

Belief about Belief

Hiu Chuk Winnie Sung

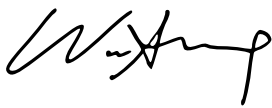
University College London

MPhilStud – Masters in Philosophical Studies

Declaration

I, 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.

Signed:

A handwritten signature in black ink, appearing to read 'Winnie Sung', written in a cursive style.

ABSTRACT

This thesis is concerned with the self-ascription of belief. It raises the possibility that one can have a higher-order belief that she believes that p without having the lower-order belief that p . This possibility not only poses a challenge to accounts of self-ascription that model self-ascription in terms of a constitutive relation between higher-order and lower-order belief, it also raises the puzzling question as to how a subject from a first-person point of view is able to distinguish whether in believing that she believes that p , she is forming a view about the world or forming a view about her own mental state. In Chapter 1, I bring to light a commonly endorsed assumption that there is a lower-order state embedded in self-ascription. In Chapter 2, I introduce possible cases and draw on the phenomenology of surprise to show how it is possible for a higher-order belief and a lower-order belief to come apart. In Chapter 3, I discuss the questions such a possibility raise for us and focus especially on its impact on the transparency account. Chapter 4 builds on preceding discussions to show the absurdity about self-ascriptions of belief that Moore's paradox reveals, and suggest that any attempts to solve Moore's paradox should first recognise that both conjuncts in Moorean sentences rest at the higher-order level.

CONTENTS

1 Introduction	5
1.1 Preliminaries	6
1.2 Situating the Puzzle in Context	10
2 BBp without Bp	19
2.1 Sam's Surprise	19
2.2 The Surprise Principle	21
2.3 The Testimony Principle	35
3 BBp and the First-Person Perspective	38
3.1 Indistinguishability between B[Bp] and BBp	38
3.2 Impact on the Transparency Account	42
4 Moore's Paradox	49
4.1 Moorean sentences	49
4.2 Common Assumption: BBp linked to Bp	53
4.3 One Level Up	50
5 Conclusion	60
Bibliography	62

1. INTRODUCTION

A central question in discussions of self-knowledge is how we come to have knowledge of our mental states. One way to approach this question is to broaden our consideration of knowledge to beliefs about mental states, and narrow our consideration of mental states to the beliefs we have. This approach thus begins with the question: How do we come to have beliefs about our beliefs? The way we answer this question directly bears on the way we answer the question concerning knowledge of our mental states. This essay will not take a stand on either of these two questions. Rather its primary goal is to raise a puzzle about the beliefs we have about our own beliefs. Without a satisfactory solution to this puzzle, we will not be able to answer the question about how we have beliefs about beliefs nor how we have knowledge of our mental states. This puzzle will be raised via the suggestion that it is possible for a subject to believe that she believes that p without believing that p . The point of this puzzle is not about how we can be wrong about our own mental states but how, from a first-person perspective, a subject can ever tell if she is merely forming a belief about her own mental state or if she is forming a belief about the way she takes the world to be. Although a subject could be mistaken about both, the error involved with each is very different. In a case where the subject believes that she believes that p but in fact she believes that $\text{not-}p$, the problem is one of fallibility. She makes a mistake about the way she takes the world to be. In a case where the subject believes that she believes that p but in fact does not have a belief about p , the problem is not one of fallibility. Rather, she mistakenly takes her belief about her own mental state to be her belief about the way the world is. Unlike worries with fallibility cases, which are mainly concerned with whether a subject is accurately reporting or expressing her lower-order belief, the worry raised by the putative puzzle has to do with whether a subject is able to tell the difference between forming a belief about her own mental state and forming a belief about p . This presents a challenge to any view that assumes that the first-person perspective is constituted by the subject's lower-order belief. In the first half of this introductory chapter, I first set the scope for my discussion and make clear the assumptions I hold about first-person self-ascription of belief. In the second half of the chapter, I situate my question in the context of philosophical debates and explain why common accounts of introspective belief that assume a

constitutive link between higher-order belief and lower-order are threatened by the putative puzzle.

1.1. Preliminaries

This thesis is only concerned with present-tense self-ascription of belief. By ‘present self-ascription of belief’, I mean an assertion or thought of the form ‘I believe that p ’. As Wittgenstein points out, the verb ‘to believe’ can be used differently, as first-person present indicative or as first-person past indicative (*PI* II x, §89). In saying, ‘I believed then that it was going to rain’ for example the subject takes herself to be saying something about her state of mind. In saying, ‘I believe that it is going to rain’, the subject takes herself to be saying something about whether it is raining. The topic of concern here is the present-tense self-ascription. Another distinction worth noting is the distinction between an assertion about oneself and a self-ascription of belief. It is possible that one makes a sincere assertion about p without believing that p . 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 a proposition about others but does not hold a corresponding lower-order belief about herself. Suppose there is a kind of mental disorder, M , the nature of which is such the agent who has M does not believe that she has M . Suppose further that an agent can be convinced by the doctor that she has M on the basis of compelling medical evidence available to her. If the subject then asserts that ‘I have mental disorder M ’, this is a case in which the subject affirms a proposition about herself in a third-personal way but does not believe in the affirmed proposition in a first-personal way. 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 at least *prima facie* different from her experience of affirming that the higher-order belief 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 high-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 in the medical report that says she has M , she is not in a position to form the belief that she has M . The affirmed proposition ‘I have disorder M ’ will not constitute

a lower-order belief in the proposition. The topic of concern here is self-ascription of belief, meaning that when a subject asserts that ‘it is raining’, she has to sincerely believe that she believes that it is raining.

Throughout my discussion of self-ascription of belief, I 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 p obtains. This sets my use of ‘higher-order belief’ and ‘lower-order belief’ apart from some contemporary discussions that use the two terms respectively to track beliefs that result 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. Keith Frankish, for example, uses ‘level 1 belief’ to refer to states that operate at the lower-order level and ‘level 2 belief’ 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 while the latter to behavioural dispositions. This is not the way I use ‘higher-order belief’ and ‘lower-order’ belief, and I can remain neutral on whether there are two kinds of beliefs that are generated by different systems of cognitive functioning. Even if there are these two kinds of beliefs that mandate different ascriptive constraints (e.g. we may need different constraints for ascribing beliefs to creatures with language and creatures without language), it should only concern cases where we ascribe a belief to others. First-person ascriptive constraint, whatever it is, is unaffected by this sort of two-system view because whenever a subject is in a state that allows her to self-ascribe a belief, that state must already be a conscious one in which she is capable of judging and assenting. Hence, it does not make a difference to the first-person perspective whether she is ascribing to herself a ‘level 1 belief’ or a ‘level 2 belief’. Whenever I think about what I believe, that is, the way the world is from my perspective, I am already at one level above my belief about the world, regardless which system generates this belief. It is in this sense I mean our beliefs about our own beliefs are at a higher-order level, and it is in this

¹ Cf. Dennett 1978, Frankish 2004, 2009, Kahneman 2011.

² Frankish 2004.

strict hierarchy sense that I speak of higher-order and lower-order beliefs. To facilitate discussion, I will also use 'Bp' to refer to a subject's belief that p and 'BBp' to refer to a subject's belief about her own belief that p .

For the kind of self-ascription of belief that is relevant to our consideration here, I assume that it meets two conditions: sincerity and identification. The first condition is

Sincerity: if the subject believes that she believes that p , the subject believes that it is the case that she believes that p .

The *sincerity* condition stipulates that the subject is being sincere in her self-ascription, whether an assertion or thought. This prevents confusion in cases where one may suppose that she believes that p , and so she can use Bp as a premise in her reasoning and decision-making. It is worth noting that we should be more careful with self-ascription in thought. A subject could have many thoughts constantly running in the background of her mind with some of them fainter than others. There might be a thought that a 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 'I miss her' or 'Canberra is the capital of Australia'. These thoughts are like assertions, only not spoken aloud. The second condition is

Identification: if a subject believes that she believes p , she is conscious of the belief she ascribes to herself as her own belief.

The identification condition stipulates that the subject identifies with the belief she ascribes to herself. This prevents confusion in cases where one may ascribe Bp to herself but feel alienated from Bp. There are at least two possible ways in which *identification* fails. One possible way in which identification fails 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. Although the sufferers of schizophrenia are

mostly conscious of a thought, it is possible 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. In an alienated case like this, the subject will not ascribe the lower- inserted-belief to herself.³ In our consideration of standard cases of BBp, we assume a necessary connection between a subject's being conscious of Bp and a subject's identification of Bp 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 very lower-order belief itself does not occur in the subject's consciousness. This usually occurs when subjects come to accept certain unconscious lower-order beliefs that she has in a third-personal way, for example, through testimony. An oft-mentioned example in the literature is a person being 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 is such that 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 Richard 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 other people.⁴ Although the subject has a sense of ownership over the lower-order belief, she does not have access to the lower-order belief. More needs to be said about precisely how these two kinds of beliefs are different. For present purposes, however, it suffices to note that such alienated self-ascriptions are possible, but they are not the kind of first-person self-ascriptions that concern us here. In the kind of standard BBp cases that are of interest to us here, we assume that the lower-order belief ascribed by the subject to herself has to figure in the subject's consciousness. It is not enough if the subject holds Bp but is not conscious of Bp. What concerns us is the kind of higher-order belief that is formed on the basis of

³ A similar line of thought is suggested by Peacocke (1999). According to Peacocke, some sufferers of schizophrenia describe certain thoughts that occur in their consciousness as being 'inserted'. Even though they are conscious of these thoughts, the inserted thoughts are not what the subjects are conscious of as her own. It is in this sense that these subjects of thought-insertion do not have a sense of ownership over the inserted thoughts (243-245).

⁴ Moran 2001: 85. I follow Moran here in taking BBp formed from a theoretical stance to be beliefs that are formed in a way that parallels how we form beliefs about others' beliefs.

one's own conviction, not on the basis of external evidence for her beliefs. It is a different matter as to what the basis of her conviction is. Her conviction may well be found on the basis of external evidence. Here, we are only concerned with belief with which the subject identifies. Stipulating this identification condition is crucial for my subsequent discussion because the possibility I put forward, that one can have BBp without Bp, is not a case where the subject fails to identify with a 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.

1.2 Situating the Puzzle in Context

While various models have been proposed in the literature to account for the nature of self-ascription, discussions of self-ascription of belief are often framed in terms of the question of how, in having the higher-order belief that 'I believe that p ' (BBp), I can come to know that I have the lower-order belief that p (Bp). Philosophers tend to take it for granted that there are two components in self-ascription of belief: a higher-order and a lower-order belief. A common assumption is that there is a constitutive link between a higher-order BBp and a lower-order belief (Bp or $B\neg p$). On this assumption, discussion about self-ascription of belief often focuses on how Bp can intimate itself into BBp or how BBp can bring with it into the existence a Bp. A serious threat to such accounts would be the possibility that there are not always two elements in self-ascription of belief or there is no constitutive link between high-order belief and lower-order belief. If it is possible to have BBp without Bp, then the problem with many existing accounts for self-ascription is not simply that they tell a wrong story about the relation between BBp and Bp, but that they have not quite captured the nature of self-ascription of belief in the first place. This is altogether a different kind of threat to accounts of self-ascription from the possibility that subjects may be fallible and make mistakes about their mental states, such as cases of self-deception and confabulation. Accounts that model self-ascription in terms of the constitutive link between BBp and Bp can accommodate the possibility of fallibility by specifying the conditions under which our self-ascriptions are authoritative, such as the presence of agency, rationality or sincerity, so that one may say that under normal circumstances, the constitutive link between BBp and Bp is intact. However, such accounts cannot accommodate the possibility of BBp without Bp, for such a

possibility runs against their fundamental assumption that there is a lower-order belief constituting a higher-order belief. In the following, I will survey two dominant accounts of self-ascription of belief. While these two accounts disagree about the ways in which BBp and Bp are related, they share the assumption that a lower-order belief is constitutive of a higher-order belief in sincere self-ascriptions. I group them under the heading of “constitutive account”. By bringing to light the difficulties with each version of the constitutive account, I attempt to show that it is at least reasonable to question whether it is possible to have BBp without Bp. Since I will not be able to provide an exhaustive summary of all the relevant constitutive accounts here, I limit my discussion to two representative accounts from each side, one by Sydney Shoemaker and the other by Jane Heal.

One version of constitutive account can be broadly labelled as the ‘factualist account’ of self-ascription, which holds that a subject can only self-ascribe a lower-belief when there is a lower-order belief present.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. Ryle rejects the Cartesian model and turns the psychological outward. On the Rylean account, the accuracy of the statement ‘I believe that *p*’ is determined not by the subject’s having inner access to her mental items but by observing her outward behaviour. Although the Rylean account rejects the claim that mental states are inner items to which the subject has privileged access, it retains the assumption that self-ascriptions of belief are supposed to accurately describe the mental states one has, and claims that 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.⁶ Although the difficulties with both the Cartesian and Rylean accounts have been widely addressed in the literature, the assumption that self-ascription of belief is a matter of describing one’s own mental state is still adopted in more recent accounts of belief.

Realist accounts that have neither inner sense nor behaviourist commitments may appeal to a subject’s rationality or a subject’s possession of the concept of belief.⁷ Shoemaker maintains that a subject’s second-order belief that she believes

⁵ I owe this labelling to Mike Martin. See also Peacocke 1992 for another factualist position.

⁶ Cf. Ryle 1949.

⁷ See also Peacocke 1992 for another version of factualist position.

that p is necessarily constituted by her first-order state of believing that p because there is a rationality-based connection between first-order beliefs and self-ascriptions of belief. A subject knows what she believes as long as she is rational and possesses the relevant concepts of belief. 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.⁸

Shoemaker maintains that the requirement of full human rationality has one revise and update one's belief system such that when one encounters a new piece of evidence that contradicts p , one will adjust one's first-order belief that not- p . Since being rational means responding to evidence about one's beliefs, and since readjustments are rational only if they are made on the basis of one's being aware of the contents of one's own attitudes, being rational requires a subject to have second-order attitudes in order to rationalise an adjustment of belief. Hence, it is necessary to postulate second-order beliefs to explain how a rational subject can revise and update her belief system. This led to Shoemaker's later view that further explains the constitutive relation between first-order beliefs and second-order beliefs. In a more recent article, Shoemaker argues that:

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 .⁹

⁸ Shoemaker 1988: 208.

⁹ Shoemaker 2009: 42.

Shoemaker acknowledges that not all of a subject's first-order beliefs are accompanied by higher-order beliefs. The first-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'.¹⁰ On this picture, available first-order beliefs partially constitute standing second-order beliefs. The available beliefs are parts of the second-order beliefs that self-ascribe them. Shoemaker emphasises that this is very different from saying that available beliefs can cause second-order beliefs. What the available first-order belief can cause is the subject's affirming the belief that p when p is considered—but not the second-order beliefs. Second-order beliefs come into being with first-order beliefs when the subject is aware of her first-order beliefs. It is in this sense that available first-order beliefs 'intimate' themselves to the subject and ground self-ascription of beliefs.

A major difficulty with Shoemaker's factualist view is the case of disbelief. Shoemaker claims that disbeliefs are partly constitutive of standing second-order beliefs. It is unclear how a subject can be aware of disbelief and for that belief at the same time to be self-intimating. On one hand, Shoemaker says that:

There is no first-order manifestation of disbelief corresponding to affirmation or assent as a manifestation of belief—disbelieving that p need not go with affirming not- p , for one may have no opinion on the matter, and so believe neither p nor not- p .¹¹

This suggests that there is no content in disbelief. But if that is the case, it is unclear how disbelief can intimate itself in a second-order belief. On the other hand, Shoemaker speaks of disbelief as if it has a content, suggesting that such a belief can intimate itself into a second-order belief. The subject ascribes to herself a negative second-order belief, namely 'I do not believe that I believe P ', and has these disbeliefs as a part of her higher-order belief. According to Shoemaker:

The only disposition to affirm that goes with having an available disbelief that p is the disposition to make the second-order affirmation that one does not

¹⁰ *Ibid.*: 40.

¹¹ Shoemaker 2009: 45.

believe that p . The disbeliefs that are self-intimating are the ones that are available; negative second-order beliefs that ascribe these disbeliefs have them as parts.¹²

Shoemaker's way of discussing disbelief is confusing, for he seems to have taken disbeliefs to be entities or a state that can intimate itself. But 'disbelief' can also be taken to mean that a state of belief is absent i.e. $\neg Bp$.¹³ It is unclear if Shoemaker uses the term "disbelief" exclusively to mean $\neg Bp$ or $B\neg p$. While it is possible for $B\neg p$ to intimate itself to become an available belief because it is a belief state, it is odd to think that $\neg Bp$ can intimate itself because there is no lower-order state to begin with. My suspicion is that Shoemaker has conflated two levels of beliefs. While $B\neg Bp$ can 'intimate' itself into $BB\neg Bp$, $\neg Bp$, as a non-existent state, cannot intimate itself into a higher-order state. What Shoemaker could have said is that $B\neg Bp$ can 'intimate' itself into $BB\neg Bp$ but what he wants to establish is that $\neg Bp$ can intimate into $B\neg Bp$.

Another version of the constitutive account may be labelled as the 'non-factualist' account of self-ascription, which holds that the lower-order belief cannot be understood independently of the higher-order belief in self-ascription.¹⁴ 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. Although it is unclear if Wittgenstein himself endorses a non-factualist account of self-ascription, he quite 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?
(*PI* II x, §86).

¹² *Ibid.*: 45.

¹³ See p.25 for discussion of the distinction between disbelief and nonbelief. †

¹⁴ See also Wright 1989 for another version of the non-factualist view.

Here 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. Following Wittgenstein, 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’ (*RPP* I, 750). The point here is that by prefixing ‘I believe’ or ‘I do not believe’ to ‘*p*’, the subject has not therefore changed the subject matter from *p* to her own state of mind.

Jane 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 of belief that does not presuppose that prior first-order mental states are necessary for self-ascriptions of belief. On this account, a second-order belief can help to constitute the first-order state that it is about. A subject who sincerely utters ‘I believe *p*’ can cause her mental state of believing *p* to exist. Heal draws a distinction between a 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. In Heal’s words, ‘the 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. An implicit underlying assumption of this view is that the first-order mental state is somehow built into the structure of self-ascription of belief. This suggests that non-factualists like Heal also subscribe to the view that there is a lower-order belief embedded in self-ascription. It is just that the lower-order belief does not have to exist prior to the higher-order belief in order for the subject to self-ascribe.

One difficulty with Heal’s account has to do with her view on authority of self-ascriptions. For Heal, ‘belief’ has three features. (A), beliefs are often attributed to people on the basis of their behaviour. (B), self-ascriptions are criterionless. One does not have to observe one’s own behaviour in order to determine whether it is the case that she believes that *p*. In this sense, her believing that she believes that *p* is just her believing that *p*. (C), in making criterionless self-ascriptions, the speaker or thinker is presenting herself as satisfying the behavioural criteria.¹⁷ Heal is emphatic that it is only when the self-ascriptions are sincerely made that they become

¹⁵ Heal 2001: 16.

¹⁶ *Ibid.*: 4.

¹⁷ Heal 1994: 20-21.

authoritative and constitute lower-order belief. The claim about sincere pronouncement is an attractive element in Heal's account. However, it is difficult to see how the notion of sincerity is doing the critical work that Heal supposes it does. Can't I sincerely ascribe a belief to myself and describe myself as satisfying the behavioural conditions without really satisfying the behavioural conditions? We can think of certain kinds of self-ascriptions that have a structure that does not bring about Bp. Self-ascriptions that resemble the Liar's paradox is a case in point. One might sincerely ascribe to herself the belief that 'this is a self-deceived belief'. How is this higher-order belief supposed to bring about a lower-order belief on Heal's account? If the self-ascribed belief is true, then we get a denial of the lower-order belief. If the self-ascribed belief is false, we get an affirmation of the belief but then the subject will lose her authority. Unfortunately, there are not enough resources in Heal's account to address this worry, for Heal has not provided a clear epistemological account of how we come to acquire the higher-order belief such that it guarantees the presence of lower-order belief. But the worry at least suggests that even if lower-order beliefs and higher-order beliefs have the same metaphysical status, it is possible that the higher-order and lower-order beliefs are formed in different manners. Hence, Heal's account has not ruled out the suspicion that lower-order beliefs and higher-order beliefs can be formed from different mechanisms that might make them come apart.

As the above discussion shows, both factualist and non-factualist accounts have their difficulties and restrictions in their explanations of self-ascription. My suspicion is that such difficulties and restrictions arise in part from their shared assumption that there is a constitutive link between lower-order and higher-order beliefs. 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 tackle 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 but systematically favours local graduates when making hiring decisions.¹⁸ On Shoemaker's account, this is a case where the administrator mistakenly ascribes to herself the belief 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

¹⁸ Shoemaker 2009: 43.

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'.¹⁹ 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 the cases of self-deception.

Heal would likely say 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 for sincerity.'²⁰ But let us slightly modify Peacocke's example. Suppose we are considering the same administrator but she works at a different university. She is very complacent and tends to make decisions that are in line with her superior's judgements. It so happens that her new superior also 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 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 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. But we may say 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 deep down 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 her pattern of behaviour is also consistent with the belief she ascribes to herself. It is difficult to locate the ground on which Heal can charge the administrator for being insincere, or that her self-ascriptions are false. Heal thus faces a dilemma: she either must accept that the administrator in the modified case is making authoritative self-ascriptions, or she must abandon the claim that sincere pronouncement on belief itself is sufficient to bring to existence the beliefs in question. My conjecture is that the dilemma in Heal's account is a corollary of the

¹⁹ Peacocke 1998.

²⁰ Heal 1994:21

inseparability between authority and infallibility of self-ascriptions. I take ‘authority’, for Heal, to mean that whenever one believes that she believes p , it is not up for others to challenge whether she is accurately reporting what she takes her mental states to be; and ‘infallibility’ to mean that whenever one believes that she believes that p , then it must be the case that she believes that p . 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 BBp . 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 . Authority is preserved. But this is still a step away from saying that since her belief about her belief is authoritative, it must also be the case that she believes that p . Fallibility is still possible.

It seems that a much more parsimonious explanation for the self-deceived administrator case would be to simply say that the self-deceived administrator has BBp and $B\neg p$ or she has BBp and $\neg Bp$. Such a route is unavailable to the constitutive view, but I do not see why we should not entertain the possibilities that BBp and Bp may come apart. There are at least two different ways in which they may come apart. First, both the higher-order belief and lower-order belief are present but there is a mismatch between the two. This possibility turns on fallibility and is already addressed in various ways.²¹ A second way in which the two can come apart is that the higher-order state is present but the lower-order belief is absent. This is a possibility that I will raise and defend in the following chapter.

²¹ See Carruthers 2011 for detailed discussion of cases where the subjects are fallible about their own mental states.

2. BBp without Bp

This chapter attempts to question whether a self-ascribed belief necessarily has a lower-order belief embedded in it. I seek to do so by setting up hypothetical cases in which the subjects believe that they believe that p without believing that p . The cases presented in the following should be familiar to our everyday experience. The goal of introducing these cases is a modest one. The point is not to establish any positive philosophical position on self-ascription of belief. Rather, I have the limited goal of suggesting that it is possible for a subject to have a higher-order belief that she believes that p without having the lower-order belief that p . In the first section of this chapter, I present the putative puzzling case of higher-order belief without lower-order belief. In the second section, I introduce the ‘surprise principle’ and rely on it to argue that a subject’s surprise at the obtainment of p could indicate a prior absence of first-order belief concurrent with her high-order belief that she believes that p . In the third section, I address how the putative puzzling case is not threatened by our ordinary conception of testimony.

2.1 Sam’s surprise

Let us first consider a hypothetical case in which the subject, call him Sam, sincerely ascribes to himself the belief that p but is later surprised by a state of the world in which p obtains. Sam grew up in a town where there are many white swans but no black swan has ever been sighted. Sam never gave any thought to whether black swans exist until one day, on d_1 , Sam’s friend told Sam that she saw some black swans in her trip to Australia. Sam did not think his friend was lying. He accepted his friend’s testimony and ascribed the belief ‘there are black swans in the world’ to himself. He could have avowed, ‘I believe that there are black swans in the world’ or simply said the sentence to himself in thought. The point is that Sam came to believe that he believes that there are black swans. On a later day, d_2 , when Sam was in Australia and walking around a lake, he was surprised to see a black swan. Assuming that Sam has not forgotten the ascription he made, the surprise Sam experiences is indicative of the absence of Bp in Sam’s self-ascription. Given my stipulation that BBp has to involve sincerity and identification on the part of the subject, this story precludes the possibility that Sam had merely accepted the proposition that there are

black swans and was using that proposition as a premise in his reasoning but had not really believed that he believes that proposition before d_2 .

I concede that there could be many different ways of telling Sam's story. For instance, it could be said that Sam had contradictory beliefs (Bp and $B\neg p$) before d_2 . One may say that while Sam's Bp intimates into BBp , $B\neg p$ has somehow failed to intimate into a higher-order belief. Since Sam has both Bp and $B\neg p$ Sam's surprise results in a clash between his $B\neg p$ and the state of world in which p obtains, and so Sam is still right in believing that he has Bp . Another way of telling the story is to draw on the assumption that belief is a binary notion with qualified content. It may be said that Sam actually has an unqualified belief that there are black swans but the content of Sam's belief is qualified such that Sam only believed that there are black swans to a very low degree. The subject will be surprised if the proposition to which she assigns a low probability does obtain. For instance, if Sam only believes that there is 0.1% chance that there are black swans, Sam will be surprised when he sees a black swan.²² This way of telling Sam's story, however, faces difficulties such as where the threshold for generating surprise is set or how qualitative flat-out belief can influence behaviour. One may also think that belief itself is gradable in degree such that we can believe things to a stronger or weaker degree. So, if Sam only very weakly believed that there are black swans, he might still be surprised when he saw the black swan. I shall bracket the question as to whether lower-order belief is a binary state or a partial state and whether these two states are distinct. My present concern is how we come to believe that we have a certain lower-order belief. Regardless of whether lower-order belief is binary or partial, we can still ask the following question: if we have a belief that we believe that p , is it necessarily the case that we also have a lower-order belief about p ? The many different ways of explaining how Sam could be surprised does not affect my purpose. As long as I can show that my way of telling Sam's story is a possible one, I will be able to put forward the possibility that a subject has a belief that she believes that p in the absence of the belief that p . So even if surprise could be modelled on credence, it will not affect the point of my story. What concerns me is whether the subject can be surprised in a way that reveals that she lacks a lower-order belief that she ascribes to herself.

²² See Frankish 2004, Chapter 2 section 1.3, for detailed discussions of the distinction between binary (or flat-out belief) and partial belief.

2.2 The Surprise Principle

The principle I rely on to tell Sam's story may be labelled the 'surprise principle', which may be formulated as follows:

Surprise principle: if a subject *S* is surprised by a state of the world at *t* in which *p* obtains, then *S* did not take the world to be in a way in which *p* would obtain before *t*.

It is on the basis of the surprise principle that I say Sam's surprise reveals that Sam did not believe that there are black swans even though he sincerely ascribes to himself the belief that there are black swans. Since it is Sam's surprise that is doing the work in the story, Sam's behaviours between d_1 and d_2 , whether or not they are consistent with the belief that there are black swans, does not affect the point of the story. We can suppose that between d_1 and d_2 , Sam had not encountered any situation that would prompt him to act one way or another that suggests the presence or absence of his belief that there are black swans. We may also allow Sam to have acted consistently or inconsistently with the belief that there are black swans. When asked on a quiz whether black swans are fictional animals, Sam could have answered that they are not fictional.²³ Regardless of how Sam had acted between d_1 and d_2 , as long as he was surprised by the black swans, then by the surprise principle, Sam did not believe that there are black swans. To keep the story simple, we do not need to speculate about Sam's behaviour between d_1 and d_2 , for it is not by Sam's behavior that we come to decide whether Sam had *Bp*.

The claim that Sam's surprise shows that Sam did not believe that there are black swans in this case is not by itself controversial. It is a common experience that when we are surprised by a state of world in which *p* obtains, e.g. when an old friend showed up at your door without notice or when you opened the fridge and could not find the cake you believed was there, we do not believe that *p*. The difficulty with telling Sam's story lies in another claim in the story, that is, Sam also accepted his

²³This will involve an assumption that high-order flat-out belief can also motivate actions. A more difficult question concerns what Sam's betting behaviour would be like if he is offered bets on the truth of the proposition that black swans exist. How we answer this question depends on whether we can model belief on degree of confidence. Since I have bracketed this question and focus only on my point of introducing Sam's story, I will not try to speculate Sam's betting behaviour here.

friend's testimony. There seems to be an apparent contradiction between Sam's surprise and Sam's acceptance of his friend's testimony. I suppose we have the intuition that if Sam has sincerely accepted a piece of testimony that p , he must have also believed that p . We may label the principle behind this intuition the 'testimony principle', which states that:

Testimony principle: if a subject S accepts testimony that p , then S believes that p .

I anticipate the reluctance of the reader to acknowledge the possibility of Sam's case lies in this apparent contradiction between the surprise principle and the testimony principle. On the surprise principle, since Sam was surprised by the black swans, he must not have believed that black swans exist before d_2 . On the testimony principle, since Sam accepted his friend's testimony that black swans exist, he should have also believed that black swans exist on d_1 . In the following, I will first turn to defend and work through the implications of the surprise principle. I will then be able to apply the surprise principle to Sam's case and argue that the higher order belief (BBp) and the lower-order belief that p may come apart. Once I have shown that it is possible to have BBp without Bp, we will see how the testimony principle is formulated vaguely, for it is not clear which level of belief the principle is supposed to target. Hence, the contradiction between the surprise principle and the testimony principle, if there is any, might not be as worrying as it first appears to be.

Before I turn to consider the surprise principle, there are a few distinctions to be made about surprise. Jenefer Robinson helpfully draws our attention to the non-cognitive aspects of emotional responses based on the startle response. According to 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."²⁴ A novel or intense stimulus is sufficient to trigger startle; it does not take a belief or disbelief to be disconfirmed in order to trigger startle. Since surprise is a developmentally later form of startle,²⁵ surprise shares many important features with startle.²⁶ For example,

²⁴ Robinson 1995: 59.

²⁵ According to Klaus Scherer, startle is present at birth and surprise tends to appear between one and three months. See Scherer 1984: 293-317. Charlesworth (1969) argues that surprise cannot occur until five to seven months of age.

surprise is also characterised by distinctive facial expressions.²⁷ Like startle, surprise also has the biological function of preparing the subject to deal with a new or sudden situation. 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.²⁸ This suggests that surprise can be triggered by any stimulus that interrupts an ongoing activity. It does not necessarily take a belief to be disconfirmed to trigger surprise. But if this is the notion of surprise that we are working with, it is difficult to see what significant role surprise has to play in Sam's story. Even if Sam was surprised, it could simply be because Sam has not seen a black swan before. When he encounters the black swan as a novel and sudden stimulus, the surprise response simply breaks the ongoing program of his nervous system for adjustment to a novel situation. Even though this way of understanding the surprise response is compatible with the claim that Sam did not believe that there are black swans, it is no longer clear how Sam's story tells us something interesting about Sam's doxastic state. Sam's surprise would just be a way of Sam's body telling him that something new has occurred in the environment and readies Sam to deal with the situation.²⁹ Hence, in order to defend the surprise principle, there must be something more in the surprise response than just preparing an organism for a new situation.

One of the additional elements that distinguishes startle and the kind of surprise that interest us here has to do with the cause of surprise. While an intense and sudden stimulus is sufficient to trigger startle, the stimulus that triggers surprise has to be one that is unexpected by the subject.³⁰ 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.³¹ It is this unexpectedness of the stimulus that distinguishes surprise and startle and allows us to assess the doxastic state a surprised subject is in. In light of this consideration, I propose to draw a distinction between the kind of surprise that

²⁶ Izard (1977), for example, does not draw a distinction between startle and surprise and has used the terms "interchangeably."

²⁷ See Robinson 1995: 58 for the facial expressions characteristic of startle and Izard 1977:277 for the facial expressions characteristic of surprise.

²⁸ Izard 1977: 281.

²⁹ A similar point is also acknowledged by Charlesworth, who notes that if surprise is largely construed in terms of its biological taxonomies, we risk ignoring how surprise can be diagnostic of the presence or absence of cognitive structures. Charlesworth 1969, see especially pp. 258-60.

³⁰ Both Plutchik 1980 and Robinson 1995 note that surprise involves the unexpectedness of the stimulus but startle does not.

³¹ Photographs of these trained marksmen show that they still blinked and exhibit facial expression characteristic of startle when they fired a pistol. Robinson 1995: 55.

is closer to startle and the kind of surprise that involves expectation and requires a certain degree of cognitive activity. I concede that it is unclear whether a sharp distinction can be drawn, for the difference between startle and surprise could be separated by a continuum of cognition complexity.³² For my present purpose, however, it suffices to say that we can at least approximate a distinction between startle and surprise, with the latter being a more complex kind of psychological response and involves cognitive activity.³³ The point of drawing attention to this distinction is not to suggest that basic surprise and complex surprise are structurally or functionally different but to acknowledge that it is the more complex kind of surprise that allows us to attribute epistemic states. The kind of surprise that Sam experiences must be complex enough such that the fact that he is surprised says something significant about the epistemic condition he is in. Throughout my discussion of Sam's surprise, it is this more complex kind of surprise that I have in mind.

One more distinction that needs to be made is between the event that triggers surprise and the ground for surprise. It should also be noted that there is a distinction between the surprise response itself and the conditions that precedes surprise (the cause, trigger, or elicitor). What concerns us in Sam's story is the latter. One may argue that sometimes a subject may still be surprised by a certain event even though she believed that the event would obtain. For example, if my friend told me that she would bake a cake for me and if I believed her, I could still be surprised when I see my friend's cake. It should be noted that when a subject was surprised by a certain event E, it is not necessarily the case that she did not believe that E would obtain. Rather, it could be some other salient feature in E that triggers the subject's surprise. When I was surprised upon seeing the cake my friend baked for me even though I believed she would bake a cake for me, I was not surprised by the fact that my friend made a cake. Instead, my surprise could be triggered by some other features of this event. I could be surprised by how beautiful the cake looks or how fast it took my

³² 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 1960s. 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.

³³ What I mean here is close to the kind of "developmentally sophisticated" emotions Robinson discusses. According to Robinson, the basic or primitive emotional responses do not involve cognitive activity but the more developmentally sophisticated emotional responses do. Robinson 1995: 64.

friend to bake a cake. My surprise would reveal I did not believe that my friend could make such a beautiful cake or that she could make it so efficiently. The surprise itself does not always reveal prior disbelief in the event that triggers the surprise, but a prior disbelief in certain salient features in the event that triggers surprise. Such salient features in E that triggers the subject's surprise is the ground of the subject's surprise. Hence, when the same event might trigger surprise in different individuals, the grounds of surprise could be different for those who are surprised. My contention is that a necessary condition for p to be the ground of the subject's surprise is that the subject could not have believed in p before the surprise. For one to be surprised, there needs to be a determinate link to one proposition rather than another. If a subject had the belief that p but was nevertheless surprised when she saw the black swan, then the ground of her surprise must be something other than the fact that black swans exist. For instance, she could be surprised by the fact that she was lucky enough to see a black swan in person. In this scenario, her surprise would reveal that she did not believe that she would get to see a black swan in person. Since the grounds for surprise vary, there is no problem in accepting that one could still be surprised when she sees a black swan even if she believes that there are black swans. In Sam's case, my aim is to say how Sam can be surprised in a way that tells us that he did not believe that there are black swans.

So far, I have been speaking of the subject not believing that p in an ambiguous manner. It is now appropriate to draw a distinction between disbelief and nonbelief, that is between believing that not- p , and not believing that p . The distinction between disbelief and nonbelief is helpfully defined by Quine and Ullian:

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.³⁴

For convenience, I will sometimes refer to “disbelief” and “nonbelief” respectively as “ $B\neg p$ ” and “ $\neg Bp$ ”.

The claim that surprise reveals certain $B\neg p$ that a subject has is not a particularly contentious one. If I believe that there is no coin in my pocket, I will be

³⁴ Quine and Ullian 1978:12.

surprised if I find a coin in my pocket. But it would be too weak for my purpose to simply argue that Sam's case is one where Sam has BBp and $B\neg p$. It could be argued that Sam has contradictory beliefs Bp and $B\neg p$ with only the former being intimated into BBp . When Sam was surprised upon seeing the black swan, his surprise was generated by the clash between the obtainment of p and his disbelief in p . This leaves intact the link between lower-order belief about p and BBp . Those who follow Shoemaker's account will propose something like this.³⁵ Even if one does not posit the presence of contradictory beliefs, it may be argued that Sam was fallible about his own mental states. But the point of my initial puzzle is not about whether we could hold contradictory beliefs or whether we could be fallible when it comes to accessing our lower-order beliefs, for questioning along these lines would lead us back to the assumption that there is a lower-order belief (Bp) embedded in higher-order belief (BBp). The more puzzling question I would like to raise is whether it is possible for a subject to self-ascribe a belief when the belief in question is not present at all. It is in this sense I question whether it is possible to have BBp without Bp . In order for me to set up the puzzle, Sam's story has to show that Sam's surprise is generated by the clash between $\neg Bp$ and Bp so that we have a case where Sam BBp but $\neg Bp$. This will require me to show that surprise does not only reveal $B\neg p$ but may also reveal $\neg Bp$. The claim that $\neg Bp$ can generate surprise is a contentious one.

Donald Davidson and Daniel Dennett, for example, have respectively argued that it is not possible to be surprised without having a prior belief being upset. The claim can be construed as a general one that says it is necessary for some belief to be upset in order for one to be surprised. It is unclear if Davidson thinks that one has to believe that $\neg p$ in order to be surprised that p . Davidson gives the example of someone who is surprised that there is no coin in his pocket:

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 first believe there is a coin in my pocket, and after emptying my pocket I no longer have this belief. Surprise requires that I beware of a contrast between what I did believe and what I come to believe. Such awareness, however, is a

³⁵ See discussion of Shoemaker's view in Chapter 1.2.

belief about belief: If I am surprised, then among other things I come to believe my original belief was false.³⁶

This example suggests that there has to be a contrast between a subject's prior belief that not- p and her later belief that p in the case of surprise. Before checking her own pocket, the subject believed it is false that there is no coin in her pocket ($\neg p$) and was later surprised that there is no coin in my pocket (p). And then Davidson adds that it is by coming to believe that she believed that not- p and that she now believes that p that she can be surprised. It seems to me that Davidson is laying down two necessary conditions for surprise: (1) the subject has $B\neg p$ on d_1 and Bp on d_2 and (2) the subject's belief on d_2 that she believed $\neg p$ on d_1 and the belief that she now believes that p on d_2 . In the following, I will first turn to consider (1) the question whether $B\neg p$ is a necessary condition for surprise. How we answer this question will bear on whether (2) is true. Before we have a clear sense of whether $B\neg p$ for surprise could be a necessary condition, it is difficult to determine whether one has to be aware of a contrast between her old and new beliefs in order to be surprised.

A commonly accepted claim about the cognitively more complex kind of surprise is that it is a psychological response to an upset belief. And I take it that it is a lower-order belief that is being upset. However, it is not obvious to me how one has to believe a certain proposition to be false in order to be surprised. There are many everyday examples that seem to suggest surprise is possible even when it does not upset a belief. I would be very surprised if I go to work one day and suddenly find a colleague, whom I just saw in her office the day before, has left her job and emptied her office. But it does not seem necessarily the case that when I came in for work this morning, I believed that this colleague would not leave her job. It could have never occurred to me that she would even consider leaving her job. Or suppose I returned home on a regular day and was surprised to see that my friends had spontaneously organised a party for me, just to surprise me. Given that such a random surprise party never occurred to me before and that this evening is not close to a special occasion, it is unlikely that I believed that there would be a surprise party for me tonight before I stepped into my flat. There is a vast number of ways that the world could turn out to surprise us. If a subject's surprise at p implies that she believes not- p , we must also be

³⁶ Davidson 1982, p.326.

assuming that we are constantly holding beliefs that rule out all the possible states the world could turn out to be. As subjects with cognitive limitation, there must be some possible states of the world about which I had not formed a belief, whether consciously or unconsciously, before the surprise. In the above cases, we simply did not have a prior belief about the states of the world that obtain.

An immediate worry with saying that a surprise can be generated when p obtains and the subject does not believe that p has to do with the lack of a robust link between $\neg Bp$ and one's being surprised that p . When one is surprised that p , she also learns that p . That she learnt that p potentially suggests that she did not have the belief that p . But we often learn something without being surprised. For example, I do not have a belief about whether my neighbour's phone number ends with an even digit but I will not be surprised if it does. There must be something more than just $\neg Bp$ in order to generate surprise. More needs to be said about what this additional element is before we can decide whether one can also be surprised when one non-believes that p .

In order to decide whether a belief is necessarily upset when one is surprised, we should first decide on what is necessary for surprise. It seems difficult not to accept that surprise is only possible if it upsets an expectation, a point that is often raised in both philosophical and psychological discussion on surprise.³⁷ But when we talk about unexpectedness, we also need to draw a distinction between not expecting that p and expecting that not- p . If we just say that one will be surprised that p when one does not expect p will lead us back to the same problem we encountered with $\neg Bp$. It is difficult to see how one can be surprised that p when there is no expectation that p would obtain. There are many things about which we did not form an expectation, such as the last digit of my neighbour's phone number, and would not be surprised if the last digit turns out to be an even number. As Charlesworth points out, surprise should be defined as a function of misexpected events, as opposed to unexpected events.³⁸ Indeed, both conceptual and empirical considerations point toward the necessity of "misexpectedness" for surprise. In studies that suggest infants can be surprised, the infant has to be first habituated 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. And when the infant spends longer time looking at a

³⁷ Davidson (1982), for example, argues that surprise is only generated when it violates an expectations; Dennett (2001) describes surprise as the "betrayal" of expectation.

³⁸ Charlesworth 1969:257-273.

novel, inconsistent stimulus, the infant is arguably surprised because the new stimulus violates an expectation that is formed through habituation.³⁹

In light of the above considerations, it seems that even if I want to argue that surprise can arise when one non-believes that p , I have to at least accept that surprise can only arise when an expectation is violated. The challenge for me is to explain how there can be a mismatch between one's expectation and the way the world is in a case of surprise without falling back to a mismatch between a lower-order belief and the state of the world. It may be argued that one can be surprised that p even if one does not believe that p because expecting that not- p is not reducible to or does not entail forming a belief about p . This is possible if we assume that certain forms of conscious cognition, such as expecting that p , are not reducible to believing. As Norman Malcolm suggests, conscious cognition is more than just thinking of propositions. We may 'consciously think' 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 could consciously think that the path is slippery without thinking of the proposition, 'This path is slippery'.⁴⁰ If we apply this line of thought to expectation, we could say that Sam was expecting that all the swans he is going to see are white without the belief that there are only white swans.⁴¹ On this view, Sam's surprise results in a clash between the way Sam expects the world to be and how the world turned out to be. Since Sam grew up in a town where only white swans have been sighted, Sam must have been habituated or acquainted with the world being in a state where there are only black swans. He does not need to have a propositional attitude that 'there are only black swans in the world'. But the world suddenly turned out for Sam to be in a state in which there are also black swans. It is this clash that causes Sam's surprise reaction.

Yet, the suggestion that one can expect something without a belief has its difficulties. First of all, it may be argued that Sam has what Frankish would call 'level 1 belief' that there are only white swans and therefore there are no black swans.⁴²

³⁹ Casati and Pasquinelli 2007: 174.

⁴⁰ Norman 1972-3:6.

⁴¹ This reading differs from that of Davidson. Davidson seems to assume that 'consciously 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 ; but in Malcolm's original paper, his focus is always on conscious thinking and has not discussed belief in a technical sense. Davidson 1982: 102.

⁴² For Frankish, 'the term "belief"' can also track states that are 'multi-track behavioural dispositions, which are non-conscious, passive, graded, and holistic'. Frankish 2012: 24.

However, even if there is no such thing as belief that is generated by a lower system of reasoning, we are still faced with the difficult intuition that expecting something is more than just taking the world to be a certain way. Imagine a young child sees an object she has never seen before and she puts it in her mouth. The way she acts suggests that she thinks that the object is edible but she may not be surprised when she had a bite and realized it is not edible. But suppose the child has seen a similar object of the same size and shape many times before and each time the object has been an edible one. Then, the child in this instance would be surprised that this particular object that looks just like what she ate before is not edible. The difference between these two cases is that the child in the former case has some experience with a similar situation and is therefore able to make a prediction about a similar situation. The fact that she is able to make such a prediction suggests that she is drawing on some general belief in expecting a certain state of affairs to obtain.

It may be argued that a distinction between expectation and belief cannot be drawn. Dennett 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.⁴³

Dennett's point is that if one is surprised by a situation in p obtains, then she must have expected that not- p . And her expecting that not- p shows that she has 'ideological commitment' to not- p , which is a subclass of belief. Hence, one's surprise also reveals one's belief that not- p . 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. For example, if subjects are surprised by experimental demonstrations of change blindness, which show we do not have a snapshot-like visual experience that

⁴³ Dennett 2001: 982, italics original.

represents a visual scene in high resolution and detail, then they must have believed beforehand that visual experience is like snapshots.

Alternatively, one may argue that such a distinction may still be drawn but maintains that an expectation has to be generated by a different belief. This implies that there is still some belief being upset in the case of surprise. Alva Noë suggests that certain beliefs are upset in cases where subjects are surprised by results of change blindness but the beliefs that are being upset do not have to be the belief that visual experience is like snapshots. The idea here is that one may be surprised at a situation in which p obtains, not because one believes that not- p , but because one believes q . According to Noë:

But one need not attribute to them (to us) a commitment to the snapshot conception. The surprise is explained simply by supposing that we tend to think we are better at noticing changes than in fact we are, or that we are much less vulnerable to the effects of distracted of distracted attention than we in fact are.⁴⁴

Let p be the proposition that ‘visual experience is snapshot-like’ and q be the proposition that ‘we are bad at detecting changes in visual scene’. Recall our earlier discussion of the distinction between event that triggers surprise and the ground for surprise. Noë seems to be suggesting that the subjects who are surprised by the results of change blindness are not necessarily surprised that they are change blind. The salient feature of the event that triggers surprise could be q rather than p . One is surprised because one’s belief that not- q clashes with a situation where q obtains. If this is what Noë has in mind, then he is also working with the assumption that it takes some prior belief to be upset to be surprised. For Noë, Sam does not have to believe that there are no black swans in order to be surprised. Instead, Sam’s surprise can be explained by some of Sam’s other beliefs being upset, such as the belief that he will not get to see a black swan in person or the swan looks different than he imagined what black swans are like. But this route will not work for us, for I have already stipulated that Sam is surprised that there are black swans, it cannot be other upset beliefs that generate the surprise.

⁴⁴ Noë 2002: 7.

Even if it is not a well-defined belief that is upset, as Casati and Pasquinelli point out, one's surprise that p is still generated by volatile expectations, which are in turn generated by more general dispositional or ideological beliefs.⁴⁵ On Casati and Pasquinelli's account, my surprise at a spontaneous house party is not caused by a violation of the belief 'that this particular party that takes place in this particular evening will not take place' but by a more general belief like 'that there will not be a random special event at my flat'. This general belief generates the volatile representation that my flat is more or less in the same state as I left it and the volatile representation is violated when I find out that there is a party organised in my flat. This route allows us to say that a subject could be surprised by p without believing that $\text{not-}p$. But drawing such a distinction between belief and expectation is not particularly helpful for Sam's case because it is still committed to the thought that some belief has to be violated. Any general or ideological belief in Sam's case will get us close to saying that Sam does have a belief about black swans. It cannot be an ideological belief so general as one that says all animals are not black, for Sam could have seen other black swans before. It also cannot be a too specific belief that says no swan will turn out to be a different colour, for this will get us close to saying that Sam did believe that swans are not black.

So far, our considerations of the phenomenology of surprise seem to be against the possibility of my puzzle; but, in fact, we are getting close to showing how Sam could be surprised in a way that reveals that he does not believe that there are black swans. We have seen how a surprise is inevitably bound up with a violated belief. Building on this suggestion, we can refine the surprise principle to:

Surprise Principle (refined): A subject is surprised that p at t if and only if she takes the world to be in a way in which p would not obtain at t .

We can leave vague what 'taking the world to be in a way in which p would not obtain at t ' means. It can mean one's thinking that $\text{not-}p$ or one's belief that $\text{not-}p$ or one's ideological belief that is incompatible with $\text{not-}p$. The crucial point here is that if one is surprised, one is implicitly or explicitly committed to a view about the world. In fact, Noë has already suggested, but has not pursued, the idea that a subject's lack

⁴⁵ See Casati and Pasquinelli 2007.

of surprise may indicate her lack of commitment to p .⁴⁶ This suggestion that lack of surprise can indicate a lack of belief is now developed and substantiated with our discussion on surprise.

Let us turn to consider a case in which a subject is not surprised by $\text{not-}p$ even though she self-ascribes the belief that p . Suppose Leonard grew up in a town where no black swan had ever been sighted. But a major difference between Leonard's town and Sam's town is that Leonard's town is a very religious community in which one of the creeds is that all swans are white. Leonard is a devout disciple and ascribes to himself the belief that all swans are white. One day, Leonard travelled to Australia and when he was walking around a lake, he saw a black swan. But Leonard is not surprised. How should we account for Leonard's lack of surprise then? On the surprise principle, this is a case where both the antecedent and consequent are false. Leonard's lack of surprise indicates that Leonard does not take the world to be one in which there are no black swans. There are at least three possible ways to explain Leonard's not taking the world to be one in which there are no black swans. (1) It is possible that Leonard in fact believed that it is not the case that there are only white swans. (2) It is possible that Leonard had contradictory lower-order beliefs, that is, he believed both p and $\neg p$. Somehow he is only aware of his Bp but is not aware of $B\neg p$. (3) It is possible that Leonard did not have a view about whether it is the case that there are only white swans in the world ($\neg Bp$). (1) and (2) are standard cases of fallibility in which Leonard makes a mistake about his lower-order mental states. The difference lies in the explanation for how that mistake is being made. An explanation for (1) has to say how it is possible for introspection to fail. An explanation for (2) has to say how it is possible for a subject to hold contradictory beliefs and why one of the contradictory beliefs occurs in consciousness whereas the other does not. Unlike (1) and (2), Leonard in (3) has not made a mistake about his lower-order belief. Rather, he makes a mistake about the direction of self-ascription. While he might want to track his lower-order belief about whether it is the case that there are only white swans, he in fact was tracking whether he believes that he believes that there are only white swans. Hence, he forms the belief that he believes that there are only white swans but in fact he does not have a view on this matter at all. An explanation for (3)

⁴⁶ Noë 2002:6-7.

has to say how it is possible for a subject to distinguish whether her self-ascription of belief is directed at the way the world is or at the way her mental state is.

Through a long detour, we have a case where the subject believes that she believes that p without believing that p . I had to first model surprise and show that surprise necessarily reveals a violation of the way the subject takes the world, and then provide a lack of surprise case to establish how it is possible for a subject to have BBp without Bp. Now that I have established that it is possible to have BBp without Bp, Sam's case also becomes a possibility. But how should we account for the surprise that Sam experiences if Sam does not have the lower-order belief? I suggest that Sam's surprise is generated at a higher-order level between his current BBp upon seeing the black swan on d_2 and his awareness that he did not have the belief $B\text{---}Bp$ before d_2 . Sam's surprise does involve the violation of a belief but the belief that is being violated is at a higher level, that is, his belief that he believes that there are black swans. So the surprise is directed at his belief about his own mental state instead of his belief about the world. Here, Davidson's insight into surprise becomes extremely helpful (recall condition (2)). Davidson has already noted 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.⁴⁷

Given our earlier discussion of surprise, it does not seem plausible to say that all cases of surprise involve the subject's awareness of a contrast between what she believed and what she comes to believe. But Davidson's insight is applicable in this case of Sam. Sam was surprised because he comes to believe that his initial belief that he believes that there are black swans was false. How did Sam come to believe that his initial BBp was false then? Is there something special about direct experience or about the mental phenomenology of actually having a Bp? These questions will be taken up in the next chapter. In the following, I will return to the testimony principle and argue that it does not conflict with the surprise principle.

⁴⁷ Davidson 1982: 326.

2.3 The Testimony Principle

Our initial difficulty in establishing the plausibility of Sam's story is that the case cannot be told in a way that is compatible with both the testimony principle and the surprise principle. Now with the possibility of BBp without Bp established, it becomes clear how the principle no longer poses a threat to the surprise principle. To recall, the testimony principle holds that if one accepts testimony that p , then she believes that p . The testimony principle captures the intuition that when we accept a piece of testimony, we formed a belief about the world. When we read from a reliable newspaper report that there was a submarine earthquake in the Atlantic Ocean, we formed a belief about what happened in the Atlantic Ocean. In everyday situations, we often perceive an intimate connection between accepting testimony that p and coming to believe that p . But it is ambiguous as to what level of belief the testimony principle means to capture. In view of the consideration the higher-order and the lower-order beliefs may come apart, we may formulate two versions of the testimony principle:

Weak version: if a subject S accepts the testimony that p , S believes that she believes that p .

Strong version: if a subject S accepts the testimony that p , S believes that she believes that p and S believes that p .

The weak version of the testimony principle is consistent with Sam's story. And since the practice of giving and accepting testimony occurs at a level that requires linguistic competency, if one has consciously accepted the testimony that p , one must have also believed that she believed that p . It is the strong version of the testimony principle that is in tension with the hypothesis that Sam did not believe that there are black swans. If we are to preserve the surprise principle in telling a coherent story about Sam, it is this strong testimony principle that we have to give up. But since we have already established that there could be a gap between BBp and Bp, it is unclear why when one accepts a testimony and BBp, she must also have the lower-order belief that p . Without a satisfactory account of how one's BBp necessarily brings about Bp or tracks Bp, there is no ground for us to accept the strong testimony principle.

Once we allow the possibility that BBp and Bp are distinct states, the everyday examples of receiving and accepting testimonies offer little help to the defender of the strong testimony principle. These examples can be interpreted as merely saying that if one accepts the testimony that *p*, then one believes that one believes that *p*. At most, these examples indicate a correlation between BBp and Bp. It is possible that BBp and Bp are formed independently at different levels for different reasons. Given the possibility of BBp and Bp being two distinct states, the burden is on the defender of the strong version of testimony principle to provide a satisfactory account of how accepting testimony and coming to believe that you believe that *p* also guarantees the presence of Bp.

Without a satisfactory account of epistemology of belief that tells us how our beliefs about beliefs necessarily guarantee the existence of a lower-order belief, the testimony principle alone cannot establish that one's linguistic acceptance of testimony that *p* necessarily entails a belief that *p*. All that can be established is that if a subject accepts the testimony that *p*, it is the case that she believes that she believes that *p*. Until we have an account of self-ascription that shows how a subject's belief about her belief guarantees the existence of the lower-order belief in the case of accepting testimony, there is no apparent contradiction in saying that one accepts testimony that *p*, believes that she believes that *p*, and does not believe that *p*.

In this chapter, I suggested the possibility for a subject to have BBp without Bp. Before Sam saw the black swan, he believed that he had formed a mental state that is directed at the world, but it turned out that he had only formed a belief about his own belief. It should be obvious that such a possibility undermines both the factualist and the non-factualist accounts we discussed in the previous chapter, for they both assume that a lower-order state is constitutive of self-ascriptions of belief. To the factualist, we have a case of self-ascription where the ascribed lower-order state is absent. This challenges their main claim that self-ascription is about describing certain lower-order states that one already possesses. To the non-factualist, we have a case where the higher-order belief fails to bring with it into existence a lower-order belief. This challenges the non-factualist claim that a higher-order belief can determine, or at least bring about, the existence of a lower-order belief.

This possibility puts pressure on any constitutive account that assumes a necessary link between higher-order belief and lower-order belief. If the constitutive link can be severed, then our present self-ascriptions are not necessarily expressive of

a lower-order belief. The more general sceptical concern then is that our first-person perspective is never directed at the lower-order level where we take the world to be in a certain way but at a higher-order level where one takes herself to take the world to be in a certain way.

In the next chapter, I will turn to consider an alternative account, the transparency account, which does not assume a constitutive link between lower-order mental state and self-ascriptions. I will use the putative phenomenon of one's having BBp without Bp to raise a general sceptical concern about knowledge of our own beliefs. It will be argued that since we have not answered the question as to how, from the first-person perspective, one can tell the difference between forming a belief about the world and forming a belief about her own mental state, the transparency account also cannot provide a satisfactory account of self-ascription of belief.

3. BBp and the First-Person Perspective

In the previous chapter, I have used Sam's case to show the possibility of a subject's having BBp without Bp. Such a possibility threatens any account that holds that a lower-order belief is constitutive of a higher-order belief. It is not the task of this thesis to propose an alternative account of self-ascription, and explain how the subject can generate higher-order beliefs when the first-order belief in question is absent. In this chapter, I aim to focus on highlighting the puzzling questions that the possibility of BBp and Bp could be distinct states raises. The first section of this chapter raises a philosophical puzzle: How can a subject, in ascribing a belief to herself when the ascribed belief does not have to be present, is not only entitled to say that she believes that she believes that p , but is also entitled to say that she believes that p ? This puzzle forces us to make clearer which level of belief we are concerned with before we can engage in meaningful discussion of a subject's belief. A more general sceptical concern brought out by this puzzle is whether a subject is ever able to distinguish between her attending to the way the world is and her attending to the way her mind is. The second section uses the sceptical concern to make trouble for the transparency account of self-knowledge.

3.1. Indistinguishability between B[Bp] and BBp

It should be noted that even if the link between the lower-order belief that p and the higher-order belief that p may be severed under some circumstances, this does not automatically imply that self-knowledge is not possible. It is highly likely that under normal circumstances, a subject's self-ascriptions are accurate and hence has Bp embedded in her Bp. The challenge rather is this: if it is possible for a subject to have BBp without Bp, then BBp and Bp must be distinct and independent states. This raises the question as to whether there is any evidential basis for BBp and where we should locate it. Suppose there is evidential basis for BBp. Since BBp and Bp are distinct and independent states, it is possible that BBp is formed independently of the lower-order state. Accordingly, it is not necessarily the case that Bp is an evidential basis for BBp. Even if Bp is an evidential basis for BBp in good cases, there are bad cases where there is no lower-order belief embedded in BBp. In these bad cases, there must be some evidential ground other than Bp that enables BBp is formed. Hence, we

need to further investigate the way(s) in which accurate self-ascriptions can be made, sometimes independently of the lower-order state, on an evidential ground different from the lower-order belief.

Or suppose no evidential basis is needed for BBp. We are still left with the question of how self-knowledge is possible. This worry is in line with a compelling remark made by Paul Snowdon in his discussion of Wright's constitutive account. Snowdon points out that just because the subject does not have evidence for his mental state, it does not mean that the question of how we have knowledge of our mental states is not askable.⁴⁸ The question that can be raised here is similar to that of Snowdon, but at one-level up. Snowdon's point is that we can ask how one knows she is in a certain lower-order state. My point is that in light of the possibility of one's BBp without Bp, we can also ask how one knows whether she is forming a belief about the way the world is or a belief about the way her mind is. The former question concerns the lower-order state, while the latter concerns the higher-order state. In the following, I will show why one will have difficulty in answering the question that concerns higher-order belief.

It is apparent to a third-person perspective that a necessary condition for one's self-ascriptions to amount to knowledge is that one in fact believes that *p* if she believes that she believes that *p*. However, from a first-person perspective, how can the subject herself tell whether her higher-order belief that *p* has the corresponding lower-order belief that *p*? For brevity, let 'BBp' stand for a subject's belief that she believes that with the belief that *p*; let 'B[Bp]' stand for a subject's belief that she believes that *p* without believing that *p*.

Let us suppose that Sam's sister, who heard the testimony about black swans at the same time as Sam, did come to believe that there are black swans. From a third-person perspective, we can tell that the difference between Sam and his sister lies in how the former has B[Bp] whereas the latter BBp. But from a first-person perspective, if both are being sincere in their self-ascription and presumably have strong conviction that they do believe that there are black swans, there does not seem to be good reasons for not conferring authority to Sam's B[Bp], for it is the case that Sam believes that he believes that there are black swans.

⁴⁸ Snowdon 2012: 260.

One may argue that Sam's BBp and B[Bp] are formed on different grounds. Sam's BBp is authoritative, but his B[Bp] is not because the phenomenology of actually having a Bp is present in BBp. 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 at d_2 , he learned 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 a lower-order belief.

Let us assume that whenever a subject has a direct experience with a state of world in which p obtains, then the subject will form the belief that p . The mental phenomenology of actually having the belief that p must be different to Sam and contrasts with how the mental phenomenology Sam had when he did not have the belief that p . The contrast in mental phenomenology between B[Bp] and BBp is what enables Sam to tell the difference. However, there are a number of scenarios in which a subject is not able to experience the contrast in mental phenomenology that enables her to tell that she does not have the lower-order belief she ascribes to herself.

One such scenario is when a subject never gets a chance to directly experience with a state of world in which p obtains. If Sam has never seen a black swan, he will not have the direct experience that allows him to learn that there are actually black swans and therefore realise that he did not really have a view on whether black swans exist prior to his seeing the black swan. Even though he has never seen a black swan, he would continue to believe that he believes that there are black swans. From a third-person perspective, we may say that Sam's B[Bp] is mistaken. But from Sam's first-person perspective, it is unclear why his B[Bp] is made less authoritative by the fact that he never had the direct experience. We often ascribe beliefs to ourselves without ever having the relevant direct experiences, for example, 'I believe that there was a First World War'. The presumption is that our self-ascriptions are not made less authoritative just because we did not have the relevant direct experience. A different scenario is when a subject B[Bp] where p is in fact false. In this case, even if there is a direct experience, the direct experience will be one in which not- p obtains. Suppose the subject comes to acquire B¬p, the difference in phenomenology the subject could be aware of, if any, would be one between B[Bp] and B¬Bp. The subject might

believe that she has revised her belief about p but she is not able to tell from her first-person belief that she never had the lower-order belief that p .

Another scenario is when the change of mental phenomenology is so subtle that the subject fails to notice a contrast between $B[Bp]$ and BBp . In this scenario, the subject has direct experience with a state of world in which p obtains and comes to acquire Bp but is not sensitive enough to be aware of the difference in phenomenology between $B[Bp]$ and BBp . Hence, she is not able to tell that she did not initially have Bp . A subject like Sam, call her Samantha, could have started off with $B[Bp]$. But before Samantha saw a black swan at the lake, she was strolling in the town and saw some cartoon drawings of black swans in the area's restaurants. Later at a souvenir shop, she came across a counter that sells black swan figurines and black swan photographs.⁴⁹ When she finally saw a black swan, she was excited but not surprised. It is possible that at some point during her stroll in town, she acquired the belief that there are black swans. But since her experience with the world accumulates in an incremental way in this case, the change in phenomenology of having Bp is not vivid or intense enough for her to tell the difference. From the perspective of the subject, she continues to ascribe Bp to herself without noticing the difference between $B[Bp]$ and BBp . From her first-person perspective, no change has occurred at the higher-order level. She is not conscious of the phenomenology of having Bp . If Samantha herself were to answer this question, her answer to the question of 'How do you know' in both $B[Bp]$ and BBp cases will be appealing to her own conviction, to nothing more than the fact that she sincerely feels that she is in Bp . This suggests that from a higher-order first-person point of view, it is indistinguishable to the subject whenever she is conscious of a belief that she has, whether that belief is directed at the world or at her own mental state. If we are willing to grant authority to Samantha on non-epistemic grounds, it is unclear why we should not grant authority to Sam's self-ascription. It is unclear why the presence of a certain phenomenology of Bp should be the basis upon which we grant authority to the subject's BBp but not her $B[Bp]$.

Without a clear picture of the ground(s) for self-ascription, it is difficult to answer the question of how a subject is able to tell whether in believing that she believes that p she actually has a view on p or she only has a view on her belief about

⁴⁹ I owe this example to Mike Martin.

her belief. This intensifies a puzzle M. G. F. Martin has aptly raised in response to Peacocke. For Martin, ‘The puzzle was to explain why in attending to the world I should be given reason to self-ascribe my state of mind.’⁵⁰ This puzzle is now intensified because it seems that from a first-person point of view, the subject cannot even distinguish whether she is attending to the world or attending to her state of mind. Moreover, a puzzle may also be raised in the other way: if BBp and Bp are distinct states, how can we say that when the subject comes to form a mental state that is directed at her own mental state (BBp), she must also come to form a mental state that is directed at the world (Bp). Let me now turn to discuss how this worry makes trouble for the transparency accounts of self-knowledge.

3.2. Impact on the Transparency Account

Proponents of first-person authority of avowals often draw on the following two properties of second-order beliefs to establish the authoritative and transparent character that makes self-knowledge special:

- (1) If Bp, then BBp.
- (2) If BBp, then Bp.

Akeel Bilgrami, for example, takes both of these two properties of second-order beliefs to be the properties that make self-knowledge special.⁵¹ My suggested possibility that one can have a higher-order belief without the lower-order belief threatens (2) directly. But before we go on to discuss the implication of this threat to (2), let me first address how the possibility of a subject’s having BBp without Bp also bear on the way we should think about (1).

Alex Byrne, for example, introduces a broad perceptual model and argues that a subject can know her own beliefs by moving from a lower-order belief to a higher-order belief by ‘reasoning without the perception of anything mental.’⁵² 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

⁵⁰ Martin 1998: 120.

⁵¹ Cf. Bilgrami 2012.

⁵² Byrne 2005: 93.

not need to look for perceptual evidence of one's own mental states or behaviours. Byrne maintains that this is not a version of the inner sense model because the procedure involves the faculty of reasoning rather than some special inner perceptual mechanism that detects one's mental states. According to Byrne, a subject may know her own beliefs by following a rule that he calls BEL:

If P, believe that you believe that P.⁵³

Byrne claims that a subject who follows BEL will necessarily have knowledge of her beliefs, because BEL is a self-verifying rule; which means that if P does express the content of a subject's mental state and if BEL is followed, then the subject necessarily comes to believe that she believes that P. Hence, the lower-order belief that derives from BEL is guaranteed to be true.

I take it that for the antecedent of BEL, Byrne means 'if Bp'. In light of the distinction between BBp and Bp discussed above, we may now see that this rule can be understood to mean either:

BEL (i): If Bp, then BBp.

BEL (ii): If BBp, then BBBp.

For Byrne, a failure to follow BEL involves a failure to recognise Bp. This presupposes the presence of a first-order belief that is available for the subject to affirm. Sam's case is an example in which (i) is not followed. But Sam's failure to follow *p* is not due to his failure to infer the proposition "I believe that *p*" from his lower-order belief that *p*. Rather, it is because there is no available premise from which Sam can infer. It is unclear how Sam can follow BEL (i) when Bp does not figure in his belief system. It might be insisted on Byrne's behalf that following BEL is a necessary condition for acquiring justified first-person second-order beliefs. For a rational subject to make knowledgeable self-ascriptions of belief, she must have followed BEL. However, it is open to objection whether following BEL is necessary for a rational subject to make authoritative self-ascriptions if the subject is prevented from accessing the grounds for her ascription. If a subject is not able to tell whether

⁵³ *Ibid.*: 95.

she is actually using Bp as a premise or only BBp as a premise, and if we are still to recognize BBp is a case of authoritative self-ascription, then there is no obvious reason why B[Bp] cannot also be a case of authoritative self-ascription.

The weakness of Byrne's account in explaining Sam's case lies in the way BEL is ambiguously formulated, namely that it conflates (i) and (ii). For Byrne, BEL is an epistemic rule that has the following structure: if conditions C obtain, believe that *P*. Byrne then uses the example of the link between a ringing doorbell and someone standing at the door to illustrate how the rule DOORBELL works. If one is following DOORBELL, then, if the doorbell rings, she believes that someone is at the door.⁵⁴ However, it is questionable whether the case of DOORBELL is analogous to BEL. In the case of DOORBELL, the obtainment of condition C is an either/or matter—that is to say, either the doorbell rings or it does not. But the case for belief is more complicated. To say that condition C does not obtain in the case of belief, that is, if it is not the case that *p*, this may either mean that the subject does not believe that *p* or that the subject negates *p*. So, a condition under which *p* does not obtain can mean:

B¬*p*, but BB*p*.

¬B*p*, but BB*p*.

Since Byrne appeals to safety to determine whether a belief amounts to knowledge, and since it is ambiguous what counts as a false proposition when following BEL, it is unclear whether a case in which the subject does not believe *p* but believes that she believes that *P* counts as knowledge.

Assuming that Byrne is right and it is the case that when one believes that *p*, she believes that she believes that *p*, an account of self-ascription that only draws on the conditional (1) is inadequate. From the perspective of the subject, she is conscious only of the fact that she believes that she believes that *p* when she ascribes to herself the belief that *p*. (1) fails to account for the special element in self-ascription that explains why it is possible to have BB*p* without B*p*. Any satisfactory account of self-ascription should broaden its guiding question in a way that does not presuppose the presence of B*p* in the first place. A starting place for an account of self-ascription should be how a subject comes to believe that she believes that *p*, rather than how a

⁵⁴ Byrne 2005: 94.

subject moves from believing that p to believing that she believes that p . In this regard, the transparency account seems to be more attractive because it does not assume in the first place that one has a lower-order belief.

The conditional a transparent account seeks to defend is:

(2) If BBp , then Bp .

Here I shall restrict my focus to Richard Moran's account. For Moran, when a subject forms the belief that she believes that p , she should treat the question 'Do I believe p ' as equivalent to 'Is p true?'. So when one comes to form a belief about one's own belief, one is required to answer such a question by reference to how things are in the world, rather than to herself or her own beliefs. As Moran notes, the question of whether I believe that p is a self-directed question, whereas the question of whether it is the case that p is a world-directed question. We can imagine cases where the two questions are not being answered in the same way: 'someone may want to know whether it is resentment *that* he feels [...] whether it is what he is *to* feel.'⁵⁵ Moran thinks that these two questions can be treated as equivalent by the subject because of transparency. For Moran, transparency is a feature of self-knowledge that holds as long as one addresses to the question about her own belief from a deliberative stance. Moran contrasts a 'deliberative stance' with a 'theoretical stance'. From a theoretical stance, the subject provides a description of her state by answering 'What do I believe?';⁵⁶ from a deliberative stance, the subject forms or endorses an attitude of hers by answering 'What am I to believe?'. Moran writes:

In characterizing the two sorts of questions one may direct toward one's state of mind, the term 'deliberative' is best seen at this point in contrast to 'theoretical,' the primary point being to mark the difference between that inquiry which terminates in a true description of my state, and one which terminates in the formation or endorsement of an attitude. (2001:63)

For Moran, 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

⁵⁵ Moran 2001:62, emphasis mine.

⁵⁶ *Ibidi*:63.

can easily fail under such circumstances. Moran gives the example of a person who is convinced by the therapist that she feels betrayed by her sibling.⁵⁷ Judging from the evidence available to her, she is convinced that she has to ascribe this 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.’⁵⁸ It is from a deliberative stance that one answers the question ‘Do I believe that *p*’. The deliberative stance is one from which one forms an attitude in answering the question ‘What am I to believe?’ or ‘Is this what I really want?’.⁵⁹ We can take the world-directed question about *p* and the self-directed question about my belief about *p* as equivalent because the ‘outward looking’ character of first-person belief is rooted in one’s deliberating and forming her own state of mind.⁶⁰ As rational agents, we have the capacity to make up our mind and also review and revise our attitudes.

On this deliberative approach, one may reject my hypothesis about Sam by saying that if Sam is a rational agent and if his avowal is authoritative, Sam must have the corresponding lower-order belief that there are black swans because in determining what his own belief is, i.e., whether he believes that there are black swans, he has already formed a view about the world, i.e., there are black swans. This line of objection is premised on two claims: (1) the question ‘What do I believe?’ and ‘What am I to believe?’ can be treated as equivalent by the subject; (2) the subject is able to distinguish, from a first-person point of view, whether she is answering ‘Is *p* true?’ and ‘Do I believe that *p* is true.’ Even if Moran is right in holding (1), he has not fully satisfactorily argued for (2). And as the above discussion suggests, the subject is not able to distinguish whether she is answering a question about herself or a question about the world. This leaves room for us to say that even if the subject is

⁵⁷ *Ibid.*:85.

⁵⁸ *Ibid.*:85.

⁵⁹ *Ibid.*:63.

⁶⁰ *Ibid.*:64.

capable of entertaining a world-directed question and a self-directed question, it does not automatically entail that when she answers a question about what she believes, she is able to tell which question she is answering. Hence, Sam might think that he formed an attitude about whether black swans exists, but what he in fact did was form an attitude about his own attitude. Since the mental state captured by the self-ascription is a higher-order state about a first-order state, it seems that the subject's first-person perspective always begins at the higher-order level of BBp. The kind of transparency a subject enjoys, at best, is a reflective kind of transparency such that if one has BBBp, then it is the case that one has BBp.

The preceding discussion shows that both Bp and BBp can figure as content in self-ascriptions of belief and sometimes a subject might only have formed a belief about her own mental state without forming a belief about the world, resulting in BBp without Bp. This puts us in a better position to appreciate Evans's emphasis on looking out to the world in making a self-ascription, for it is only through forming a belief about the way the world is that one avoids just forming a belief about the way her mind is. Evans famously claims that:

[I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outwards—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?'⁶¹

However, in light of our preceding discussion, we also come to see that the subject might not be in a position to tell whether her eyes are directed outward upon the world or inward upon her mind. It is puzzling how self-ascription of belief is made. Moreover, even if we can come up with a theory that explains the ways self-ascriptions can be made, the subject from a first-person perspective might not be able to tell what evidential ground she has for her lower-order state when she self-ascribes a belief. This further raises a general sceptical question about self-knowledge: Is the first-person point of view is ever tied to lower-order belief? In the next chapter, I will

⁶¹ Evans 1982: 225.

discuss how this sceptical concern about first-person perspective is highlighted by Moore's paradox.

4. Moore's Paradox

In the previous chapter, we saw how it is always at a higher-order level that we ascribe a belief to ourselves and it seems that our first-person perspective only begins at the higher-order level of BBp. This prevents us from knowing whether in believing that we believe that p , we are actually forming a belief about p or forming a belief about what we believe that we believe. Even if it is the case that the subject does have the lower-order belief that p , from the first-person perspective, it is unclear how she can tell whether her eyes, so to speak, are directed outward to the world or inward to her own mind. In this chapter, I argue that any attempts that seek to explain Moore's paradox by assuming that lower-order belief is constitutive of higher-order belief are bound to fail, given the possibility of B[Bp]. In the first section, I focus on discussing Jaakko Hintikka's classic solution to Moore's paradox and show that his strategy of explaining the paradoxical nature of Moorean sentences by starting at the level of BBp is undermined by the possibility of B[Bp]. In the second section, I discuss other attempts to solve Moore's paradox, which also rely on the principle Hintikka adopted, and suggest that they are also undermined by the possibility of B[Bp]. In the third section, I draw on my previous discussion about higher-order first-person perspective to suggest an alternative 'higher-order' approach to Moorean statements. The point of this chapter is not to offer a new solution to Moore's paradox. It only seeks to suggest that the paradoxical nature of these Moorean sentences could be revealing a central problem highlighted by this thesis, namely, whether it is even possible for the first-person perspective at the level of Bp.

4.1 Moorean Sentences and Hintikka's Solution

G. E. Moore notes that it is absurd for someone to assert sentences such as 'Though I don't believe it's raining, yet as a matter of fact it really is raining'⁶² and 'I believe that he has gone out but has not.'⁶³ What is absurd about these sentences is that there is nothing wrong with just uttering the words. There is also no absurdity in asserting these sentences in the past tense or in the second - or third-person counterparts of

⁶² Moore 1942:543.

⁶³ Moore 1944: 204.

these sentences. But it is absurd to assert these sentences in first-person present tense. Another way in which these sentences are puzzling is that there is no apparent contradiction in the conjunction. It may well be the case that it is raining but I don't believe that it is raining and yet it is absurd for me to assert both conjuncts at the same time.⁶⁴

It is pointed out by philosophers that there are two forms of Moorean sentences, one being ommissive and the other commissive. They are respectively:

(1) p but I do not believe that p

or

(2) p but I believe that not- p ,

In the following, I shall first examine a few proposed solutions to Moore's paradox and point out their problem: assuming that BBp entails Bp .

Jaakko Hintikka draws on doxastic logic to show that a contradiction can be derived in (1). The gist of Hintikka's argument is something like this: (1) is closely related to the following form of sentence:

(3) I believe that the case is as follows: p but I do not believe that p ,

which has the form

$B(p \ \& \ \neg Bp)$

By doxastic distribution principle $B(p \ \& \ q) \supset Bp \ \& \ Bq$, which says if a subject believes that p and q , then a subject believes that p and believes that q , (3) has the form

$Bp \ \& \ B\neg Bp$

⁶⁴ Cf. Moore 1944.

In Hintikka's view, BBp can be reduced to Bp ; $B\neg Bp$ can be reduced to $\neg Bp$ and then further reduced to $B\neg p$. Hence, we get a contradiction:

$$Bp \ \& \ B\neg p$$

Formally, Hintikka's proof is as follows:

Assume the contrary: ' $B_a(p \ \& \ \neg B_a p)$ ' $\in \mu$. μ stands for model set of some model system Ω (with respect to the subject), μ^* stands for the second model set, μ^{**} stands for the third model set.

By using a reductive strategy, Hintikka demonstrates that the counterassumption can be reduced into a contradiction:

1. ' $p \ \& \ \neg B_a p$ ' $\in \mu^*$ from the counterassumption by C.b*;⁶⁵
2. ' $B_a(p \ \& \ \neg B_a p)$ ' $\in \mu^*$ from the counterassumption by C.BB*;⁶⁶
3. ' $\neg B_a p$ ' $\in \mu^*$ from (1) by (C.&)⁶⁷
4. ' $C_a\neg p$ ' $\in \mu^*$ from (3) by (C. \neg B)⁶⁸
5. ' $\neg p$ ' $\in \mu^{**}$ from (4) by (C.C*)⁶⁹
6. ' $p \ \& \ \neg B_a p$ ' $\in \mu^{**}$ from (2) by (C.B*)⁷⁰
7. $p \in \mu^{**}$ from (6) by (C.&)

⁶⁵ Since Hintikka claims that the notion of belief may be discussed in much the same way as the notion of knowledge, conditions C on belief may be formulated on the basis of conditions on knowledge. C.b*: if ' $B_a p$ ' $\in \mu$, and if μ belongs to a model system Ω , then there is at least one alternative μ^* to μ (with respect to a) such that $p \in \mu^*$ (Hintikka: 36).

⁶⁶ C.BB*: if ' $B_a p$ ' $\in \mu$, and if μ^* is an alternative to μ , (with respect to a) in some model system, then ' $B_a p$ ' $\in \mu^*$ (Hintikka: 35).

⁶⁷ C.&: If ' $p \ \& \ q$ ' $\in \mu$, then $p \in \mu$ and $q \in \mu$ (Hintikka: 33).

⁶⁸ C. \neg B: If ' $\neg B_a p$ ' $\in \mu$, then ' $C_a\neg p$ ' $\in \mu$ (Hintikka: 34-5).

⁶⁹ C.C*: if ' $C_a p$ ' $\in \mu$ and if μ belongs to a model system Ω then there is at least one alternative μ^* to μ , (with respect to a) such that $p \in \mu^*$ (Hintikka, p. 35).

⁷⁰ C.B*: if ' $B_a p$ ' $\in \mu$, and if μ^* is an alternative to μ , (with respect to a) in some model system, then $p \in \mu^*$ (Hintikka, p. 36).

(5) and (7) contradict the rule that says if p belongs to μ^{**} , then it is not the case that “ $\neg p$ ” belongs to μ^{**} . Therefore, Moorean sentence (1) cannot be believed by the subject.⁷¹

The important moves are (1) and (4), where Hintikka respectively draws on the condition $C.b^*$ and $C.\neg B$.

$C.b^*$: if ‘ $B_a p$ ’ $\in \mu$, and if μ belongs to a model system Ω , then there is at least one alternative μ^* to μ (with respect to a) such that $p \in \mu^*$.

$C.\neg B$: If ‘ $\neg B_a p$ ’ $\in \mu$, then ‘ $C_a \neg p$ ’ $\in \mu$.

For Hintikka, 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. Since 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 . This shows how $C.b^*$ holds for Bp . If a subject believes that Bp , then the subject believes that for all possible worlds compatible with p , it is the case that p . In informal terms, $C.b^*$ is the principle that says if a subject’s belief that she believes that p , then it is the case she believes that p ($BBp \rightarrow Bp$). $C.\neg B$ is the principle that if a subject believes that it is not the case that p , then, for all that the subject believes, it is possible that not- p i.e. $BBp \rightarrow B\neg p$. We may label rules $C.b^*$ and $C.\neg B$ under the heading of ‘belief-reduction’ principle.

I agree with Hintikka that $C.\neg B$ holds for commissive $BB\neg p$ cases, for we can assume that rejection of every p is transparent to the subject such that the subject rules

⁷¹ Hintikka pp. 53-4.

out possible worlds in which p is true. However, it is unclear why Hintikka thinks that $C.\neg B$ also holds for omissive $B\neg Bp$. Since we have already established the possibility that BBp is not necessarily reducible to Bp , there might be cases where the compatible set of possible worlds is indeterminable. In the case of $BB\neg p$, one has the belief that she does not believe that p . Even if BBp entails Bp , $B\neg Bp$ is not an instance of the iterated belief operator. For Hintikka, for $B\neg Bp$ to satisfy $C.\neg B$ means those worlds in which it is the case that p are ruled out. But this is open to counterexamples that suggest for some instances of $B\neg Bp$, it is indeterminate what worlds can be ruled out. This can happen when the agent never entertained p so that the compatibility of the worlds cannot be determined. If $C.\neg B$ does not hold for $B\neg Bp$, then $\neg p$ cannot be derived in step (5).

If Hintikka's belief-reduction principle does not hold, then his attempt to reduce the counterassumption ' $Ba(p \ \& \ \neg B_a p)$ ' *ad absurdum* will fail, leaving the absurdity of Moore's paradox unexplained. If I am right in suggesting that the reason Hintikka's solution to Moore's Paradox fails is that BBp does not necessarily entail Bp , then any approach to Moorean sentences that employs the belief-reduction principle will face similar problems. In the following, I discuss more recent approaches to Moore's paradox that essentially rely on the same belief-reduction that Hintikka has adopted and argue such an assumption is what renders their explanations unsatisfactory.

4.2 Common Assumption: BBp linked to Bp

One criterion for a satisfactory solution to Moore's paradox is that it has to explain what is paradoxical about Moorean sentences. We may begin our discussion with a functionalist approach to Moore's Paradox. What is problematic about a functionalist approach is that it seems to make the absurdity of Moorean sentences disappear. As Jane Heal points out, a problem with the functionalist way of solving Moore's Paradox is that all versions of functionalism are committed to the view that says 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 .'⁷² If we set out this functionalist conception of belief in detail, we come to see that the

⁷² Heal 1994: 13.

oddity of Moore's paradox disappears on functionalism. 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 knowledge 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. But since 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. It may be argued by the functionalist account, 'a belief can be attributed only if *all* behaviour is unified under the control of one representation.'⁷³ 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 to reach out to pick up A. Since 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.

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

⁷³ Heal 1994: 14-15.

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 could be 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.

It is unclear if the second Muller-Lyer case that Heal considers works against the reformulated version of functionalism she considers. Shoemaker’s account of conscious belief, for example, seems to provide a way to preserve at least the oddity of the omissive version of Moore’s paradox. Recall that Shoemaker takes the relation between first-order belief and second-order belief to be a constitutive one, and his account acknowledges that not all of a subject’s first-order beliefs are accompanied by higher-order beliefs. The first-order beliefs that are constitutively ‘self-intimating’ are ‘available’ beliefs’; but there are also some first-order beliefs whose tendency to become available beliefs are ‘blocked’.⁷⁴ Shoemaker therefore can postulate a ‘divided mind’ to explain the phenomenon of self-deception. According to Shoemaker, a ‘divided mind’ is one such that:

One part of a person’s mind believes something, and another does not, and the part that does not believe it ascribes the belief to the part that does. (2009: 44)

On this account, Shoemaker concedes that there is no self-contradiction in a commissive Moorean sentence if the subject believes in the proposition because it is a matter of the subject herself having contradictory beliefs, not a matter of the sentence itself being self-contradictory. The two conjuncts of commissive Moorean sentences are both available beliefs in two parts of the divided mind of the subject.⁷⁵ But if a subject believes in an omissive Moorean sentence, it cannot be the case that the subject herself has contradictory beliefs, for a single belief cannot consist in believing the proposition and not having a belief about the same proposition. Hence, there are still some contradictory elements in omissive Moorean sentences that do not disappear on functionalist account but also cannot be unexplained by functionalism. However, although the oddity of omissive Moorean sentence is preserved, the

⁷⁴ Shoemaker 2009: 40

⁷⁵ Shoemaker 2009: 46-7.

functionalist cannot provide a good explanation for the oddity of omissive Moorean sentences.

That a functionalist can still preserve the oddity in omissive Moorean sentence suggests that Heal might have overlooked some elements of self-ascriptions in her criticism of functionalism. Perhaps this has to do with her underdeveloped treatment of the epistemology of belief. What the putative phenomenon of a subject's having BBp without Bp sheds light on is that, from the first-person perspective, our belief about belief is not necessarily expressive of a lower-order belief. If this is right, then another way of thinking of the problem with the functionalist approach to Moore's paradox is not that the absurdity of the paradox disappears but that it assumes that one's higher-order belief is expressive of one's lower-order belief. For example, the functionalist still assumes that the first conjunct of the Moorean sentence 'I believe that A is longer than B, but B is longer than A' is an expression of one's belief about the way the mind is whereas the second conjunct is an expression of one's belief about the way the world is. But suppose one cannot tell the difference between these two movements of her mind, then it remains odd how someone would be in a position to utter a Moorean sentence.

Without a clear account of epistemology of belief, Heal's own solution to Moore's paradox also suffers. Heal considers two opposite lines of strategy to solving Moore's paradox.⁷⁶ One line of approach attempts to expand the '*p*' part in a Moorean sentence to 'I believe that *p*'. By contrast, the other line reduces the "I believe that *p*" in a Moorean sentence to '*p*'. For convenience, we may label them as 'belief-expansion' approach and 'belief-reduction' approach. Heal defends the belief-reduction strategy by appealing to the constitutive nature of self-ascriptions. 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*.⁷⁷

⁷⁶ Heal 1994.

⁷⁷ *Ibid*: 22.

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 to say that I believe that p . And if I add ‘but I believe not- p ’, I have thereby contradicted myself. But 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 BBp, there is also a Bp. Heal’s approach to Moore’s paradox faces a similar problem as the one we saw with Hintikka’s solution, that is, it is premised on the questionable assumption that BBp entails Bp.

4.3 One Level Up

So far, we have seen that a general problem with the above approaches to Moore’s paradox has to do with the assumption that the first-person perspective is necessarily tied to a lower-order belief. But if the first-person perspective only starts at the level of BBp as this thesis hypothesises, then any approach to Moore’s paradox has to take the Moorean sentences to be starting at the level of BBp. On this approach, we take the first conjunct in a Moorean sentence, ‘ p ’, to be a claim about one’s own belief so that we get:

(1’) I believe that p and I believe that $\neg p$

and

(2’) I believe that p and I do not believe that p

respectively for commissive and omissive sentences. Heal thinks that since ‘ 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 ’. But Heal might have conflated two levels of beliefs. Given our earlier observation that we are always in a higher-order state when we ascribe to ourselves and that we are not able to tell whether we actually have a belief that p or whether we just believe that we believe that p , a Moorean sentence, whether it is commissive or omissive, is always at the level of a higher-order belief. Hence, Moorean sentences can be unpacked as having the following forms:

$BBp \ \& \ BB\neg p$ (commissive)

and

$BBp \ \& \ B\neg Bp$ (omissive).

It is beyond the scope of this thesis to offer a solution to Moore's paradox here. The main proposal is that, in light of the possibility that a self-ascribed belief is not necessarily linked to a lower-order belief, a satisfactory solution to Moore's paradox has to work at the level of higher-order belief. In this regard, Moore's paradox reveals the absurdity of self-ascription because one of the conjuncts sounds like it is expressing a higher-order belief while the other is expressing a lower-order belief. Perhaps what makes Moorean sentences seem odd is a clash between our common assumption that first-person perspective expresses both our higher-order and lower-order mental states and the putative claim that our first-person perspective only starts at the higher-order level.

Moore's paradox motivates us to have an account of the manner in which our higher-order belief is formed. It will not suffice to simply say it is in virtue of consciousness that we have a higher-order belief about a lower-order belief, for it is possible for consciousness to play different roles in getting us to arrive a higher-order beliefs. It is possible that there are different grounds for one's higher-order beliefs. Before we have a clearer account of how higher-order belief is formed and what their grounds are, it is difficult to decide whether it is even possible for a subject to consciously have both BBp and $BB\neg p$ (or BBp and $B\neg Bp$) in a higher-order state. If it is not possible, then there could be something like a 'blindspot' in Moore's Paradox. This blindspot is different from the way Sorensen understands it, however. For Sorensen, all 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.⁷⁸ On my proposal, the reason the subject cannot at the same time both believe that she believes that it's raining and believe that she believes that it is not raining has nothing to do with the nature of the set of propositions but has to do

⁷⁸ Sorensen 1988: 52.

with the nature of higher-order belief, which might prevent her from having both BBp and $BB\neg p$ or BBp and $B\neg Bp$ at the same time.

Conclusion

The significance of this thesis lies in its suggestion that it is possible for one to believe that she believes that p without believing that p . In Chapter 1, I discuss two dominant accounts of the self-ascriptions of belief in the literature and make explicit the common assumption that a lower-order belief is constitutive of higher-order belief. In Chapter 2, I present a few cases that involve the subjects' experiencing surprise to show that it is possible for one to self-ascribe a belief that p without having a view on p , thereby challenging the common assumption. In Chapter 3, I make a case for the philosophical significance of the putative phenomenon described in Chapter 2. Two related questions are raised. One question concerns the appropriate grounds for self-ascription and the mechanism in virtue of which self-ascription is made; the other concerns the movement of the mind that enables the subject to tell the difference between forming a view about the world and forming a view about her own mental states. Even for those who have an answer to the first question, they will still need to address the second question. Proponents of the transparency method of belief, for example, would argue that no transition is involved between the Bp and BBp in authoritative self-ascriptions. However, without an answer to the second question, proponents of transparency accounts can at most claim that first-person authority and immediacy in self-knowledge is of a reflective kind. That is, if one believes that she believes that she believes that p , then it is the case that she believes that she believes that p . There is still no guarantee that the higher-order belief is linked to first-order belief about p . In Chapter 4, I try to use Moore's paradox to press the main sceptical worry about self-knowledge raised by the putative phenomenon, namely, whether our first-person point of view is necessarily tied to a lower-order belief. I suggest that the absurdity of Moore's paradox could be revealing something peculiar about self-ascriptions. While the audience is inclined to take one's self-ascription of belief to be an expression of one's lower-order belief, one's self-ascription of belief does not in fact express one's lower-order belief, but only one's higher-order belief. An investigation into the philosophical questions raised in this thesis will require another project of its own. The main achievement of this thesis lies in gesturing towards a promising direction for our investigation into the nature of self-awareness and high-order thought.

ACKNOWLEDGEMENTS

I am indebted and extremely grateful to Mike Martin. The few footnotes of acknowledgements are far from doing justice to all the help he has kindly given me. I thank Mike for helping me conceptualise this project from the start and for all his critical comments, patient mentorship, and generous support along the way. I am also grateful to Lucy O'Brien, Paul Snowden, and Kwong-loi Shun for stimulating discussions that help me think through issues related to this topic. I thank all participants at Thesis Preparation seminar for their helpful questions and comments. I would like to especially thank Vanessa Carr, Ben Fardell, Pete Faulconbridge, Harry Phillips, and Ashley Shaw, whose comments prompted me to revise parts of this thesis.

Bibliography

Bilgrami, A. 2012. The Unique Status of Self-knowledge. In Annalisa Coliva (Ed.), *The Self and Self-knowledge*. Oxford: Oxford University Press.

Byrne, A. 2005. Introspection. *Philosophical Topics* 33.1: 79-104.

Casati, R. and Pasquinelli E. 2007. How can you be surprised? The case for volatile expectations. *Phenom Cogn Sci* 6: 171-183.

Carruthers, P. 2011. *The Opacity of Mind*. New York: Oxford University Press.

Charlesworth, W. R. 1969. The role of surprise in cognitive development. In D. Elkind & J. H. Flavell (Eds.) *Studies in Cognitive Development*. Oxford University Press.

Davidson, D. 1982. Rational Animals. *Dialectica* 36:318-327.

Dennett, D. 2001. 'Surprise, surprise.' Commentary on O'Regan and Noe. *Behavioral and Brain Science* 24.5:982.

Dennett, D. 1978. *Brainstorms*. Brighton: Harvester Press.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press (ed. J. McDowell).

Frankish, K. 2001. Systems and levels. In J. Evans and K. Frankish eds., *In Two Minds*, New York: Oxford University Press.

_____. 2004. *Mind and Supermind*. New York: Cambridge University Press.
systems and levels

_____. 2012. Delusions, Levels of Belief, and Non-doxastic Acceptances. *Neuroethics* 5: 23-27.

Heal, J. 1994. Wittgenstein and Moore's Paradox. *Mind* 103: 5-24.

_____. 2001. 'On First-Person Authority'. *Proceedings of the Aristotelian Society* 102: 1–19.

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

Izard, C. 1991. *The psychology of emotions*. Plenum, New York.

Klaus Scherer 1984. 'On the Nature and Function of Emotion: A Component Process Approach,' in Scherer and Ekman eds, *Approaches to Emotion*, pp. 293-317, esp. p. 314.

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

Malcolm, N. 1972-3. Thoughtless Brutes. *Proceedings and Addresses of the American Philosophical Associations* 46:13.

Martin, M. G. F. 1998. An Eye Directed Outward. In C. Wright, B. Smith, & C Macdonald (Eds.), *Knowing Our Own Minds*. Oxford: Oxford University Press.

McDougall W. 1923. *Outline of psychology*. Scribner's, New York.

Moore, G. E. 1942. Reply to My Critics. In P. Schilpp ed, *The Philosophy of G. E. Moore*. New York: Tudor.

_____. 1944, 'Russell's theory of description', in P. A. Schilpp (Ed.), *The Philosophy of Bertrand Russell*, La Salle, Ill.: Open Court.

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

Noë, A. 2002. Is the Visual World a Grand Illusion? *Journal of Consciousness Studies* 9. 5-6: 1012.

- Ortony A, Clore GL, Collins A.1988. The cognitive structure of emotions. Cambridge University Press. Cambridge: Massachusetts.
- Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.
- _____. 1998. Conscious attitudes attention and self-knowledge. In B. Smith and C. Wright (Eds.), *Knowing Our Own Minds*. McDonald Press
- _____. 1999. *Being Known*. Oxford: Oxford University Press.
- Plutchik R. 1980. Emotion: theory, research and experience, vol 4, The measurement of emotions. New York: Academic.
- Quine, W. V. and J. S. Ullian. *The Web of Belief*. McGraw-Hill.
- Robinson, J. 1995. Startle. *The Journal of Philosophy* 92.2: 53-74
- Ryle, G. 1949. *The Concept of Mind*, London: Hutchinson. Page references are to the 2000 republication, London: Penguin Books.
- Shoemaker, S. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.
- _____. 1998. On Knowing One's Own Mind. *Philosophical Perspectives* 2:183-209. Page reference to the reprint in Shoemaker 1996.
- _____. 2009. Self-Intimation and Second Order Belief. *Erkenntnis* 71.1: 35-51.
- Snowdon, P. 2012. How to Think about Phenomenonal Self-Knowledge. In Annalisa Coliva ed, *The Self and Self-Knowledge*. Oxford: Oxford University Press.
- Sorensen, R. 1988. *Blindspots*. Oxford: Clarendon Press.
- Sumitsuji N. 2000. The origin of intermittent exhalation (A! Ha! Ha!) peculiar to human laugh. *Electromyogr Clin Neurophysiol* 40.5:305–309.
- Vanhamme J (2000) The link between surprise and satisfaction: an exploratory research on how best to measure surprise. *J Market Manag* 16:565–582

Wittgenstein, L. 1958. *The Blue and Brown Books*. Oxford: Blackwell.

_____. 1980. *Remarks on the Philosophy of Psychology* (RPP), Vol. 1, edited by G.E.M. Anscombe and G. H. von Wright, translated by G. E. M. Anscombe.

_____. 2009. *Philosophical Investigations* (PI), 4th edition edited by P. M. S. Hacker and Joachim Schulte, translated by G. E. M. Anscombe, P. M. S. Hacker and Joachim Schulte. Oxford: Blackwell.

Wright, C. 1989. Wittgenstein's Rule-Following Considerations and the Central Problem of Linguistics. In A George ed. *Reflections on Chomsky*. Oxford: Blackwell.

_____. 2001. 'The Problem of Self-Knowledge I & II', *Rails to Infinity*. Cambridge: Harvard University Press, 319–373.