p*. The subject's self-ascriptive belief amounts to self-knowledge because it is true that, in virtue of judging that *p*, one believes *p*. Peacocke might read Sam's case as one in which Sam is self-deceived; however, we should resist this reading. The lack of practical considerations—the fact that Sam is not specially invested in swans—makes turning to a failure of the mechanism of self-ascription undermotivated.

Now to the second point of clarification. To keep the case simple, I will assume that Sam has not encountered any situation that suggests the non-existence of black swans. In other words, in the interval between his deliberation and seeing the swans, he did not encounter anything that prompted him to change his mind or made him confused. For example, we are not to suppose that, in the interim, Sam has visited a swan museum in which he has only seen exhibits of white swans and has later received a postcard with a picture of a black swan. If he is confused, Sam will not be surprised. He might experience some other feelings, such as relief, as seeing the black swans settles for him that black swans do exist. If Sam is surprised, then it must be the case that Sam has an attitude committing to a proposition that clashes with the existence of black swans. To avoid saying that there are possible situations that might have confused Sam, we stipulate that Sam simply reasoned that there are black swans, the topic never came up again, and then he saw the black swans in Australia.

---

[22] *Ibid*.
[23] *Ibid*.

The third point of clarification is this: Sam's surprise is what shows that Sam holds an attitude that commits him to the non-existence of black swans. I will call such an attitude a 'contrary attitude'. We do not need to speculate about Sam's behaviour, whether or not it is consistent with his self-ascribed belief that black swans exist. We may suppose that, before Sam saw the black swans, he had not encountered any situation prompting him to act one way or another that suggests the presence or absence of his belief that there are black swans.

Although Sam is fallible about his own mental states and has made a mistaken self-ascription, this is not what is of central importance. The point of introducing Sam's case is not to show that people can be fallible about their own mental states, but to help probe the nature of self-ascriptive belief formation. More specifically, I use Sam's case to introduce the possibility that one's false self-ascriptive belief is due to there being two distinct ways of fixing belief: one that fixes belief directly at the lower-order level, and one that fixes belief starting from the higher-order level. It is worth stressing that this explanation does not require us to make any substantive commitments concerning how the method of self-ascription is supposed to work; nor does it require us to hold that Sam is irrational. It is also worth stressing that, as described, there are many ways of explaining Sam's surprise. My claim is that my explanation is one of them, and this will suffice for the purposes of this thesis. I am not claiming that my account provides the best explanation or captures the closest possible world to the actual world where the events of Sam's surprise take place. It will suffice if this is at least one possible way of describing his surprise.

2.2 The Surprise Principle

What is crucial to our understanding of Sam's case is that it is Sam's surprise that reveals Sam's contrary attitude about black swans. The principle I rely on to tell Sam's story may be labelled the 'surprise principle':

> Surprise principle: if a subject S is surprised that *p,* at time, *t*, then S has an attitude that is contrary *to p* prior to *t*.

The surprise principle is in line with our ordinary conception of surprise. For example, I am surprised to see my friend at the restaurant because I believe that she is travelling overseas. I will rely on the surprise principle to tell Sam's story. For those who do not find the principle obvious, my defence of the surprise principle can be found in the Appendix.

The surprise principle can accommodate a broad spectrum of surprising events, involving different levels of cognitive complexities. The question important to us is in what way Sam is surprised. For one to be surprised, there needs to be a determinate link to one proposition. In order to say that Sam's surprise reveals that he only believed that he believes that black swans exist but did not have the belief that black swans exist, Sam's surprise has to be linked to the proposition that black swans exist. In setting up Sam's story, I need to say Sam is surprised that black swans exist to prevent a reading of the case where Sam is surprised by something other than the existence of black swans. If this were the case, his surprise won't be very interesting. He could be surprised that black swans have red bills. It is only when Sam's surprised is generated by his encounter with black swans that I can say Sam did not believe that black swans exist. Sam either believed that black swans do not exist or had some belief whose content is inconsistent with the existence of black swans.

On the surprise principle, one can only be surprised if one holds an attitude that is contrary to *p*.[24] Hence, Sam's surprise reveals that he held a contrary attitude before he saw the black swans. Since we are assuming that Sam is rational, he is not holding both an attitude contrary to *p* and the belief *p*. Hence, we may say he was mistaken to believe that he believes that black swans exist.

With these clarifications made, I claim the following is a possible account of what happened: Sam concludes, from reasoning, that there are black swans; yet he fails to acquire the belief that there are black swans. As a result, he rationally believes that he believes that there are black swans. Later, when Sam saw the black swans, he was surprised because he lacked the lower-order belief that there are black swans, while holding an attitude that is contrary to the proposition that there are no black swans. The disconfirmation of his prior belief is contrary the belief that 'black swans exist' that generates Sam's surprise. If my

[24] The surprise principle only lays down the necessary condition for surprise that says if one is surprised that *p*, one must have a prior attitude that is contrary to *p*. It is silent on what to say about cases where one lacks surprise. Noë suggests that a subject's lack of surprise that *p* may indicate her lack of commitment to *p* (2002, p. 6-7). This is possible by the surprise principle, but the principle also allows one to have a commitment to *p* but is not surprised. One might be emotionally overwhelmed or underwhelmed to feel surprised. Or the stimulus is not strong enough to generate a contrast that is sharp enough to trigger surprise. I might believe that there are sixty sweets in my bag and when it turns out there are sixty-one sweets, I am not surprised. In this case, the stimulus might not be strong enough to generate a sharp contrast between old belief and new belief. Hence, by the surprise principle, it is not particularly informative when one does not experience the surprise that *p*.

explanation is correct, then Sam's case is an example of a subject forming a higher-order belief without forming the corresponding lower-order belief.

2.3 Possible Objections

I anticipate two main objections to my claim that Sam's case is possible. The first might come from those who uphold the incorrigibility thesis—that is, if one believes that one believes that *p*, then one believes that *p* (BB*p* → B*p*).[25] On the incorrigibility thesis, since Sam has BB*p*, then Sam must have B*p*. Those who accept the incorrigibility thesis might want to explain Sam's surprise in terms of inconsistent beliefs. When Sam believed that he believes that there are black swans, he did believe that there are black swans. However, he also believed that there are no black swans, or believed some other proposition that is inconsistent with there being black swans that generates the surprise. This way of explaining Sam's surprise comes with its own theoretical burdens, for it has to also posit something like a divided mind. We have to explain why a rational subject holds inconsistent beliefs in divided parts of the mind. Moreover, suppose we grant that it is possible for a subject to hold contradictory or inconsistent beliefs; still, it is not obvious why Sam would be surprised. Since Sam is surprised, it is not his belief that there are black swans that is generating the surprise. But why should we think that Sam's belief that there are no black swans generates surprise, while his belief that there are black swans is suppressed? It is unclear why the belief that there are no black swans can generate surprise when the belief that there are black swans is also present. In any case, even if this is one possible way of explaining the case, it does nothing to threaten the possibility of my explanation.[26]

Those favouring the incorrigibility thesis might try to rule out my explanation by claiming that Sam must lack the self-ascriptive belief. He cannot, they might argue, sincerely believe that he believes that black swans exist. However, to deny that Sam can be sincere in his self-ascriptions is to deny the sincerity of subjects in all cases of mistaken self-ascriptions. I have already discussed this problem in the previous chapter.[27] There could be subjects who, like Sam, ascribe to themselves the belief that black swans exist but never got a chance to see black swans. For these subjects, surprise reveals their contrary attitude. Many of us learn

[25] See Chapter 6.2 for further discussion of the incorrigibility thesis.
[26] See Chapter 6.2 for discussion of why Sam could not have believed that there are black swans prior to his surprise.
[27] Heal 1994, p. 21.

about different aspects of the world through reasoning without having the chance to directly experience those aspects of the world, yet we do at least believe that we have certain beliefs about it. For example, I believe that I have the belief that there are pink river dolphins. I do not know if I will be surprised when I see a pink river dolphin, or there might never be an occasion that requires me to act in a way that reveals my lower-order belief about pink river dolphins. However, these are not reasons for discrediting my belief that I have the belief that there are pink river dolphins.

Alternatively, those favouring the incorrigibility thesis might reject the stipulations of the case. They might deny that Sam could be surprised. If they take this route, the burden is on them to say why BB*p* entails B*p*. I have presented a case in which the subject's BB*p* seemingly does not entail B*p* and provided my explanation for it. My explanation for Sam's story seems possible. It does not seem to rest on some conceptual, logical, or metaphysical confusion, thus the burden is for those who defend incorrigibility to say that it is impossible in one of these senses for Sam to be surprised by the fact that black swans exist.

Another group of my main opponents would argue that, since Sam concludes from reasoning that there are black swans, it must be the case that he also believed that black swans exist, and therefore cannot be surprised. This line of objection relies on the assumption that, when a rational subject concludes from reasoning that *p*, necessarily one also believes that *p*. I will, in the next two chapters, provide an elaborate argument for why a rational subject does not necessarily believe that *p* even though she concludes from reasoning that *p*. For now, I will further elaborate on my proposed explanation of Sam's case, which avoids the suggestion that Sam necessarily believes that there are black swans because he concludes from reasoning that there are black swans.

2.4 Two Levels of Fixing Belief

Most philosophers would agree that perception and reasoning are two ways in which one's higher-order belief is fixed; we may call this standard view the 'two-system' view. On the two-system view, when a subject perceives that *p*, she comes to believe that *p*. On the basis of her belief that *p*, she comes to believe that she believes that *p*. Similarly, for a rational agent, when she concludes that *p* in reasoning, she comes to believe that *p*. On the basis of her belief that *p*, she believes that she believes that *p*. However, if my explanation of Sam's case is possible, the two-system view will not be able to say why Sam did not believe *p*. If Sam concluded that *p*, he should have also believed that *p*. However, his surprise reveals that

he held an attitude contrary to *p*. Those that accept the two-system view might then resort to other motivations—e.g., practical considerations—that might have intervened, preventing the normal route of higher-order belief formation through reasoning. However, because we assume that this is not a motivated case, this route seems unavailable. Alternatively, they might appeal to the idea the Sam is irrational. It will be further developed in later chapters why it is possible for Sam to be rational, given the stipulations of the case.

I agree with the standard view that there are different systems that fix one's lower-order belief, but disagree that these systems fix beliefs from the same level. My reading of Sam's case motivates the idea that there are two 'levels' of belief formation. One starts, as in the case of perception, at the lower-order level. The other starts, as in the case of conscious reasoning, at the higher-order level. Corresponding to the two levels of belief formation are two ways of fixing lower-order belief and two ways of fixing higher-order belief. With regard to fixing higher-order belief, there is the 'directly-top way' that directly fixes higher-order belief at the reflective level (e.g., I infer from conscious reasoning that there is an apple and I believe that I believe that there is an apple). I will refer to it as the 'reflective way' for ease of discussion. There is also the 'bottom-up way' that fixes higher-order belief from a lower-order level (I see that there is an apple, I believe that there is an apple and I believe that I believe that there is an apple). With regard to fixing lower-order belief, there is the 'directly-bottom way' that directly fixes lower-order belief (e.g., I see an apple and I believe that there is an apple). I will refer to it as the 'unreflective' way. There is also the 'top-down way' that fixes lower-order belief from a reflective level (e.g., I infer from conscious reasoning that there is an apple and I believe that there is an apple). Since there are two levels from which belief can be fixed, it is possible that top-down fixation of lower-order belief fails but reflective fixation of self-ascriptive belief is successful. In such a case, one ends up believing that one believes that *p* without believing that *p*, as depicted in the following table.

	Reasoning	Perception
Fixing Lower-order belief	Top-down ✗	Directly-bottom
Fixing Higher-order belief	Directly-top ✓	Bottom-up

On this 'two-level' view, some methods, such as reasoning and perception, fix one's belief directly at different levels.

In this thesis, I will for the most part ignore the Perception column. I take it for granted that perception can directly fix lower-order belief and that there is the unreflective

way of fixing lower-order belief and the bottom-up way of fixing higher-order belief. These two claims are relatively uncontroversial and are in line with the orthodox view. My focus is on the Reasoning column. I will focus on explaining the top-down way of fixing lower-order belief (in Chapters 3 and 4) and the reflective way of fixing higher-order belief (Chapter 5). I will argue that the conclusion of ratiocination, which is self-consciously directed reasoning, is in the form of *I ought to believe* p. For an ideal subject, if she concludes at the end of ratiocination that *p* is the case, she will come to believe *p* and believe that she believes that *p*. However, for an average rational subject, it is possible that her higher-order belief that she ought to believe *p* fails to bring about her belief that *p*. However, it is also rational for her to believe that she believes that *p* because of transparency. In such cases, it is possible that one has a false self-ascriptive belief even though there is no failure in the mechanism for self-ascription nor failure of rationality.

Chapter 3
Ratiocination

Our beliefs can be fixed by different methods, such as perception, reasoning, testimony, and memory. Suppose two people both believe that there is an apple in the fridge. Percy stares at the apple in the fridge; it appears to him that there's an apple in the fridge. Rae reasons that there were two apples in the fridge earlier but then she took one, and no one else has access to the fridge, so there must still be one apple in the fridge. Both Percy and Rae believe that there is an apple in the fridge. Yet crucially they rely on different methods to acquire the belief that there is an apple in the fridge. Percy relies on perception whereas Rae relies on reasoning. Are beliefs that get fixed by different methods fixed from different levels?

That there is little discussion of this topic suggests that philosophers tend to assume that the level of belief fixation does not vary according to different methods. But, as we saw from the previous chapter, Sam's case invites us to reconsider this assumption.

In this chapter, I analyse the nature of a particular form of theoretical reasoning – ratiocination. Ratiocination is self-consciously directed reasoning. Here, I am only concerned with how a subject's mind moves when she ratiocinates. Note that I am not concerned with what it is to ratiocinate well or correctly. I will argue that the conclusion a subject draws in ratiocination is in the form of *I ought to believe p*. Those who work in the vicinity of this topic could easily confuse what I am trying to establish here with some familiar debates in the literature. Hence, a large part of this chapter will be spent on clarifying what ratiocination is and what it is not.

In section 1, I suggest that the way a reasoner's mind moves in ratiocination is different from the way her mind moves in non-ratiocinative reasoning. Such a difference should motivate an analysis that focuses just on ratiocination. In section 2, I provide a general characterisation of ratiocination and sketch my argument for the claim that the conclusion of ratiocination is in the form of *I ought to believe p*. In section 3, I separate my concern from concerns about whether there are reasons to be rational. I explain that my claim is compatible with the claim that reasons to be rational are transparent to reasons to believe *p*. I separate my concerns from concerns about normative requirements for reasoning, including the concern about whether logic is normative for reasoning. I explain that the form of one's conclusion in ratiocination is characterised by the form of one's response to normative requirements, not the form of normative requirements for reasoning. I separate my concern from concerns

about whether belief can be motivated by non-epistemic considerations. Finally, I explain that my claim is compatible with the claim that belief is truth-governed. This will set the stage for my argument in Chapter 4 for why ratiocination fixes lower-order belief in a top-down way.

3.1 Reasoning and Self-Consciously Directed Reasoning

Theoretical reasoning is concerned with what to believe. Philosophers sometimes use 'reasoning in a formal sense' to refer to the process of drawing out the consequences of premises and 'reasoning in an informal sense' to refer to the process of revising one's beliefs.[28] In this thesis, I use 'theoretical reasoning' to mean the latter kind of activity or as Harman calls it, 'the reasoned changed in view.'[29] Theoretical reasoning is a key method by which we form beliefs. There are many truths, for example, truths about the things that do not exist, that can only be grasped by reasoning.

It is standardly assumed in the literature that the conclusion of theoretical reasoning is a proposition about the first-order subject matter of reasoning (*p*). Suppose we hear someone reasoning to herself, 'It is raining. If it rains, the grass is wet. So, the grass is wet,' it is tempting to think that what she concludes is 'The grass is wet.' Although 'The grass is wet' is what she says or thinks to herself, it does not mean that 'The grass is wet' is the conclusion of her reasoning. To better understand the precise conclusion of her reasoning, we have to understand the nature of the activity. I will argue that given the nature of the mental activity that ratiocination is, the conclusion is a normative judgment about what the subject ought to believe (*I ought to believe that p*).

In existing discussions of reasoning, philosophers are generally sensitive to the differences between practical and theoretical reasoning.[30] Even for those who think that practical and theoretical reasoning can be given the same treatment in certain respects, the burden is on them to justify why they think practical and theoretical reasoning can be treated the same in those respects.[31] The default view is that there are significant differences between

[28] MacFarlane 2004, p. 4.
[29] Harman 1986, p.1.
[30] For a summary of main views about the distinctions between theoretical and practical reasoning, see Streumer 2010.
[31] Marusic 2018, Rinard 2019, and the collection of papers in McHugh, Way, and Whiting 2018 are examples of recent attempts to explain certain similarities between theoretical reasoning and practical reasoning.

practical reasoning and theoretical reasoning. And any account of reasoning will need to capture these differences.

However, within the domain of theoretical reasoning, philosophers are not generally sensitive to the possible differences between self-consciously directed reasoning, ratiocination, and non-ratiocinative reasoning. Indeed, the distinction between non-ratiocinative reasoning and ratiocination often goes unnoticed. In some discussions of theoretical reasoning, reasoning seems to be used interchangeably with ratiocination. For example, when Boghossian discusses reasoning, he says that what he means by reasoning is 'System 1.5 reasoning'. It is somewhere between what Kahneman calls 'System 1' reasoning, which is 'sub-personal, sub-conscious, involuntary and automatic,' and 'System 2' reasoning, which is 'personal-level, conscious, attention hogging and effortful.'[32] This suggests that in Boghossian's view, even though reasoning is not necessarily effortful and slow, it has to be at the 'person-level' which is what I mean by the conscious level. This point gets clearer later when Boghossian explicitly says that '[reasoning] is something we do, not just something that is done by sub-personal bits of us.'[33] And when Boghossian goes on to discuss the 'Taking Condition', he seems to imply that the reasoner is not only conscious of what he is doing, he also has to guide himself in the activity of reasoning. The 'Taking Condition' states that 'Inferring necessarily involves the thinker *taking* his premises to support his conclusion and drawing his conclusion *because* of that fact.'[34] Thus what Boghossian means by 'reasoning' is very close to what I mean by 'ratiocination'.[35]

In other discussions, reasoning is used for both ratiocination and non-ratiocinative reasoning. For example, Grice says that a reasoner has to entertain 'in thought or in speech of a set of initial ideas (propositions)'.[36] This suggests that Grice thinks that reasoning is conducted self-consciously and is close to what I mean by ratiocination. But it is not entirely clear if Grice means ratiocination because he later goes on to say that the steps made by a reasoner in reasoning are '*either* validly made or are thought to be validly made.'[37] The claim that a reasoner could make a valid step without thinking that she makes a valid step suggests

[32] Boghossian 2014, p. 2.
[33] *Ibid*.
[34] *Ibid*., p. 5.
[35] If I am right in saying that there is a substantive difference between ratiocination and non-ratiocinative reasoning, then Boghossian's debate with many of his critics talk past each other because they are talking about different kinds of activities.
[36] Grice 2005, p. 5.
[37] *Ibid*., p. 6.

that perhaps Grice includes some forms of reasoning that would not qualify as ratiocination, in my sense.

For still others, even though the difference between ratiocination and non-ratiocinative reasoning is noted, the focus is on developing a general account of reasoning that can cover both.[38] Broome, for example, says that reasoning is not necessarily explicitly conducted in language. His account of reasoning can apply to both explicit and non-explicit cases of reasoning. Kolodny also makes it clear that his discussion of reasoning covers both cases where 'reasoning is explicit, or deliberation' and reasoning is 'implicit and not voluntarily directed.'[39]

Those who attempt to offer a single account of theoretical reasoning, covering both ratiocinative and non-ratiocinative, might defend this strategy as follows: 'Even though the ratiocinative and non-ratiocinative cases are different, in that one involves consciousness and the other does not, that difference does not matter for theoretical reasoning, as such. The difference between ratiocination and non-ratiocinative reasoning is not like the, at least apparent, differences between practical and theoretical reasoning. In the practical case, one is trying to determine what to do, whereas in the theoretical case, one is trying to determine what to believe. Acting and believing could be very different. However, whether one is self-consciously engaged in theoretical reasoning or not self-consciously engaged in theoretical reasoning, it still is the case that one is engaged in theoretical reasoning. Depending on the kind of activity and what we try to capture, it might not be problematic to give the same treatment to both consciously directed and non-consciously directed activities. It is true that some kind of activities precludes self-conscious directing or the presence of self-conscious directing brings a stop to the activity. For example, suppose one is trying to articulate what her 'gut feelings' are about *p*. Given the nature of gut feelings, the moment when one starts to self-consciously direct her gut feelings about p, she is no longer articulating her gut feelings about *p*. But for some kinds of activities, the presence of self-conscious directing does not change the nature of activity itself. For example, if we try to explain the nature of, say, swimming, whether one is self-consciously directing her movements or non-self-consciously swimming, it does not change the nature of the activity that she is performing. In both cases, she is still swimming. There is of course a difference between one who is consciously monitoring and directing her own movements when she swims and one who is 'in the flow'

[38] See, for example, Broome 2013, Kolodny 2005, Scanlon 2007, p. 91, Streumer 2007, p. 2.
[39] Kolodny 2005, p. 520.

and does not consciously direct her movements, but that difference does not prevent us from giving a general, uniform account of swimming. The additional presence of self-conscious directing might affect how well the swimmer performs this activity. Perhaps the presence of self-conscious directing will prevent the swimmer from finding a good rhythm, but it does not change the nature of the activity she is performing. Similarly, we do not have to think that reasoning is the kind of activity that precludes self-conscious directing or that its nature is changed by the presence of self-conscious directing. The presence of self-conscious directing does not immediately bring the activity of reasoning to an end. If anything, reasoning might be the kind of activity that is better performed when done self-consciously. Whether it is self-consciously directed theoretical reasoning or non-self-consciously directed theoretical reasoning, what the reasoner tries to do is to figure out what she is to believe. As long as our target is theoretical reasoning, not self-consciousness, we can just focus on the reasoning part.'

But is theoretical reasoning the kind of activity the nature of which is not affected by self-conscious directing? Is it like swimming such that whether or not self-conscious directing is involved, the basic characteristics of this activity remain unchanged such that we can still give a uniform account of this activity? For the swimming example, we can all agree that there is a difference between self-consciously directed swimming and non-self-consciously directed swimming, namely, the way the swimmer's mind moves. The self-conscious swimmer has thoughts like 'I should push my chest a little down,' 'I need to kick three times with my left leg' and so directs her movement in a way that the non-self-conscious swimmer does not. We may imagine that the self-conscious swimmer ends up performing exactly the same bodily movements as the non-self-conscious swimmer. The similarities in their movements preserve the target 'swimming' for analysis. It is in this way that the nature of swimming is not affected by self-conscious directing. But note that, though their physical movements are the same, the mental movements of the swimmers when they are involved in these physical movements are different. One's mind is consciously monitoring her movements and directing her movements. The other's mind is not consciously monitoring how her body is moving. If the target of our analysis is the swimmers' mental movements, then we cannot give the same analysis. This is the main disanalogy between physical activities and theoretical reasoning.

Theoretical reasoning is a mental activity. How the reasoner's mind moves is how her reasoning goes. We cannot think, as is plausible with swimming, of self-consciously directed theoretical reasoning simply as one part of the mind doing the directing and the other part

doing the reasoning. With ratiocination we cannot simply bracket the self-conscious directing part and focus on the reasoning part. The mind of ratiocinative reasoner and the mind of the non-ratiocinative reasoner are different. Let's say the way a ratiocinative reasoner's mind moves is *a* and the movement ends at a point *x*. And the way a non-ratiocinative reasoner's mind moves is *b* and the movement ends at a point *y*. We need to recognise that at least theoretically *x* and *y* could be different end points. In other words, we need to recognise the theoretical possibility that the conclusion of ratiocination and non-ratiocinative reasoning take different forms. We cannot unquestioningly assume that the conclusion of ratiocination is the same as that of non-ratiocinative reasoning.

Here, my focus is on the conclusion of ratiocination. I will argue that the conclusion of ratiocination takes the form of *I ought to believe p*. I leave open what form the conclusion of non-ratiocinative reasoning takes. But if I am right, I can at least reject the standard assumption that the conclusion of all cases of theoretical reasoning is simply *p*. In the following, I will turn to consider more closely how a reasoner's mind moves when she ratiocinates.

3.2 Ratiocination

By 'ratiocination,' I mean the kind of theoretical reasoning that is self-consciously directed by the reasoner. Some earmarks of ratiocination are, for example, when one says or thinks to herself, 'How should I think about all of this…,' 'Let's reason this out…,' 'What should I believe here?' Descartes's inner dialogue in the *First Meditation* is a good example of ratiocination. But it is not always the case that one deliberately initiates the process of ratiocination. Sometimes people simply start ratiocinating.

There is no constraint on the subject matter of ratiocination. Perhaps certain subject matters – e.g., whether God exists and whether it is morally permissible to eat meat – are more frequent subject matters of ratiocination. But we do not just ratiocinate when we are about to write an opinion piece or make an important decision. It is perhaps better to consider a mundane case to show the pervasiveness of ratiocination. We can imagine an everyday case where someone ratiocinates and her inner dialogue sounds like this: 'Do I believe that my wallet is at the restaurant? The last time I used it was at the restaurant when I paid the dinner bill. And I drove straight home after dinner. I just checked, my wallet is not in the car. I left my wallet at the restaurant'.

This mundane case tempts us to consider what goes on in the subject's head. But we should resist this temptation. When writing the above example, I am forced, to communicate effectively, to present a kind of inner dialogue. Yet, even if there were such a dialogue, it may be a mistake to focus on it. We have to remind ourselves that our task is to describe her mental activity, and the description might distort what actually goes on. It is possible that a certain mental process and the output of that process depart. To be sure, the processes might be different from how the subject herself might describes it, if she were pressed. We have to be open to the idea that theoretical considerations cannot always defer to the subject's own take on what's going on in her mind.[40] For example, one might have the thought that 'I had a haircut last week' but as describer, to give a fuller description, we have to say she is remembering that she had a haircut last week. Pinning down a mental activity is difficult because we have to both consider how things seem to the subject and also what the mental process is. In the existing literature, there is a tendency to either pay attention just to the sentences one produces in one's head or to a theory of reasoning that is devoid of the first-personal view. This thesis stands somewhere in the middle. I will try to capture how a subject's mind moves when she ratiocinates, both in terms of the subject's perspective and the nature of the activity itself.

When one ratiocinates, one reasons self-consciously. Reasoning self-consciously is not the same as reasoning consciously. For one to reason self-consciously is for one to be conscious of the fact that she is reasoning. This is more than just reasoning consciously. A child might reason consciously that the playground is wet because it just rained without being conscious of the fact that she is reasoning. One who is engaged in ratiocination, by contrast, is necessarily conscious of the fact that she is engaged in reasoning. Being self-conscious of a state or process that one is in alone does not amount to self-consciously directing that state or process. I can be self-conscious of the fact that I am blushing without being able to self-consciously direct my blush to fade. But reasoning is the kind of process which, when self-consciously done, is self-consciously directed. To see this point better, we need to first remind ourselves what reasoning, as in reasoned change of view, is about. In Harman's words:

[40] I am grateful to Rory Madden for helping me articulate this point.

> Reasoning in the sense of reasoned change in view should never be identified with proof or argument; inference is not implication. Logic is the theory of implication, not directly the theory of reasoning.[41]

An argument is logically valid when the conclusion is entailed by the premises. Entailment is logically necessary consequence. To say that *B* is entailed by *A* is to say that, if *A* is true, there is no way for *B* to be false. An inference is not guaranteed to be true by the premises. It is a consequence one draws out on the basis of evidence. To say that p is an inference of *r* is to leave open the option that *p* can be false.

Deduction is the process in which one draws out the implication of premises.[42] There is just one way to do deduction. Let us assume that *A* entails *B*. Imagine a student in her logic exam. She deduces *B* from *A* but does not need to believe *A* nor *B*. It might only be using *A* as a hypothesis to draw out an implication or use it to investigate the implicative relationship between propositions. Even if she accepts *A*, there is no change in view when she deduces *B* from *A*. She simply draws out the logical implication from *A*. This is not to say that the logic student will always succeed in drawing out the implication. She might get stuck in the process of deduction, not knowing how to apply the rules, or make a mistake in her application of the rules. One who engages in deduction is like being in a one-way tunnel, there is just one path, it is a matter of whether she can make it to the end or not. If she fails to make it to the end, we do not say that she takes the wrong turn. The process of moving to the end of the tunnel stopped. When one makes a mistake in deduction, the mistake is not characterisable as deduction done poorly; it simply is not deduction at all.

Theoretical reasoning is the process in which one draws out an inference on the basis of evidence. One reasons when one thinks that the evidence for *p* does not immediately settle the question whether p. Even if we assume that the evidence for *p* is in fact conclusive, the evidence is not wearing *p* on its face for the reasoner. The reasoner has to take the evidence to be evidence for her in her reasoning. She has to make sense of the evidence and draw a conclusion about *p* on the basis of the evidence she possesses. Hence, two reasoners' minds will move differently and change their minds differently even though they both believe *q* and believe *r*. The courtroom, for example, is a place where different reasoners take the same set of evidence in different directions. One who engages in theoretical reasoning is like finding a

---

[41] Harman 1986, p.10.
[42] Rumfitt 2015, for example, uses 'deduction' to refer to the activity in which 'a thinker engages in the task of tracing out the implications of some premises' (p.35).

destination, even if we assume there is in fact one path to the destination, the reasoner herself sees many possible paths available to her. If she in fact fails to find the destination, she might still take herself to have reached the destination. We say that she might have made a wrong turn somewhere. The process of finding the destination never stopped; it's just that one made a wrong turn. When one makes a mistake in theoretical reasoning, the mistake is characterisable as reasoning done poorly. In short, there can be no mistakes *in* deduction but there can be mistakes *in* theoretical reasoning.

In everyday life, we often apply rules of deductive logic in our reasoning. For example, we may imagine a subject ratiocinating: 'Is the grass wet? It is raining. If it is raining, the grass is wet. The grass is wet.' Philosophers usually call this deductive reasoning. This is a merely terminological difference. This still counts as theoretical reasoning in the way I use the term. For the way the reasoner's mind moves when she applies the *modus ponens* rule mirrors an inferential relation, not an entailment relation. I draw on Harman's distinction between induction and deduction.[43] According to Harman, 'The "conclusion" [of reasoning] is not a deductive consequence of the premises; it can only be "made probable" by them. Adding premises in this case *can* undercut the argument.'[44] The conclusion 'The grass is wet' is an inference she made based on her evidence from the world, that it is raining and that if it is raining, the grass is wet. An inference drawn from evidence is not guaranteed to be true. Suppose she checks the grass and realises that it is not wet, she will change her view to 'The grass is not wet' and revise her belief that if it is raining, the grass is wet. By contrast, the validity of a deductive argument cannot be undercut by adding premises.

1. If it is raining, the grass is wet.
2. It is raining.
3. The grass is wet.

If we add a premise 'The grass is not wet,' the argument is still deductively valid. But a rational reasoner will recognise that the conclusion is absurd and will not believe that the grass is wet.

The unsettledness of *p* in the reasoner's mind opens up many possible ways for her to move her mind. Let us assume that *q* and *r* in fact provide good reasons to believe *p*. It could

[43] Harman 1986, pp.3-6.
[44] *Ibid*., p.4.

nevertheless turn out that not-*p* is in fact true. Even if a reasoner believes *q* and believes *r*, she still has the option to believe *p* or believe not-*p*. If she does come to believe *p*, then the reasoner acquires a new belief. From the perspective of the reasoner in theoretical reasoning, the conclusion is what she believes to be a fact, which is inferred from what she believes to be another fact. We may correct her belief that *q* or belief that *r* or both, we can also correct her belief that *p*. If for some reason, she thinks that she should infer not-*p* from *q* and *r*, we do not immediately consider the person to have failed to do theoretical reasoning even if *p* in fact turns out to be true. We might say she has reasonably confused falsehood for truth. Hence, theoretical reasoning is in the realm of beliefs. There is room for reasonably forming a false belief. But in deductive reasoning, there is no room for ~*B* if *B* is entailed by *A*. One either successfully completes deductive reasoning and gets to the entailment or fails. Again, the failure here is not a case of bad deductive reasoning, rather it is simply not deductive reasoning. In the following, I will use 'reasoning' to refer to theoretical reasoning.

The validity of an argument does not map onto the way a reasoner's beliefs stand in relation to one another in ratiocination. The difference is that a reasoner forms beliefs by moving her mind through options. This still is the case even if logic is supposed to constrain the way her mind moves. In the realm of logic, there are no options. A logic machine follows rules and moves towards entailments. The logic machine can break down but it cannot take a different turn to a different conclusion other than what is entailed. Suppose someone's mind is equipped with a logic machine which the reasoner can use to help form beliefs. Logic can help a reasoner eliminate options. Nonetheless, the reasoner's mind is moving in a way that eliminates options. This is still in the realm of reasoning. A reasoner's mind that is equipped with a logic machine still applies rules and makes inferences. She still is moving through options and her mind could have taken different turns. Because of this difference, even though one can make mistakes in both deduction and reasoning, the nature of the mistakes are different. To a logic machine, it will not make the kind of mistake that results from being confused by other options. Accordingly, when the logic machine gives the wrong output, we do say that it is wrong, but we do not make the additional, normative claim that it is wrong in the sense that it made a wrong turn. When the logic machine gives the correct output, we say that the output is correct without making the additional, normative claim that it made the right turn. To the reasoner, there are options presented to her, even if there is in fact only one acceptable way for her to reason her way to *p*. When a reasoner makes a mistake, she makes a mistake in the midst of options. It is the presence of other options that have 'distracted' her from what she should have inferred. The mistake results from exercising wrong discretion.

When a reasoner forms a false belief, we do not just say it is false, we make the additional normative claim that it was the product of making a wrong turn. When the reasoner forms a true belief through reasoning, we make the additional, normative claim that her mind made the right turns.

Since there are different possible ways for a reasoner to change her mind, then the question 'How should I change my mind' arises for the reasoner. There are some standards the satisfaction of which counts one as having changed her mind in a way that is appropriately responsive to apparent reasons. Here, I refer to these standards as 'requirements of reasoning'.[45] A reasoner, qua reasoner, is subject to the requirements of reasoning. Since one's reasoning has to satisfy the requirements of reasoning, reasoning has to be a controlled process. What I mean here is similar to what Burge says about 'rational control':

> As a critical reasoner, one not only reasons. One recognizes reasons as reasons. One evaluates, checks, weighs, criticizes, supplements one's reasons and reasoning….A non-critical reasoner reasons blind, without appreciating reasons as reasons. Animals and small children reason in this way. But reasoning under rational control of the reasoner is critical reasoning.[46]

It is possible that the controlled reasoning process occurs without the reasoner self-consciously controlling her reasoning. Imagine how a seasoned swimmer's bodily movements are in a sense controlled. Her arms and legs can move with precision even though she does not self-consciously control her movements. Or imagine how the actions of an ideally virtuous person is in a sense controlled. On one understanding of this notion, the ideally virtuous is so perfectly harmonised with moral requirements that she effortlessly acts in a way that complies with the demands of morality in any given situation. She does not need to deliberate and then try to get herself to comply with the demands of morality. But for a less than ideally virtuous agent, she needs to deliberate about what morality requires of her and then direct her actions accordingly. Likewise, for an ideally rational reasoner, her reasoning process is in a sense controlled. She reasons in a way that complies with the requirements or reasoning. Yet, for less than ideal reasoners, like many of us, we need to

[45] I use 'requirements of reasoning' instead of 'rational requirements' to avoid the impression that ratiocinator must have rationality as her goal. See discussion in 3.3.
[46] Burge 1996, pp.98-9.

deliberately control our reasoning. To exercise deliberate control, one has to be conscious of the way she reasons. Hence, we may speak of two senses of control involved in reasoning. There is a lower-order sense of control. One's reasoning counts as under lower-order control if it complies with the requirements or reasoning. There is a higher-order sense of control. One's reasoning counts as under higher-order control if the reasoner is deliberately controlling how reasoning goes. When one is reasoning non-ratiocinatively, we can still think that there is agential control involved. We do not have to suppose that non-ratiocinative reasoner is zombie-like or machine-like. But there is no deliberate control. Ratiocination necessarily involves deliberate control. In the following, I will elaborate on the way in which a ratiocinator is deliberately controlling her reasoning.

There are three related features of reasoning that are important to our understanding of ratiocination. First, to the reasoner's mind, *p* is unsettled. The reasoner has to work out whether *p* is the case on the basis of evidence. Second, reasoning is not something that just happened to the reasoner. The evidence does not immediately tell her whether *p* is the case. The reasoner has to actively exercise agency to make something of the evidence. If there is a chip built into someone's brain such that she will always produce the most rational answer, but she cannot control how her reasoning goes, then she cannot be considered as having engaged in reasoning. Third, reasoning is a norm-governed activity. One cannot take anything to be evidence that supports *p*. There are some restrictions on what can be taken as evidence that gives one reason to believe that *p*. In light of this, there are a set of normative requirements that need to be met for an activity to qualify as reasoning. The normative requirements are, in Grice's words, 'directives (the precise kind of which remains to be determined), the observance, or non-violation, of which is a desideratum.'[47] Because of the first and second feature, it is up to the reasoner which transitions she will make to reach her conclusion when reasoning. Because of the third feature, when the reasoner makes transitions in a way that complies with the set of normative requirements, what she is doing is considered reasoning.

3.2.1 Directing oneself to follow requirements

Let us now focus on the mind of a ratiocinator, *S*. Something is unsettled for her and she needs to make up her mind. *S* ratiocinates in order to work out whether *p* is true. She thinks to

[47] Grice 2005, p. 22.

herself, '*q*, *r*.' As mentioned earlier, reasoning does not just happen to *S*. It is an activity she has to perform. As Grice notes:

> Reasoning is characteristically addressed to *problems*….A *mere* flow of ideas minimally qualifies as reasoning, even if it happens to be logically respectable. But if it is directed, or even monitored (with intervention should it go astray, not only into fallacy or mistake, but also into such things as irrelevance), that is another matter.[48]

S's mind is at a crossroads, so to speak. She can take *q* and *r* in different directions. But she cannot just arbitrarily pick a way. *S* thus must consider which direction she should take the evidence. Since reasoning does not just continue by itself, if she does not direct her mind, her conscious reasoning will come to a halt. Hence, as soon as a subject becomes self-conscious of her reasoning process, she has to take charge of what she is doing.

An ideally rational reasoner is someone – whether she is self-consciously directing her reasoning or not – who gets to a conclusion rationally. But suppose that the ideal reasoner is self-conscious that she is reasoning. In that case, she is not simply observing how her reasoning unfolds in front of her mind's eye. If we think that thoughts will just stream in her mind and she is merely self-conscious of the streaming, then what she is doing cannot be considered as reasoning. As long as she is reasoning, she must be trying to sort out a problem that is not settled for her. And if she is self-conscious that she is reasoning, she has to self-consciously sort out the problem. Even though we know in fact there is just one way her mind will move, from her perspective, she still has to sort out how to move her mind. She still has to subject herself to normative requirements in the process of working out whether *p* is the case. So, even for an ideally rational reasoner, if she becomes self-conscious of her reasoning, she still will have to direct how her reasoning goes.

And even if she knows that she is ideally rational, she still has to direct her reasoning. Consider the difference between the following two cases: Mary knows that she is ideally virtuous. She is considering whether she should give an extension to a student and she becomes self-conscious of practical reasoning process. The fact that she is ideally virtuous does not help. She still has to consider the grounds for giving an extension and make a call on whether this student's request for extension should be granted. She then decides to grant the

[48] Grice 2001, p.16.

extension. Knowing that she is ideally virtuous, she has the reassurance that she did the right thing. Next, let us assume that a god knows that whatever she does makes that action right. If she grants an extension to the student, then it is right to grant the extension to the student. Then this god does not need to direct her action. She can act in whatever way she fancies and still end up doing the right thing.

The ideally rational reasoner parallels Mary's case. Even if she knows that she is ideally rational, as long as we assume that the requirements of reasoning are not made true by the ideally rational reasoner's reasoning process, then the ideally rational reasoner still has to direct her reasoning. Since the reasoner has to consider how to direct her reasoning, the sense of self is bound up in ratiocination. It does not mean that the reasoner has to be explicitly thinking to herself 'What should *I* believe? What should *I* make of the evidence?' This goes back to the point that what one takes her mental process to be and the nature of that process might not always agree. The reasoner in her mind could just be attending to the evidence. But the theorist should bear in mind the nature of the activity; in this case, the activity is one that is self-consciously performed and requires self-direction.

Grice seems to suggest that it is possible to monitor one's reasoning. I think the kind of monitoring he meant in the quote above implies the presence of self-directing, for he suggests that as soon as the ratiocinator realises that one's reasoning is not going in accordance with the requirements, she will direct it to what she deems as the correct way. This is compatible with what I have said. The kind of monitoring that is not possible on my account is monitoring as a descriptive enterprise – like a heart monitor. Just because an ideally rational reasoner's reasoning cannot in fact go astray, it does not mean that her reasoning is something that passively happens to her. It is not akin to monitoring her heart's beating. Reasoning is a controlled activity. Once it becomes self-conscious, one has to take over this control at the self-conscious level. And since reasoning is a norm-governed activity, one has to self-consciously control it in a way that she deems as complying with the requirements of reasoning. This is not to say that a ratiocinator cannot monitor her reasoning. I only mean to say that the kind of monitoring involved is a normative enterprise like a judicial monitor. Such monitoring is checking whether the activity of reasoning is carried out correctly against normative requirements.

Our next question then is how the reasoner directs herself. I assume that reasoning is a norm governed activity. At this point, we need only say that reasoning is governed by a set of requirements; we need not commit to the details concerning their precise content. Since the activity of reasoning requires the reasoner to be governed by the requirements of reasoning,

as long as she wants to continue the activity of reasoning, she has to direct herself to comply with these requirements. She has to at least think to herself that she directs herself the right way, 'right' in the sense that it is in compliance with what she takes to be the requirements of reasoning. Hence, ratiocination is the mental activity in which the ratiocinator directs herself to follow the requirements of reasoning. As long as one is ratiocinating, necessarily, she takes herself to be in compliance with the requirements of reasoning. This mental activity concludes when the ratiocinator thinks that she complied with the requirements and she has arrived at an endpoint. The end point is where one thinks one is required to arrive. Hence, the conclusion of ratiocination is of the form: *it is in compliance with requirements of reasoning to believe that* p.

To say that *S* is directing herself in accordance with requirements is not to say that *S* is consciously rehearsing the requirements of reasoning to herself. It might not even be clear to her what the content of the requirements are. Not everyone is trained to do reasoning, but this does not prevent one to direct themselves in a way that she thinks is required of her. Every bit of mental movement that is under self-directed reasoning is ordered under the reasoner's seeing the normative force of the requirements. Let R be requirements of reasoning in the generic sense and $r_1, r_2, r_3 \ldots r_n$ be the specific requirements of reasoning. There is a difference to one's mind between taking oneself to be following R and taking oneself to be following $r_1, r_2, r_3 \ldots r_n$. One does not necessarily consciously rehearse to herself 'because of r_5' to reach her conclusion. But it is from her conclusion that we can say she has appealed to r_5. A case will help illustrate the point.

A professor asks her graduate student to be in charge of the in-class test in her course. The professor is quite lax so she has not specified what the student is supposed to do. All the professor has said is that 'You need to help me with the in-class test.' The professor has not given much thought to what she wants the student helper to do. The only thing she has thought of is to have the helper grade the tests. But the student helper is very diligent. On the day of the test, as soon as the professor signals that the test has begun, the student starts walking up and down the aisle checking for cheaters, counting the number of students present, announcing the time, sorting the test papers in alphabetical order, and the like. The lax professor herself has never thought of these measures. But when she sees what the graduate student is doing, she is also not opposed to any of the things he does. From the perspective of the helper, when he was announcing the time, for example, he was not rehearsing to himself any requirement that he has to announce the time, for there was no such pre-existing requirement. Still, he took what he did to be what is required of him. He does not

reason from 'there is a requirement that I have to announce the time' to announcing the time. Rather, he simply reasons from 'I am required to announce the time' to announcing the time. It is from his announcing the time that we, the observers, can say he takes announcing the time as part of what he is required to do.

Similarly, a ratiocinator does not have to consciously think about which rule of reasoning she should apply when she ratiocinates. For some ratiocinator, she might think that every move she makes in ratiocination complies with R without thinking that she is applying certain specific rules. She does not have to reason from 'according to r_5, I should believe *p*,' to thinking '*p*'. She simply reasons from something like, 'I am required by R to believe *p*,' to thinking '*p*'. This is similar to someone who might think that morality requires her to φ without thinking what moral rule requires her to φ. Or imagine that someone might think that being a parent requires her to spend time with her child, but it does not occur to her that she is subjecting herself to a certain rule of parenting that says she has to spend time with her child. The point here is that the ratiocinator is self-consciously directing a normative activity. Some ratiocinators might rehearse rules of reasoning to themselves and guide themselves with those rules, some might not. Even for those who are not generating and guiding themselves with rules, they still are directing themselves in a way that they see as complying with R. So, broadly speaking, a ratiocinator has to subject herself to requirements of reasoning. 'Requirements of reasoning' can mean both the generic and specific senses, R and $r_1, r_2, r_3 \ldots r_n$. 'Subjecting herself', in the sense of a generic requirement, means the ratiocinator actively takes on what she regards as a normative activity; in the sense of a specific requirement, means applying a specific rule to herself. Since a ratiocinator is someone who self-consciously subjects herself to requirements of reasoning, her conclusion is in the form of *it is in compliance with requirements of reasoning to believe* p.

It is helpful here to highlight the distinction between a reasoner subjecting herself to requirements and a reasoner being subject to requirements. I mean, when it comes to ratiocination, to be talking about the former. To help illustrate the difference, it is worth considering the 'error constraint' proposed by Lavin. The error constraint states: 'a reasoner is subject to a principle only if the reasoner can go wrong in respect of it.'[49] I remain neutral about this constraint if Lavin is here talking about a reasoner being subject to a principle. What I am talking about is a reasoner *subjecting herself* to requirements, in which case the constraint does not hold. If we consider how a reasoner is self-consciously directing herself to

[49] Lavin 2004, p.425. It is unclear whether Lavin makes a distinction between these two.

be subject to R, it is not necessary that she in fact violates the requirements. Consider the practical case. By the error constraint, there cannot be a rule that says one ought not to run faster than the speed of light because no human being can go wrong in respect of it. Suppose an agent mistakenly takes there to be a rule that says she ought not to run faster than the speed of light. We can imagine that this agent is running faster and faster, and then cautions herself that she ought not to exceed the speed of light. Similarly, in the reasoning case, it is not possible for a ratiocinator who is ideally rational to go wrong. But as long as the ideally rational ratiocinator (who does not know she is ideally rational) is self-consciously directing her reasoning, she still has to subject herself to what she takes to be R. She does not recognise that she cannot violate R and will subject herself to R. In this case, the form of the ideally rational ratiocinator's conclusion is still in the form of *it is in compliance with requirements of reasoning to believe* p. (I will say more about this below in 3.3.)

It does not concern us whether a ratiocinator actually gets the requirements of reasoning right. One who ratiocinates might in fact get certain requirements wrong or fail to comply with others, but from her first-person perspective, she takes the principles she is applying to be the right principles and takes herself to be applying them correctly. A ratiocinator might not be able to answer the question 'Which rule have you applied?' but will answer positively to the question 'Are you complying with the requirements of reasoning in believing *p*?'

In this regard, there is a parallel between ratiocination and practical reasoning. To the practical reasoner's mind, there are courses of action available. Usually, what prompts practical reasoning is when the practical reasoner is not sure which course of action to pursue. Because the possibility of pursuing the wrong course of action is available—she might make a mistake—she has to work out which course of action she ought to take. Suppose one deliberates about what to do and decides to donate to charity. She might not be able to spell out the normative rule she is applying that moves her to donate, rather than spending the money on Christmas shopping. But this does not prevent her from thinking that what she is doing is what she is required by morality to do. Even if she just thinks to herself 'Let's donate,' the conclusion of her practical reasoning is in the form of *I ought to donate*. She might not be able to answer the question 'Which moral principle have you applied?' but she will answer positively to the question 'Are you required by morality to do this?'

Similarly, in ratiocination, the ratiocinator is in fact trying to work out which way her reasoning should proceed. What's vexing about understanding ratiocination is that we have to describe, from a non-first-person perspective, what goes on from the first-person perspective.

In one sense, we have to consider how a reasoner's mind moves, what it is like for her to ratiocinate. But we are also in the business of describing her mental movement. We have to consider what she takes herself to be doing but we cannot be 'in her head'. We can imagine *S* ratiocinating and her inner dialogue run something like this: 'The bank is closed on holidays. Today is a holiday. Therefore, the bank is closed today.' As we hear *S* producing these sentences, we might be led to think that the inference she draws is 'the bank is closed today', not 'I ought to believe that the bank is closed today.' However, we should not conflate what one says out loud, or in her head, with how her mind moves. It is one thing to describe what *S* is thinking to herself, it is another to describe the movement of her mind. From her first-person perspective, she might just think '*q*, *r*, therefore *p*.' But to describe what is going on in her mind, we have to remind ourselves that reasoning is in the realm of beliefs and that *S* is self-consciously directing herself to comply with the requirements of reasoning. Such self-direction is only successful when she takes herself to have complied with R. Hence, even if *S* might just think to herself at the end of her ratiocination 'therefore, *p*,' as describers of her mind, we say that her conclusion is still in the form of *I am required to believe that p*.

We may next consider the sense in which *S* takes herself to have complied with R. To head off a confusion, we are not talking about the actual scope of R. Rather we are interested in how *S* directs herself to comply with R.[50] We can first rule out: 'If I am required by R to believe *q* and believe *r*, I am required by R to believe *p*.' Imagine *S* is directing her reasoning to work out whether *p* is the case. If this is how S applies R to herself, then, in order to work out what she is required to believe with respect to *p*, she would first need to work out what she is required by R to believe with respect to *q* and *r*. This will lead to a regress. *S* would never be able to change her mind in a reflective, reasoned way.

S could direct herself to comply with R in the following ways:

UNRESTRICTED: R requires me to see to it that if *q* and *r*, then *p*.

RESTRICTED: *q*, *r*, R requires me to believe *p*.

[50] What I am saying here is not intended to weigh in on the current debates concerning deontic scope. Debates concerning the scope of the deontic operator concerns the norms themselves. But I am concerned with the ratiocinator's response to norms. Arguments against narrow deontic scope does not apply here.

When we hear the sentences that a ratiocinator produces when she ratiocinates, what she said might sound like UNRESTRICTED. But this is not helpful because she could be using 'R requires me to see to it that if *q* and *r*, then *p*' as a premise in one's reasoning. She could also be saying '*q*, *r*, therefore *p*.' We have to understand how she is directing the ratiocination process. On UNRESTRICTED, the normative requirement is attached to the relation between beliefs and inference. However, as explained before, we need to separate what normative requirements look like and what it takes for a reasoner to guide herself to comply with normative requirements. Let us assume that the deontic operator for R takes a wide scope such that it is R (if *q* and *r*, then *p*).[51] When a ratiocinator brings herself to comply with R, she still has to think along that '*q* is true, *r* is true, and given R, then I am required to believe *p*. She has to give R, the normative directive to herself. So, even if R has wide scope, when the ratiocinator follows R, she cannot take R to come in an UNRESTRICTED form. By her own lights, she takes R to be RESTRICTED. *S* believes *q*. Then given *q*, she is required by R to believe that *p*. RESTRICTED allows her to work out what she should believe.

3.2.2 Ought to believe p

I have explained how a ratiocinator's conclusion is in the general form of *I am required to believe that p*.

Let us now turn to analyse the specific form of one's conclusion in ratiocination. I follow MacFarlene and consider three types of deontic operator:

PERMISSION: *q*, *r*, I have the option to believe *p*.

[51] Those who think that logic is normative for reasoning typically reject narrow deontic scope. As Broome points out, a belief that *p* is not self-justifying (Broome 1999, p.405). The same point is reiterated in MacFarlane 2004, p.9. The objection is something like this: logic provides normative guidance to the reasoner, which has $A \vDash A$ as a theorem. Suppose the deontic operator has a narrow scope, then if one has the belief that the moon is made of cheese, one ought to believe that the moon is made of cheese. However, it is absurd to say that rationality requires one to believe that the moon is made of cheese. Another problem with thinking that the deontic scope of normative requirements is narrow is that if logic is normative for reasoning is that it licences one who has contradictory beliefs to believe in anything. On the side of classical logic, *A* and ~*A* imply anything. However, it is absurd to say that if one who believes that it is raining and believes that it is not raining, she ought to believe that the moon is made of cheese. For a defence of a narrow-scope account, see Schroeder 2014, Chapter 11.

REASON: *q*, *r*, I have a reason to believe *p*.

OUGHT: *q*, *r*, I ought to believe *p*.

We can eliminate PERMISSION because it is too weak.[52] If PERMISSION, then *S* thinks that given *q* and *r*, it is not against rationality to believe *p*. It does not say whether rationality requires her to believe *p*, nor does it say that it is against rationality to believe not-*p* or believe some other propositions. Since *S* is directing herself to conform to normative requirements, to be effective in providing such directives, she has to find what rationality requires of her. Sorting out what rationality permits her to believe does not terminate the process of ratiocination because the ratiocinator still has to find out whether it is the case that *p*.

REASON is also too weak. It implies that she has a defeasible reason to believe *p*. This still leaves open the possibility that *S* has some other reason not to believe *p* or believe not-*p*. Suppose *S* is the author in the preface paradox case. She sees that she has a reason to believe *p*: each of the propositions in the monograph is true. She also sees that she has a reason to believe not-*p*: at least one of the propositions in the monograph is false. REASON gets the odd result that *S* is rational whether she believes *p* or believes not-*p*. For *S* who is trying to satisfy normative requirements, the normative constraint cannot be so indiscriminate that whether she changes her mind one way or the other, she is still rational. We assume that *S* ratiocinates in order to work out whether *p* is true. Her ratiocination cannot conclude if she thinks that she has a reason to believe *p* and also a reason to believe not-*p*.

This brings us to OUGHT. OUGHT is strong enough such that *S* thinks that she is under a normative requirement. It does not leave a range of possibilities among which *S* will still have to figure out how she is to revise her view. It also is not indiscriminate between possibilities. It says *S* ought to believe *p*. *S* in the preface paradox case can work out what she ought to believe. Since we have bracketed what the content of normative requirements are, we cannot determine what the norms will require *S* to believe in the preface paradox case. It might be 'I ought to believe *p*.' It might be 'I ought to believe not-*p*.' It might even be 'I ought to believe both *p* and not-*p*' (note this is what S *takes,* perhaps mistakenly, R to require of her). Regardless, to *S*'s mind, there is a definite proposition she ought to believe. OUGHT

[52] I assume that we do not mean uniquely permitted.

eliminates all other possibilities for *S* and tells *S* what she should believe. It does not leave open some other option that *S* could have believed and remains rational. Hence, the kind of normative requirement that *S* has to take herself to have satisfied has to be in the form of OUGHT.

Next consider whether the reasoner takes herself to be required by norms of reasoning to believe or merely not to disbelieve (i.e. believe not-*p*).

BELIEVE: *q*, *r*, I ought to believe *p*.

DISBELIEVE: *q*, *r*, I ought not disbelieve *p*.

For reasons similar to those that motivated the rejection of PERMISSION, a response in the form of DISBELIEVE is too weak and does not capture how a ratiocinator moves. If *S* is operating with DISBELIEVE, she would think she is not violating the requirement of rationality if she does not believe that *p*. She will only be violating the requirement of rationality if she believes that not-*p*. Suppose *S* believes that *q*: it is not raining. She looks out the window and sees that it is raining. She then comes to believe that *r*: she sees that it is raining. Imagine after seeing that it is raining outside, *S* by DISBELIEVE thinks that she is required not to believe that it is not raining. But this means that she still cannot direct herself in conformity with normative requirements to work on what she is to believe. She might just decide to suspend belief about whether it is raining. But that just means ratiocination is not concluded. One has not yet worked out whether *p* is the case. The ratiocinator's response to normative requirements has to take a form that directs herself to a view about *p*. If after seeing that it is raining, she thinks that she is required not to believe that it is not raining, then she is not engaged in ratiocination. If she wants to continue ratiocination, she still has to work out whether it is raining. She will need to look at the evidence and direct herself in some way that will allow her to draw a conclusion about *p*.

BELIEVE is to be preferred because it requires *S* to respond to evidence and change her belief accordingly. If *S* sees that it is raining, then she ought to believe that the grass is wet. This will mean that requirements of reasoning require a reasoner to at least treat something as evidence for whether something else is true. It requires the reasoner, in accordance with her evidence for *p*, to take a stand on *p*.

To summarise, in ratiocination, *S* self-consciously directs herself to follow the normative requirements to arrive at a view about *p*. Ratiocination ends when *S* thinks that she

has arrived at an endpoint in a way that complies with requirements of reasoning. Her conclusion is in the form of *I ought to believe p*. We may schematically represent the conclusion as OB*p*.

It might seem counter-intuitive to claim that the conclusion is in the form of *I ought to believe p*. We can dampen the counter-intuitiveness of this claim by clarifying the OB-operator. *I ought to believe p* can be rewritten as *p is what I ought to believe*; but it cannot be rewritten as *it ought to be the case that p*. As explained earlier, the sense of self is bound up in the activity of ratiocination. *S* is directing her mind to make the transitions. These transitions flow from her beliefs. Hence, the conclusion is still about what *I* ought to believe. That said, the emphasis of the 'ought' in OB*p* is on *p*, not on believing. The conclusion gives content to what *S* ought to believe. It is not saying she ought to believe *p*. The focus is not on 'I' in the sense that, to maintain my subjective rationality, I ought to believe that *p*.

'Is it the case that *p*?' and 'What ought I to do?' are not the only two questions that can be asked. One can also ask 'What should the object on which an act perform?' And this, I suggest, is the question a ratiocinator asks when she is trying to consciously direct herself *to* sort out whether *p is the case*. Due to the nature of ratiocination, one has to guide one's mind to move in a way that complies with the requirements of reasoning, but the subject's focus is on whether *p* is the case. To help illustrate the difference, consider:

(1) 'What ought I to do?' 'I ought to eat something.'
(2) 'What ought I to eat?' 'I ought to eat a sandwich.'

The focus of the 'ought' in (1) is about doing something, namely, eating. The focus of the 'ought' in (2) is about the object, namely, the sandwich. (1) prescribes an action but (2) does not. (2) tells me what to eat on the condition that I eat something. In (1) if I do not eat, I violate the 'ought' but, on some reading of (2), if I do not eat at all, I do not violate the 'ought'. We can think of the answer to (1) 'What ought I to do' as necessarily including a verb that is about an act; whereas the answer to (2) 'What ought I to eat' involves a noun about a specific object. I only violate the 'ought' in (2) if I eat something but that thing is not a sandwich. Once we come to see the difference, then we can see the difference between:

(1*) 'What ought I to do?' I ought to eat a sandwich.
(2) 'What ought I to eat?' I ought to eat a sandwich.

It is easy to confuse (1*) and (2). (1) and (1*) ask the same question. As I explained above (1) and (2) are different inquiries, so (1*) and (2) are different inquiries. However, the answers to (1*) and (2) could be written in the same way or appear in the same way in a subject's thought. The sentences are not very helpful. What is important is to see whether the 'I ought to…' is the output of which kind of inquiry. (1*) is the output of an inquiry about what action to take, (2) is an output of an inquiry about the object on which an act performs. If one has sorted out (1*), one has also sorted out (2) but not vice versa. Similarly, there is a difference between

(3) 'What ought I to do?' I ought to believe *p*.
(4) 'What ought I to believe?' I ought to believe *p*.

The 'ought' in (3) is about believing whereas the 'ought' in (4) is about the object of belief. I violate the 'ought' in (3) if I do not go on to believe *p*. But my conclusion takes the form of (4), for in ratiocination my focus is the object of belief. I try to work out in ratiocination what it is that I should believe. I do not violate the 'ought' in (4) if I do not go on to believe that *p*. I only violate the 'ought' in (4) if I form the belief not-*p* or a belief that is inconsistent with *p*. It tells me what to believe but it does not tell me to believe. It is possible that rationality requires me to form true beliefs if I were to form beliefs yet permit me not to form a belief in some cases. If I am going to form a belief about *p* at all, then *p* should be the content of my belief. But it might not violate rationality if I am not motivated to form a belief about *p*.

In summary, the inference a ratiocinator makes at the end of ratiocination is *I ought to believe p*. The ratiocinator does not necessarily have an inner dialogue running that says, 'I ought to believe *p*.' She could simply be thinking '*p*.' But given the nature of ratiocination, we have to recognise when she thinks '*p*' she in fact is in a state of believing or judging that she ought to believe *p*.[53] Someone might be motivated to work out what she ought to believe concerning *p* simply because she wants to preserve her rationality. I am not concerned with such cases. I am concerned with cases of ratiocination where the subject wants to find out whether or not *p* is true. In the following, I will address some possible objections to my account and try to head off some potential confusions.

[53] I remain neutral on whether judging is believing. *Believing that I ought to believe p* should thus be read as consistent with both.

3.3 Possible Objections

3.3.1 Ratiocination and rationality

One of my main claims is that a ratiocinator directs her reasoning in a way that she sees as satisfying the requirements of reasoning. We should not take this to mean that one who ratiocinates always must have the goal to be rational or the goal to comply with requirements of reasoning. It is true that some ratiocinators might deliberate by thinking explicitly, 'What would a rational person believe?' 'What am I rationally required to believe?' in a way similar to how some agents might deliberate by thinking explicitly, 'What would a moral person do?' 'What am I morally required to do?' But ratiocinators need not consciously have the goal of being rational. Some ratiocinators can simply deliberate 'Is it the case that *p*?'

Not every ratiocinator wants to be what Kolodny calls 'subjectively rational' or what Scanlon calls 'structurally rational'.[54] Subjective rationality is when certain relations between one's own attitudes holds. Suppose one believes that *p* and believes that if *p*, then *q*, then it will be subjectively rational for one to believe that *q*. But if we consider the everyday psychological phenomenon of ratiocination, one does not normally ratiocinate in order to form a belief that is consistent with pre-existing beliefs. Often, it is precisely because one worries that she might not arrive at a true belief that she wants to be extra cautious in her reasoning. Imagine someone has the pre-existing belief that her friend has never committed any crime. Then allegations emerge suggesting that her friend committed a crime in the past. She wants to figure out whether the allegations are true, and she is worried that she could be biased towards her friend. Hence, she ratiocinates with additional caution, making sure she attends and responds to the evidence. If she merely wants to be subjectively rational, she can easily believe that the allegation is false, which is consistent with her pre-existing belief that her friend has never committed a crime. Maybe some people ratiocinate simply in order to maintain consistency with their pre-existing beliefs. But I think there are also many cases where one ratiocinates to work out whether something is true. Hence, we do not need to think of ratiocinators as those who are 'fetishising' rationality.[55] It is possible that for some of

[54] Kolodny 2005; Scanlon 2007.
[55] Kolodny 2005 discusses the worry that having the goal of being rational can seem 'fetishistic' (pp.546-7). Sections 1 to 3 of the paper provide arguments for why there are no reasons to be rational.

them, rationality is not the goal. If they were presented with the options of either taking a pill and forming the true belief *p* or ratiocinating about whether *p* is the case, some would choose to take the pill and form the true belief.

We do not need to speculate at this point why one is motivated to ratiocinate. We should not confuse what motivates *S* to ratiocinate with what motivates *S* in ratiocination. Ratiocination might occur more often in instances where the reasoner worries that she might reason poorly. For example, a department chair might be worried that she is biased because a job candidate is also a friend. She therefore ratiocinates about whether the job candidate is qualified. However, ratiocination is not limited to cases where non-theoretical reasons might distort one's judgment. One may also ratiocinate out of pure theoretical interest. For example, one might happen to have a thought about Canberra and wonder what season it is in Canberra in July. She can ratiocinate about the weather of Canberra given its location and infer that it is winter in Canberra in July. Some might think that rationality is the way to get at the truth, some might be in a context where there is some practical payoff for being rational, or some might have the self-conception that she is a rational person and feels she supposed to live up to this self-conception. Different situations might trigger different people to ratiocinate. Some might also ratiocinate more frequently than others because of differences in circumstances or dispositions. Our focus is not on why one is motivated to ratiocinate, but what it is for one to ratiocinate. My account remains open to different possible ways in accounting for why a subject can be motivated to ratiocinate.

It is not as though ratiocinators are more interested in being rational than non-ratiocinative reasoners and therefore direct themselves to comply with the requirements of reasoning. The point is that as long as a reasoner is self-consciously reasoning, necessarily, she has to direct her reasoning. And when she directs, she will direct in a way that she takes to be the way she ought to move her mind. Hence, when she reaches the conclusion, it must be the case that she thinks her mind has moved to the end point in compliance with certain requirements. (Whether she has landed on the correct requirements is another matter; one I am not concerned with here.) By contrast, a non-ratiocinative reasoner does not self-consciously direct herself. This does not prevent her from being governed by requirements of reasoning. So, when she arrives at a conclusion, she does not have to have the additional thought that her mind has moved as it should. But again, we do *not* need to suppose that the ratiocinator has set herself the task of trying to be rational. The ratiocinator is instead directing her reasoning in a way that she takes to be in compliance with requirements of reasoning.

One might object: In order to monitor and direct one's reasoning, one has to recognise the evidence she has and use it in her reasoning. Hence, one's reasons must be explicit, in the sense that she can consciously reflect on her reasoning. For example, if one were to ratiocinate about whether the ground is wet, she not only looks out and sees that it is raining, she also has to be conscious that she is using her belief that it is raining as a reason for her to think the ground is wet. Doesn't this qualify as having set herself the task of being rational?

Unlike a non-ratiocinative reasoner who is not self-conscious that she is using *q* and *r* as evidence, the ratiocinator has to critically use *q* and *r* in her reasoning and see if they should be revised. The ratiocinator has to be self-aware of the propositions that she uses. And recall that unlike a thinker in deductive reasoning who does not have to believe the premises but only uses the premise as a hypothesis to investigate implicative relationships between propositions, a ratiocinator has to believe the propositions she uses. So, it is true that a ratiocinator has to be aware of the propositions that she believes. However, this does not mean that she has to be self-aware that she is using her beliefs in her reasoning. Again, what we are doing is to describe the mind of the ratiocinator. It is a description of a first-personal activity. In the head of the ratiocinator, she does not have to make explicit to herself that 'I believe that *q* and I believe that *r*'. She can simply be thinking to herself '*q*, *r*, therefore *p*'. So, from her perspective, she may not be *trying* to maintain the consistency of her beliefs. But, from the theorist's perspective, we might say that *S* is working at the level of higher-order beliefs. If we were to describe what goes on in her mind, we may write out the process as: *S* believes that she believes *q*, *S* believes that she believes *r*, *S* concludes that she ought to believe that *p*.

Those who hold the transparency account of rationality might raise this question: If we assume that *S* is a ratiocinator who does not have the goal of being rational but is interested in working out whether *p* is the case and the 'ought' of rationality is transparent, why does *S* need to be directing herself to comply with the requirements of reasoning? Can't she simply be attending to the reasons for and against *p*?

My account is compatible with transparency account of rationality. Because by attending, in a self-conscious directed fashion, to the reasons for and against *p*, *S* is directing herself to comply with what she *takes* the requirements of reasoning to be. Moreover, the requirements of reasoning do not have to be the same as rational requirements if one holds that the latter are requirements for maintaining one's rationality. For example, some might think that if I believe *q* and if I believe that *q* entails *p* then to be rational I have to believe *p*. I might have no interest in whether *p* is true. Rational requirements require you to have the

beliefs that maintain your rationality. Requirements of reasoning require you to respond to reasons. Some think that there can be additional requirements to be subjectively rational that go beyond responding to reasons. Even if there are, it does not affect my claim.

Consider, for example, Kolodny's view that the 'ought' of rationality from the first-person perspective is transparent to the 'ought' of reasons. As he writes:

> From the first-person standpoint, the 'ought' of rationality is transparent. It looks just like the 'ought' of reasons. It is only from the second- or third-person standpoint that the 'ought' of rationality and the 'ought' of reasons come apart. For it is only from a standpoint other than the subject's that it is possible to distinguish what attitudes he has reason to have from what attitudes, as it seems to him, he has reason to have.[56]

It is possible that from *S*'s perspective, she is just responding to the 'ought' of reasons. She thinks that she is required by *q* and *r* to believe *p*. Even if she is in fact not required by *q* and *r* to believe *p*, she will subject herself to what seems to her to be the ought of reasons. *S* does not necessarily have the further thought that 'because of requirements of reasoning, I have to believe that *p*.' As explained above, *S* could simply attend to the evidence and think that the 'ought' of reasons mandates her belief that *p*. However, her conclusion is still not *p* because she brings herself to be governed by what she sees as the 'ought' of reasons. Hence, her conclusion is still in the form of OB*p*.

But an objector may continue to press the point: If *S* is not thinking something like 'because of requirements of reasoning, I have to believe *p*,' then can't we just say that one responds to reasons? It seems redundant to say that *S* is directing herself to comply with the requirements of rationality. A ratiocinator is just a reasoner who is self-conscious of the reasoning process. We do not have to think that she is doing the extra work of directing herself to comply with requirements of rationality. It is just like seeing that it is raining and believing that it is raining. Person A might not be self-conscious of her acquiring the belief that it is raining from seeing that it is raining whereas person B might see that it is raining, form the belief that it is raining and is self-conscious of the process from seeing that it is raining to believing that it is raining. The belief formation process goes on the same way. But B's being self-conscious of this process does not change the nature of the process. Similarly, while a non-ratiocinative reasoner reasons and forms the belief that *p*, a ratiocinator is

[56] Kolodny 2005, p.558.

someone who reasons and forms the belief and is self-conscious of the process. One's being self-conscious of the reasoning process does not change the nature of the reasoning process. So can't we say that when one attends to the evidence and thinks that the evidence supports *p*, she concludes that *p*. This process is the same for both non-ratiocinative and ratiocinative reasoning.

Again, the reason why the conclusion of ratiocination is in the form of OB*p* has nothing to do with whether the ratiocinator, from her perspective, is responding to reasons or the demands of rationality. It has to do with the nature of reasoning. Reasoning is the kind of activity that is norm-governed. Because reasoning is an activity, when one is self-consciously engaged in it, one has to think about how to direct it. Because it is norm-governed, one has to think about how to direct it in a way that she takes to satisfy the norms (i.e., the requirements of the activity). In short, once the reasoner become self-consciously engaged in this norm-governed activity, she has to direct herself in accordance with normative requirements. Since the ratiocinator is directing herself to satisfy the requirements of reasoning, the conclusion is in the form of OB*p*.

3.3.2 Ratiocination and normative requirements

The next possible objection is this: what if logic is normative for reasoning? Can't it be the case if *S* apprehends the inference '*q,* therefore *p*' as having the formal structure of the schema $q \vDash p$, then in believing *q*, *S* just believes *p*?

There is an ongoing debate about whether logic is normative for reasoning. Harman, for example, thinks that logic is separate from reasoning. Consider a valid argument $q, q \supset p \vDash p$. As Harman argues, even if one believes *q* and believes if *q* then *p*, it does not follow that one ought to believe *p*. It does not even follow that one may believe *p*.[57] Or consider the valid argument *q* and $\sim q \vDash p$. A contradiction implies any conclusion. But this does not mean that a reasoner who has contradictory beliefs may believe that a square is round. Or consider $p \vDash p$. While *p* implies *p*, the belief that *p* does not make the belief that *p* true. For example, one's existing belief that the moon is made of cheese cannot be what makes it the case that she should believe that the moon is made of cheese.[58]

[57] Harman 1984, p. 107.
[58] See Steinberger 2017 for a clear summary of objections to the claim that there is a normative relation between principles of deductive logic and reasoning.

Some philosophers will have no trouble accepting the distinction between entailment and inference. However, they argue, logic is normative for reasoning. They argue that logic puts normative constraints on a subject's beliefs about the relevant propositions. MacFarlane, for example, suggests that we can find a bridge principle from logic to normative requirements on believing. A bridge principle can come in different forms. The basic form of a bridge principle is one that says if a logical consequence holds between a set of propositions such that $q, r \vDash p$, then we can make some normative claim about a reasoner's beliefs with respect to q, r, p.[59] MacFarlane's view is that we should think of formal validity as 'a property of inference *schemata*' instead of a property of inferences. If a schema is formally valid and *S* apprehends the inference '*q, r*, therefore *p*' as having the formal structure of the schema, then *S* ought to see to it that if she believes *q* and she believes *r*, she believes *p*. [60] Those who agree with MacFarlane, or any view that holds that logic is normative for reasoning, might disagree with my claim that a reasoner's mind can move in different ways. They might say that if one is rational, if she believes *q* and *r* and apprehends the formal validity of '*q*, *r* therefore *p*,' then there is just one way for her mind to move, namely, believing *p*.

My account is neutral on whether logic sets requirements of reasoning. Even if it is the case that logic is normative for reasoning, it only affects how we understand the nature of R. If R is constrained by logic, then a rational reasoner will be in some way constrained by logic in her reasoning. For an ideally rational reasoner, suppose she believes *q* and believes *r*, and she apprehends "*q*, *r*, therefore *p*" as having the formal structure of a schema that is formally valid, then when she non-ratiocinatively reasons, her mind will move in accordance with R and believe that *p*. However, if we consider the case where an ideally rational reasoner is ratiocinating, then necessarily, for reasons explained above, she still has to direct

---

[59] MacFarlane 2004, p.6.

[60] MacFarlane 2004, p.22-24. The preface paradox poses a challenge to the view that logical validity normatively constrains beliefs (McKinson 1965). Consider an author of a well-researched book. She believes each of the propositions $p_{1\ldots}\, p_{n}$ she puts forward in her book. Since $p, q \vDash p \,\&\, q,$ if logic is normative for reasoning, she ought to believe in the conjunction of propositions $p_{1,\ldots,}p_{n}$. However, given that she is epistemically fallible, it seems that she should also believe that at least one of the propositions she believes is false; in other words, she should disbelieve the conjunction of propositions $p_{1,\ldots,}p_{n}$. MacFarlane's response to the preface paradox is that we have to accept that our logical obligations sometimes conflict with epistemic obligations, the obligations to believe that some of our beliefs are false.

herself in accordance with R. She has to reflect on *q* and *r* and the formal structure of a schema and then decides what conclusion she should draw.

The debates about whether logic is normative for reasoning raises questions about the normative relations between beliefs.[61] Our focus here is the activity of directing oneself in accordance with normative requirements. I am not concerned here with the content of normative requirements. I am concerned with the fact that a ratiocinator must take requirements to bind her in reasoning. Even if logic is normative for reasoning, at the level of ratiocination, it still takes the ratiocinator to subject herself to this norm. For it is this fact that makes it the case that the conclusion of ratiocination is OB*p*. At this point, we can bracket whether the 'ought' in OB*p*, one that the ratiocinator issues to herself after subjecting herself to R, is the same as the ought of reasons or whether it flows from logic.

Note that we do not have to normatively separate logic from reasoning, we only need to separate logic from one's consciously guiding oneself to satisfy the demands of reasoning. In ratiocination, one has to work out *p* by guiding oneself to follow normative requirements. Consider an analogy: suppose there is just one way to cook a dish. Alex knows exactly how to make the dish. When he gets to the final step, he adds a teaspoon of sugar. He does not need to consult the recipe and he is doing exactly what he should be doing. When he gets to the last step, he thinks, 'add a teaspoon of sugar.' Amy has not made the dish before and has to follow the recipe to the letter. When she gets to the final step, she looks at the recipe and also thinks, 'add a teaspoon of sugar.' We may assume that both Alex and Amy are interested in cooking the same dish, both have the same inner dialogue such as 'add a teaspoon of sugar,' and both perform the same sets of bodily actions. But, crucially, their minds are directed differently. Amy is directing herself to follow the recipe while Alex is not, even though Alex is in fact doing exactly the same as what the recipe prescribes. Even though Amy's inner dialogue might be exactly the same as Alex, she is directing herself to follow the recipe. If we are to describe Amy's mind, what she was thinking is in fact in the form of *the recipe says that I ought to add a teaspoon of sugar*.

It is important to separate the question of whether logic is normative for reasoning from how a ratiocinator's mind moves when she ratiocinates. We are not trying to figure out how her theoretical reasoning to a belief ought to be constrained. Rather, the point I want to stress is this: If one is engaged in an activity of bringing oneself to comply with requirements

[61] For example, Broome (1999) sections 2 and 3 discuss 'detaching' and 'non-detaching' relations.

of reasoning, the activity concludes when one believes that one has reached the endpoint in a way that is in compliance with the requirements of reasoning. Hence, the conclusion of this activity is in the form of *I ought to believe that p*. The norms of reasoning, whatever they are, govern reasoning. In the case of ratiocination, the reasoner self-consciously directs herself to be guided by the norms of reasoning. The norms of reasoning do not additionally direct the reasoner to direct herself to be governed by the norms.

The point I am making here is different from Steinberger's point about whether logic provides first-personal directives. Steinberger argues that there are three distinct ways in which logic is normative for reasoning: in its role as first-personal directives to the reasoner, as third-personal evaluative standards, or as third-personal appraisals on which praise and blame are attributed.[62] Steinberger might agree that MacFarlene's bridge principles cannot provide first-personal directives to a reasoner. However, our target of analysis is different. My focus is not on what normative requirements look like, whether they are first-person directives or not. My focus is on what a ratiocinator's response looks like, one that is made under the condition that she is directing herself to comply with normative requirements. Even if logic is normative for reasoning and provides first-person directives to the reasoner, such that if she believes *q* and *r*, it directs her to believe *p*, it does not by itself direct the ratiocinator during ratiocination. When one ratiocinates, one's mind does not mechanically follow the directives. It does not work like a function such that if it processes *q* and *r*, it generates the output *p*. In order for the norms to guide the ratiocinator during ratiocination, the ratiocinator has to direct herself to be subject to what she takes to be normative requirements.

3.3.3 Believing that I ought to believe p is not a stage of ratiocination

Many hold that if a ratiocinator thinks that rationality requires her to believe that *p*, but she herself does not see any reason to believe *p*, then she will not form the belief that *p*. This motivates them to think that a higher-order normative belief must be involved at some point of reasoning such that one moves from recognising that she ought to believe *p* to believing that *p*.[63] They might raise the following objection to my account: instead of saying that the higher-order normative belief is the conclusion, can't we say that the higher-order normative

[62] Steinberger 2019.
[63] Broome 2013, p.209.

belief is involved at some intermediate stage of the reasoning process? Won't this will allow us to make the less controversial claim that a conclusion of ratiocination is in the form of *p* and that the reasoner is in a state of believing that *p*?

There are at least two possible ways in which a higher-order normative belief is involved in some intermediate stage of ratiocination. One possibility is that the higher-order normative belief is recognised as a reason in ratiocination. A second possibility is that the higher-order belief operates in the background. In the following, I will argue that neither of these options are compatible with my account. My account is not a higher-order account of reasoning, and thus does not succumb to the problems associated with such accounts.

3.3.4 I ought to believe p is not a premise in ratiocination

The first option is held by what Broome calls the 'higher-order account of reasoning.' According to Broome, a higher-order account is 'any account of reasoning in which the content of a normative higher-order belief serves as a premise at some stage [of the process of reasoning].'[64] What Broome means by 'premise' is what I mean by reason. To avoid confusion between deductive reasoning and theoretical reasoning, I will continue to use the term 'reason'. The problems with the first possibility—that the conscious normative belief recognised as a reason in reasoning—is discussed in detail by Broome. I will not rehearse this discussion.[65] I will not assess Broome's criticisms of the higher-order accounts. Here, I would like to focus on explaining how my account is different from these higher-order accounts. What's crucial is explaining why the difficulties faced by such accounts are not faced by my account.

When Broome talks about 'higher-order normative belief', he means that the reasoner consciously believes that she ought to believe that *p*. This question is difficult to answer due to the ambiguity of 'higher-order'. Usually, when 'higher-order belief' is being discussed in this context, it is taken to mean a subject's belief about her own belief.[66] This gives rise to the main question: How is the conscious belief that she ought to believe *p* supposed to figure in one's reasoning? My account does say that a ratiocinator has to eventually come to be in the

[64] *Ibid*. Broome thinks the higher-order account is implicitly adopted, for example, by Korsgaard 2009.
[65] Broome 2013, Chapter 12.
[66] The 'higher-order' component in my explanation of ratiocination is a higher-order state but that does not always amount to a belief about one's belief. I will explain more in Chapter 5.

higher-order state in which she believes that she ought to believe that *p*. But my account is not a higher-order account of reasoning in two important respects: First, it does not hold that the ratiocinator treats her belief that she ought to believe *p* as a reason in her ratiocination. The higher-order normative belief is not involved in her ratiocination process as a reason that the ratiocinator reasons with; instead, it is the *conclusion* of the ratiocination process. As explained in the previous chapter, as soon as the ratiocinator worked out what she ought to believe, the activity of ratiocination ends. The mind's movement to believe *p* starts with the ratiocination but the final movement is not part of the ratiocination process.

Second, my account does not say that the ratiocinator has to consciously believe that she ought to believe that *p*. As explained before, to the ratiocinator's mind, she need not consciously think that 'I ought to believe that *p*.' To her mind, she might simply attend to evidence about *p* and consciously conclude: '*p*.' Therefore, the worry that she will from her own conscious perspective go on to reason with her conscious belief that she ought to believe that *p* does not arise. The nature of theorising invites a confusion that we need to avoid. When theorising we need to bear in mind that the ratiocinator is engaged in a conscious normative activity. If we are to give an accurate description of the state she is in, the state that she is in should be described as a higher-order state in which she believes that she ought to believe that *p*. To say that *S* is in a higher-order state in which she believes that she ought to believe *p* is not the same as saying that *S* has a conscious belief that she ought to believe *p*. On the higher-order account, a conscious belief that she ought to believe *p* is derived from what she takes to be reasons to be rational. On my account, *S*'s higher-order normative belief is not derived from premises. Rather, *S*'s higher-order normative belief is the result of the conscious normative activity that she's engaged in. The occurrence of the higher-order normative belief is explained by the nature of ratiocination. When ratiocination ends, one ends in a higher-order state in which she believes that she ought to believe *p.* Because of these differences between my account and the higher-order account, it cannot be the case that the ratiocinator is using her higher-order belief that she ought to believe *p* as a reason in her reasoning. So, on this front, my account does not face the problems faced by the higher-order account.

Another problem associated with the higher-order account concerns how one's consciously believing that *p* can bring about one's aiming to believe *p*. To answer this question, the higher-order account might have to say that by enkratic reasoning the reasoner

believes that if she believes that she ought to believe that *p*, then she aims to believe that *p*.[67] The problem with this, as Broome points out, is that it will require her to also believe that she believes that she ought to believe *p*. One will then have to believe that if I believe that I believe that I ought to believe that *p*, then I have to aim to believe that *p*. This will lead to a regress. My account has the advantage of avoiding the regress problem because it does not hold that one has to consciously reason with her belief that she ought to believe that *p*.

Another potential problem associated with the higher-order account is that it also has to explain how one's higher-order belief that *p* directly brings about one's believing that *p* in the process of reasoning. According to Broome, although it is conceivable how one's belief that she ought to believe that *p* will cause her to believe that *p*, such as 'enlisting the help of a hypnotist or by undertaking a programme of self-persuasion,' that movement from believing that she ought to believe that *p* or from aiming to believe *p* to believing that *p* cannot be a process of reasoning.[68] One has to rely on other means other than reasoning to come to believe *p*. My account holds that ratiocination ends with one's being in the state of believing that she ought to believe that *p*. The movement from believing that she ought to believe *p* to believing that *p* is not part of ratiocination. So, it does not face the difficulty of explaining how one can move from believing that she ought to believe *p* to believing *p* within the process of ratiocination. I will discuss more about how one can move from believing that she ought to believe *p* to believing *p* in Chapter 5.

One more difficulty that my account avoids but the higher-order account faces is how one arrives at the higher-order normative belief in the first place. To explain how one gets from *q*, *r* to consciously believing that *I ought to believe p*, it is difficult for the higher-order account to avoid saying that the subject has to attend to her reasons to be rational. Borrowing Broome's example, a reasoner who reasons with her higher-order normative beliefs will reason like this: 'It is raining. If it is raining, the snow will melt. Rationality requires of me that, if I believe it is raining and I believe that if it is raining the snow will melt, I believe the snow will melt. I ought to believe that the snow will melt…'[69] This invites the worry that the reasoner's higher-order normative belief is not brought about by object-given reasons (reasons that have to do with whether snow will melt) but by state-given reasons (reasons that have to do with whether I am rational). The reasoner is attending to what belief she should have to be rational, rather than whether the snow is melting.

---

[67] Broome 2013.
[68] *Ibid*., p. 213.
[69] *Ibid*., p. 219.

On my account, it is not necessary that the reasoner has to reason with 'Rationality requires of me that…' As explained previously, the ratiocinator could be just attending to evidence about *p* and sees herself as governed by the 'ought' of reasons. Probably for most everyday cases of ratiocination, one just wants to find out whether something is true, such as whether the bank is closed, not whether it preserves her rationality to believe that her neighbour committed a crime. A subject like *S* does not have the goal of being subjectively rational. She just wants to work out whether *p* is the case by reasoning. As long as she is conscious of the reasoning process, she has to also direct her reasoning in a way that she thinks is right. She does not even have to have specific rules in her mind. My claim is only that, for every transition she makes, the transition seems to her to be required. Once she thinks she has made all the transitions she is required to make, her ratiocination ends. Hence, *S*'s belief that she ought to believe that *p* is not 'derived' from a conscious belief about what rationality requires of her. Rather, it is in virtue of the conscious normative activity that she undertakes that the endpoint of her activity is a higher-order state in which she believes that she ought to believe that *p*.

S's ratiocination process in her head could sound just like this, '*q*, *r*, therefore *p*.' But for the theorist, we do not just pay attention to what she said in her head. Bearing in mind the nature of the activity she engages in—a higher-order self-directed normative activity—we describe the endpoint of this activity as a higher-order state in which *S* believes that she ought to believe *p*. Consider an analogy. Someone is thinking to herself, 'Since I had a haircut last week, I don't need a haircut this week.' This is not the same as thinking to herself 'Since I remember that I had a haircut last week, I don't need a haircut this week.' The latter raises the question as to how the state of remembering brings her to believe that she does not need a haircut this week, but the former does not. She is only considering whether she is due for a haircut. She does not have to be conscious of the fact that she is remembering that she had a haircut and treat the state of remembering as a reason. Similarly, *S* does not have to be conscious of the fact in thinking that '*p*' she is in the higher-order state of believing that she ought to believe *p*. It is for our theoretical purposes that we have to be precise about the state she is in.

3.3.5 The belief that I ought to believe that p cannot be a background linking belief in ratiocination

For those who do not want to say that the ratiocinator has to consciously believe that she ought to believe that *p* and treat that as a reason, they might suggest the second option: one's belief that she ought to believe *p* operates in the background of ratiocination.[70] Those who favour this option might say: since we assume that *S* is just attending to *q*, *r*, we can agree with you that a ratiocinator is at some stage of ratiocination in the state of believing that she ought to believe that *p* in virtue of the activity of ratiocination; however, the state that she believes that she ought to believe that *p* operates in the background through a rule or linking belief that looks like 'If I ought to believe that *p*, then infer *p*.' Understood in this way, the conclusion of ratiocination is still *p*.

I leave it open that there could be some linking belief operating in the background when one ratiocinates; however, the linking belief cannot be the belief that *I ought to believe p*. The reason for rejecting this possibility for my account is not that it requires a ratiocinator to have concepts such as belief or ought, although this could be another reason for rejecting this possibility. It is also not because the ratiocinator might arrive at the belief that *I ought to believe p* through some 'weird theory of rationality' such that her linking belief is 'weirdly-grounded.'[71] A ratiocinator is directing herself in what she regards as the requirements of reasoning. So, to her, the linking belief or background inferential rule would not be 'weirdly-grounded.'[72] This would allow my opponent to say that at least for some ratiocinators, they might use the belief that *I ought to believe* in a rule linking belief and therefore, for some ratiocinators, their conclusion is in the form of *p*.

Now turning to the second option. Whatever background rule operates in ratiocination, there cannot be a background rule that says something like, 'If you ought to believe *p*, infer *p*.' To follow such a rule, one has to work out what she ought to believe. But as soon as the ratiocinator works out what she ought to believe, the process of ratiocination ends. She has concluded the activity the nature of which is that she is consciously directing herself to work out what she is required to believe. So, there cannot be a background rule like 'If I ought to believe that *p*, then infer *p*' in ratiocination.

3.3.6 Ratiocination and truth

[70] This option is motivated by Carroll 1895 and Broome's discussion of linking belief in Broome 2013, Chapter 13.
[71] See Broome 2013, pp. 228-9 for discussions for these two worries.
[72] Broome 2013, p. 228.

A final objection. Even if we restrict the cases to those of theoretical reasoning that are self-consciously directed, it still sounds odd to say that the conclusion of theoretical reasoning is OB*p*. Theoretical reasoning is supposed to aim at what is true. Adler, for example, writes:

> Theoretical reasoning is directed to the content or proposition believed, and only directed secondarily or derivatively, to the agent or to his attitude. This structure is obscured if the conclusions of theoretical reasoning are taken to be of the form 'I ought to believe **p**'. But this is an error. Theoretical reasoning aims to answer whether **p** is the case, not whether I ought (ethically?, prudentially?) to believe it.[73]

The worry is about what question is being answered in ratiocination. It is often assumed that since theoretical reasoning is supposed to answer whether *p* is the case, the conclusion of theoretical reasoning is *p*. My account does not require us to reject the thought that theoretical reasoning is supposed to answer whether *p* is the case; it only requires us to unlink these two claims. It opens up the possibility that one, in trying to answer whether *p*, answers in the form of OB*p*. But OB*p* is not answering the question.

Those who agree with what I have said in section 3.3.1 will accept that it is not necessary for a ratiocinator to respond to the 'ought' of rationality. Still, some might worry that on my account, the ratiocinator answers the wrong question. And if the ratiocinator answers the wrong question, then she is responding to the wrong kind of reasons. This worry is motivated by Hieronymi's suggestion that the right kind of reasons bear on the question of whether *p*.[74] I am not sure if the question 'Is it correct for me to believe *p*?' counts as the kind of question on which wrong kind of reasons bear for Hieronymi, but I can see how some might argue that one who tries to answer this question is perhaps thought to be responding to the wrong kind of reasons. Let me try to articulate the worry. It might be put like this: Even if we assume that *S* is not trying to be subjectively rational and is only responding to the 'ought' of reasons, *S* is still trying to monitor and check her reasoning. For example, *S* is thinking that 'The sky is looking grey. It is very cloudy outside. But wait, these are not clouds. The weather forecast mentioned that there would be haze today. It is the haze that is making the sky look grey.' The 'but wait' indicates the moment when *S* reminds herself that she could have made a mistake in reasoning and prompts herself to check her reasoning. When she

[73] Adler 2002, p.4. See Owens 2002, Hieronymi 2005, Boyle 2009 for similar line a thought.
[74] Hieronymi 2005.

checks, she still has to respond to the 'ought' of reasons. We do not have to suppose that she has to consult a different set of evidence. Nevertheless, she is answering the question 'Is it correct for me to believe *p*?' instead of 'Is *p* true?'

This worry arises from a wrong conception of what it means for a ratiocinator to check her reasoning. The objector mistakenly thinks that the ratiocinator is checking her own reasoning as in the checker is verifying the accuracy of something. Imagine you put down your passport number on a form. Suppose further that a checker checks if you have put down your passport number correctly by looking at your original passport. She looks at your passport and then at what you wrote on the form. When you are copying down the numbers, you are answering the question, 'What are the numbers?' In assessing whether you are 'correct,' the checker attends to the same set of evidence as you do, namely, your passport. This is consistent with the thought that a ratiocinator only responds to what seems to her to be the 'ought' of reasons. But when the checker is checking what you put down, she is answering a different question, namely, 'Have you put down the numbers correctly?'

This way of understanding checking as verifying might mislead one to think that the difference between the copier and the checker's mental movements does not lie in whether they are attending to the same set of evidence, but in whether they are answering the same questions. When one is ratiocinating, one is one's own checker of reasoning and is trying to answer the question 'Is it correct for me to believe *p*?' One might think the 'ought' in OB*p* is a positive response to this conclusion and then worry that my account fails to capture what is characteristic about theoretical reasoning, which answers the question 'Is it the case that *p*?'

It is true that my account leaves it open that some ratiocinators might be in the checking mode and try to answer the question 'Is it correct to believe *p*?'. However, on my account, it is not necessary for a ratiocinator to answer the question 'Is it correct to believe *p*?' The checking-as-verifying model is misleading. When one monitors one's own reasoning, there is not a checking part of the mind looking into another part of the mind, seeing if the answer produced matches the evidence. Recall that reasoning occurs when *p* is unsettled for the reasoner. Even if the mind prompts itself to check, there is no answer printed somewhere for it to check against. So, eventually, the ratiocinator still has to answer the question 'Is it the case that *p*?' A better analogy is this: a student is asked in an exam to multiply without a calculator 1623 and 1932. The first time she calculated she got 3,135,636. But before she reports her answer, she thought she should double check. She has no answer sheet or calculator to check. All she can do is to apply the rules she thought she knows and calculate again. The question she is answering is still 'What is the product of 1623 and 1932?' rather

than the question ‘Is 3,135,636 the correct answer?’. Since *p* is unsettled and S in ratiocination is trying to work out whether *p* is the case, *S* still has to answer the question whether it is *p* even if she is prompting herself to check her reasoning.

In the checking passport example, the checker’s answer has to make reference to what you put down against the passport. The ‘ought’ involved in her answer concerns checking whether what you produced matches with what is in the passport. But in the checking math example, the checker’s answer does not make reference to what she already put down and then check it against something else. She is working out the answer by ensuring that the rules are followed correctly. The ‘ought’ involved in her answer concerns whether she has followed the rules correctly. Ratiocination parallels the math example in that the ‘ought’ part of OB*p* is about whether the ratiocinator has successfully brought herself to comply with what she takes to be the requirements of reasoning. It is not the case that the ratiocinator is making reference to what she believes and then checks it against *p*.

In this chapter, I have argued that a ratiocinator could be answering the question whether *p* but her conclusion of ratiocination is in the form of OB*p* because ratiocination is a self-consciously directed activity. An immediate worry one might have is that believing that *I ought to believe p* is not believing that *p*. At the end of ratiocination, a ratiocinator only has the belief that *p* is what she should believe. One has not yet changed her mind about *p* when ratiocination concludes. In the next chapter, I will discuss how believing that *I ought to believe* fixes the belief that *p* top-down and the possibility for a rational subject to conclude that *I ought to believe p* without believing that *p*.

Chapter 4
Top-Down Fixation of Belief

In the previous chapter, I argued that the conclusion of ratiocination is *I ought to believe p*. There still is a gap between believing that *I ought to believe p* and believing that *p*. In this chapter, I argue that ratiocination fixes beliefs in a top-down way. If the top-down fixation process is successful, one acquires the belief that *p*. However, sometimes this process fails. This chapter has two main sections. In section 1, I will explain in detail how ratiocination fixes lower-order belief from the top-down. I argue that one is motivated by her belief that she ought to believe *p* to form a belief that *p* and thereby intends to believe that *p*. However, since one cannot believe at will, one has to adopt some indirect strategy to bring about the belief that *p*. I will explain why my account does not conflict with the view that belief is truth governed nor with the view that we lack voluntary control over belief. In section 2, I argue that it is possible for the top-down fixation process to terminate before it brings about one's belief that *p*, and I suggest that sometimes this sort of termination is rational.

4.1 Top-Down Fixation of Belief

Let us start by considering how one forms the belief that *p* in response to her concluding that she ought to believe *p*. An immediate worry one might raise is that theoretical reasoning, including ratiocination, is supposed to lead to a change in view. On my account, even though one has reached a conclusion about what one ought to believe about *p*, one still has not acquired an attitude towards that conclusion. To be more precise, one has not acquired the relevant kind of attitude towards that conclusion. We can say that one does acquire an attitude towards that conclusion, namely, believing that she ought to believe that *p*, but that is not the relevant attitude. The relevant attitude is the belief that *p*. Generally, this process – from considering evidence to reaching a conclusion to acquiring a belief – is all part of the reasoning process. Yet, on my account, ratiocination stops when the ratiocinator arrives at the conclusion that she ought to believe *p*. The rest of the story, leading to the acquisition of the belief that *p*, is not part of the ratiocination process. This gives rise to the difficulty of explaining how one can come to believe *p* after she believes she ought to believe *p*. Broome succinctly captures the worry:

> But if there were these two stages, at the end of the first stage the conclusion would be parked somewhere in your consciousness, without your having any particular attitude towards it. We would have to explain how you then come to take up the attitude.[75]

Assuming that one cannot believe at will, at least not in the standard sense of believing whatever one pleases for whatever reasons one sees fit, we need to explain how *S* can move from believing that she ought to believe *p* to believing *p*.

We can explain how *S* can come to believe by appealing to *S*'s ability to respond to reasons (or apparent reasons). Recall that we assume *S* is an average rational person. Let us also assume that reasons (or what appear to her to be reasons) require her to believe *p*. (I will now leave the apparent reasons aside implicit). If *S* can attend to *q* and *r* and work out that she ought to believe *p*, then she must be responsive to reasons. As long as she continues to be responsive to reasons, after concluding she ought to believe *p*, under normal circumstances, she will go on to believe *p*. This is not the same as saying that *S* is responding to her believing that she ought to believe *p*. First, even though we say that *S* is in the higher-order state in which she believes she ought to believe that *p*, to *S*'s mind, she could just be thinking '*p'*. Suppose, however, she agrees with my account of ratiocination. She might have the additional thought that 'I am in fact believing that I ought to believe *p'*. Still, what she has worked out at the end of ratiocination is, by her lights, what facts give her reasons to believe what. However, that 'I believe I ought to believe *p*' is not one of the facts that give her reasons to believe *p*. After she has consciously worked out that epistemic reasons are on the side of *p*, which is the endpoint of her ratiocination, as long as she remains responsive to reasons, she will go on to believe *p*. So, this process from conclusion to believing *p* at least does not invite the problem that she is not responding to reasons that are unrelated to the truth of *p*.

What I have said is slightly different from what Scanlon means by 'judgement-sensitive' attitude.[76] For Scanlon, a judgement-sensitive attitude is sensitive to beliefs about the reasons there are for that attitude. Belief is a judgement-sensitive attitude. If a rational subject believes that she has a decisive reason to believe *p*, she believes *p*. On Scanlon's account, one has to recognise the reason-providing facts as reasons. My account is non-

[75] Broome 2013, p. 243.
[76] Scanlon 1998, p. 20.

committal on this front. It only says that the ratiocinator has to be responsive to reasons. As Parfit points out: 'We respond to reasons when we are aware of facts that give us these reasons, and this awareness leads us to believe, or want, or do what these facts give us reasons to believe, or want, or do.'[77] *S* only needs to be aware of the facts and respond to them. She does not have to further believe that these facts are her reasons for believing that *p*.

In the practical case, suppose I notice that I incorrectly typed 'ratiocination' as 'ratiocnation'. I will then correct my spelling. That I have misspelt this word gives me the reason to correct my spelling, but I do not have to further believe that my misspelling gives me a reason to correct my spelling. Similarly, in theoretical reasoning, suppose *S* is ratiocinating about whether the bank is closed today. She only attends to the fact that today is a public holiday and the fact that banks close on public holidays. If she is rational, in the face of these two facts, under normal circumstances, she will come to believe that the bank is closed today. She does not have to further believe that the fact that today is a public holiday and the fact that all banks close on public holidays give her reasons to believe that the bank is closed today. We only need to assume that beliefs can be formed by rational agents on the basis of what they take to be the facts that serve as the reasons for or against them.

Having worked out, in ratiocination, that the facts decisively support *p* is just for a rational subject to conclude that she ought to believe *p*. Her mind is in the state of believing that she ought to believe *p*. This belief that *I ought to believe p* then initiates a process to bring about the belief that *p*. I call this process 'top-down fixation' because a subject starts reasoning about *p* at the reflective level. Then, after concluding that she ought to believe *p*, she goes on to acquire the lower-order belief that *p*. The question, then, is this: How does one's mind move from believing *I ought to believe p* to believing that *p*.

One possibility we can rule out is that in concluding that *I ought to believe p*, one is simultaneously in the state of believing that *I ought to believe p* and the state of believing that *p*. Believing that *I ought to believe p* and believing *p* can happen so swiftly that it seems to *S* that there is no transition, but it does not mean that there in fact is no transition. Since *S* is engaged in ratiocination, she has to first work out what she ought to believe, and only after does she move to the corresponding belief.

We cannot say that the belief that *p* is necessitated by the belief that *I ought to believe p*. Believing that *I ought to believe p* is not the same as believing that *p*. Believing *I ought to believe p* has to do with what one thinks one is supposed to think about a subject matter,

[77] Parfit 2011, p. 493.

whereas *believing p* has to do with what one thinks about a subject matter. There is a gap between what one ought to believe and what one believes. Hence, in ratiocination, one has to first believe that she ought to believe that *p* and then move on to believe *p*.

One might ask: Is it possible that the belief that *I ought to believe p* directly brings about the belief that *p*? If *S* is responsive to reasons, can we not just say that her mind is moved by reasons to the state of believing *p*? In the practical case, one needs to sort out which course of action to pursue and deploy some strategy to get herself to do what she ought to do. In most cases, one has to have a plan to make the world fit one's mind. However, in the theoretical case, the belief formation process is involuntary. There is no strategy a subject can undertake to get herself to believe p. One does not pursue a strategy to get the belief 'it is raining' in their head. Even if we accept that OB*p* is the conclusion of ratiocination, can the explanation for how one moves from OB*p* to B*p* not just be what we normally think of how one comes to form a belief? Can we not say that as long as one is willing to let the world impact her mind, and that one takes ratiocination to be one of the ways she can find out about the world, then if her conclusion is *p* is what she ought to believe, she just believes that *p*?

As long as we assume that one cannot believe at will, then a ratiocinator cannot directly move from believing that *I ought to believe p* to believing that *p*. It is true that, on my account, I only say that when one is engaged in ratiocination, one is actively bringing herself to comply with requirements of reasoning. Once ratiocination has concluded, one might think that one is not consciously directing her mind anymore. Upon concluding, for a rational subject, a mechanism takes over and moves her mind to believing that *p*. However, even if we are just considering those who focus on *p* instead of those who focus on the state of their subjective rationality, '*p* is what I ought to believe' is not the kind of belief that will automatically move one's mind to believing *p*. Reasons for *p* might move the subject's mind to believe that *p*, but OB*p* is not a reason for *p*. So, one's belief that she ought to believe *p* by itself cannot move a subject's mind to believing *p*. If this much is right, our question is this: How does a ratiocinator's mind move from concluding OB*p* to believing *p*?

Recall from section 3.2.2 that the emphasis of the 'ought' in OB*p* is on *p*, not on believing. The 'ought' prescribes a particular object of the attitude formed, but whether you form the attitude is another matter. OB*p* only says that, on the condition that you form a belief, *p* has to be the content of your belief. If one does not go on to believe *p*, one does not necessarily violate the 'ought' in 'I ought to believe *p*'. For *S* might not go on to form a belief at all. This suggests that for *S* to move from OB*p* to B*p*, she has to be motivated to believe *p*. By 'motivated', I do not mean in the sense that one can motivate oneself to believe *p* for

practical considerations. I also do not mean that *S* wants to be rational, so she is motivated to believe *p* because forming this belief is what she takes to be the rational thing to do. There might be cases where one wants to be rational; so, if she believes that the balance of reasons favours her believing *p*, she will be motivated to believe p. However, I am only concerned with cases where the subject ratiocinates to work out whether *p* is the case. We do not have to suppose that *S* has the extra motivation to be rational. It suffices to say that her believing that she ought to believe *p* will have an impact on her mind. It is telling *S* that if you want to have a view on *p*, then reasons tell you that *p* is the case. For a rational subject who is disposed to conform to reason, she will be motivated to change her mind. The motivation here is the product of her desire to find out whether *p* is the case in a way that conforms her mind to reason. I am also not suggesting that belief generally requires one to be motivated to be in that state. I only mean that in the case of fixing belief by ratiocination, one has to also be motivated to believe *p*.

Perhaps we might say that *S* has a rational mechanism that is sensitive to reasons for not believing *p* which moves her mind directly to believe *p*. We could then say that this mechanism is sensitive to her believing that she ought to believe *p* and moves her mind directly from believing that she ought to believe *p* to believing *p*. On this explanation – appealing to a rational mechanism – one does not have to go through an intermediary state. Yet the appeal to a rational mechanism cannot do the required work. Assuming that we cannot believe at will, one cannot believe *p* simply because she is motivated to believe *p*. We cannot say that one will automatically go on to believe *p* because of some rational mechanism at work. Being motivated to believe *p* is itself not a reason to believe *p*. Hence, even if there is a rational mechanism that automatically moves one to believe *p*, the rational mechanism cannot be sensitive to one being motivated to believe *p* and move one to the state of believing *p*. We still need an answer to the question: How does a ratiocinator's mind move from concluding OB*p* to believing *p*?

My suggestion is that for *S*'s mind to move from being motivated to believe *p* to believing *p*, she has to adopt a cognitive strategy to bring herself to believing *p*. The particulars of this cognitive strategy might vary from person to person. Though the details might differ, all strategies will need to use an intermediary state of intending to believe *p*. At this point, we can leave the exact nature of this intention unspecified. This state of intending to believe *p* does not have to be conscious. All I want to claim is that a ratiocinator has the goal of believing *p,* and because some strategy is needed to get into the state of believing *p*,

one has to go through the intermediary state of intending to believe *p*. I will call this way of forming B*p* via OB*p* the 'top-down way'.

This explanation avoids the difficulty of saying that one's mind can move directly from believing she ought to believe *p* to believing that *p*, for it does not presuppose that one can believe at will. Moreover, it is relatively uncontroversial to say that intention is also reasons responsive. If I believe there is conclusive reason to drink more water, then, so long as I am rational, I form the intention to do so. Once *S* has concluded that she ought to believe *p*, she has reason to intend to believe *p*. Here we only need to assume the weaker claim that, generally, 'ought' provides reason for intention. If I ought to attend a meeting, I have a reason to aim to attend the meeting. One's believing that *I ought to believe p* provides reason for *S* to be motivated to form a belief about *p*. Given *S* is responsive to reasons, she will want to conform her mind to reasons (when nothing else tells her whether *p* is the case). If she believes that she ought to believe that *p*, then, barring the complications I will address in 4.2, she will normally come to believe that *p*.

It is difficult to say whether top-down fixation is at root practical or theoretical. Although there is a practical aspect, namely, intending to believe *p* and adopting a strategy to achieve the goal, the practical aspect is theoretically driven. The practical part is downstream. It flows from the subject's trying to sort out through reasoning whether *p* is the case. On my account, there are some overlaps between practical reasoning and ratiocination. One such overlap is that when ratiocination fixes belief top-down, the subject has to go through the state of intending to believe *p*. Another overlap is that one has to undertake some cognitive strategy, such as conjuring up image of *p*, to believe *p*. In this sense, there is some control the subject has to exercise to undertake the cognitive strategy, and the undertaking of the cognitive strategy no longer deals with evidence.

Nonetheless, there are important differences between top-down fixation and practical reasoning. A key difference is that a subject who engages in ratiocination is concerned with whether *p* is true. She is motivated to believe *p* because she is disposed to or wants to conform to reason. It is certainly not because one wants some practical benefits that one is motivated to believe *p*. In ratiocination, she tries to work out whether *p* is true. Ratiocination ends when one believes that she ought to believe *p*. If she is rational, her mind conforms to reasons. This initiates a top-down fixation process. She then intends to believe *p*. Since fixation is downstream of ratiocination, the focus is still on whether *p* is true. Thus, the goal is downstream of the question 'Is *p* true?', not the question 'What should I do?'. It is only when a proposition *p* is believed that it counts as her having worked out whether *p* is true.

Hence, the goal 'to believe *p*' is different in kind from the goal to perform some action, 'to *A*'. The goal 'to *A*' is the output of inquiring 'What should I do?', whereas the goal of 'to believe *p*' is the output of inquiring 'Is *p* true?'. Even though *S* takes on a cognitive strategy to believe *p* instead of just responding to evidence about *p*, the way her mind moves is downstream of her working out whether *p* is true.[78]

4.1.1 *Top-down fixed Belief and Truth*

It might seem, on the account on offer, that the ratiocinator is guided by her commitment as a rational agent to form the belief that *p* instead of a commitment to the truth of *p*. For, if correct, it seems that our beliefs are not always formed directly in response to the evidence. However, this does not mean that belief is not truth-governed. Those who defend doxastic exclusivity argue that one can only be motivated by truth considerations in deliberation over what to believe.[79] When one deliberates over whether to believe that *p*, one can only be motivated by whether *p* is true. Below, I will explain how my account is compatible with accounts that hold that belief is truth-governed.

The doxastic exclusivists maintain that deliberation can only be motivated by truth. One objection that defenders of exclusivism might level against my account is that the ratiocinator's belief that *p* is not motivated by truth but by what she thinks she ought to believe. Yet, what I have said so far is compatible with the view that the ratiocinator is guided by the truth of *p*. One way for one's belief that *p* to be motivated by truth is when one attends to evidence for *p* and directly infers *p*. But I disagree that it is the only way for one's belief to be motivated by truth. Being guided by requirements of reasoning is consistent with the ratiocinator's being motivated by considerations for whether *p* is true. As I have explained in the previous chapter, the conclusion that *I ought to believe p* is drawn on the basis of evidence for *p*. There might be cases where one is ratiocinating in order to work out

[78] Boyle (2009), for example, holds the view that we exercise agential control over our belief 'in believing' in a way parallel to how we exercise agential control over our actions because our actions are our doings. I agree with Boyle that we have control over our belief to the extent that in ratiocination we direct our reasoning. However, I disagree with Boyle that answering the question whether *p* necessarily amounts to believing that *p*. I also disagree with Boyle that the control one exercises over believing is rational control. As explained earlier, in the top-down fixation process, we cannot appeal to a rational mechanism doing the work. One has to adopt some indirect strategy that hopefully will bring about the belief that *p*.
[79] E.g. Owens 2003, Archer 2015, Sullivan-Bissett 2017.

what beliefs she should have to maintain coherence with her existing beliefs without being concerned about whether *p* is the case, but in many cases of ratiocination, one is just trying to work out whether *p* is the case.

The distinction between a ratiocinator and a non-ratiocinative reasoner does not lie in whether they are interested in the truth of *p*; rather, the distinction lies in their minds moving in different ways. Both the non-ratiocinative reasoner and the ratiocinative reasoner could be deliberating about whether *p* is the case. A non-ratiocinative reasoner's conclusion is arrived at from her mind considering evidence for *p* without monitoring how she is considering the evidence; a ratiocinative reasoner's conclusion is arrived at from her mind monitoring whether she is considering evidence for *p* in a way that complies with what she takes to be the requirements of reasoning. Since the form of the conclusion depends on the movement, the conclusion of a movement that brings itself to comply with normative requirements is in the form of *I ought to believe p*. Recall the cooking example from section 3.3.2. Both Amy and Alex are interested in cooking the same dish, and they performed the same bodily actions. Yet Amy has to direct herself in accordance with the recipe. As a result, Amy's and Alex's minds move in different ways. Since Amy is self-consciously directing her movements in accordance with the recipe, we have to describe everything she does as trying to bring herself in accordance with what the recipe requires.

My account stays neutral on the debate about whether belief has an aim and whether the truth aim can interact with other aims.[80] You might be motivated by non-epistemic reason to believe that your neighbour is innocent, for example, that it will be very awkward to see her again if you believe she is guilty but she turns out to be innocent. But my account does not say that the subject can deliberately be motivated by non-epistemic factors to believe. Of course, being motivated by 'I ought to believe that *p*' is different from the consideration 'I ought to avoid awkward situations with my neighbour'. The ratiocinator is not deliberately acknowledging non-epistemic factors as reasons to believe. One's interest in being rational can be truth governed in a way that one's interest in being a good neighbour is not.

---

[80] For a helpful summary of the debate about whether belief has an aim, see McHugh 2011. For arguments for belief as an aim, see for example, Williams 1973, Steglich-Petersen 2006, McHugh 2011, Velleman 2000. For arguments against belief has an aim see, for example, Owens 2003, Archer 2017. McHugh formulates demanding as: you cannot deliberatively form an outright belief in a proposition if you regard your evidence for that proposition as less than sufficient, where sufficiency involves more than having better or stronger evidence for the proposition than for its negation. You require what you take to be some high degree or strength of evidence, or some particular kind of evidence, for the proposition (2015, p. 1120).

Since being motivated by truth is compatible with a ratiocinator directing her reasoning in a way she deems to be in compliance with normative requirements, if there is a lingering worry about the conclusion of ratiocination being *I ought to believe p*, it cannot be that the ratiocinator is not motivated by truth. If there is a worry, then, it must have to do with the fact that the ratiocinator is directing herself to satisfy the requirements of reasoning. This strikes me as unproblematic. We may find fault with the content of the norms she applies, but that is a different issue.

In practical cases, we might say that it is less than ideal for an agent to *infer* that she ought to *A*, and then intend to *A*. We might want to give more credit to someone who can cook delicious dishes without having to follow a recipe or one who decides to save his wife without having what Williams calls 'one thought too many'.[81] Perhaps what is appealing about these practical cases is that the agent exhibits a special kind of virtue, that they can unreflectively do what they ought without having to think too much about it. However, it is unclear what this worry amounts to for theoretical reasoning. Given that we are fallible creatures, the questions whether something is the case constantly confront us and the conclusion is unsettled. We have to reason it out. Moreover, given we are self-conscious creatures, we sometimes are aware of our possible irrationalities and want to make sure we reason properly; at other times, we might not be able to help but become self-conscious of reasoning. As soon as we become self-conscious of it, we are in control of the reasoning process, and we have to direct it in a way that we think is in compliance with requirements of reasoning. We should not conflate the claim that theoretical reasoning aims at what is true with the claim that the conclusion of theoretical reasoning is '*p*' rather than 'I ought to believe that *p*'. A reasoner can be governed by truth and yet seek the truth in different ways. Some can attend to evidence and believe that *p*. Some can attend to evidence and first conclude that she ought to believe that *p*. After concluding that she ought to believe that *p*, she then comes to believe that *p*.

If the foregoing is correct, we should be more precise in discussions about reasoning. There is a substantial difference between ratiocination and non-ratiocinatively reasoning, and that difference bears on our understanding of how we form beliefs in response to evidence. While some form beliefs directly in response to evidence, some form beliefs indirectly in response to evidence via first concluding that I *ought to believe p*. Both ratiocinative and non-ratiocinative reasoner could be interested in finding out whether *p* is the case. Both could be

[81] Williams 1981, p. 18.

following the requirements of rationality. The difference lies in that the ratiocinative reasoner is self-consciously directing her reasoning to comply with the requirements of rationality. Again, I am not suggesting that the conclusion of theoretical reasoning is always in the form of *I ought to believe p*. My argument only applies to ratiocination. I leave it open what the conclusion of non-ratiocinative theoretical reasoning is.

4.1.2 *Top-down fixation and control*

That we sometimes adopt a cognitive strategy to fix our lower-order beliefs does not amount to a claim that we have voluntary control over our belief.[82] It is true that, on my account, there is some control involved in the process of forming belief. The ratiocinator has to actively direct herself in accordance with the requirements of reasoning, and as a result of that self-direction, we conclude that we ought to believe that *p*, and then attempt to go on to believe that *p*. I also agree that there is some parallel between ratiocination and action. In both cases, we are faced with possibilities and limited evidence, and we have to work out what belief or action is required. However, just because one's mind is not automatically moved to believe *p* and that she has to motivate herself to believe *p*, it does not mean that voluntary control over belief is necessary for getting oneself to believe *p*. One adopts a strategy to believe *p* because one conforms to reason, not because one is moved by practical considerations. No matter how attractive the practical payoff is, as long as the subject does not think that reasons require her to believe *p*, she is not able to motivate herself to believe *p*.

Since I assume that one does not have control over one's belief in a way that parallels how a rational agent can control her acts, there is no straightforward strategy that will invariably get one to believe *p*. In the practical case, if I intend to turn on the light, then the strategy for me is to flick the light switch. Under normal circumstances, this is the straightforward strategy that will enable me to achieve my goal. However, in the case of belief, the cognitive strategies available, such as conjuring up images in my mind or repeating *p* to myself, are not straightforward strategies that will get one to believe *p*. Even under normal circumstances, the strategy of conjuring up a feathered dinosaur might not work to get one to believe there are feathered dinosaurs. It might work for one subject but does not work for another. Since there are no straightforward strategies to get oneself to believe *p*, it is difficult to outline the form of the cognitive strategy that one will take to

[82] E.g. Broome 2013, Bennett 1990, Hieronymi 2006, McHugh 2012, Williams 1970.

believe *p*. One possible strategy is that one might try to get herself to believe *p* by believing that she believes *p*. I will discuss this possibility more in the following section.[83]

4.2 Early Termination of the Top-Down Fixation Process

4.2.1 *Incompetence*

I have now explained how one gets from believing that *p* is what she ought to believe to believing that *p* in normal cases of ratiocination. I now turn to a possibility that arises on my account: Since one has to move from believing *p* is what she ought to believe to believing that *p*, there is a possibility one rationally fails to believe *p*. In the following, I will consider three possible ways in which top-down fixation could fail even though the ratiocinator is not irrational.

One possible way in which top-down fixation might fail has to do with the transition from believing *I ought to believe p* to believing *p*. Since there is a procedure involved in moving from believing that she ought to believe *p* to believing *p*, the procedure might stop short of completion. In such cases, one's higher-order belief that she ought to believe *p* fails to fix the lower-order belief that *p*. My opponent might say: You are supposed to give us an account of how someone who is rational can fail to believe *p* if she believes that she ought to believe *p*. *S* has done all the reasoning work to get to the conclusion that *p* is what she ought to believe. If not for irrationality, how could this last step fail?

In response, we can start with the distinction between incompetence and irrationality. Being irrational is one way of being incompetent, but there are ways to be incompetent without being irrational. Here, I will only talk about incompetence that is rational. We do not have to think of incompetence as being unable to attain certain level of fineness, such as playing performance level piano or expert-level chess. Generally, one is incompetent when one fails to do something one aims to do and that can include cases where one fails to do something that one normally can successfully do. In the practical case, sometimes we think we ought to act in some way, we normally can act in this way, but when we go to perform this act, we fail. For example, one might want to pour water into the glass but spills it on the table. One might want to make a reservation at a restaurant but forgets to do so. In these

[83] I leave it open that it is possible that one can form belief directly under some circumstances, but I am only concerned with cases where the endorsement of content *p* is theoretical. Frankish 2007, for example, argues for direct activism.

cases, the agent is simply incompetent in the general sense. They do not fail because they lack some skills, but still, they fail to do what they aim to do. Just like how one can be incompetent in the practical domain, one might also be incompetent in the theoretical domain.

One might object that there is an important disanalogy between the practical and theoretical domain. For example, one might have a spasm and spill the water. There might have been a gap between aiming to make a reservation and making the reservation. One has to look up the phone number, call the restaurant, and wait on the line. Many things could have happened in between that could have interrupted or distracted the agent from carrying out the action successfully. The world does not always cooperate. However, in the theoretical case, one does not need such cooperation. One does not have to move across space and time to believe something. Why is competence relevant?

This objection is predicated on a mistake. Mental movements, assuming minimally that our physical state and mental state are related in some way, require our brain's cooperation. To make the move from believing that she ought to believe *p* to believing *p* possible, one's movements across time and space take place. As Snowdon points out, whenever there are two physical realisations and the second physical realisation is causally related to the first physical realisation, in principle, it is possible that something breaks down in the causal process such that the second physical realisation does not occur. All we need are some weak naturalistic assumptions that believing *p* is what she ought to believe and believing *p* requires different physical realisations.[84] So, just as my muscles might not go along with my attempt to pour water into the glass, something about the physical state might not go along with my attempt to believe *p*. In such cases, I am incompetent, but not irrational.

Cases where one fails to believe in accordance with one's conclusion of ratiocination are not restricted to physical breakdown. Another possibility is that something short-circuits the process from believing *p* is what she ought to believe to believing *p*. Recall that one ratiocinates only when her mind is unsettled as to whether *p* is the case and that, by her lights, her evidence is limited and can be taken in different directions. For a non-ideally rational subject, reasoning is one way to get to the truth, but it is not the only way. Suppose *S* is ratiocinating whether there are any physical objects and concludes that she ought to believe that there is no physical object. After she finishes ratiocination and looks around, she cannot help but see physical objects around her. In a case like this, it is possible that even though *S*

[84] Snowdon 2012, pp. 257.

wants to conform her mind to reason and intends to believe *p*, her strategy to get herself to believe *p* might fail. Her perceptual experience might short circuit the top-down fixation process and directly fix her belief that there are physical objects around her. Suppose *S* never revisited this question, then she might end up with the higher-order belief that she ought to believe that *p* but still believes that not-*p*. It is true that in this case, *S* fails to conform her mind to reasons, but this does not mean that she is being irrational. Her perceptual experience also gives her a reason for believing that there are physical objects. As long as we are considering everyday human experiences, reasoning is one but not the only way to fix one's belief. Most of the time, when the conclusion of one's reasoning clashes with what one perceives, it is rational to have one's belief fixed by perception. For example, if I conclude from reasoning that the grass must be wet because it just rained but I see and feel that the grass is not wet, it is rational for me to believe that the grass is not wet.

4.2.2 *Lack of Interest*

Another way one might not believe *p* after concluding that she ought to believe *p* is that she is uninterested in *p*. In ratiocination, one tries to work out whether *p* is true. Just because one tries to work out whether *p* is true does not mean that the subject is interested in *p*. It is possible that one is not interested or loses interest in *p* after concluding that she ought to believe *p*. Suppose S has a free afternoon and decides to attend a trial as an audience member and is ratiocinating about whether the suspect has committed a horrible crime. As the trial proceeds, she pays close attention to the evidence being presented. She then concludes that she ought to believe that the suspect has committed a crime. Let us also assume that she does not know anything about the suspect, and she was not invested in finding out whether the suspect has committed a crime. She just happened to be there that day because she was curious to see what an actual trial looks like. In this case, she might conclude that she ought to believe *p* but fails to be motivated to form a belief about *p*.

Even if one recognises that she has theoretical reasons to believe *p*, it might not violate rationality to not believe *p*. Given our limited cognitive capacity, it will be too much to believe every true proposition. Ratiocinating about *p* itself does not mean that the ratiocinator is interested in whether *p* is true. It only requires the ratiocinator to think that *p* is unsettled, and that she tries to work out whether *p* is true by self-consciously directing her reasoning. There are many true propositions – for example, over 395,000 babies are born on this day, the restaurant's phone number ends with an even digit, and there are seven blue pens and four red

pens in my colleague's office – about which we do not have a belief. We do not have a belief about these propositions, not because we cannot find out the truth about them nor because we are irrational, but because we are not interested in them. I accept the general idea behind Harman's Principle of Clutter Avoidance, which states:

> Clutter Avoidance: One should not clutter one's mind with trivialities.[85]

Given the limits of our cognitive capacities, even if we can believe infinite number of things, it does not violate rationality even if we do not believe all true propositions. I am making a weak claim here. I am not suggesting that one should form a belief about *p* only if one is interested in *p*.[86] I am only suggesting that if one is not interested in *p*, it is rational not to form a belief about *p*. We have a practical reason to be selective and focus on processing things that are more important.

It should be noted, however, that Clutter Avoidance is applied differently on my account. On Harman's account, Clutter Avoidance is a 'metaprinciple that constrains the actual principles of revision [of belief]'.[87] It applies directly to believing by allowing or discouraging one to believe certain propositions. If one is not interested or ought to not be interested in whether *p* is true, by Clutter Avoidance, one does not believe p. Harman's account does not have the resources to explain how in standard cases of ratiocination one does not go on to believe *p*. Since Harman does not think that the conclusion of ratiocination is in the form of *I ought to believe p* but rather in the form of *p*, he thinks that as long as one ratiocinates and concludes that *p*, one will still believe *p* even if it is just for a moment.[88] Clutter Avoidance cannot be figured in one's reasoning. Then how does Clutter Avoidance discourage one to believe *p* on Harman's account? Perhaps one way is to discourage one from ratiocinating about trivial matters. However, this is to assume that one should only ratiocinate about matters that interest them. This assumption is odd because one does not only ratiocinate in order to find out what interests them. Sometimes, one is put on the spot to

---

[85] Harman 1986, p.12.
[86] This is weaker than Harman's *Interest Condition (on theoretical reasoning)*, which states: 'One is to add a new proposition *P* to one's beliefs only if one is interested in whether *P* is true (and it is otherwise reasonable for one to believe *P*)' (1986, p.55).
[87] Harman 1986, p.15.
[88] In Harman's words: 'Once one is explicitly considering whether or not to accept a conclusion, one cannot decide not to [believe] on such grounds. One might rationally decide not to try to remember it, perhaps, but one cannot decide not to believe at least for the moment' (Harman 1986, p.15).

ratiocinate, for example, in class discussions, in playing games, or in conversations with friends about matters that do not necessarily interest them. Even though the subject matter might not interest them, they might have other reasons to complete the ratiocination process, for example, to be a good participant in class, to advance to the next stage in the game, or to help a friend think through something.

Another way, on Harman's account, is to say that Clutter Avoidance discourages one to retain the belief *p*.[89] This will require us to posit two changes in view, namely, forming a belief about *p* and dropping a belief about *p*. If the main motivation behind *Clutter Avoidance* is to avoid overloading one's cognitive capacities, then this seems to create more mental work. Moreover, it also invites the worry that practical considerations can motivate a change in view. Alternatively, Harman could say that dropping a belief that one is no longer interested in does not count as one instance of change in view. This does not seem to account for cases where one's dropping beliefs about things they are no longer interested clearly counts as change in view. Imagine a student who learns about the wires in a plug in a science class. She forms true beliefs about the wires and is able to fix a plug based on what she has learnt. Many years after the course, she is no longer interested in wires in plugs and has forgotten what she learnt. When she encounters a plug that is not properly wired, she does not know how to fix it. The simple explanation here is that there has been some change in her view. It is difficult to see how her forgetting does not count or initiate a change. Another alternative is for Harman to say that momentarily believing something does not clutter the mind as much. This explanation is also unsatisfactory because, first, it might have to posit momentary believing as its own state, which is not parsimonious; second, it is difficult to determine the duration of momentary believing, which does not create significant cognitive burdens; third, trivial things can add up. Believing two trivial things momentarily might not burden cognitive capacities but believing twenty thousand trivial things, even momentarily, probably will.

With regard to how *Clutter Avoidance* is applied, my account has a few advantages. First, on my account, one can still ratiocinate about trivial matters and observe *Clutter Avoidance* by not trying to form the belief that *p*. This can account for situations where one is dragged into ratiocination so to speak. There could be cases where one agrees to serve as an adjudicator of a debate competition and has to listen to debates about a topic she is not interested in or a secondary school student is asked in an exam to write an essay on a topic

[89] Harman 1986, p.61.

that she is not interested in, for example, whether Tyrannosaurs were feathered. These topics could very well be important to some people, but for some others, they might not hold any interest. They might ratiocinate and get to the conclusion but fail to be motivated to form a belief about *p*. Even if it is the case that, in fact, every true proposition requires one to believe it, but in practice, not everyone is able to believe all true propositions. We might fault them for something else, for example, not being interested in the things they should have practical interest in (let us assume here that the trivial belief does not facilitate theoretical reasoning in this particular instance of ratiocination), but we do not have to fault them for being irrational in not forming the belief that *p*.

A second advantage of my account is that it gives a more parsimonious explanation of the relevance of *Clutter Avoidance* than Harman's. It avoids positing two changes in view, one coming to believe *p* and one dropping the belief that *p*. It also avoids the difficulty of explaining how momentary believing is less taxing on cognitive capacities than believing. On my account, one simply does not go on to believe *p* because one is not motivated to form a belief about *p*.

A third advantage is that, on my account, *Clutter Avoidance* is applied to the motivated state, not to the state of believing. This avoids debates about whether practical considerations can encroach on theoretical considerations and motivate one to change one's view. Many of the above-mentioned difficulties associated with Harman's account have to do with *Clutter Avoidance* applying to the state of believing. Although, as a rational subject, *S* has a reason to be motivated to believe what she ought to believe, there could be other reasons for not acquiring a belief about *p*. By *Clutter Avoidance*, even if one concludes that one ought to believe *p*, if *p* is trivial to *S*, then *S* has a reason not to form the belief that *p*. In some cases, even when S has rightly concluded that *p* is what she ought to believe, all things considered, it is rational for *S* not to form a belief about *p*. Recall that the 'ought' is in the form of '*p* is what I ought to believe'. It does not prescribe an act. It only prescribes an object of an act, but whether you take the act is another matter. Perhaps most of the time, it is rational to go on to form a belief once one concludes that she ought to believe *p*, and the content of the belief formed is that *p*. However, in cases where *p* is trivial to *S*, it is also rational for *S* not to be motivated to form a belief about *p* at all. Given our limited cognitive capacities, we can rationally ignore some propositions that I am in a position to believe so as to better allocate our mental resources. One might not be able to be motivated by practical considerations to believe *p*, but one might be motivated by practical considerations not to form a belief.

4.2.3 *Believe that I believe that p*

Since there is no straightforward strategy to get oneself to believe *p*, different subjects might adopt different cognitive strategies. It is possible that some subjects might adopt the strategy of believing that she believes that *p* to get herself to believe *p*. This strategy is in fact a self-defeating one. The cognitive strategy that tries to get one to believe *p* is likely to be sensitive to the subject's registration of whether the goal of believing *p* is met. As soon as the subject believes that the goal is obtained, that is, she has the belief that *p*, she will stop trying to believe *p*. Hence, even if believing that one believes *p* might work in other cases to bring about the belief that *p*, in the case of top-down fixation, believing one has the belief that *p* will terminate the top-down fixation process. A subject who uses the strategy of believing that she believes *p* has the belief that she believes *p* without the corresponding belief that *p*. I will discuss more about the implications of this in Chapter 6.

In this chapter, I have argued that it is possible for one to be in a higher-order state in which she believes that she ought to believe *p* but does not end up believing *p*. It is possible that such a failure of higher-order fixation is not due to irrationality. The kind of phenomenon I am concerned with is close to discussions about doxastic *akrasia*.[90] The kind of doxastic *akrasia* standardly discussed in the literature is mostly concerned with akratic believing, that is, one believes that she ought to believe that *p* but believes that not-*p*. Doxastic *akrasia* may also be more broadly understood as one believes that she ought to believe that *p* but fails to believe *p*.[91] On my account, it is both possible that one believes that she ought to believe *p* but believes not-*p* and that one believes that she ought to believe *p* but does not believe *p*. However, the word 'akrasia' might be misleading, for it is usually used in the literature to mean weak-willed. Yet a ratiocinator like *S* might fail to believe *p* upon concluding that *p*, not because her will is weak nor because she has other desires that override her judgement. Rather *S* could fail to form the belief she believes she ought to have simply because she has certain psychological limitations that all human reasoners have. In some cases, as explained above, due to these psychological limitations, it would not be irrational to fail to form the

[90] Since I am only concerned with belief in this thesis, I will only discuss doxastic *akrasia* here. Since discussions of epistemic *akrasia* in the literature are primarily about belief, the term 'epistemic *akrasia*' is often used in the same sense as doxastic *akrasia* (e.g. Adler 2002, Greco 2014, Chislenko 2016).
[91] Owens 2002.

belief that *p*. To avoid confusion with discussions of *akrasia*, we may understand the kind of cases where top-down fixation process does not go through more generally and more neutrally as 'termination of top-down fixation' to include cases where a rational subject does not go on to believe *p* without assuming that there is irrationality or weakness of will on the subject's part.

Chapter 5
Reflective Transparency

In the previous chapter, I explained the top-down fixation process and argued that it is possible for this process to fail. In this chapter, I attempt to show why the possibility of top-down fixation failure also opens up the possibility that one believes that she believes *p* without believing *p*. To do so, I will draw on the transparency account of self-knowledge. I develop two main lines of argument. I argue (in sections 1-3) against the transparency account. I argue that one's consciously endorsing *p* does not necessarily lead to one's believing that *p* because top-down fixation could fail.[92] But this does not lead to a wholesale rejection of the transparency account. I then argue (in sections 4-5) that one's consciously endorsing *p* could lead to one's rationally believing that she believes that *p*. Putting these arguments together, we get the possibility that one might believe that she believes *p* because she consciously endorses *p* but does not believe *p* because one's conscious endorsement fails to fix the lower-order belief from the top-down.

In section 1, I discuss Moran's transparency account and argue that it is not clear what Moran means by answering a question about one's belief. There can be cases where one says 'I believe *p*' without being in the state of believing that she believes *p*. What Moran wants to say is that in answering *p* is the case, one is in the state of believing that one believes that *p*. However, as I argue in section 2, in answering that *p* is the case, one is not necessarily in the state of believing *p*. This shows that consciously while endorsing *p* is transparent to believing *p* from a subject's own perspective, it does not mean that the subject is right in believing that she believes *p*. In section 3, I argue that Byrne's version of the transparency account is also threatened by the possibility that one consciously endorses *p* without believing *p*. In section 4, I argue that the transparency account does still capture something important: namely, that in consciously endorsing *p*, it is rational for one to go on to believe that she believes that *p*. This is what I call the 'reflective' way of fixing self-ascriptive belief. In section 5, I argue that the transparency holds at the reflective level.

[92] My challenge to the transparency account does not apply to all cases of self-ascriptions, such as self-ascriptions of desires and intentions. It only applies to the self-ascriptions of beliefs that are made after the subject has deliberated about whether *p* is the case.

5.1 Moran's Transparency Account

Evans famously claimed that one makes a self-ascription of belief by looking to the world and answering a question about the world:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?'[93]

The basic idea is that, in answering a question about the world, one also answers a question about one's belief. Let us call this the Outward Looking method.

There are various ways of attending to outward phenomena. When someone asks, 'Is it raining?' one might apply the Outward Looking method by looking out the window to see whether it is raining. When someone asks, 'Did you see Uluru on your trip?' one might try to recall or look through photographs from the trip. When someone asks, 'Are there moral facts?' one might answer by reasoning. Among those who try to answer by reasoning, some might self-consciously direct the reasoning process. In other words, ratiocination is one way of applying the Outward Looking method. Here, we are only interested in ratiocination as one of the ways in which the Outward Looking method can be applied. So, if we only consider the case of ratiocination, we could imagine Evans saying something like this: Someone asks *S* 'Do you believe that *p*?' *S* ratiocinates about whether *p* is true and concludes '*p*' out loud. (I emphasise that she says '*p*' out loud because one could consciously conclude '*p*' but the conclusion is in the form of *I ought to believe* p.) S then comes to believe that *p* and ascribe to herself the belief that *p*. At this point, let us assume that the top-down fixation is successful. One does come to believe *p*. Still, how does concluding '*p*' not just bring about *S*'s belief that *p* but also her belief that she believes that *p*?

Moran's transparency account attempts to provide an explanation. According to Moran, when one self-ascribes a belief, one answers the question 'Do I believe *p*' by reference to how things are in the world, rather than to herself or her own beliefs. Moran is

[93] Evans 1982, p. 225.

not saying that the 'world-directed' question 'Is *p* true?' and the 'self-directed' question 'Do I believe *p*?' are equivalent.[94] He is saying that a rational subject should treat them as equivalent. This is so because the 'world-directed question' is to be answered from a deliberative stance in which the subject forms or endorses an attitude of hers. When a rational subject endorses an attitude of hers, she is also answering 'What am I to believe about *p*?' To say that 'I am to believe *p*' is to say that 'I believe *p*.'

If one tries to only answer the question 'Do I believe *p*' without making reference to whether *p* is the case, then one assumes what Moran calls a 'theoretical stance.'[95] In that way, one is alienated from the assertion 'I believe that *p*.' A subject who ascribes to herself a belief from a theoretical stance parallels the way in which she would ascribe a belief to someone else. Transparency can easily fail under such circumstances. Moran gives the example of a person who is convinced by the therapist that she believes that her sibling betrayed her.[96] She ascribes the attitude to herself. However, when she considers the 'world-directed' question as to whether her sibling did betray her, she is unwilling to assent to its content. In this case, transparency fails because there is a discrepancy between the way the 'world-directed' question and the 'self-directed' question is answered. The analysand is merely 'reporting' on a belief or 'describing' herself as feeling betrayed but she does not affirm the content that she is betrayed. In Moran's words: 'When the belief is described, it is kept within the brackets of the psychological operator, 'believe'; that is, she will affirm the psychological judgement 'I believe that *p*,' but will not avow the embedded proposition *p* itself.'[97] In such cases, the subject is alienated from her self-ascriptions. On Moran's account, the threat to transparency lies in one's taking a theoretical stance towards the 'Do I believe that *p*?' question.

There are two relevant questions here: (1) How does one normally come to self-ascribe a belief? (2) Why does the method identified in (1) work? Moran's answer to (1) is the Outward Looking method, and his answer to (2) is the Transparency Condition. The Outward Looking method describes how a rational agent comes to form a belief about *p*. As rational agents, we have the capacity to make up our minds and also to review and revise our attitudes. One comes to form a belief about *p* by deliberating about whether *p* is the case. The Transparency Condition says that one can then answer a question about what she believes the same way she answers a question about whether or not *p* is true. The question 'Is it the case

[94] Moran 2001, p.63.
[95] *Ibid*., p. 63.
[96] *Ibid*., p. 85.
[97] *Ibid*., p. 85.

that *p*?' is transparent to 'Do I believe *p*?' as long as one addresses the world-directed question from a deliberative stance. When one addresses the world-directed question from a deliberative stance, one's belief about her own belief has an 'outward looking' character that is rooted in one's deliberating and forming her own state of mind.[98] Moran's formulation of the Transparency Condition is as follows:

> A statement of one's belief about X is said to obey the Transparency Condition when the statement is made by consideration of the facts about X itself, and not by either an 'inward glance' or by observation of one's own behavior.[99]

Moran's account captures the way in which one normally comes to ascribe a belief to oneself. When one says that 'I believe that it is raining' one is normally not trying to make a statement about one's mental state but on whether it is raining. However, Moran's claim is not just saying that this is how one comes to make a self-ascription by employing the Outward Looking method. He is making the stronger claim that, when the Transparency condition is met, one gains self-knowledge. The transparency condition is supposed to explain the 'immediacy' of self-knowledge, that is self-knowledge is not based on any observations or inferences that the subject makes.[100] Why is that so? Martin raises a general worry that is applicable to our considerations of Moran's account here:

> …in relation to cases where the subject is self-ascribing beliefs which form part of how the world is for her, which mirror her point of view on the world, the subject is aware of aspects of the world, and her attention is drawn out into the world. Yet, by directing her eyes outward, so to speak, she gains knowledge of her own mind. Why should this be so?[101]

Moran's explanation for why looking outward entitles one to have a belief about her mind appeals to the questions that a subject tries to answer. According to Moran, the question 'Do I believe?' and the question 'Is it the case that *p*' are indistinguishable to a rational subject from her first-person perspective. Moran quotes Edgley in saying:

---

[98] *Ibid*., p. 64.
[99] Moran 2001, p. 101.
[100] Moran, 2001, Chapter 4.
[101] Martin 2000, pp. 117–118.

> [My] own present thinking, in contrast to the thinking of others, is transparent in the sense that I cannot distinguish the question 'Do I think that P?' from a question in which there is no essential reference to myself or my belief, namely 'Is it the case that P?' This does not of course mean that the correct answers to these two questions must be the same; only I cannot distinguish them, for in giving my answer to the question 'Do I think p?' I also give my answer, more or less tentative, to the question 'Is it the case that P?'[102]

As the questions 'Do I believe *p*?' and 'Is it the case that *p*?' are indistinguishable to a rational subject when she considers them, when she consciously believes that '*p*' she also consciously believes that 'I believe that *p*.' Moran wants to say that because the two questions are the same to the subject's mind, answering one, 'Is it the case that *p*?', just is answering the other, the 'Do I believe *p*?'

Considering the questions one tries to answer is sometimes helpful, but it does not always tell us the whole story. Let us assume that the questions 'Do I believe *p*?' and 'Is it the case that *p*?' are indistinguishable to the subject such that, for the subject, she answers the two questions in the same way. However, just because a subject answers the two questions in the same way it does not mean that she gives the right answers to both questions. Let us consider what amounts to an answer to 'Do I believe that *p*?' The question 'Do I believe that *p*?' can be phrased in two ways. It can be rephrased as a world-directed question 'Is it the case that *p*?' as Moran suggested; it can also be rephrased as a mind-directed question 'Am I in the mental state of believing that *p*'? The difference between the two is obvious when the question is asked in the past tense. If the question: 'Did I believe that God exists three years ago?' is posed to a recently converted theist, the question is clearly a question about the state of her mind instead of a question about the world. In saying, 'I believed then that it was raining,' for example, the subject takes herself to be saying something about her state of mind. In saying, 'I believe it is raining', the subject takes herself to be saying something about whether it is raining.[103]

When the question is asked in the present tense, the world-directed question and the mind-directed question seem to become one. This has to do with the nature of belief as truth

---

[102] Edgley 1969, p. 90
[103] Wittgenstein *Philosophical Investigations* II x, section 89.

governed. One believes what she takes to be true. One's taking *p* to be true is just one's believing *p*. However, just because a subject's mind cannot distinguish the world-directed question from the mind-directed question, it does not mean that they are identical questions. When the question is put forward in the present tense, we will need more context to tell whether the 'Do I believe *p*?' question is asked as a world-directed question or a mind-directed question. Compare a case where *S* considers 'Do I believe that it is raining?' when she sees that her flatmate is about to go out of the house without an umbrella with a case where *S* considers 'Do I believe that everyone else is a better parent than me?' when she is sitting at the doctor's office for postnatal depression screening. Intuitively, although questions are in the form of 'Do I believe *p*?' there is an important difference between the two. When she answers, 'Do I believe that it is raining?' her attention is on the world. If *S* says, 'I am in the mental state of believing it is raining,' her flatmate might respond, 'I asked about the rain, not about your state of mind!'[104] When she answers, 'Do I believe that everyone else is a better parent than me?' her attention is on her mind. If *S* answers, 'There is no evidence that shows everyone else is a better parent than me,' the doctor might respond: 'I am not asking about whether you are in fact the worst parent in the world. I am just asking about your state of mind.' The difference between *S*'s saying, 'I believe it is raining' and affirming that 'I do believe that everyone else is a better parent than me' is that in the latter case, she is ascribing a belief to herself.

As there are different ways of answering the 'Do I believe *p*?' question, an answer to 'Do I believe *p*?' does not always amount to a self-ascriptive belief. When our focus is on self-knowledge, we are interested in how one answers the 'Do I believe *p*?' question in a way that amounts to a self-ascriptive belief. We are interested in why a subject's rationally endorsing *p* makes true her belief about her mind. Note that I am not saying that self-ascriptions cannot be made by attending to evidence about the world. I am also not saying that we have to understand self-ascriptions of beliefs as reports on one's mental states. I am also not challenging Moran's claim that taking a theoretical stance towards one's mental state will result in alienation. All I am suggesting is that what counts as answering 'Do I believe *p*?', for the purposes of understanding the nature of self-knowledge, is that a subject has to be

[104] This example is inspired by Wittgenstein. Wittgenstein writes to Moore: 'To call this, as I think you did, "an absurdity for *psychological* reasons" seems to me wrong, or *highly* misleading. (If I ask someone "Is there a fire in the next room?" and he answers "I believe there is" I can't say: Don't be irrelevant. I asked about the fire, not about your state of mind!)' (Wittgenstein 1995, pp. 315–316).

in the state of believing that one believes that *p*. The following table can help to illustrate this point:

Question	**What is in fact asked**	**Answer**	**The state of a subject when answering**
'Do I believe *p*?'	'Is *p* true?'	'I believe *p*.'	Consciously endorsing *p*.
	'Am I in the state of believing *p*?'		Believe that I believe that *p*.

A subject can answer the question 'Do I believe *p*' in saying 'I believe *p*' but this does not mean she is in the state of believing that she believes *p*. It could just be the way we are taught to use the word 'believe'. Saying 'I believe *p*' is another way of saying '*p* is true.'[105] So, just because a subject answers, 'I believe *p*' it is not indicative of whether a subject is, in fact, in a state of believing that she believes *p* or the state of consciously endorsing *p*. For Moran's transparency account to work, it is not enough for him to say that one is in a position to answer, 'I believe *p*.' He needs to show that when one answers in the sense 'I believe I have the belief that *p*,' one is entitled to believe that one has the belief.

Moran holds that one is entitled to believe whether or not she has the belief that *p* if one has concluded deliberation about *p* and that one's self-ascriptive belief amounts to knowledge. For Moran's account to succeed in delivering the immediacy of self-knowledge, it is not enough for Moran to tell us why a subject makes a statement about one's belief. He has to tell us why such a statement about one's belief is an expression of self-knowledge. Even if he is right to say that our beliefs are determined by reasons, that is not enough. He has to show our reasons always determine our beliefs. It is only when our reasons always bring about belief that *p* that we can say one's endorsing *p* always brings about belief that *p* and hence his self-ascriptive belief must be true. Moran himself is aware of this. He relies on the claim that one is 'entitled to assume, or in some way even obligated to assume, that his considerations for or against believing P (the outward-directed question) actually determined in this case what his belief concerning P actually is (the inward-directed question), then he would be entitled to answer the question concerning his believing P or not by consideration of the reasons in favour of P.'[106]

[105] Heal 1994.
[106] Moran 2004, p.457.

We can question the assumption that the considerations for or against believing *p* determines one's belief about *p*. Moran has not gone into detail about what counts as deliberating about *p*. He uses various terms such as 'making up your mind,'[107] 'the thought that concludes my deliberation,'[108] 'not an open question for me,'[109] that suggests concluding deliberation is just believing that *p*. The following passage makes this point explicitly:

> One must see one's deliberation as the *expression and development* of one's belief and will, not as an activity one pursues in the *hope* that it will have some influence on one's eventual belief and will. Were it generally the case…that the conclusion of his deliberation about what to think about something left it still open for him what he *does* in fact think about it, it would be quite unclear what he takes himself to be *doing* in deliberating. It would be unclear what reason was left to *call* it deliberation if its conclusion did not count as his making up his mind.[110]

Moran's suggestion that the conclusion of deliberation is just believing that *p* is, as we have seen in the previous two chapters, too quick. Moran seems to have relied on the claim that concluding deliberation *p* is just believing that *p*. Since consciously endorsing *p* entails believing *p*, one is entitled to believe that she believes *p* if she consciously endorses *p*. It is only with this assumption that Moran can say we 'have a kind of access to one's beliefs that is not based on evidence of any kind.'[111] In the following, I argue that consciously believing *p* does not necessarily bring about the belief *p*.

5.2 Endorsing without Believing

I assume that ratiocination qualifies as one of the forms of deliberation that Moran has in mind.[112] When a subject utters '*p*' at the end of ratiocination, '*p*' is her thought that concludes deliberation. From the subject's perspective, she consciously endorses *p*. 'Consciously endorse' is only a loose expression here. We may also understand it as consciously affirming that *p*, consciously avowing that *p*, or consciously judging that *p*. As argued in Chapter 3,

---

[107] *Ibid.*, p.92.
[108] *Ibid.*, p. xix.
[109] *Ibid.*, p.74.
[110] *Ibid.*, p.94.
[111] Moran 2001, p. 84.
[112] *Ibid.*, p. 63.

from the ratiocinator's perspective, she could just be thinking 'It is the case that *p*.' Yet we the theorists are bearing in mind the nature of the activity she is engaging in and therefore note that she is, in fact, in the state of believing she ought to believe *p*.

When ratiocinating, one is concerned only with whether *p* is the case. She is not trying to rationalise her attitude nor trying to decide what belief she should have to maintain her own rationality. For example, let us assume that Rae is rational. Rae lost her wallet and she ratiocinates about where her wallet is. When she concludes her ratiocination, she sincerely endorses the proposition 'My wallet is at the restaurant.' In her inner dialogue, she could just be saying to herself that 'My wallet is at the restaurant.' Even though she in fact is in the state of believing that she ought to believe that her wallet is at the restaurant, this does not mean that she is giving a 'normative assessment or rational recommendation' of her mental state.[113] She is only making a rational assessment of whether *p* is the case. Hence, one way of being in the state of consciously believing that *p* is to be in the state of believing that *I ought to believe p*.[114]

I have also explained in Chapter 4 why a rational subject might not go on to believe *p* and how such a failure of top-down fixation is not necessarily irrational. This means that it is possible for a rational subject to sincerely endorse *p* but not believe *p*. In cases where *S* consciously endorses *p* after ratiocination but fails to believe *p*, one will be falsely believe that she believes *p*. This shows that even in cases where one is ratiocinating about *p* and where one's rationality is not compromised, one's conscious endorsement of *p* does not always lead to true self-ascriptive belief.

Moran also acknowledges that there are cases where one deliberates and concludes that *p* but fails to believe that *p*. However, he seems to think that these are either cases where the agent has not exercised authority over her deliberation from the first-person point of view or cases where the agent's rationality is compromised. If, for some reason, the conclusion of her deliberation says *p* but she does not believe *p*, then her rationality must be compromised. Moran writes:

---

[113] Moran 2004, p. 466.

[114] Finkelstein (2012) suggests that on Moran's account, when one takes the deliberative stance one is considering what she ought to believe. I do not claim that the ratiocinator has to consider what she ought to believe. It is simply by virtue of the activity she engages in that she ends up in the state of believing that she ought to believe *p* at the end of ratiocination.

> In that sort of case, all the deliberating or critical reflection may be so much rationalization, a well-meaning story I tell myself that has little or nothing to do with what my actual belief is. This is a familiar enough situation of compromised rationality.[115]

An agent exercises a particular authority over her deliberation in that she does not just make a 'normative assessment or a rational recommendation' of her belief. Rather she takes responsibility for her beliefs and determines, through deliberation, what to believe.[116]

To properly engage with Moran's account, we do not have to consider cases where one is not exercising authority over one's belief or cases where one's rationality is compromised. I only claim that it is possible that a ratiocinator takes the deliberative stance, and answers 'I believe *p*' without going on to believe *p*. The ratiocinator is doing everything Moran describes a rational agent does when she makes a self-ascription. She considers whether *p* is the case through careful reasoning, thinks to herself that 'It is the case that *p*,' and then goes on to make a self-ascription saying 'I believe *p*' in the sense that I believe *p*. However, the top-down fixation process fails to fix her lower-order belief that *p*. It will be wrong for her to believe that she believes *p*. The failure cases with which I am concerned result from the early termination of the top-down fixation process that is compatible with rationality. Although what Moran means by 'deliberation' is broader than what I mean by ratiocination, as long as we agree that in the case of ratiocination, conscious endorsement does not always bring about the corresponding belief, this still creates a problem for Moran's account because it is not clear how in considering a question about the world one also has a true belief about her mind. Minimally, this shows that one could fail to know one's mind by adopting the Outward Looking method in the case of ratiocination.

One might say in Moran's defence that even though ratiocination does not always succeed in fixing the lower-order belief, it is still the case that ratiocination generally succeeds in fixing lower-order belief. However, the reason one is then entitled to believe that she believes *p* cannot be that one must believe *p*. One knows one's beliefs because the top-down fixation process is generally reliable, not because one is entitled to form a belief that she has the belief *p* in the same way she forms a belief about *p*. One would have to say that since the top-down fixation process is generally reliable, one is entitled to *infer* that she

---

[115] Moran 2004, p. 466.
[116] *Ibid.*

believes *p* if she consciously endorses *p*. However, Moran wants to say that one's knowledge that she believes *p* is not based on any inference. Moran's transparency account thus lacks the resources to explain cases where ratiocination fails to bring about the relevant belief.

In the following section, I will consider another version of the transparency account, which holds that one comes to know what one believes by inference.

5.3 Byrne's Transparency Account

Byrne offers an alternative explanation for why the Outward Looking method works, which also lies in the transparency of the 'Do I believe *p*?' question to 'Is it the case that *p*?' question. According to Byrne, a subject may know her own beliefs by following an epistemic rule that has the general form:

If conditions C obtain, believe that *p*.

In the case of epistemic rules for belief about belief, C will be conditions that obtain in the external world, and '*p*' will be a proposition about one's belief. Byrne calls the epistemic rule an agent follows to form a belief about her beliefs

BEL: If *p*, believe that you believe that *p*.[117]

Byrne uses the example of the link between a ringing doorbell and someone standing at the door to illustrate how an analogous rule, which he calls DOORBELL, works.

DOORBELL If the doorbell rings, believe that there is someone at the door.[118]

If one is following DOORBELL, then, if the doorbell rings, she believes that someone is at the door.[119] Similarly for BEL if one has 'recognised' that *p*, then one is in a position to

[117] Byrne 2018, p. 103.
[118] *Ibid.*, p.101.
[119] *Ibid.*, p. 105.

believe that she believes that *p*. According to Byrne, '[r]ecognising that *p* is (inter alia) coming to *believe* that *p*.'[120] This also explains why BEL is a '*self-verifying*' rule.[121]

It is unclear what the precise extension of 'recognising that *p*' is supposed to be. But we can focus on belief. That is, we can assume that Byrne is saying that what it means for conditions to obtain in BEL is that one believes that *p*. So, when one comes to believe that she believes that *p*, her second-order belief is necessarily true. In this way of forming a belief about one's belief, the subject moves from a lower-order belief to a higher-order belief by 'reasoning without the perception of anything mental'.[122] The procedure involves the faculty of reasoning rather than some special inner perceptual mechanism that detects one's mental states. When one tries to find out whether one believes that it is raining, for instance, one's perceptual evidence about the weather will give one reasons to affirm that it is raining. One does not need to look for perceptual evidence of one's own mental states or behaviours.

Critics of Byrne's account point out that the fact that *p* is not a reason for one to infer that she believes that *p*. There are many facts in the world about which one does not have a belief.[123] To be charitable to Byrne, for C to obtain, that is, for one to recognise *p*, one necessarily believes *p*. Byrne is unclear as to whether the recognition has to occur at the conscious level for the subject. But, in any case, his account should apply to cases where a subject consciously recognises that *p*. For a subject to follow BEL is for her to recognise *p* and then come to believe that she believes that *p*. And according to Byrne, if she follows BEL, her second-order belief is necessarily correct. But is this so?

Consider Rae again. She ratiocinates and concludes in her inner speech 'My wallet is at the restaurant.' Byrne might say that since Rae is in the state of believing that she ought to believe that her wallet is at the restaurant, it does not count as her recognising that her wallet is at the restaurant. Yet if this does not count as recognising a proposition, then what Byrne has to say is that one has to be in the state of believing a proposition for C to obtain. Then it is unclear how a subject can follow BEL because it requires to recognise what she believes first. Yet the whole point of the rule is supposed to help a subject form a belief about what she believes.

Perhaps Byrne would say that BEL does not apply to ratiocination. It only applies to methods that directly fix the belief that *p*. However, Byrne thinks that the transparency

---

[120] Byrne 2018, p. 104.
[121] *Ibid.*, p. 104.
[122] *Ibid.*, p. vii.
[123] See Boyle 2011 for challenges along this line.

account is supposed to explain how all mental states are transparent.[124] It cannot single out the beliefs that are fixed in a top-down way and say that the epistemology of these particular top-down fixed beliefs is different from other beliefs to which the transparency account applies. Presumably, there's nothing about the beliefs themselves that are different, only the ways they are fixed are different.

Suppose *S*'s thinking 'My wallet is at the restaurant' counts as recognising that her wallet is at the restaurant. Following BEL, she goes on to believe that she believes that her wallet is at the restaurant. Yet imagine that the top-down fixation process in this case fails. She does not acquire the belief that *p*. Then, she will end up with a false second-order belief about her belief. Byrne therefore is wrong to think that a subject who follows BEL will necessarily have a true belief about her belief.

We may try to weaken Byrne's transparency account and say that by following BEL, one is entitled to the belief that she believes that *p*. Even though her second-order belief is not necessarily true, the top-down fixation process is generally reliable enough. Most of the time, when one concludes that she ought to believe *p* from ratiocination, she goes on to believe *p*. Because top-down fixation is generally reliable, one is entitled to the belief that she believes that *p* is she recognises that *p* is the case in ratiocination. In this sense, self-knowledge is still preserved. However, even if Byrne can say that one reliably goes on to believe *p* in cases of ratiocination, it is not clear that one's belief that she believes *p* is safe.

To say that a ratiocinator's belief that she believes that *p* amounts to knowledge, Byrne would agree that the second-order belief has to be safe.[125] However, since top-down fixation involves an indirect strategy to bring about the belief that *p*, there is no guarantee what strategy works and in which case a particular strategy works. Whether the top-down fixation process is successful might vary from subject to subject and also from subject matter to subject matter. A strategy *x* might work for *S* but does not work for another subject *T*. A strategy *x* might work for *S* with regard to subject matter *m* but not for subject matter *n*. A strategy *x* might work for *S* with regard to subject matter *m* in context C but not in context C'. It is possible that when *S* believes that her wallet is at the restaurant, it is the case that she believes that her wallet is at the restaurant. However, the strategy could have easily failed in this particular case so that *S* does not end up believing *p*. Or imagine that *S* after ratiocination does come to believe that there are feathered dinosaurs but *T,* who has gone through the same

[124] Byrne 2018, p. 23.
[125] *Ibid.*, p. 106.

ratiocination process, does not acquire the belief that there are feathered dinosaurs. Both *S* and *T* are rational. They follow BEL and come to believe that they believe that there are feathered dinosaurs. *S*'s second-order belief is true, but T's second-order belief is false. Hence, even if we assume that the top-down fixation process is generally reliable in fixing lower-order beliefs, it does not mean that one's particular belief that she believes that *p* is safe. She could easily have not formed the belief that *p*. Byrne's transparency account does not provide a satisfactory explanation for why *S*'s second-order belief formed following BEL amounts to knowledge.

5.4 The First-Person Perspective

Although I have criticised the transparency account's assumption that endorsing *p* necessarily leads to believing *p*, I do agree that a subject's mind could move from rationally endorsing *p* to believing that she believes *p*. However, this is just a claim about how one's mind could move. One who consciously endorses *p* does not have to go on to become aware of a belief she has or ascribe a belief to herself.

Consider again the case where *S* at the end of her ratiocination thinks 'My wallet is at the restaurant.' *S* does not initially have any thought about her own mind. Yet suppose *S*'s spouse asks, 'Do you believe that your wallet is at the restaurant?' naturally *S* will say 'Yes, I do believe that my wallet is at the restaurant.' Let us bracket whether this answer about her belief is correct. My point is that if we consider things from *S*'s first-person perspective, since she concluded at the end of her ratiocination that 'My wallet is at the restaurant,' then what else should she believe? From her first-person perspective, since she consciously endorses 'My wallet is at the restaurant,' given she is rational, she will go on to believe that she believes that her wallet is at the restaurant. It would be absurd for her to respond, 'Let me look into my mind and see ….'

The point that it is rational for a subject who consciously endorses *p* to go on to self-ascribe the belief that *p* is accepted by a number of existing accounts.[126] However, their focus tends to be on the a rational connection between one's endorsement of *p* and one's belief about what one's belief is. And the rational connection is built on one's believing that *p*. I am suggesting something different. One does not have to make a self-ascription after deliberating about *p*. But if one does, since one consciously endorses *p*, it will simply be irrational for one

[126] E.g. Peacocke 1999, Burge 1996, Bilgrami 2006, Boyle 2009.

to believe she does not believe p. It would be absurd for her to say that '*p* but I do not believe *p*' or '*p* but I believe not-*p*.' This, I take it, is one of the lessons from Moore's paradox.

Moore notes that it is absurd for someone to assert sentences such as 'Though I do not believe it is raining, yet as a matter of fact it really is raining'[127] and 'I believe that he has gone out but has not.'[128] It is generally agreed that there are two forms of Moorean sentences, one being omissive and the other commissive. They are respectively:

(Omissive) '*p* but I do not believe that *p*'

and

(Commissive) '*p* but I believe that not-*p*.'

What is puzzling about these sentences is that neither is, on the face of it, a contradiction. Both conjuncts could be true. It may well be the case that it is raining but I do not believe that it is raining, and yet it seems absurd for me to assert both conjuncts at the same time in first-person present tense.

We should keep in mind that these are sentences that are uttered by the speaker. They are not propositions. One thinks to oneself or says out loud that, '*p* but I do not believe that *p*.' When the speaker utters the Moorean sentence, her mind is already at the conscious level. On my account, if the ratiocinator at the end of ratiocination concludes in her mind '*p*,' then to her mind, since to affirm that *p* is the case just is to believe that *p*, if she makes a claim about her belief to herself at all, she will say she believes that *p*. It will be absurd for her to then say she does not have a belief about *p* or she believes not-*p*. For a rational subject like *S*, a Moorean sentence is unthinkable from her first-person perspective after she concludes ratiocinating.

What we can learn from Moorean sentences is that the subject's first-person perspective on her belief starts at the conscious level. There are no lower-order beliefs she can peek into. Even if there is an inner sense mechanism, the subject herself is only aware of the output the inner sense mechanism delivers. She has to rely on the inner sense mechanism to bring these inputs to her awareness. She cannot look beyond what is available to her consciousness to see what the input is and check whether the inner sense mechanism has correctly scanned it. The point of positing an inner sense mechanism is to explain how

[127] Moore 1942, p. 543.
[128] Moore 1944, p. 204.

our self-ascriptions are reliable even though we do not directly 'see' the lower-order states. If one can directly see the lower-order states, there is no need to posit an inner sense mechanism.

My explanation for why it is unthinkable for *S* is different from Sorensen's explanation in terms of 'blindspots'. For Sorensen, Moorean sentences are belief 'blindspots' in the sense that the Moorean propositions are consistent but the agent cannot have an attitude towards such inaccessible propositions. On my proposal, the reason *S* will not utter a Moorean sentence after ratiocination has nothing to do with the nature of the set of propositions; instead, it is because *S*'s first-person perspective starts at the conscious level. It is absurd for a rational subject to believe that she does not believe what she consciously endorses.

Suppose a rational subject read this thesis and is convinced by my account. She ratiocinates about whether there are feathered dinosaurs and comes to the conclusion that there are. When she is being asked 'Do you believe that there are feathered dinosaurs?' She might qualify her statement a bit, 'If you ask me, at least consciously, I believe that dinosaurs have feathers. But do I really have the belief that there are feathered dinosaurs? Is the top-down fixation process successful? Maybe I am actually not interested in whether dinosaurs have feathers, so I didn't acquire the belief. I can't be sure that I have the belief that p. But what I can tell you now is that after deliberating, I do believe that there are feathered dinosaurs.' This is the closest a subject can get to a Moorean sentence: '*p* but maybe I don't have the belief that *p*' but it is not a genuine Moorean sentence. One can assert that *p* is the case but be hesitant about whether she actually has the belief that p. She might say, 'There are feathered dinosaurs but I am not sure if I believe there are no feathered dinosaurs.' She will not say 'There are feathered dinosaurs but I believe there are no feathered dinosaurs' if the second conjunct is asserting the state of her mind. If the second conjunct is asserting the state of her mind, then perhaps in the first conjunct she means 'I ought to believe *p*' instead of '*p*'. In either case, what she utters is not a genuine Moorean sentence.

It is difficult to imagine how a rational subject could believe that she does not believe *p* when she consciously endorses *p*. What it means for a subject to get to the conclusion of ratiocination is for her to close the inquiry concerning *p*'s truth. It would be irrational for her to at the same time believe that she does not believe p. For this would suggest that she has not worked out whether p is true. In that case, she should ratiocinate again. It is not that a Moorean sentence is unthinkable; it is just that a rational subject will not think a Moorean sentence.

My explanation avoids difficulties with the two general approaches to Moorean sentences. One general approach is a functionalist approach. All versions of functionalism are committed to the claim that for someone to believe that *p* is for him or her to be 'in a state which, together with his or her desires, will normally cause behaviour which satisfies those desires only if *p*'.[129] As Jane Heal points out, if we set out this functionalist conception of belief in detail, we come to see that the oddity of Moore's paradox disappears. Heal invites us to consider an extended Mueller-Lyer illusion case. In a standard Mueller-Lyer illusion with two lines A and B, an agent visually perceives A as longer than B when in fact A is equal to or shorter than B. In Heal's extended case, the agent not only has the visual illusion that A is longer than B, her bodily behaviour also fails to register with her belief that A is in fact shorter than B. She will, for instance, reach out to A when she desires to pick out the longer stick. On the functionalist account of belief, which says that having a belief that *p* is equivalent to a state apt to cause behaviour appropriate to its being the case that *p*, the behaviour of the agent suggests that the agent has the belief that A is longer than B. Yet because the agent also realises this is a case of Mueller-Lyer illusion, she also believes that B is longer than A. Under such circumstances, the agent is entitled to assert the Moorean sentence: 'I believe that A is longer than B, but B is longer than A.' This example shows that the oddity of Moore's paradox seems to have disappeared on the functionalist account of belief and therefore fails to satisfy the condition that a solution to Moore's paradox must identify a contradiction or something contradiction-like in Moorean sentences.

Heal considers two possible ways in which functionalism can be reformulated to preserve the oddity of Moore's paradox. One way is to rule out the possibility of contradictory beliefs. That is, it may be argued by the functionalist that, 'a belief can be attributed only if *all* behaviour is unified under the control of one representation.'[130] For the extended Mueller-Lyer case, the functionalist could say that the agent also produces the verbal behaviour of uttering 'B is longer than A,' suggesting that she must be under the control of some other representation of the world than the one that causes her to reach out to pick up A. As the agent's behaviour is not unified under the control of one representation, we cannot attribute a belief to her. Heal finds this approach implausible for she thinks that people clearly do have contradictory beliefs.

[129] Heal 1994, p. 13.
[130] *Ibid.*, pp. 14–15.

An alternative way of reformulating functionalism is to deny that the agent has full belief in the extended Muller-Lyer case. On this alternative formulation, to say that someone believes that *p* is to say that he or she is in a particular relation to *p*. This preserves the absurdity of the paradox in Moorean sentences because the agent would be saying that she has a particular relation to *p* and at the same time denying that she has this relation to *p*. Heal argues that this approach is also unsatisfactory because we can extend the previous Muller-Lyer case even further to suppose that the agent cannot stop herself from having thoughts that A is longer than B and the control of these thoughts over her bodily behaviour is very extensive. If the agent could only retain the control over her voice, then she is still entitled to say: 'I (really) believe *p* but (really) not-*p*.' The paradox again disappears.

Heal considers two explanations of Moore's paradox.[131] One approach attempts to expand the '*p*' part in a Moorean sentence to 'I believe that *p.*' The other approach reduces the 'I believe that *p*' in a Moorean sentence to '*p*'. Heal defends the belief-reduction strategy by appealing to the constitutive nature of self-ascriptions.[132] According to Heal, when an agent comes to sincerely believe that she believes that *p*, then it is the case that she believes that *p*. Heal writes:

> I am entitled to pronounce on my beliefs not because I have some privileged epistemological access to an independent state but because when I come to think that I believe that *p* then I do, in virtue of that very thought, believe that *p*.[133]

According to Heal, if I sincerely believe that 'I believe that *p*', this higher-order belief by itself constitutes in me the first-order belief that *p*. To say that I believe that I believe that *p* is just an alternative way of saying that I believe that *p*. And if I add 'but I believe not-*p*,' I have thereby contradicted myself. However, as we have already seen, such a constitutive account of self-ascription is problematic because it is not always the case that when there is BB*p*, there is also B*p.*

Heal's approach to Moore's paradox rests on the questionable assumption that *BBp i*s reducible to B*p*, an assumption that goes back to Hintikka's explanation of Moore's paradox.

---

[131] Heal 1994.
[132] Heal's reservation with the belief-expansion strategy is that '*p*' cannot by itself generate into a proposition 'I believe that *p*'; some event or state that has '*p*' as its content must have occurred that expands '*p*' to 'I believe that *p*.'
[133] Heal 1994, p. 22.

According to Hintikka, the omissive sentence is closely related to a sentence: 'I believe that the case is as follows: *p* but I do not believe that *p*,' which has the form:

B (*p* & B¬*p*)

This combined with the doxastic distribution principle – if a subject believes that *p* and *q*, then a subject believes that *p* and believes that *q* < B (p & q) ⊃ Bp & Bq > – we get the result that an omissive Moorean sentence has the form:

B*p* & BB¬*p*

In Hintikka's view, BB*p* can be reduced to B*p*.[134] Hence, we get a contradiction:

B*p* & B¬*p*

In informal terms, if a subject believes that she believes that *p*, then it is the case that she believes that *p* (BBp ⊃ Bp).

Both Hintikka's and Heal's explanations rely heavily on the belief-reduction principle. Yet if BB*p* is not reducible to B*p*, the belief-reduction strategy fails, leaving the absurdity of Moore's paradox unexplained. I will discuss in the next chapter why BB*p* is not

[134] Hintikka 1962, pp. 33–36. Hintikka's formal doxastic logic proof works better for the commissive sentence but it is not clear how it explains the omissive sentence. For Hintikka, in the omissive case, B ¬ B*p* can be reduced to ¬ B*p* and then further reduced to B ¬ *p* (p. 53–54). This is because the set of possible worlds is divided into those that are compatible with what the subject believes and those that are incompatible with what she believes. These worlds would be the doxastic alternatives with respect to the subject. As a doxastic alternative, μ* should be compatible with the informational resources the subject has at time *t*. It is in this sense that μ* is accessible from μ. Suppose the subject believes that 4 February is a Tuesday. This means that in all possible worlds that are compatible with what the subject believes, she considers the proposition '4 February is a Tuesday' true. For the subject to believe a proposition is to consider the proposition in question to be the case; this suggests that the subject's believing that *p* rules out possible worlds in which *p* is not true. A world in which it is not the case that 4 February is a Tuesday cannot be a possibility for *a*. However, for some instances of B¬B*p*, it could be indeterminate what worlds can be ruled out. This can happen when the agent has never entertained *p* so that the compatibility of the worlds cannot be determined. We may bracket these considerations for now and just focus on Hintikka's assumption that BB*p* is reducible to Bp.

reducible to B*p*. Here, I simply want to make clear what sets my explanation apart from explanations that rely on the belief-reduction strategy.

I do not think that we have to expand the first conjunct '*p*' to 'I believe *p*' nor reduce 'I believe *p*' in the second conjunct to '*p*'. The absurdity of a Moorean sentence can be explained by how a subject, from her first-person perspective, is unable to tell the difference between thinking '*p* is the case' and 'I believe that *p.*' This is so because the first-person perspective on what one believes about *p* starts at the higher-order level. The Moorean sentences suggests to us that if one consciously endorses that '*p*', and if one makes a self-ascription at all, then she will believe that she believes that *p*. If she consciously endorses that *p*, assuming she does not see any reason to revise her conclusion and yet goes on to say, 'Let me introspect whether I believe *p*' or 'Let me infer from my behaviour whether I believe *p*,' it is difficult to resist the thought that she exhibits some signs of irrationality. Those who are sympathetic to Moran's transparency account should find this point about how a subject's mind moves to self-ascriptions after ratiocination acceptable. As the question 'Is it *p*' is indistinguishable to the subject's mind from the question 'Do I believe *p*?' if she answers 'Yes' to one, she has no reason, it seems to her, to answer 'No' to the other. To one's mind, the move from consciously endorsing *p* to believing that she believes that *p* is natural and rational.

I am only suggesting that if a subject wants to ascribe a belief to herself, her mind will move from consciously endorsing *p* to believing that she believes *p* given the indistinguishability between the two questions from her first-person perspective. I am not suggesting that the state of consciously endorsing *p* and the state of believing *p* are the same states. For convenience, I will label the state in which a rational subject *S* who consciously endorses *p* after ratiocination as 'BOB*p*' and the state in which *S* believes that she believes *p* as 'BB*p*'.

At the end of her ratiocination, *S* is in the state of BOB*p*. Recall from section 4.1 that in the top-down fixation process, *S* has to adopt an indirect cognitive strategy to get herself to believe *p*. If this strategy is successful, *S* will then go on to believe that *p*. It is very likely that in many cases one is successful in bringing about the belief *p* after ratiocination. This suggests that being in BOB*p* does not prevent *S* from coming to believe *p*. Being in BBp, however, could potentially prevent *S* from coming to believe *p*. In the case of top-down fixation, if *S* believes that she believes that *p*, then she mentally registers that her goal of believing *p* obtains. Once she mentally registers that her goal has been achieved, the process of top-down fixation will stop. This is like in the practical case, where if I am motivated to

take my medicine, then I will adopt some strategy to take my medicine. If I believe that I have already taken my medicine, I will stop trying to take my medicine.

BB*p* does not always prevent one from believing that *p*. In a normal case, one comes to consciously believe *p*, one comes to believe *p* through top-down fixation, and comes to believe that she believes that *p*. This process can proceed through various stages so quickly that it is not detectable from the first-person perspective. It is like when one watches a short clip on TV, it seems that it is a seamless flow of movements, but the screen has already taken on a number of different states without the viewer recognising. In the case where top-down fixation is successful, BB*p* is formed after one has come to believe *p*. A good case will be something like this: One is in the state of BOB*p*, the top-down fixation process begins, and she adopts a cognitive strategy to believe *p*. The strategy succeeds and she believes *p*. She then moves to BB*p* and has a true second-order belief.

However, before *S*'s cognitive strategy is completed, if one forms a belief about her belief too soon, it is possible that she falsely registers that her goal has been achieved and, as a result of this false registration, the top-down fixation process terminates before it fixes B*p*. A bad case will be something like this: One is in the state of BOB*p*, the top-down fixation process begins, and she adopts a cognitive strategy to believe *p*. While her mind is still working to believe *p*, she believes that she believes *p*. The top-down fixation process terminates because of her mental registration that the goal of believing *p* has been achieved. A bad case could also be something like this: *S* is in the state of BOB*p* and she simultaneously believes that she believes *p*. Since she believes that she already believes *p*, she does not intend to believe *p*. The top-down fixation process is never initiated.

Since being in BOB*p* does not discontinue the top-down fixation process but being in BB*p* could potentially discontinue the process, BOB*p* and BB*p* are not the same states. In good cases, *S* is necessarily in BOB*p* after ratiocination but not necessarily in BB*p* after ratiocination. If BOB*p* and BB*p* are the same states, the top-down fixation process will never get initiated.

A lesson from this is that, in the case of ratiocination, one should keep her eyes focused on the world and not turn to her mind too soon. If one comes to believe that she believes *p* too soon, she risks stopping the top-down fixation process before it successfully fixes her belief that *p*. This does not mean that one cannot immediately answer the question 'Do I believe *p*?' As discussed in Chapter 5, sometimes we provide a world-directed answer (e.g., 'My wallet is at the restaurant') sometimes a mind-directed answer (e.g., 'Everyone else is a better parent than me'). As long as one is not in the state of believing that she has a

belief, she will not discontinue the top-down fixation process just because she answers, 'Do I believe *p*?' One can give a world-directed answer. This is why considering what one answers is not always informative.

5.5 Reflectively Fixed Self-Ascriptive Belief

The transparency account is right in saying that one's reasons for endorsing *p* are also reasons for one to self-ascribe the belief that *p*. We need not think that self-ascriptions of belief must await the deliverances of an inner sense mechanism. However, that there are reasons to believe that they believe that *p* does not guarantee that the self-ascriptive belief is true.

One might ask: does the self-ascriptive belief have to be sensitive to the presence of Bp? Even though the transparency account does not rely on inner sense, one might say that the presence of belief that p is a reason for one to self-ascribe.

What Moore's paradox teaches us is that the first-person perspective starts at the conscious level. When we try to explain the Moorean sentence, if we assume that the first-person perspective starts at the lower-order level, then we will have trouble explaining the absurdity. For, on such a reading, the first conjunct 'p' can be saying something about the world and the second conjunct 'believe not-p' – where the perspective is at the higher-order level – can be saying something about her mind. But if the first-person perspective starts at the conscious level, then both conjuncts have to be expressive of what she takes the world to be.[135]

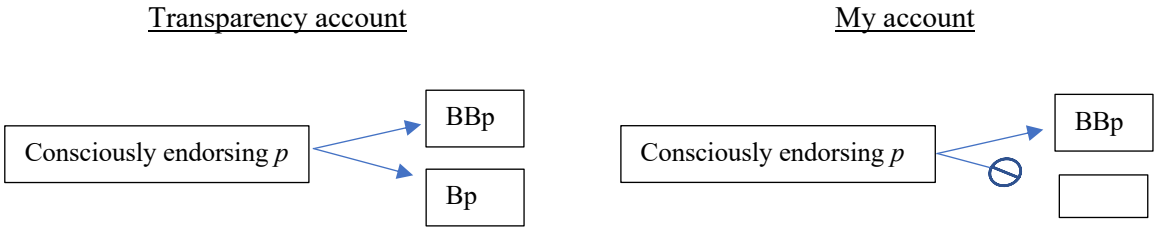
I agree with accounts that hold that the belief that one ascribes to oneself is immediately sensitive to judgements one makes.[136] But the judgements about reasoning for believing *p* can be formed via ratiocination. This does not mean that the belief that *p* is always immediately fixed by judgments about reasons for believing *p*. In the case of ratiocination, one judges that there are reasons for believing *p,* but one is in the state of believing that *I ought to believe p*. One might not go on to form B*p*. Hence, one's BB*p* is sensitive only to the judgements, not to some lower-order belief p. What I am suggesting is

[135] Note all I am saying is how Moorean sentence cannot be uttered by someone who completes ratiocination. Maybe with some like 'one line longer than the other, but I believe the two lines are of the same length' it is not as problematic because the first conjunct is not something she has to make up her mind about. But for the kind of avowal that is made after deliberation, then it is absurd why one will say that.

[136] E.g. Peacocke 2000, Moran 2001, Shah and Velleman 2005, Parrott 2017

that judgement can fail to fix the lower-order belief, but it will be irrational of someone not to ascribe to herself a belief the content of which is what she judges.

Once we come to realise that the first-person perspective starts at the conscious level, then when a rational subject answers a question about the world by looking outward, she does come to BBp. What else could she rationally believe she believes?



The transparency account shows us that considerations for *p* are transparent to considerations for whether I believe that *p*. But it has not satisfactorily explained why a subject must be correct in believing that she believes that *p* by basing her self-ascriptive belief on deliberation about *p*. All it has shown is why she must be right in believing that she believes that she believes that *p*. Transparency, in other words, holds at a reflective level, the world-directed question 'Is it the case that *p*?' is transparent to the mind-directed question 'Do I believe that I believe *p*?'

Both Moran's and Byrne's transparency accounts try to provide explanations for how one gains knowledge of one's own mind by attending to evidence. Both accounts depend heavily on the assumption that to consciously endorse *p* is to believe *p*. I reject this assumption by arguing that there are cases where a rational subject *S* ratiocinates about *p*, which also counts as following the Outward Looking method, comes to consciously endorse *p* but does not go on to believe *p*. I also argue the first-person perspective where one will naturally move from consciously endorsing *p* to believing that she believes that *p*. *S* does not necessarily form a belief about her belief when she consciously endorses *p* from ratiocination. Yet should *S* form a belief about her belief, she will believe that she believes *p*, for the question about whether *p* is the case and a question about whether one believes *p* is indistinguishable from the first-person perspective.

Since a ratiocinator's endorsement of *p* might not bring about her belief that *p* and since one will on the basis of her conscious endorsement come to believe that she believes *p*, it is possible that one might believe that she believes *p* even though she does not believe *p*. In

more schematic terms, it is possible that one's mind moves from BOB*p* to BB*p* but the top-down fixation process fails to bring about B*p* so that one ends up with BB*p* without B*p*.

That a false self-ascription is possible after one has rationally considered a question about *p* suggests that the way one answers a question about the world is not transparent to the way one answers a question about one's belief. What is transparent is the way one answers a question about the world to the way she answers a belief about her belief. Hence, what the transparency accounts have shown us is that transparency holds in a more limited way. That is, we should opt for a more qualified view. I will call this account *reflective transparency*: if one consciously endorses that *p*, then it is the case that one believes that she believes *p*.

We do not have to worry that the problem we have been exploring will occur at the reflective level. That is to say, we do not need to worry that *S* might believe that she believes that she believes that *p* but in fact she does not believe that she believes that *p*. All the higher-order beliefs (BBBp, BBBBp…) are reducible to the second-order belief because the first-person perspective on one's mind starts at the conscious level. From the second-order belief upward, the subject is aware of what she believes she believes. This is unlike the problem with second-order belief about first-order belief. The second-order belief is about a belief about one's mental state that is connected to the world. The possibility of error comes in because the first-person perspective on one's belief starts at the second-order belief; but one's perspective on the world starts at the lower-order level. One might mistakenly take oneself to have a view on the world when in fact—when she believes that she believes *p*—she merely has a view on her own mind. But from the second-order upward, the beliefs are all about one's mind. For a rational subject, if she consciously believes that she believes *p*, she will not also consciously believe that she believes that she believes not-*p*.

To conclude, I agree that from the first-person perspective, the question of whether *p* is the case and the question of whether one believes *p* are indistinguishable. However, the transparency account makes too much of this. We should not think that to consciously endorse *p* just is to believe *p*. When the target of our analysis is self-ascriptions of belief, we cannot be satisfied with an account of how one answers the question 'Do I believe *p*?'. We also need an account that tells us why one's answer is correct. The right answer to 'Do I believe *p*?', for the purposes of understanding self-knowledge, turns on whether the subject has the belief that *p*. Cases of ratiocination are examples where one consciously endorses *p* but might not believe *p*, and yet it is possible that one still rationally goes on to believe that she believes *p*. In such cases, one has a false belief about her belief. This shows that the state of rationally endorsing *p* is not transparent to the state of believing that one believes *p*.

Hence, the transparency account cannot explain how a subject gains knowledge of her belief when she adopts the Outward Looking method. All it shows is that there is reflective transparency such that if one consciously endorses *p*, it is the case that she believes that she believes that she believes *p*.

Although the indistinguishability between a world-directed question and the self-directed question itself does not entitle one to believe that she has a certain belief, the point about indistinguishability is an important insight the transparency account provides. It can account for how one's mind moves from consciously endorsing *p* to believing that she believes *p*. But Moran intended to provide us with an explanation of the immediacy of self-knowledge. Moran writes:

> 'Transparency' stands for the claim that a person answers the question whether he *believes* that P in the same way he would address himself to the question whether P itself. From the first-person point of view, the one question is treated as 'transparent' to the other. And to say the knowledge that I believe that P is 'immediate' is to say that it is not based on any observations or inferences that I make. I argue that such immediacy is best explained as a consequence of what I call the 'Transparency Condition' on first-person statements of belief and other attitudes. What that means, however, is that immediacy and transparency, understood as describing the conditions for first-person authority, stand or fall together.[137]

As we have seen, being able to explain how one comes to self-ascribe does not mean we can explain why one has self-knowledge. What I have shown is that the transparency condition and immediacy do not have to stand and fall together. It is possible that while transparency holds, immediacy fails. And the reason for this is that when one answers a question about whether she believes that *p* she has every reason to answer that way. Nevertheless, the answer is incorrect. While the subject treats the world-directed question and the mind-directed question as transparent to the other, these two questions are in fact not transparent to each other. These two questions are only reflectively transparent to each other.

The transparency account is motivated by the natural thought that we are well positioned to form beliefs about our beliefs if we have concluded deliberation concerning whether or not *p* is the case. It seems that I am in a better position to say what I believe if I

[137] Moran 2004, p. 457.

have carefully deliberated about whether there will be a third world war than just reporting on what my mental state is.

I argued in this chapter that at least one kind of deliberation – the self-consciously directed kind – makes us vulnerable to a special kind of mistake: one might end up only with a view on the state of her mind when in fact she has no view on the world. If this is correct, then where we would have expected to find ourselves in the best position to believe truly about our own minds – after careful reflection – we are liable to error. In the next chapter, I will discuss the implications of the possibility of one's believing that she believes *p* even though she does not believe *p*.

Chapter 6
Belief about Belief

In the previous chapter, I drew on the transparency account to suggest that from the first-person perspective, it does not make a difference to a subject's mind whether she believes *p* or whether she believes that she has the belief that *p*. This gives rise to the possibility that a rational subject believes that she believes *p* but does not believe *p*. In this chapter, I will discuss the metaphysical implications that this psychological phenomenon has for our understanding of second-order belief and the epistemological puzzles that arise as a result. In section 1, I argue that an implication of the kind of false self-ascriptive belief cases that I have been suggesting is that higher-order belief and lower-order belief are distinct states. In section 2, I discuss how my account is different from existing discussions of fallibility of self-knowledge. In section 3, I return to Sam's case and raise more puzzles about self-knowledge. In section 4, I offer a sceptical challenge to self-knowledge about belief. The main worry is that, from the first-person perspective, a subject is not able to tell whether she has a view on the world or whether she only has a view on her own mind. Without the ability to discriminate, how can a subject know that she has the belief that *p*?

6.1 Distinct States

We can start with the constitutive view discussed in Chapter 1. Recall that the constitutive view is characterised by its commitment to the following thesis:

> Constitutive thesis: Given certain background conditions, one believes that *p* if and only if one believes that one believes that *p*.[138]

The factualist takes the lower-order belief to be more fundamental whereas the non-factualists takes the higher-order belief to be more fundamental. Both factualist and the non-factualist versions of the constitutive account hold that the higher-order belief that *p* and the lower-order belief that *p* are not distinct states. Both versions would agree that there could be cases where one's higher-order belief is in fact false. It is possible that one believes that she

[138] Coliva 2016, Chapter 7 provides a helpful summary of the constitutive account.

believes *p* when in fact she does not believe *p*. But these mistaken self-ascription cases are ones where one's rationality or sincerity are compromised. If the argument of the previous chapters is sound, then one important implication is that the Constitutive thesis is false. It is possible for there to be cases of mistaken self-ascriptions where neither one's rationality nor sincerity is compromised.

Here is a brief reminder of what gives rise to this possibility. Some mistaken self-ascriptions are explained by different levels at which the higher-order beliefs and the lower-order beliefs are fixed. One level starts at the bottom level, which directly fixes the lower-order belief *p*. The lower-order belief that *p* then in turn grounds the higher-order belief. The higher-order belief is fixed in a bottom-up way. (I leave it open whether it is possible for bottom-up fixation of higher-order belief to fail). Another starts at the top or reflective-level, which then fixes the lower-order belief in a top-down way. It is possible for the top-down fixation to terminate before it brings about the belief that *p*.[139] One's ratiocination at the reflective level directly fixes one's higher-order belief that *p*.[140] However, top-down fixation can terminate, without the failure of rationality, prior to the formation of the corresponding lower-order belief. We thus have a case where a rational subject believes sincerely that she believes *p* without believing *p*.

If this much is correct, we can straightforwardly reject the factualist version of the constitutive account. For the factualists, it is impossible for a rational subject to have higher-order belief without the corresponding lower-order belief. But I have shown that this is possible in the case of ratiocination.

The argument against the non-factualist will need more. Some factualist might say that even if top-down fixation fails, because of transparency, when a rational subject comes to believe that she believes *p*, then in sincerely believing that she believes *p*, she will come to believe *p*. Heal, for example, says that 'When I come to think that I believe that *p* then I do, in virtue of that very thought, believe that *p*.'[141]

I take it that Heal is not saying that one has B*p* simply by having BB*p*. Rather *S* is in BB*p* because she consciously endorses that *p*. Her BB*p* is anchored to her consideration of the evidence for *p*. It is conscious endorsement of *p* that is doing the work. But the case of ratiocination shows that conscious endorsement of *p* does not always succeed in believing *p*.

---

[139] See Chapter 4. 2
[140] See Chapter 5.5.
[141] Heal 1994, p.22.

Let us assume that ratiocinator Rae's pronouncement is sincere, and she meets all the background conditions for thinking that she believes *p*, e.g. she is not deranged, not under stress, not under influence, etc. We can imagine that Rae carefully made up her mind in ratiocination and sincerely answers, 'Yes, I believe that there are feathered dinosaurs.' However, it is difficult to see how just in believing that she believes that there are feathered dinosaurs,' Rae must at the same time also believe *p* and be in the state of believing *p*. Thinking 'I believe *p*' does not always entail one is in the state of believing that *I believe that p*. As explained in my discussion of the transparency account in Chapter 5, just because a subject utters 'I believe *p*,' it does not mean she is in the state of believing that *p*. She can say 'I believe *p*' because she learns that it is a linguistic convention to say 'I believe *p*' as a substitute for '*p* is the case.' One can affirmatively answer the question 'Do I believe *p*?' by saying 'Yes, it is the case that *p*,' 'I believe *p*' while still being in the state of consciously endorsing p, and in the case of ratiocination, to be in the state of consciously endorsing *p* is to be in the state of believing that she ought to believe *p*. So, it is possible that Rae utters, 'I believe that there are feathered dinosaurs' and is still merely in the state of believing she ought to believe that there are feathered dinosaurs. So, sincere avowal itself does not suffice for the state of believing that *I believe p*.

The state of consciously endorsing *p* and the state of believing that *I believe p* are not the same states. As discussed before, the state of consciously endorsing *p* in the case of ratiocination is the state of believing that *I ought to believe p* (BOB*p*). BOB*p* motivates one to believe *p* and adopt a cognitive strategy to form the belief that *p*. However, if for some reason one decides to self-ascribe and enters the state of believing that *I believe p* (BB*p*), one mentally registers that she has obtained the goal of believing *p* and therefore stops her cognitive strategy deployed to achieve believing at the lower-order level. Hence, BB*p* might prevent one from believing *p*. Since BOB*p* does not stand in the way of top-down fixation but BB*p* could terminate the top-down fixation process, BOB*p* and BB*p* cannot be the same states.

The non-factualists might want to say that sincere avowals amount to believing that *I believe p*. Here, the non-factualists may make the weaker claim that being in the state of believing that *I believe that p* necessarily brings about the state of believing that *p.* Or, they may make the stronger claim that being in the state of believing that *I believe that p* just is being in the state of believing that *p*. On my account, BB*p* and B*p* cannot be the same states. For the same reason mentioned above, since it is possible for BB*p* to terminate the top-down

fixation process, it is not necessary that BB*p* brings about B*p*.[142] To reject something like Heal's constitutive account, it suffices to show that, for rational subjects who are sincere, BBp can sometimes fail to bring about B*p*. Even if it is true that in most cases BB*p* can bring about B*p* or comes with B*p*, as long as we accept that it is possible for one to have BB*p* without B*p*, then BB*p* and B*p* must be distinct states.

Heal thinks that sincerely avowing ensures both BB*p* and B*p*. I argue that one can sincerely avow without BB*p* nor B*p*. One can sincerely pronounce 'I believe *p*' without being in BB*p*. In cases where one avows and is in BB*p*, it is not necessary that one is in B*p*. I bracket cases where there are purported practical considerations involved. In the kind of ratiocination cases that we have been considering, a subject does not voluntarily believe that she believes *p* because she thinks believing *p* will lead to some payoff. But even if we assume that BB*p* might succeed in bringing about B*p* when there are practical considerations – e.g., Pascal's wager – BB*p* is being used as part of a strategy of fixation, and there is still a chance this strategy will fail. If it is a sure-fire strategy, then it just means one can believe at will. Since I assume that we cannot believe at will, BB*p*, as part of a practically motivated strategy, does not always succeed in bringing about B*p*.

6.2 Corrigibility

Moving on from the constitutive view, some who hold that BB*p* and B*p* are distinct states can disagree that it is possible to have BB*p* and B*p*. It may be argued that, even though BB*p* and B*p* are distinct states, BB*p* necessarily entails B*p*. Those who hold this view tend to think that self-ascriptive belief is incorrigible, that one's self-ascriptions are always true. Stoneham, for example, advances this line of argument. Stoneham's incorrigibility thesis states:

> **Incorrigibility thesis**: If someone believes that he believes that p, then he believes that p.[143]

The incorrigibility thesis can allow for the claim that BB*p* and B*p* are distinct states because it can allow one to believe *p* without believing *p*. Stoneham argues that it is a conceptual truth that the state of believing that one believes *p* contains the state of believing. To say why it is

[142] If I am right, then any view that holds that being in the state of endorsing that *p* is also the being in the state of believing *p* cannot be right (e.g. Peacocke 2016).
[143] Stoneham 1998, p.128.

a conceptual truth, Stoneham argues that the state of believing that one believes *p* 'has at least all the consequences and commitments' of the state of believing that *p*. What it means to have consequences and commitments is to 'think and act as if p is true.'[144] Stoneham, relying on a weak form of functionalism, then argues that since being in the state of believing one believes *p* 'meets all the conditions of conceptual grasp and involvement or activation for believing that p, and also has all the consequences and commitments of believing that p,' then being in the state of believing that one believes *p* is just the state of believing *p*.

Again, appealing to the conclusion of the argument in the previous chapters, I claim that, in believing that she believes *p* – i.e., in mentally registering that her goal of believing *p* is already fulfilled – a ratiocinator could terminate the top-down fixation process and therefore fail to believe *p*. Hence, it cannot be the case that it is a 'logical property' of BB*p* that it contains B*p*.[145]

The incorrigibilists might say one of the following about Sam:

(1) Sam does not have genuine self-ascriptive belief.
(2) Sam cannot be surprised.
(3) Sam has contradictory beliefs.

They might say that (1) Sam's statement about his belief is not a genuine self-ascriptive belief. Even if we assume a functionalist view, it is difficult to tell whether BB*p* in fact has all the consequences and commitments of B*p*. It is possible that Sam has not encountered any situation that allows himself or anyone else to tell whether he acts or thinks as if it is true that black swans exist. If Sam were put in a situation, for example, when he has to circle the animals that have black feathers in his class exam, he will not circle swans. It is not obvious to me that this must be the case. Even if it is the case that his self-ascriptive belief is false, why should we think that Sam acts in a way that suggests he does not believe black swans exist? Suppose Sam has studied hard for the exam, and during the exam, he remembers vividly how he arrives at the conclusion that there are black swans, it seems at least possible that he will circle swans as one of the correct answers.

And even if Sam does act or think as if it is true that black swans exist, I do not see why we have to deny that Sam has a genuine self-ascriptive belief. Suppose a Buddhist

[144] *Ibid.*, p. 134.
[145] *Ibid*., p. 133.

practitioner believes in the Buddhist doctrine that nothing exists. Yet she does not think and act as if it is true that nothing exists. She puts what she takes to be cups on what she thinks of as tables, and tries to make sure the incense is laid properly so that it does not burn the tablecloth. One might question if she really believes that nothing exists. But it might be too strong to deny one's self-ascription as a genuine self-ascription of belief. My point is that it is possible that the commitments and consequences of BB*p* and those of B*p* do not line up.

The incorrigibilists might then say that (2) if Sam really has a self-ascriptive belief, then he cannot be surprised. However, students often deliberate and do well in classroom exercises. And yet, when they encounter the course content in the real world, they are surprised. And their surprise, by the surprise principle, reveals that they held an attitude contrary to p. The incorrigibilists might then maintain (3) that Sam must have contradictory beliefs. He both believes that there are black swans and believes in a proposition that is inconsistent with there being black swans. It might be said that the surprise principle only shows that Sam's surprise reveals that he holds an attitude contrary to the existence of black swans. He could have also believed that there are black swans and therefore, his self-ascriptive belief is still true.

In the following, I will discuss another necessary condition for generating surprise. I call this

Acquisition Requirement: if one is surprised that *p*, one acquires the belief that *p*.

According to this requirement, at the moment of being surprised that there are black swans, Sam acquires the belief that there are black swans. If this is correct, then it cannot be the case that Sam also believed that there are black swans right before he was surprised.

Now, I am suggesting another necessary condition for surprise: the subject acquires the belief that *p* at the moment when she is surprised that *p*. When surprise occurs, a subject must monitor the compatibility between her pre-existing beliefs with newly acquired beliefs about the world.[146] This requirement is what marks cases of surprise as different from cases

[146] Self-monitoring here does not entail that one has to monitor one's own surprise reaction. Suppose I am at the doctor's office and am told by the nurse that the doctor will come and perform some tests on me. When the nurse said "doctor", she briefly gestured towards two people standing in the hallway, one a man and the other a woman. When the doctor walked in, I was surprised to see that the doctor is the woman I saw earlier. I was too nervous about the tests to dwell in my surprise. Few days later, I read an article on gender bias. Having read

where one simply changes her beliefs. We may imagine a jury member, A, who is easily swayed by others and has unstable beliefs about the defendant. When juror A is with one group of jurors, she believes that the defendant is innocent; when she is with another group of jurors, she believes that the defendant is guilty. She might not be even aware that she is wavering. This example suggests that a mere change of belief from believing not-*p* to believing *p* does not necessarily generate surprise. Something more is needed.

It may be argued that if a subject is surprised by *p*, not only does the subject have to have a change in her beliefs, she also needs to be aware of a contrast between what she believed prior to the event and what she now comes to believe. Donald Davidson, for example, holds that:

> If I believe I have a coin in my pocket, something might happen that would change my mind. But surprise involves a further step. It is not enough that I have this belief. Surprise requires that I be aware of a contrast between what I did believe and what I come to believe.[147]

Davidson's view can better explain why juror A is not surprised when she changed her belief in the afternoon because she is not keeping track of how her beliefs have changed. However, something crucial is still missing in Davidson's account of surprise. Imagine another juror, B, who at the early stage of a trial believes that the suspect is innocent. After thinking more about the evidence presented, she changes her mind and believes that the suspect is guilty. Even though she has not acquired any new information or evidence, her belief changed from 'the suspect is innocent to 'the suspect is guilty' without being surprised. Unlike A, B is well aware of the change in her belief that results from elaborate deliberation. This is different from a scenario in which C changed her mind, not because her deliberation, but because she sees a newly found video record of the defendant committing the crime. If the video was

---

the article, I reflect on my own surprise reaction and am surprised that I was surprised that I had expected to doctor to be a man. The propositions link to these two instances of surprise are different. In the first instance I was surprised that the doctor is a female. It is a proposition about the world. In the second instance of surprise, I was surprised that I expected that the doctor is a male. The proposition is about my own belief. Sam's surprise needs to link to a proposition about the world, that black swans exist, in order to reveal that he does not believe that black swans exist.

[147] Davidson 1982, p. 326. Jonathan E. Adler 2008 also argued for this strong view of surprise.

released when she still believed that the defendant is innocent, it is very likely that she will be surprised. The difference between B's and C's cases cannot just lie in the difference between reasoning and seeing a video. We can imagine a mathematician who through reasoning comes to realize that a certain theorem does not follow from the axioms of a theory and is surprised. This suggests that there is some additional element that comes with surprise.

My suggestion is that this additional element is captured by the Acquisition Requirement. In order to be surprised that *p,* one has to come to believe that *p* at the time of surprise. The subject has to take herself to be confronted by a state of the world in which *p* obtains such that she has no more room to make up her mind. This does not mean that one cannot be surprised by a certain state of affairs that is arrived at through reasoning. As mentioned before, the mathematician could be surprised that the theorem does not follow from the axioms. Still, she takes this result to be dictated by the world, not up to her. In other words, she acquires the belief that the theorem does not follow from the axioms. This is different from the kind of case we were imagining for juror B. Of course, juror B understands that there is a fact of the matter as to whether the suspect is guilty; but when she eventually arrives at the belief that the defendant is guilty, she still does not take it that the world is dictating her to think that the suspect is guilty. Perhaps in the case of B, the newly self-ascribed belief is formed through ratiocination and the top-down fixation process might not have succeeded. So, B has not yet acquired the belief that the suspect is guilty to contrast with her pre-existing belief that the suspect is innocent. This explains why even though B is aware of a contrast between her self-ascribed belief and her newly self-ascribed belief, she is not surprised.

6.3 Sam's Case Again

We can now take a final look at Sam's case. As I said before, there are many ways of explaining Sam's case. I have only attempted to show that my explanation is a possible one. As long as it is, then the constitutive thesis and the incorrigibility thesis have to be rejected. My explanation of Sam's case is this: Sam has the pre-existing belief that all swans are white. Sam learns about the swans and ratiocinates and concludes that he ought to believe that there are black swans. He therefore, at his conscious level, endorses the proposition that there are black swans. Since he consciously endorses that there are black swans, he then goes on to self-ascribe the belief that there are black swans. However, the top-down fixation process fails to bring about his belief that there are black swans. Perhaps, at the time, whether there

are black swans strikes Sam as too trivial. Between the time he self-ascribes and the time he sees the black swans, he never considers again whether there are black swans. When he sees the black swans, his perceiving the black swans directly fixes his belief that there are black swans. What generates the surprise? His newly acquired belief clashes with his pre-existing belief that all swans are white. Unlike Stoneham's account, which incurs the cost of denying everyday experience and the considerations which make the surprise principle attractive, my explanation allows us to preserve the intuition that Sam genuinely believes that he believes that there are black swans and genuinely feels surprised that there are black swans when he sees the black swans.

This reading of Sam's case further gives rise to an interesting puzzle. For brevity, I will use '*t*' to denote the time Sam saw the black swans, 'B*p*' the belief that black swans exist, and 'BB*p*' the belief that he has the belief that black swans exist). We can at least say the following things about Sam.

(1) Sam has BBp before *t*.
(2) Sam has B¬*p* (or B*q* where *q* is inconsistent with *p*) before *t*
(3) Sam has B*p* at *t*
(4) Sam is aware of a contrast between his B¬*p* before *t* and his B*p* at *t*

The aspect of Sam's case that makes it theoretically interesting is (4). How can Sam be aware that there is a change in his attitude? There are two possibilities. One possibility is that by "awareness", we mean something that is less than belief. Perhaps the mechanism of surprise is one that allows Sam to automatically detect a conflict between his B¬*p* before *t* and his B*p* at *t*. On the psychological cognitive-evolutionary model of surprise, for example, the surprise mechanism is "assumed to consist at its core of an innate, hardwired information-processing device that continuously compares, at an unconscious level of processing, the currently activated cognitive schemas…with newly acquired information (new beliefs)."[148] If awareness just means Sam can detect a conflict, where the detection is made possible by some unconscious or subconscious hardwired mechanism, then Sam would simply be surprised and his surprise reveals that he held a contrary belief before he saw the black swans. This will be just another case of fallibility.

[148] Reisenzein, Meryer and Niepel 2012, p. 565.

Another possibility is that Sam is aware of the change in his attitude in that he at *t* comes to believe that black swans exist and also believes that he believed that black swans do not exist before *t*:

(5) Sam has BB*p* at *t*

Here is the puzzle: at *t*, Sam's BB*p* is true because he actually has Bp. Before *t*, Sam's BBp was false because he did not have the corresponding B*p*. On what grounds can we say that Sam knows that he believes that *p* at *t*? In other words, how can Sam's BB*p* at *t* count as knowledge, if he also had false BBp just before *t*? His BB*p* at *t* seems to fail the safety condition of knowledge, for his BB*p* could have been easily wrong.

Before *t*	At *t*
BBp	BBp
B¬p	Bp

One option is to say that the phenomenological character of actually having the belief that *p* must be different from the phenomenological character of not having the belief that *p*. It is when B*p* is now embedded in BB*p* that Sam can become aware of the contrast between his previous BB*p* that lacks B*p* and his current BB*p* that has B*p*. We can denote the true BBp with the subject possessing Bp as B[+Bp] and the kind of mistaken BBp with the subject lacking Bp as B[-Bp]. The idea is that if one is in fact in Bp, due to the phenomenology of having Bp, one is in an epistemically privileged position relative to Bp. When Sam saw the black swan, he learnt from direct experience that there are black swans and it was at that point that he actually acquired the belief that there are black swans. If this is the case, then there must be something in Sam's direct experience with the black swans that allows Sam to be aware of the difference between actually having a lower-order and not having it.

Yet this explanation has many difficulties. There could be situations where a subject is not able to experience a sharp contrast in mental phenomenology that enables her to tell that she had B[-Bp] and now has B[+Bp]. For example, the change of mental phenomenology from B[-Bp] to B[+Bp] is subtle. We may consider a similar subject, Sarah, who also ascribes to herself the belief that black swans exist without believing that black swans exist. Sarah could have started off with B[-Bp]. But before Sarah saw the black swans, she strolled in the town. She saw drawings of black swans in a restaurant. And, at a souvenir shop, she came

across a counter that sells black swan figurines and black swan photographs.[149] When she finally saw a black swan, she was not surprised. It is possible that at some point during her stroll in town, she acquired the belief that there are black swans. In this scenario, the subject starts off with B [-B*p*] and later acquires B*p* and B[+B*p*]. However, she is not sensitive enough to be aware of the difference in phenomenology between B[-Bp] and B[+Bp] because the change is gradual. From the perspective of the subject, she continues to hold BB*p* without noticing the difference between B[-Bp] and B[+Bp]. From her first-person perspective, no change has occurred. If Sarah herself were to answer this question, her answer to the question of 'How do you know' in both B[-Bp]and B[+Bp] cases will be appealing to nothing more than her conviction that she has Bp. This suggests that from a higher-order, first-person point of view, it is indistinguishable to the subject whenever she actually has the lower-order belief that *p*. Hence, we cannot turn to the phenomenology of B[-Bp] and B[+Bp], even if there is a difference, to serve as the basis of knowledge.

Turning to another proposal, we might hold that Sam's knowledge of his belief at *t* is grounded in the way in which his self-ascriptive belief is formed. Let us denote Sam's BB*p* before seeing the black swans at *t* as $BBp_{t\text{-}i}$ and the BB*p* formed at *t* as BBp_{t}. $BBp_{t\text{-}i}$ is formed on the basis of reasoning, whereas BBp_{t} is formed on the basis of perceiving the black swans.

It will be helpful to recall our earlier discussion that surprise requires the subject to take *p* as the way the world is imposing on her belief. Since Sam is surprised that there are black swans, he must have also taken *p* to be dictated by the world. When he sees the black swans, he is confronted with a state of the world in which *p* obtains, which constrains his belief about *p*. He acquires the first-order belief that *p* through perceiving the black swans. It is on the basis of his first-order B*p* that his BBp_{t} is formed. By contrast, when Sam reasons, he does not take himself to be confronted with a state of the world that dictates his belief that *p*. He reasons that he ought to believe that *p*. It is on the basis of ratiocination that his $BBp_{t\text{-}i}$ is formed.

This approach is also problematic, for there are many successful cases of top-down fixation. We can imagine a successful case in which one starts at t_1 with the state of believing that she believes *p*. She does not immediately self-ascribe a belief. At t_2, her top-down fixation is successful. She then self-ascribes the belief that *p* at t_3. In this case, she is in the state of B[+Bp] and the way in which her self-ascriptive belief is formed is also through reasoning.

[149] I thank Mike Martin for suggesting this example.

In the remaining sections, I will further develop the challenge Sam's case poses for self-knowledge.

6.4 Sceptical Challenge

Once we accept that BB*p* and B*p* are distinct states, the problems that the constitutive account avoids re-emerge. Some might worry that we will have to take the relation between B*p* and BB*p* to be a causal one in a way that parallels perceptual knowledge. This raises the question whether there is anything epistemically secured about self-knowledge. It is possible that one is self-blind and comes to know her mind in a third-personal way.[150] Accounts that defend the security of self-knowledge will have to say why a subject's belief about her belief must be true or cannot be wrong in certain ways. Here, I will not be able to offer a positive account on the assumption that BB*p* and B*p* are distinct states that defends self-knowledge in general. What I have said in this thesis only applies to the case of belief. What I attempt to do, on the assumption that BB*p* and B*p* are distinct states, is to lay out the many difficulties of offering an account that defends the view that a subject knows what she believes, once we accept that BB*p* and B*p* are distinct. The many difficulties add up to a sceptical worry about whether any special safety attaches to self-knowledge. Below, I will focus on two ways in which a subject might make a false self-ascription of belief.

First, one might believe that she believes *p* when in fact she believes not-*p* and does not believe *p*. In a case like this, one correctly believes that she has a view on *p* but incorrectly believes that her view on *p* is that *p* is the case. If BB*p* and B*p* are distinct states, it is possible for one to believe that she believes *p* when one in fact has no view on *p*. This worry assumes that there is a lower-order belief, and then questions whether one gets one's own view right. One way to address this worry is to start at the level of B*p* and explain how one's B*p* necessitates BB*p*. Parrott, for example, is a recent example of this approach. According to Parrott, it is in virtue of an agent's rationality that once one has B*p* necessarily one has BB*p*.[151] I will not directly evaluate accounts that adopt this approach that starts at the level of B*p* here. Instead, I raise a different worry that starts at the level of BB*p*. Once we

[150] Shoemaker 1994.
[151] Parrott 2017. Parrott thinks that one's BB*p* is sensitive to one's judgment for *p*. He must have assumed that judging that there are reasons for believing *p* necessarily brings about believing *p*. But in ratiocination, judging that there are reasons for p does not amount to believing that *p*.

come to see this worry about reflectively-formed BB*p*, we will also see the difficulty with the approach that starts at the level of B*p*.

Second, one might falsely believe that she believes *p* when in fact she does not believe p and does not believe not-p. In a case like this, one falsely believes that she has a view on *p* when in fact she has no view on *p* at all. Let us assume that one's BB*p* can be formed at the reflective level based on one's reasons for endorsing *p*. It does not have to be caused by B*p* or be dependent upon B*p* in any way. The three claims I have been pressing – that (1) BB*p* can be reflectively formed on the basis of deliberation about *p*, (2) BB*p* and B*p* are distinct states, and (3) deliberation can fail to bring about B*p* – jointly suggest that even if we have an account that says B*p* entails BB*p*, we cannot get from the claim that B*p* entails BB*p* to the claim that BB*p* amounts to knowledge. For, since BB*p* does not entail B*p*, it is possible that one has BB*p* without believing *p*.

We need to further investigate the nature of self-ascriptive beliefs. Suppose we want a uniform account for the nature of self-ascriptive beliefs. My account suggests that it will be very challenging to arrive at a uniform account. Some self-ascriptive beliefs are formed bottom up and some self-ascriptive beliefs are formed reflectively, independent of B*p*. The approach of relying on the claim that B*p* entails BB*p* does not apply to BB*p* that is reflectively formed. It does not tell us why a subject's belief that she believes *p* must be true. Hence, even if we assume that one's self-ascriptive belief has a special epistemic status, it cannot be because BB*p* contains B*p*. To arrive at a uniform account, we will have to resort to something else other than the presence of the corresponding belief. Perhaps we can explain the nature of self-ascriptions in terms of linguistic conventions or expressions of one's belief.[152] The problem with this is that it can only explain why self-ascriptions are authoritative. But if we still have the intuition that sometimes one does have beliefs about her beliefs, such as when answering surveys in a psychology lab or when confessing to a confidant about the 'silly' beliefs one holds, accounts that appeal to the nature of self-ascriptions in terms of linguistic conventions or expressions cannot explain why a subject has epistemic entitlement to her belief about her beliefs. As long as we agree that a subject can have a mistaken belief about her belief, no matter how infrequent such mistakes are, we still need an explanation for her epistemic entitlement.

Suppose we do not try to give a uniform account for self-ascriptions and give instead a different account for self-ascriptive beliefs that are formed reflectively and self-ascriptive

[152] For the former, see Wright 1998. For the latter, see Bar-On 2004.

beliefs that are formed bottom-up. Recall self-ascriptive beliefs that are formed reflectively are those self-ascriptive belief formed after deliberating about the truth of *p*. Self-ascriptive beliefs that are formed bottom-up are those that are formed on the basis of lower-order beliefs. For example, Percy sees that there is a cup on the table and believes that there is a cup on the table. Percy then believes that he believes that there is a cup on the table. The obvious problem for an approach like this stems from the fact that it will have to differentiate self-ascriptive beliefs on the basis of the different grounds they are formed. The ground for self-ascriptive beliefs formed reflectively are evidence from the world about *p*; the ground for self-ascriptive beliefs formed from the bottom-up are the state of mind one is in, namely, B*p*.[153] For BB*p* that is formed from the bottom-up, there might not be any evidence for *p* that gets carried over to BB*p*. A new mother might not have evidence for her belief that everyone else is better than she is, and yet she believes this and truly believes that she believes this. In such a case, her BB*p* does not include any evidential base for *p*. For BB*p* that is formed reflectively, a subject's evidence for *p* is carried over to her BB*p*. However, it is possible that one does not in fact believe *p*. Then we get the astonishing result that, for self-ascriptions made on the basis of deliberating about *p*, one might end up only with a view on her mind – that she believes *p* – but lacks a view that is connected to the world. For self-ascriptions that are made on the basis of one's lower-order belief without regard to evidence for *p*, whether her belief about *p* is false and even if her belief about her belief about *p* is false, her view is still connected to the world.

The transparency account is supposed to tell us why one is entitled to believe that she believes *p* by attending to the world. However, we come to see that when one self-ascribes a belief after one has attended to the world rather than attending to her own mind, one becomes especially susceptible to the possibility that one ends up only with a view on her mind and has no view on the world. This raises the question as to whether there is something that is particularly insecure about self-ascription that is made by adopting the Outward Looking method.

[153] This does not mean that the subject herself can offer evidence for BB*p*. If she does offer evidence, then her self-ascriptions will be third-personal and with which she lacks identification. This goes against the assumptions about self-ascriptions we made that they have to be what the subject identifies with in a first-personal way.

Those who offer non-uniform accounts for BB*p* formed bottom-up and BB*p* formed reflectively might claim that they can at least provide an account of the nature of self-ascriptive beliefs that are formed bottom-up. For nothing I have said impugns bottom-up formation. But things do not seem that optimistic. Let us use ^{B}B*p* for BBp formed reflectively, and $_{B}$B*p* for BB*p* formed bottom-up. We are considering a non-uniform account that explains ^{B}B*p* and $_{B}$B*p* differently. But note that the question now is not how we explain the difference but why one's knowledge of her belief is epistemically secured. We have to take into account a subject's first-person perspective from the inside, and she is unable to tell whether she currently has BB*p* ^{B}B*p* or $_{B}$B*p*. From her first-person perspective, she simply believes that she believes *p*. Perhaps one might say that when a subject is in BB*p* formed reflectively, she focuses on *p* when she ascribes to herself the belief. In the bottom-up case, her focus is on her state of mind. It is not *p* that she is affirming, but the feeling that she does have a certain state of mind. This will not work because if we assume that one has moved from BOB*p* to BB*p* then one also focuses on what she believes. When she answers, 'Do you believe that there are feathered dinosaurs?' the focus shifts from affirming *p* (a state of the world) to affirming a belief that she has (a state of her mind).

One also cannot tell whether she is in B[-B*p*] or B[+B*p*]. Some might want to appeal to a difference in phenomenology between B[-B*p*] or B[+B*p*]. However, let us imagine one ratiocinates and consciously endorses *p* at t_1 but top-down fixation fails to bring about the belief that *p*. If one ratiocinates again and consciously endorses *p* at t_2 and top-down fixation is completed, there is no change in phenomenology at t_2. To the subject's mind, she just continues to believe that she believes that *p*. This suggests that even if there is a change in phenomenology, the change has to be only generated by B*p* that is acquired through non-reasoning means. However, there are many things one believes that she believes without ever having any direct experience and one just has to reason it out. This is even more so in the cases of belief that something does not exist. Reasoning is the only way to fix belief about things that do not exist. One who believes that unicorns do not exist does not have any direct experience to appeal to. Suppose her belief that she believes that unicorns do not exist move B[-Bp] to B[+B*p*], there is no contrast in phenomenology for her to tell that now she really believes that unicorns do not exist.

In any case, let us assume that there is a difference in phenomenology between B[-B*p*] or B[+B*p*]. As my discussion of Sam's and Sarah's cases show, even if there is a certain phenomenology of being in BB*p* once one is in B*p*, this change in phenomenology can be so subtle that a subject is hardly in a position to notice a change from B[-B*p*] and B[+B*p*].

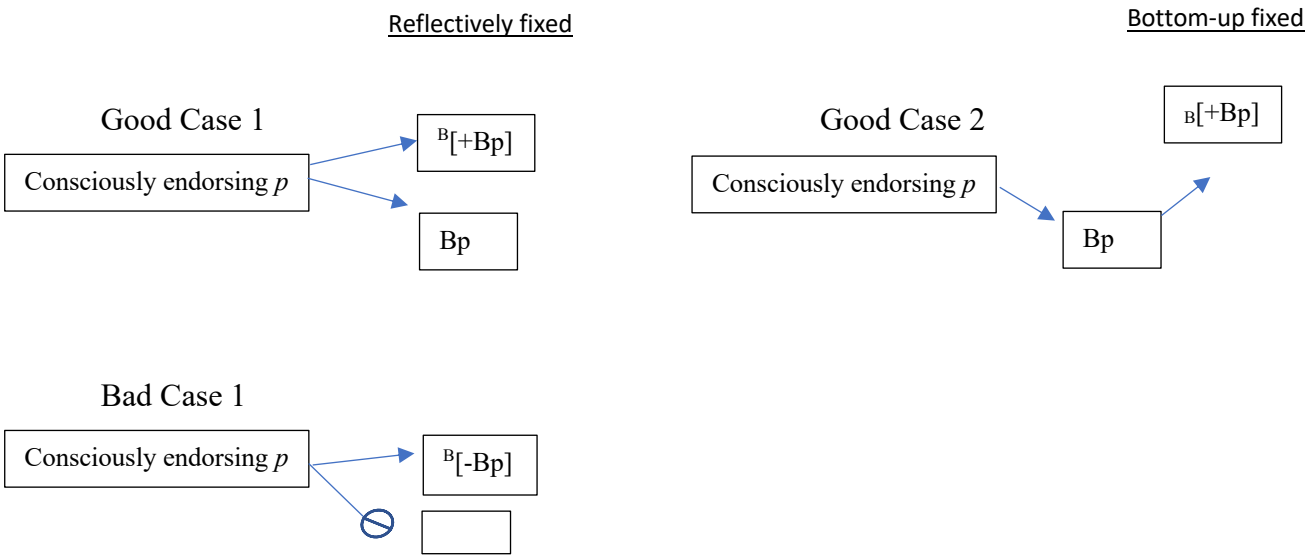
Somewhere along Sarah's stroll in town, her B[-B*p*] becomes B[+B*p*]. But all along, from her first-person perspective, she believes that she believes there are black swans. If one were to ask her before she saw the black swan figurines and postcards, she would have self-ascribed the belief that there are black swans. If one were to ask her after she saw the black swans, she still would have self-ascribed the belief that there are black swans. From Sarah's perspective, she continues to believe that she believes that there are black swans without noticing any change in the way she feels about what she believes. Sarah's answer to the question of 'How do you know?' in both B[+B*p*]and B[-B*p*] cases will be appealing to nothing more than she believes that she believes there are black swans. We should not be willing to grant Sarah's later B[+B*p*] the status of knowledge. Her B[+B*p*] could easily be wrong. Suppose, on another variant of the case, all shops were closed because of a holiday. She didn't see the black swan figurines nor postcards. In this case, she will, like Sam, go on to hold B[-B*p*].

Moreover, it is unclear why the presence of a certain phenomenology of being in Bp should be the basis upon which we grant authority to the subject's B[+B*p*] but not her B[-B*p*]. Imagine another subject, Sascha, who also ratiocinates and self-ascribes the belief that there are black swans. Unlike Sam, Sascha's top-down fixation process goes through and she comes to believe that there are black swans. From a third-person perspective, we can tell that the difference between Sam and Sascha lies in how the former has B[-B*p*] whereas the latter has B[+B*p*] before *t*. But from a first-person perspective, if both attended to the world, consciously endorse *p*, are being sincere in their self-ascription, there does not seem to be good reasons for conferring authority to Sascha's belief that she believes that there are black swans but not to Sam. From Sascha's perspective, she has no contrast in phenomenology for her to say that 'now I really believe that there are black swans'.

This problem generalises to all cases of BB*p* including the higher-order beliefs that are in fact formed bottom-up. Suppose we have an account that says that necessarily if one is in B*p*, one also is in BB*p*. However, since one can be in ^{B}B*p* without believing *p*, and since one cannot tell whether she is in $_{B}$B*p* or ^{B}B*p*, nor can she tell if she is in [+Bp] or B[-Bp], in what way can we say that when a subject believes that she believes *p*, her self-ascriptive belief amounts to knowledge? Even in a good case where a subject's true BB*p* is formed bottom-up, it could easily be the case that her BB*p* is false. Suppose Rae ratiocinates and concluded that there are feathered dinosaurs at t_1 and the top-down fixation process succeeds. She comes to believe that there are feathered dinosaurs at t_2. She then comes to believe that

she believes that there are feathered dinosaurs at t_3. Rae's $_{B}Bp$ at t_3 is formed bottom-up and is true. However, it could have easily been false had the top-down fixation process failed.

A visualization might help elucidate the worry.



Rae's self-ascriptive belief is like Good Case 2. But she could have easily been in Bad Case 1. Hence, even the non-unified account we have been exploring cannot satisfactorily explain how one knows that she believes *p*.

These puzzles are particularly difficult to address for knowledge of belief instead of other mental states for two reasons.

First, as those who defend the transparency account point out, agential involvement might be present. It could be up to us to attend to evidence and make up our mind about whether *p* is the case. And one of the dominant ways we come to make up our mind is ratiocination. But, since ratiocination starts at the conscious level and BB*p* also starts at the conscious level, we have the reflective-level worry. Other more 'passive' mental states that start at the bottom level and intimate itself to the conscious level might be immune to this worry.

Second, belief is the kind of mental state that is truth-governed. For one to believe *p* is for one to take *p* to be true. When one consciously takes p to be true after ratiocination, then it is rational for one to believe that she believes *p*. To believe *p* is to be in a mental state that is about the world. When one has a belief about one's belief about *p*, one has to be in a mental state that is about one's own mental state about the world. It is about the mind. The

subject, from her first-person perspective, is unable to tell whether she has a view on the world or a view on her mind. For the mind she is supposed to represent is also representing the world. How should we answer a question about the mental state when that mental state makes reference to the world?

In this chapter, I have discussed the metaphysical and epistemological implications of my account. I argued that BB*p* and B*p* are distinct states. I did not rely on the claim that BB*p* and B*p* are distinct states to argue for the possibility of BB*p* without B*p*. Instead, I relied on my account of ratiocination to establish the possibility of a rational subject believing that she believes *p* without believing *p*. Once this possibility is established, it follows that BB*p* and B*p* are distinct states. And since it is possible for a subject to be in BB*p* without being in B*p*, the incorrigibility thesis is also rejected.

A deeper disagreement emerges from my discussions of the constitutive thesis and the incorrigibility thesis of self-ascriptive belief. Although the constitutive thesis and the incorrigibility thesis conflict over whether BB*p* and B*p* are distinct states, they share the assumption that a subject's first-person perspective on her belief is always tied to her view on the world. Since both theses hold that BB*p* is necessarily tied to B*p*, both hold that a subject's first-person view on her belief is tied to the world. They think that the self-ascriptive belief necessarily expresses one's take on the world. I have suggested that self-ascriptive belief does not always express one's take on the world but one's take on one's own mind. The problem with constitutive accounts is not simply that they tell an incorrect story about the relation between second-order belief and lower-order belief, but that they have not captured the nature of self-ascriptive belief.

Conclusion

The main goal of this thesis is to argue that it is possible for a subject to believe that she believes *p* when in fact she does not believe *p*. My account is novel insofar as I claim that such mistaken self-ascriptions of belief need not be the product of irrationality on the part of the subject. A subject's mind can rationally move from consciously endorsing *p* to believing that she believes *p* even though she does not believe *p*. From her first-person perspective, since she rationally endorses *p*, if she were to ascribe a belief to herself and if she is rational, she believes that she believes *p*. She might from the third-person perspective doubt whether she really believes *p* but she cannot first-personally believe that she does not believe *p* first-personally. But for some propositions, even if she is rational, she might not succeed in believing those propositions through top-down fixation. The reason why she may fail to make true self-ascriptions of belief could be due to our general psychological limitations. These include the breaking down of physical processes that realise psychological states, acquiring beliefs through non-reasoning methods, and not forming trivial beliefs to preserve cognitive processing capacities. Hence, she is not necessarily irrational when she believes that she believes *p* even though she in fact does not believe *p*.

That a rational subject can believe that she believes *p* when she does not believe *p* suggests that instances of mistaken self-ascriptions might be more prevalent than expected. Pathological, irrational, or motivated self-deception cases are not the only sort that give rise to mistaken self-ascriptions of belief. It can happen in everyday cases and to average rational agents.

I have avoided loaded, for example, practically motivated, cases in this thesis to focus on ratiocination and the top-down fixation process. If my argument is sound, then we have one more alternative to explain *some* of the mistaken self-ascription cases that are often grouped under 'self-deception' or 'irrationality.' Perhaps the subject matter is not that interesting to the subject or perhaps there are other ways of fixing belief (e.g., perception, memory) that short-circuit one's reasoning process and fixes one's belief directly at the bottom level. For example, it is possible that when a husband says, 'my wife is faithful,' he believes that he ought to believe that his wife is not having an affair. But perhaps he perceives other cues of unfaithfulness and that short circuit his reasoning and fix his belief

that she is unfaithful. I do not mean to suggest that self-deception is not possible. I only suggest that, for some purported cases of self-deception, the explanation could be that top-down fixation fails.

Another difference between my account from others is that I do not assume that lower-order states are non-luminous. On my account, it is possible to have higher-order belief without the corresponding lower-order belief because the higher-order belief does not await the presence of the corresponding lower-order belief to be fixed. If one assesses the reasons for *p* and if one endorses *p*, it will be irrational for one not to believe that she believes *p*. From her first-person perspective, to answer a question about *p* is just to answer a question about what she believes. However, just because one endorses *p*, it is not guaranteed that one will believe *p*. Hence, we get the possibility of believing one believes *p* without believing *p*. My account thus shows that a subject's belief about what she believes need not be connected to her view on the world. She could simply have a view on her mind.

My account makes vivid the importance of attending to the level of fixation. When a subject thinks '*p*' at the end of ratiocination, the state she in fact is in is the state of believing that she ought to believe *p*. When a subject makes a self-ascription after ratiocination, she is in the state of believing that she believes *p*. To her mind, she cannot tell the difference between being in the state of believing that she ought to believe and being in the state of believing that she believes *p*.

Once we see that self-ascriptive belief can be fixed at the reflective level, but the deliberation process that fixes the self-ascriptive belief does not always fix the corresponding lower-order belief, worries about self-knowledge arise. The subject cannot tell whether she really holds the belief that *p* or whether she is only in a hallucinatory state where she believes that she has a view on the world when in fact she only has a view on her mind. Accordingly, even in good cases where her self-ascriptive beliefs are true, they could have easily been mistaken. Is self-knowledge epistemically special? My arguments suggest that a negative answer is perhaps more plausible than one might have thought.

My discussion is restricted to belief about belief. It might not have implications for one's knowledge of one's other mental states. But assuming that our understanding of the nature of self-knowledge depends heavily on our understanding of the nature of self-ascriptions, if we cannot say how we know our beliefs, then we might also have difficulty to say why one is entitled to believe that she is in any mental state. That is a broader question, which I will not address.

Appendix: The Surprise Principle

The 'surprise principle' states:

> If a subject S is surprised that *p,* at time, *t*, then S has an attitude that is contrary *to p* prior to *t*.

a. *Surprise and startle*

Let us first define the word 'surprise' in the surprise principle. Surprise and startle are two emotional responses that are closely linked. According to Jenefer Robinson, 'Startle is a reflex, an involuntary response that requires no prior learning and occurs too rapidly for there to be any cognitive activity at all'.[154] This suggests that startle can be triggered by any novel or intense stimulus that interrupts an ongoing activity. It does not necessarily take a belief or disbelief to be disconfirmed in order to generate startle response. If one is startled by the sound of loud thunder while working in an office, one does not necessarily believe that there would not be any thunder. The sound of the thunder could be so sudden and intense that her startle response is a way of her body telling her that something new has occurred in the environment and readies her to adjust to and deal with the situation by breaking the ongoing program of her nervous system.

Psychologists have observed that surprise is a developmentally later form of startle and shares many important features with startle. [155] For example, similar to a startle, surprise is characterised by distinctive facial expressions,[156] and has the biological function of preparing the subject to deal with a new or sudden situation.[157] Still, there are important distinctions between startle and surprise. While startle is a non-cognitive emotional response, surprise could be a cognitive emotional response. While an intense and sudden stimulus is

---

[154] Robinson 1995, p. 59.
[155] Izard (1977) for example, does not draw a distinction between startle and surprise and has used the terms interchangeably. Charlesworth (1969) argues that surprise cannot occur until five to seven months of age. According to Klaus Scherer (1984) argues that startle is present at birth and surprise tends to appear between one and three months (293-317).
[156] See Robinson 1995: 58 for the facial expressions that are characteristic of startle and Izard 1977, p. 277 for the facial expressions that are characteristic of surprise.
[157] Both Silvan Tomkins and Carroll Izard have pointed out that surprise has the function of clearing neural pathways so that an organism can respond to novelty and/or changes in the environment. See Izard 1977, p. 281.

sufficient to trigger startle, the stimulus that triggers surprise must be one that is unexpected by the subject.[158] For example, a marksman who repeatedly fires a gun would not be surprised by the sound of a gunshot but might still exhibit startle response.[159] When one is startled, her startle response only shows that she finds a certain stimulus intense or novel, but it does not tell us whether she expects the stimulus. When one is surprised, her surprise shows that she finds the stimulus unexpected. It is this unexpectedness of the stimulus that distinguishes surprise and startle. One's surprise tells us something about the doxastic state that she is in, whether she expects or does not expect something to be the case.

This thesis is only concerned with surprise that involves cognition, and I leave it open that there might be some forms of surprise that more closely resembles startle.[160] I only discuss surprise that involves cognitive activity.[161] Since startle is a non-cognitive emotional response, I will not be considering it here. I concede that it is unclear whether a sharp distinction can be drawn between startle and surprise, for the difference between the two could be separated by a continuum of cognitive complexity. However, it suffices to say that we can at least approximate a distinction between startle and surprise, with the latter being a more complex psychological response that involves cognition. The point of drawing this distinction is not to suggest that startle and surprise are structurally or functionally different, but rather, to help us focus more on the question of what surprise can tell us about a subject's doxastic state. As Charlesworth notes, if surprise is largely construed in terms of its biological taxonomies, we risk ignoring how surprise can be diagnostic of the presence or

---

[158] Both Plutchik 1980 and Robinson 1995 note that surprise involves the unexpectedness of the stimulus, but the startle does not.

[159] Photographs of these trained marksmen show that they still blinked and exhibit facial expression characteristic of the startle when they fired a pistol. Robinson 1995, p. 55.

[160] Scholars are divided on whether surprise is a basic emotion. Like other basic emotions, surprise is characterised by particular facial expressions, physiological and biological reactions, and self-reported sensations. But there are also studies that show subject reports of surprise and eyebrow movement measured by frontal EMG do not correlate. Moreover, unlike other basic emotions, surprise is a short-lasting response without definite positive or negative value. Charlesworth 1969 presents an extensive literature review on surprise from Darwin to the the1960s. See also Ortony et al. 1998, Vanhamme 2000, and Sumitsuji 2000 for more recent discussions of whether surprise can be characterised as a basic emotion.

[161] What I mean here is close to the kind of "developmentally sophisticated" emotions Robinson discusses (1995:64). According to Robinson, the basic or primitive emotional responses do not involve cognitive activity, but the more developmentally sophisticated emotional responses do.

absence of cognitive structures.[162] Our focus is on the more complex surprise response that is indicative of the doxastic state that one is in.

b. *Surprised that p*

If the surprise principle is true, there is a determinate link between one's surprise and a proposition (even though the subject herself might have trouble articulating what the proposition is). In everyday life, we are not often required to be precise about the proposition to which our surprise is linked. For example, when we surprise a friend with a gift, we normally will not press her to say exactly what she is surprised about. Conceptually, though, we can make a distinction between the event that triggers surprise and the ground for surprise. One can be surprised by an event that they believed would occur. For example, if my friend told me that she would make me a painting, I could still be surprised when she gives the painting to me. Although I believed that she would make a painting for me, I could still be surprised that the painting is beautiful, or that it took her such a short time to make a fine painting. It takes some salient features of the event to trigger one's surprise. Hence, even though the same event might trigger surprise in different individuals, the grounds of surprise could be different for those who are surprised. My partner and I may both be surprised when my friend presents her painting. While my partner is surprised that it is a painting of a beach, I am surprised that it is a painting of my favourite beach.

The surprise principle claims that, if one is surprised that *p*, then one must have held an attitude whose content is contrary to *p*.[163] In order to determine which propositions one's surprise is linked to, we need to be precise about the ground for one's surprise. Since the grounds for surprise vary, it is possible for Sam to be surprised even though he did believe that black swans exist. He could be surprised that black swans have red bills or that he is lucky enough to see black swans on his trip. However, if it is specified that Sam is surprised that black swans exist, then, by the surprise principle, Sam must have held an attitude that is contrary to the existence of black swans.

[162] Charlesworth 1969, see especially pp. 258-60.
[163] One might have inconsistent beliefs such as 'I will not publish a paper in a journal with 90% rejection rate' and 'I will publish a paper in a journal with 10% acceptance rate' (Adler 2008). It depends on how the event is described. If she sees the paper as being accepted by a journal with 90% rejection rate, she will be surprised. If she sees the event as my paper gets accepted by a journal with 10% acceptance rate, she will not be surprised.

c. *Belief and expectation*

The surprise principle is formulated generally as to accommodate a wide range of surprise cases. It does not discriminate between cases where the surprised subject believes not-*p* prior to *t* and cases where the surprised subject does not believe that *p* prior to *t*. An 'attitude that is contrary to *p*' can be understood strictly to mean a state of 'disbelief' (B¬*p*) or loosely to mean an attitude that does not necessarily amount to disbelief (¬B*p*).[164]

There are three possible ways to understand 'attitude that is contrary to *p*': (1) The belief that not-*p*; (2) an attitude that is contrary to *p* but does not necessarily amount to the belief that not-*p*; (3) an attitude that is contrary to *p* that does not necessarily amount to a belief. These three different possibilities say different things about the strength of the attitude that is necessary to generate surprise.

(1) is too strong. It is true that if I believe not-*p*, I will be surprised if it turns out to be the case that *p*. For example, if I believe that my friend is out of town, I will be surprised if I see her in town. However, it does not seem necessary for one to believe not-*p* in order to be surprised. Compare the following two scenarios. In Scenario A, Sara was nominated for an award and believed that she would not win. It turns out that she won. Sara is surprised. In Scenario B, Sara bought groceries as usual at a small local store. She was not aware that the owner decided to give one million dollars to the 101st customer of that day. She was the 101st customer and won one million dollars. Sara was surprised. In Scenario B, Sara did not have any belief about winning one million dollars prior to the owner telling her that she won that sum. However, it is conceivable that Sara was genuinely surprised. This suggests that an upset belief B¬*p* is not necessary for surprise. Instead, one can be surprised that *p* even if one does not believe that not-*p* (¬B*p*).

If a subject's surprise at *p* implies that she believes that not-*p*, we must also be assuming that we are constantly holding beliefs ruling out all the possible ways the world could turn out to be, though there is a vast number of ways that the world could surprise us. As subjects with cognitive limitation, there must be some possible states of the world that I have not even entertained before the surprise. In situations like Sara's in Scenario B, we simply do not have any view about *p* before we learn about *p*.

---

[164] As Quine and Ullian (1978) remind us: 'Disbelief is a case of belief; to believe a sentence false is to believe the negation of the sentence […] Nonbelief is the state of suspended judgement: neither believing the sentence true nor believing it false' (12).

(3) seems to be too weak. According to the psychological cognitive-evolutionary model of surprise, surprise is 'the signal that is the immediate output of the schema discrepancy detector'.[165] The surprise mechanism is a hardwired mechanism that compares and detects at an unconscious level the discrepancy between pre-existing cognitive schemas and newly acquired beliefs.[166] It is difficult to see how these pre-existing cognitive schemas are not generated by some belief.

(2) is the most plausible. As mentioned, this moderate view does not take an upset B¬*p* to generate surprise. However, an immediate worry with saying that an upset B¬*p* is not necessary for surprise has to do with the lack of a robust link between ¬Bp and one's being surprised that *p*. When one is surprised that *p*, she also learns that *p*. That she learnt that *p* suggests that she did not have the belief that *p*. However, we often learn things without being surprised. Students, for example, learn many new things in class without being surprised. It would not make sense to say that one can be surprised when one does not have any attitude towards *p*. There must be something stronger than ¬Bp in order to generate surprise.

Many philosophers—e.g., Davidson (1982) and Dennett (2001)—take surprise to be linked to the violation of expectation. There is much to be said in favour of this view. It seems reasonable to suggest that surprise requires the violation of a prior expectation. Further, to say that *p* violates an expectation is not the same as saying *p* is unexpected. There are many things about which we do not have any expectation—for example, I do not have any expectation about whether my neighbour's phone number ends with an even digit or an odd digit. When I learn that her phone number ends with an odd digit, I am not surprised. In order to be surprised, one must have a certain expectation about whether *p* is true. Hence, we may draw a distinction between the presence of expectation concerning *p* and the absence of expectation concerning *p*. It is the presence of expectation that is necessary for surprise.[167] Moreover, this connection between surprise and the violation of expectation is endorsed by psychologists. In studies that suggest infants can be surprised, the infant must be first accustomed to a repeated stimulus, and it is only when the infant's attention to a stimulus drops that the infant is considered familiar to the repeated stimulus. When the infant spends

[165] Reisenzein, Meryer and Niepel 2012, p. 656.
[166] *Ibid*., p. 565.
[167] Charlesworth 1969 suggests a distinction between misexpected and unexpected events (pp. 257-273). They correspond to what I respectively call the presence of expectation and the absence of expectation. Since 'unexpected events' can be confused with 'unexpected in the strict sense', it will be clearer to first draw distinction between having an expectation and not having any expectation at all.

more time looking at a novel, inconsistent stimulus, the infant is arguably surprised because the new stimulus violates an expectation that is formed through habituation.[168]

Some may argue that a distinction between expectation that not-*p* and belief that not-*p* cannot be drawn. Dennett, for example, suggests that expecting something implies the presence of a belief. He writes:

> Surprise is a wonderful, dependent variable, and should be used more often in experiments; it is easy to measure and is a telling betrayal of the subjects' *having expected something else*. These expectations are, indeed, an overshooting of the proper expectations of a normally embedded perceiver-agent; people shouldn't have these expectations, but they do...They are also, of course, highly reliable signs of their 'ideological commitments'[…] Surprise is only possible when it upsets belief[169]

Dennett's point is that if one is surprised by *p*, then she must have expected that not-*p*; furthermore, her expecting that not-*p* shows that she has 'ideological commitment' to not-*p*, which is a subclass of belief. For example, if subjects are surprised by experimental demonstrations of change blindness, which show we do not have a snapshot-like visual experience that represents a visual scene in high resolution and detail, then they must have believed beforehand that visual experience is like snapshots. It seems that for Dennett there is no difference in saying that one's expectation that not-*p* is violated and one's belief that not-*p* is violated. Hence, it is not possible to be surprised without having a prior B¬*p* being upset. If Dennett is right, then we will have to accept (1).

One way to defend (2) is to draw on the cognitive evolutionary model of surprise to make a distinction between 'misexpectedness' and 'unexpectedness in the strict sense'.[170] Misexpectedness occurs when one's expectation that ¬*p* is disconfirmed by a state of the world in which *p* obtains, and the expectation that ¬*p* is in turn generated by the belief that ¬*p*. 'Unexpectedness in the strict sense' occurs when one's expectation, generated by a background dispositional belief, is inconsistent with *p* and disconfirmed, even though the subject does not have the belief that ¬*p* that generates the expectation that ¬*p*. If we take 'attitude contrary to *p*' to mean both 'misexpectedness' and 'unexpected in the strict sense', then we also allow something weaker than B¬*p*—such as one's background beliefs, whose

[168] Casati and Pasquinelli 2007, p. 174.
[169] Dennett 2001, p. 982, italics original.
[170] This distinction is drawn on the cognitive-evolutionary model of surprise. See Reisenzein, Meryer and Niepel (2012).

content entails ¬*p*—to generate surprise. This will avoid the reduction of all expectations of that not-*p* into B¬*p*. It may be said that in Sam's case, for example, it is not the case that he believes that there are no black swans.[171] However, since he grew up in a town where only white swans have been sighted, he was accustomed to expecting that all swans are white. His expectation about how the world would turn out to be clashes with how the world actually turned out to be, and it is precisely this clash that generates the surprise.

If it is correct to think that surprise requires an expectation to be upset, we may ask whether the expectation requires or amounts to a belief. Even if it does not take an upset B¬*p* to generate surprise, it may still be argued that the expectation that not-*p* must be generated by some belief. Noë suggests that one may be surprised at a situation in which *p* is true, not because one believes that not-*p*, but because one believes *q*. For example, subjects who are surprised by results of change blindness do not necessarily have the belief that visual experience is like snapshots. Nonetheless, they have some belief that we may not be bad at detecting changes in the visual scene.[172] In Noë's view, it takes some prior belief *q* to be upset for one to be surprised that *p*.

One may also argue that it is too demanding to reckon that one must consciously think of propositions in order to be surprised. As Malcolm suggests, we may 'consciously think' that something is the case without having a thought in propositional terms. Malcolm gives the example of a man walking gingerly on a slippery path; the man may consciously think that the path is slippery without thinking of the proposition, 'This path is slippery'.[173] It is possible that one can be surprised that *p* without having any thought that *p*.

A weaker view, suggested by Casati and Pasquinelli, is that expectation is not reducible to a full belief, yet it still needs to be generated by some general dispositional or ideological beliefs.[174] On this view, even though it does not take my belief that there will not be any house parties to generate surprise, some general belief such as the belief that there will not be a random special event at my flat has to be violated. This general belief generates the volatile representation that my flat is more or less in the same state as I left it in. The volatile

---

[171] This reading differs from that of Davidson. Davidson seems to assume that 'conscious thinking' and 'believing' are the same for Malcolm and takes Malcolm's claim to be that: if a creature is aware that *p*, the creature believes that *p*, though in Malcolm's original paper, his focus is always on conscious thinking and has not discussed belief in a technical sense. Davidson 1982, p.102.
[172] Noë 2002, p.7.
[173] Malcom 1972-3, p.6.
[174] See Casati and Pasquinelli 2007.

representation is violated when I find out that there is a party in my flat. It is also possible that the expectation that not-*p* is generated by something like what Frankish calls 'level 1 belief'. For Frankish, 'the term "belief" can also track states that are "multi-track behavioural dispositions, which are non-conscious, passive, graded, and holistic"'[175]. 'Level 1 belief' refers to states that operate at the lower-order level and is akin to behavioural dispositions, rather than the kind of beliefs that are akin to opinion or a commitment to use *p* as a premise in reasoning. For example, if a child is presented with a new object, similar in size and shape to edible objects she is familiar with, the child might be surprised that the new object is not edible. Since the child has some experience with a similar situation, she may be able to make a prediction about a similar situation by drawing on some general belief in expecting a certain state of the world to obtain.

This being said, we do not need to settle the above controversies. At this point, it suffices to say that one's being surprised that *p* is inevitably bound up with a violated belief, though the violated belief is not necessarily the belief that not-*p*. The crucial point is that if one is surprised, one's take on the world is necessarily violated. This 'take on the world' can mean one's belief that not-*p* or something weaker, such as one's general disposition that is contrary to *p*. This allows us to leave the surprise principle basic enough to accommodate a broad spectrum of surprise cases involving different levels of cognitive complexities.[176]

[175] Frankish 2012, p. 24.

[176] Suppose an adult and a toddler both see a ball roll under the sofa and are surprised that they cannot see the ball when they look under the sofa. The adult's surprise is generated by a clash with an explicit pre-existing belief that 'the ball is under the sofa'. The toddler, by contrast, does not have the full-blown belief that the ball is under the sofa, but has some general belief about object permanence.

Bibliography

Adler, Jonathan. ‘Akratic believing?’ *Philosophical Studies* 110, no. 1 (2002): 1–27.

Adler, Jonathan. ‘Surprise.’ *Educational Theory* (2008): 149–173.

Archer, Sophie. ‘Defending exclusivity.’ *Philosophy and Phenomenological Research* 94, no. 2 (2015): 326–341.

Ball, Brian. ‘The nature of testimony: a Williamsonian account.’ *Logique et Analyse* 56, no. 223 (2013): 231–244.

Bar-On, Dorit. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Oxford University Press, 2004.

Bennett, Jonathan. ‘Why is belief involuntary?’ *Analysis* 50, no. 2 (1990): 87–107.

Bilgrami, Akeel. *Self-Knowledge and Resentment.* Massachusetts: Harvard University Press, 2006.

Boghossian, Paul. ‘What is inference?’ *Philosophical Studies* 169, no. 1 (May 2014): 1–18.

Boyle, Matthew. ‘Active belief.’ *Canadian Journal of Philosophy* 39, no. sup1 (2009): 119–147.

Boyle, Matthew. ‘Transparent self-knowledge.’ *Aristotelian Society Supplementary Volume* 85, no. 1 (2011): 223–241.

Bratman, Michael. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press, 1987.

Broome, John. *Rationality Through Reasoning*. Oxford: Wiley Blackwell, 2013.

Burge, Tyler, and Christopher Peacocke. ‘Our entitlement to self-knowledge.’ *Proceedings of the Aristotelian Society*, 96 (1996): 117–158.

Byrne, Alex. ‘Introspection.’ *Philosophical Topics* 33, no. 1 (2005): 79–104.

Byrne, Alex. *Transparency and Self-Knowledge*. Oxford University Press, 2018.

Carroll, Lewis. ‘What the tortoise said to Achilles.’ *Mind* 4, no. 14 (1895): 278–280.

Casati, Roberto, and Elena Pasquinelli. ‘How can you be surprised? The case for volatile expectations.’ *Phenomenology and the Cognitive Sciences* 6, no. 1 (March 2007): 171–183.

Charlesworth, W. R. ‘The role of surprise in cognitive development.’ In *Studies in Cognitive Development*, edited by D. Elkind & J. H. Flavell. Oxford University Press, 1969.

Chislenko, Eugene. ‘Moore’s paradox and akratic belief.’ *Philosophical and Phenomenological Research* 92 no. 3 (2016): 669–690.

Coliva, Annalisa. *On Varieties of Self-Knowledge*. London: Palgrave Macmillan, 2016.

Davidson, Donald. ‘Rational animals.’ *Dialectica* 36, no. 4 (December 1982): 317–327.

Dennett, Daniel Clement. *Brainstorms Philosophical Essays on Mind and Psychology* 1st MIT Press ed. Cambridge, Massachusetts: MIT Press, 1982.

Dennett, Daniel C. ‘Surprise, surprise.’ *Behavioral and Brain Sciences* 24, no. 5 (October 2001): 982–982.

Descartes, René. *Philosophical Writings of Descartes*, 3 vols., trans. John Cottingham, Robert Stoothoff, Dugald Murdoch, and Anthony Kenny. Cambridge: Cambridge University Press, 1984–91.

Edgley, Roy. *Reason in Theory and Practice*. London: Hutchison, 1969.

Evans, Gareth. *The Varieties of Reference*. Edited by John McDowell. Oxford: Oxford University Press, 1982.

Finkelstein, David. 'From transparency to expressivism.' In *Rethinking Epistemology* by Günter Abel and James Conant. Berlin: De Gruyter, 2012: 101–118

Frankish, Ketih. 'Deciding to believe again.' *Mind* 463, no. 116 (2007): 523–548.

Frankish, Keith. 'Delusions, levels of belief, and non-doxastic acceptances.' *Neuroethics* 5, no. 1 (2012): 23–27.

Frankish, Keith. *Mind and Supermind*. Cambridge: Cambridge University Press, 2004.

Frankish, Keith. 'Systems and levels: dual-system theories and the personal—subpersonal distinction.' In *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press, 2009.

Fricker, Elizabeth, and David E. Cooper. 'The epistemology of testimony.' *Proceedings of the Aristotelian Society, Supplementary Volumes* 61 (1987): 57–106.

Fricker, Elizabeth. 'Against gullibility.' In *Knowing from Words*, edited by Bimal Krishna Matilal and Arindam Chakrabarti, 125–161. Dordrecht: Springer Netherlands, 1994. DOI:10.1007/978-94-017-2018-2_8

Goldman, Alvin, and Dennis Whitcomb, eds. *Social Epistemology: Essential Readings*. New York: Oxford University Press, 2011.

Greco, Daniel. 'A puzzle about epistemic akrasia.' *Philosophical Studies* 167, no. 2 (2014): 201–219.

Grice, H. Paul. *Aspects of Reason*. New York: Oxford University Press, 2005.

Harman, Gilbert. *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press, 1986.

Heal, Jane. 'On First Person Authority.' *Proceedings of the Aristotelian Society* 102 (2002): 1-19.

Heal, Jane. 'Moore's paradox: A Wittgensteinian approach.' *Mind* 103, no. 409 (1994): 5–24.

Hieronymi, Pamela. 'The wrong kind of reason.' *Journal of Philosophy* 102, no. 9 (2005): 437–457.

Hieronymi, Pamela. 'Controlling attitudes.' *Pacific Philosophical Quarterly* 87, no. 1 (2006): 45–74.

Hintikka, Jaakko. *Knowledge and Belief: An Introduction to the Logic of the Two Notions.* Ithaca: Cornell University Press, 1962.

Izard, Carroll. *The Psychology of Emotions*. Plenum, New York, 1991.

Kahneman, Daniel. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

Kellerman, Henry, and Robert Plutchik. 'The measurement of emotions.' In *Emotion: Theory, Research and Experience*. Vol 4, San Diego: Academic Press, 1989.

Kolodny, Niko. 'Why be rational?' *Mind* 114, no. 455 (2005): 509–563.

Korsgaard, Christine. 'The activity of reason.' *Proceedings and Addresses of the American Philosophical Association* 83, no. 2 (2009): 23–43.

Lavin, Douglas. 'Practical reason and the possibility of error.' *Ethics* 114, no. 3 (2004): 424–457.

MacFarlane, John. 'In what sense (if any) is logic normative for thought?' Unpublished, 2004.

Malcolm, Norman. ‘Thoughtless brutes.’ *Proceedings and Addresses of the American Philosophical Association* 46 (1972): 5–20.

Martin, Michael G. F. ‘An eye directed outward.’ In *Knowing Our Own Minds*, edited by Crispin Wright, Barry C. Smith and Cynthia Macdonald. 99–122. New York: Oxford University Press, 2002.

Marušić, Berislav, and John Schwenkler. ‘Intending is believing: a defense of strong cognitivism.’ *Analytic Philosophy* 59, no. 3 (2018): 309–340.

McDowell, John. ‘Anti-realism and the epistemology of understanding.’ (1981). In *Meaning and Understanding*, edited by Herman Parret and Jacques Bouveresse. 225–248. Berlin: Walter de Gruyter, 1981.

McHugh, Conor. ‘The illusion of exclusivity?’ *European Journal of Philosophy* 23, no. 4 (2015): 1117–1136.

McHugh, Conor. ‘Epistemic deontology and voluntariness.’ *Erkenntnis* 77, no. 1 (2012): 65–94.

McHugh, Conor. ‘What do we aim at when we believe?’ *Dialectica* 65, no. 3 (2011): 369–392.

McKinson, D. C. ‘The Paradox of the Preface.’ *Analysis* 25, no. 6 (1965): 205-207.

Mele, Alfred R., and Piers Rawling. *The Oxford Handbook of Rationality*. Oxford University Press, 2004.

McHugh, Conor, Jonathan Way, and Daniel Whiting, eds. *Normativity: Epistemic and Practical*. New York: Oxford University Press, 2018.

Moore, George E. ‘A reply to my critics.’ In *The Philosophy of G. E. Moore,* edited by Paul Arthur Schilpp, 542–543. 1942. New York: Tudor.

Moore, George E. ‘Russell’s theory of description.’ In *The Philosophy of Bertrand Russell,* edited by Paul Arthur Schilpp, 175–225. 1944. La Salle, Illinois: Open Court.

Moran, Richard. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press, 2001.

Moran, Richard. ‘Replies to Heal, Reginster, Wilson, and Lear.’ *Philosophy and Phenomenological Research* 69, no. 2 (2004): 455–472.

Nagel, Thomas. *Mortal Questions*. Cambridge: Cambridge University Press, 2012.

Noë, Alva. ‘Is the visual world a grand illusion?’ *Journal of Consciousness Studies* 9 (5–6), (2002): 1–12.

Ortony, A., Clore Gerald L., and Collins A. *The Cognitive Structure of Emotions*. Massachusetts: Cambridge University Press, 1988.

Owens, David. ‘Does belief have an aim?’ *Philosophical Studies* 115, no. 3 (2003): 283–305.

Owens, David. ‘Epistemic akrasia.’ *The Monist* 85, no. 3 (2002): 381–397.

Parfit, Derek. *On What Matters: Volume Two*. New York: Oxford University Press, 2011.

Parrott, Matthew. ‘Self-blindness and self-knowledge.’ *Philosophers’ Imprint* 17, no. 16 (August 2017): 1–22.

Peacocke, Antonia. ‘Embedded mental action in self-attribution of belief.’ *Philosophical Studies* 174, no. 2 (2017): 353–377.

Peacocke, Christopher. *A Study of Concepts*. Cambridge: MIT Press, 1992.

Peacocke, Christopher. *Being Known.* New York: Oxford University Press, 1999.

Peacocke, Christopher. ‘Conscious attitudes, attention, and self-knowledge.’ In *Knowing Our Own Minds*. Oxford: Oxford University Press, 2000: 63–98.

Pears, David. *Motivated Irrationality.* Oxford: Oxford University Press, 1984.

Portmore, Douglas. *Opting for the Best: Oughts and Options*. New York: Oxford University Press, 2019.

Quine, W. V. and J. S. Ullian. *The Web of Belief*. McGraw-Hill, 1970.

Reisenzein, Meyer, and Niepel, R, W-U, M. ‘Surprise.’ *Encyclopedia of Human Behavior*. (2012): 564–570.

Rinard, Susanna. ‘Equal treatment for belief.’ *Philosophical Studies* 176, no. 7 (2019): 1923–1950.

Robinson, Jenefer. ‘Startle. (startle response as a model for philosophy of emotion).’ *The Journal of Philosophy* 92, no. 2 (February 1, 1995): 53–74.

Rödl, Sebastian. *Self-consciousness.* Cambridge, 2007.

Ryle, Gilbert. *The Concept of Mind*. London: Penguin Books, 1949.

Rumfitt, Ian. *The Boundary Stones of Thought: An Essay in the Philosophy of Logic*. New York: Oxford University Press, 2015.

Scanlon, Thomas. ‘Structural irrationality.’ In *Common Minds: Themes From the Philosophy of Philip Pettit*, edited by Geoffrey Brennan, Robert Goodin, Frank Jackson, and Michael Smith. 84–103. Oxford: Oxford University Press, 2007.

Scherer, Klaus R. ‘On the nature and function of emotion: a component process approach.’ In *Approaches to Emotion*, edited by Klaus R. Scherer and Paul Ekman: 293–317. New Jersey: Lawrence Erlbaum Associates, 1984.

Schroeder, Mark. *Explaining the Reasons We Share: Explanation and Expression in Ethics, Volume 1*. Oxford: Oxford University Press, 2014.

Shah, Nishi. 'How truth governs belief.' *The Philosophical Review* 112, no. 4 (2003): 447–482.

Shah, Nishi, and David Velleman. 'Doxastic deliberation.' *The Philosophical Review* 114, no. 4 (2005): 497–534.

Shoemaker, Sydney. 'On knowing one's own mind.' *Philosophical Perspectives* 2 (1996): 183–209.

Shoemaker, Sydney. 'Self-intimation and second order belief.' *Erkenntnis* 71, no.1 (2009): 35–51.

Shoemaker, Sydney. 'Self-knowledge and inner sense.' *Philosophy and Phenomenological Research* 54 (1994): 249–314.

Shoemaker, Sydney. *The First-Person Perspective and Other Essays.* Cambridge: Cambridge University Press, 1996.

Snowdon, Paul. 'How to think about phenomenal self-knowledge.' In *The Self and Self-Knowledge*, edited by Annalisa Coliva, 243–262. New York: Oxford Press, 2012.

Sorensen, Roy A. *Blindspots.* Oxford: Clarendon Press, 1988.

Steglich-Petersen, Asbjørn. 'No norm needed: On the aim of belief.' *The Philosophical Quarterly* 56, no. 225 (2006): 499–516.

Steglich-Petersen, Asbjørn. 'Does doxastic transparency support evidentialism?' *Dialectica* 62, no. 4 (2008): 541–547.

Steglich-Petersen, Asbjørn. 'Weighing the aim of belief.' *Philosophical Studies* 145, no. 3 (2009): 395–405.

Steglich-Petersen, Asbjørn. ‘How to be a teleologist about epistemic reasons.’ In *Reasons for Belief*, edited by Asbjorn Steglich-Petersen and Andrew Reisner. 13–33. Cambridge: Cambridge, 2011.

Steinberger, Florian. ‘Three ways in which logic might be normative.’ *The Journal of Philosophy* 116, no. 1 (2019): 5–31.

Streumer, Bart. ‘Inferential and non-inferential reasoning.’ *Philosophy and Phenomenological Research* 74, no. 1 (2007): 1–29.

Streumer, Bart. ‘Practical reasoning.’ In *A Companion to the Philosophy of Action*, edited by Timothy O'Connor and Constantine Sandis. 244–251. Oxford: Blackwell, 2010.

Stoneham, Tom. ‘On believing that I am thinking.’ *Proceedings of the Aristotelian Society* 98 (January 1, 1998): 125–144.

Sullivan-Bissett, Ema. ‘Aims and exclusivity.’ *European Journal of Philosophy* 25, no. 3 (2017): 721–731.

Sumitsuji, N. ‘The origin of intermittent exhalation (A! Ha! Ha!) peculiar to human laugh.’ *Electromyography and Clinical Neurophysiology* 40, 5 (2000): 305–309.

Ten Brinke, Leanne, Dayna Stimson, and Dana R. Carney. ‘Some evidence for unconscious lie detection.’ *Psychological Science* 25, no. 5 (2014): 1098–1105.

Vanhamme, Joëlle. ‘The link between surprise and satisfaction: an exploratory research on how best to measure surprise.’ *Journal of Marketing Management* 16, no. 6 (July 1, 2000): 565–582.

Velleman, J. David. ‘On the aim of belief.’ In *The Possibility of Practical Reason*. New York: Oxford University Press, 2000.

Williams, Bernard. ‘Deciding to believe.’ In *Problems of the Self*. Cambridge: Cambridge University Press, 1973.

Williams, Bernard. ‘Persons, character and morality.’ In *Moral Luck*. Cambridge: Cambridge University Press, 1981.

Wittgenstein, Ludwig. *Cambridge Letters: correspondence with Russell, Keynes, Moore, Ramsey, and Sraffa*. Edited by Brian McGuinness and G. H. von Wright. Oxford: Blackwell, 1995.

Wittgenstein, Ludwig. *Philosophical Investigations*. 4th ed., edited by P. M. S. Hacker and Joachim Schulte, translated by G. E. M. Anscombe, P. M. S. Hacker and Joachim Schulte. Oxford: Blackwell, 2009.

Wittgenstein, Ludwig. *Remarks on the Philosophy of Psychology*. Vol. 1, edited by G.E.M. Anscombe and G. H. von Wright, translated by G. E. M. Anscombe. Chicago: University of Chicago Press, 1980.

Wright, Crispin. ‘Self-knowledge: the Wittgensteinian legacy.’ In Crispin Wright, Barry C. Smith & Cynthia Macdonald (eds.), Knowing Our Own Minds. Oxford University Press. pp. 101-122 (1998).

Acknowledgements

I would like to thank the UCL Philosophy department for its vibrant intellectual atmosphere. I have benefitted immensely from all the talks, seminars, courses, and reading groups that it hosted. I have to thank all the students and teachers at UCL for making my time and learning experience at UCL a rewarding one. I have especially befitted from the mind- and epistemology-related courses I took with Lucy O'Brien, Mike Martin, Paul Snowdon, and José Zalabardo.

I thank Richard Edwards, Ulrike Heuer, Fiona Leigh, and Rory Madden for all the work they put in to help me sort out my overseas studies arrangement. I cannot express how grateful my family and I are for your making it possible for us to stay together. Fiona and Véroqiue Munoz-Dardé have also given me much emotional support throughout these years.

I also thank all my colleagues at NTU for covering my duties while I was studying in London. In particular, I would like to thank Alan Chan and Chenyang Li for their understanding and trust.

I thank Kwong-loi Shun for encouraging me to get formal training in analytic philosophy and for all the support he has generously given me.

With regard to this thesis, I would like to thank the following people for their very helpful comments on various parts of this thesis: Julian Bacharach, Tony Cheng, Catherine Dale, Vanessa Carr, Pete Faulconbridge, Alec Hinshelwood, Charles Jansen, Edgar Phillips, Alex Geddes, Léa Salje, Ashley Shaw, Maarten Steenhagen, and Olav Vassend. I also have to thank Yuka Kamamoto and Alexa Nord-Bronzyk for helping me with the bibliography section.

I thank Mark Kalderon and Matthew Simpson for their very helpful comments on Chapter 3 of this thesis. Our meetings helped me think through many issues related to ratiocination.

I especially thank Paul Snowdon for kindly giving me many detailed and insightful comments on earlier drafts of this thesis.

Rory Madden has also been a solid source of support and help throughout my graduate studies. I must thank him for reading many drafts and for all the critical questions and comments he raised, and for teaching me how to think and write clearly and carefully. I understand that it is not easy to supervise an overseas student. I am very grateful for all the in-person and online meetings we had.

This project will not be possible without the supervision of Mike Martin. My philosophical life would have also been drastically different if I had never worked with Mike. I thank Mike for helping me get started with this project and seeing it through, for teaching me how to do philosophy, and for guiding me to topics I did not know I would have enjoyed working on so much. I am very grateful for all his guidance and kind support.

I thank all my friends in London for their companionship. I will always be grateful to Paul and Katherine Snowdon for all the warmth and love they kindly showed us during those cold winter days in London. I thank Vincci Wong for always being there for me.

My parents gave me all the essentials in life. I can do this because they have given up a lot of things.

I thank my son Peilun for joining me on my graduate studies halfway and for adding an extra layer of meaning to this pursuit.

I thank Andrew Forcehimes for holding the fort at home, for carefully going through all my drafts, for helping me to proofread and edit, and for being stressed out about this thesis as much I am. I thank him for his patience, kindness, and love.