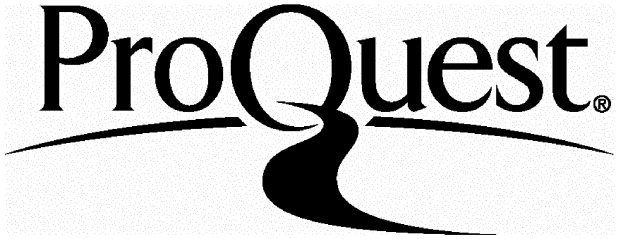
Looking Inside: An Investigation of Introspective Self-Knowledge

By
Isabella Muzio

MPhil in Philosophy
University College London
1999



ABSTRACT

Recent discussion of self-knowledge in the philosophy of mind divides the theoretical options as follows: either we know the contents of our own conscious minds inferentially, or we do so observationally through some form of 'inner sense', or we do so not on in any *way* or on any *basis*, but rather, in virtue of the holding of some constitutive link between our first-order conscious states and our second-order self-ascriptive judgements.

In this thesis, I investigate this special, immediate, authoritative knowledge we seem to have of a certain range of our thoughts, beliefs, desires and other intentional states, and argue that a close examination of the above theoretical lines of approach ultimately shows, contra all three of them, that our knowledge of our own conscious states must be based on these conscious states themselves, considered as states of primitive self-awareness.

More specifically, I argue firstly that there are conclusive reasons for rejecting all three of the above lines of approach to introspective self-knowledge; secondly, that these three lines are not in fact exhaustive; thirdly, that recent attempts made to depart from them either fail to be satisfactory given a certain explanatory aim or end up collapsing back into either the second or the third; and finally, that the only way of avoiding the obstacles faced by all three of these options is by taking on the thesis that our second-order abilities are reflected in the very nature of our phenomenally conscious states, that is, that conscious states, in appropriately conceptually equipped beings, are intrinsically (and somehow primitively) *self-conscious* states; a thesis the details and consequences of which I then outline in the final chapter.

Table of Contents

Chapter I: The Problem of Introspective Self-Knowledge

Unlike other animals, we are not only able to have thoughts, beliefs, desires, and other attitudinal states, but we are also able to *know* that we have them. Moreover, in the case of a certain range of these states, namely our *occurrent conscious* ones, the knowledge we have of them is immediate and authoritative in a way in which our knowledge of other people's thoughts is not, and indeed in a way in which even our knowledge of a wide range of our *own* thoughts is not (eg. our unconscious or repressed thoughts). As I sit here in the library thinking to myself 'animals are not self-conscious' for instance, or reflecting on the question of whether or not introspective self-knowledge should be thought of as observational in character, I am immediately able to know that this is what I am thinking or doing, in a way in which I am not able to know whether anyone else around me is having these thoughts or engaged in a similar reflection. Determining this would require my spending some time observing their behaviour, looking over to see what books they have laid out in front of them, and ultimately, in the case of complex thoughts such as these, having to ask them. Similarly, in the case of a wide range of my own attitudinal states, gaining access to them, that is, to my repressed, or otherwise unconscious beliefs, desires, hopes, fears, etc., would require paying close attention to my behavioural patterns, if not years of psychoanalysis. Nonetheless, although clearly not all our knowledge of our own minds is immediate or in any way different from our knowledge

of the minds of others, it still remains that in *some* cases we are able to know our own thoughts in a way that is distinctive, that is, in particular, in a way which is immediate, first-person authoritative, and immune to certain types of error. How is this possible?

To put things differently, the problem is this: we seem to have a way of knowing the contents of our own conscious minds which is unlike our way of knowing the minds of others, unlike our way of knowing our own unconscious minds, and indeed unlike any of the normal ways we have of acquiring knowledge whether of minds or of anything else, namely by inference and/or through our five senses. And yet, the possibility of this special kind of self-knowledge cannot be denied. I have no doubt, for instance, that I am now entertaining the thought of leaving this room to get a cup of coffee; it does not seem to me that I inferred this from any other beliefs of mine or from my behaviour - I was not looking at myself -; and surely no-one else here is better placed than myself to judge that I am entertaining this thought. Scepticism is therefore not an option,[1] and hence the problem of self-knowledge not that of explaining *whether*, but *how it is* that we are able to know a certain range of our thoughts, beliefs, desires and other intentional states immediately, non-inferentially, authoritatively, and in a way that is immune to certain types of error. In order to answer this question however, we need to first get clear about what exactly is at issue, in particular by considering the following questions: (1) In what sense, and to what extent, is our introspective knowledge of our own minds 'immediate', 'non-inferential', 'authoritative', and not subject to error? (2) What is the class of states of

[1] Even the most extreme of sceptics is going to have to provide *some* account of the distinctiveness of introspective self-knowledge by explaining, perhaps not how this knowledge can be immediate and authoritative as a *kind* of knowledge, but at least why it might be so as a matter of *degree* by comparison to our knowledge of other things (eg. the minds of others, the outside world, etc.). I will return to this view (namely Ryle's (1966)) in chapter 2 below.

which we have this special kind of self-knowledge? And (3), what theoretical options are available to us for explaining the distinctive features of this knowledge, and for thereby accounting for its possibility? Let us consider these questions in turn.

1. What are the distinctive marks of introspective self-knowledge?[2]

First of all, this knowledge is *immediate* and *non-inferential*, in the sense that coming to know introspectively what we are currently thinking or what we consciously believe only seems to require that we ask ourselves the question. For instance, if someone asks me whether I am now thinking about my work, or whether I was just wondering about whether or not to go for a drink, I only need to consider the matter in order to say which it is. I do not seem to need to consult any evidence - or at least not any evidence regarding my *mental states*. I may, though, in some cases, need to consider evidence about how the *world* is in order to make a mental self-ascription. For instance, if I am asked whether I believe that it is raining, I may need to look out the window, that is, I may need to consider whether it is or is not raining, in order to be able to say whether I do or do not believe that it is.[3] However, if no window is in the near vicinity, I am immediately able to say, not having any evidence about the weather either way, that I have no opinion on the matter. That is, although I may have doubts as to whether it is or is not raining, I will usually not have any doubts about whether or not I *believe* that it is. Upon

[2] The expression 'introspective self-knowledge' should at this stage just be taken as an intuitive label for what we are talking about, until the various accounts of what this might amount to are discussed. It may in fact turn out that it is just a form of inferential knowledge or perceptual knowledge. Or, it may even turn out that it is not actually *knowledge*.

[3] See (Evans 1982, chapter 7, especially p.225)

considering the question of whether I believe that it is raining, if I do occurrently consciously believe that it is raining, that is, if I am in a position where I would be prepared to judge that it is raining, then I am immediately able to say that this is what I would judge, or that this is what I believe.[4] In other words, our introspective knowledge of our own minds is immediate and non-inferred, at least in the sense that once we are consciously thinking that p, or in a position where we would consciously judge that p, no *further* inferential move seems to be required in order to know that this is what we are thinking or that this is what we believe.

It should be noted however, that from this non-inferential character of our introspective judgements, it does not follow that these judgements are necessarily *ungrounded* or *baseless*. In fact a judgement or belief can be both *non-inferred* from any other state, and yet *rationally based* on one. This is for instance the case of perceptual beliefs. My perceptual belief that object x is in front of me is not inferred from my perceptual awareness of x, but it is nonetheless *based* on this state of awareness. My awareness of x constitutes a *reason* for my believing or judging that x is in front of me, although my coming to this belief on that basis involved no process of inference. It would therefore make no sense to ask me to defend my belief that x is in front of me, although this belief is not rationally *ungrounded*. Similarly, it might be, with second-order belief. In other words, the mere fact that our introspective judgements about our own thoughts are not *inferred*, and therefore do not admit of nor require any defence, cannot be taken to immediately count against reason-based approaches to self-knowledge in favour of *non*

[4] There are some subtleties here to be considered about the distinction between the self-ascription of occurrent conscious attitudes (eg. occurrent thoughts, judgements, acts of assent, entertainings, pangs of desire, etc.) and that of non-conscious (but not *un*conscious) standing states such as beliefs and desires. I will return to these briefly in the next section.

reason-based ones. Claiming that our introspective self-ascriptions are rationally ungrounded is something which would require further argument.

The second distinctive feature of our introspective knowledge of our own minds is that it is *authoritative*. That is, we seem to stand in a position of authority with respect to the contents of our own minds, which we do stand in with respect to the minds of others. This however, should not be taken to mean that our judgements about our own mental states are *infallible* or *incorrigible*. In fact, both mistakes as well as complete failures to know what we believe, desire, fear, etc., are possible, as testified by common cases of self-deception, or by cases where we discover, say, by catching ourselves behaving in a certain way, or by having this pointed out to us by an analyst (who might here be in a more authoritative position than we are), that we actually have a certain attitude of which we were completely unaware. I may for instance discover, through noticing my strong reaction at the mention of someone's name, that I have strong feelings towards this person, which I was either completely unaware of having (never having even considered the matter), or which I downright believed I *didn't* have, and hence had been self-deceived about. One might also make mistakes about one's own attitudes through some form of irrationality, or even just due to a failure to fully grasp the concepts one is using. In other words, our knowledge of our own minds is clearly neither infallible nor incorrigible, nor indeed always authoritative. Nonetheless, although not *all* our self-ascriptive judgements are authoritative, and although some of them are indeed false, it still remains that there are cases where we clearly *are* in a position of authority with respect to the contents of our own minds. For example, if the thought is now occurring to me that space is not Euclidean, I seem to be far better placed to know that this thought is

occurring to me than anyone else, no matter how attentive they might be to my behaviour. So what might this authoritativeness consist in?

What the fallibility of our mental self-ascriptions suggests, is that the authoritativeness of our introspective judgements cannot be a feature of a certain category of *judgements*, or a certain category of *beliefs*, namely 'mental judgements' or 'beliefs about oneself', that is, a feature of beliefs with a certain *subject matter*, but rather, that it must be a feature of a certain *way of knowing*, since a judgement of the same form and with the same content, say, 'I believe that Jones is out to get me', may in some cases be authoritative, and in other cases *not* be authoritative (eg. when it is made inferentially on the basis of my having observed my paranoid behaviour, or on the basis of having discussed this at great length with my analyst, who may, in this case be in a far better position than myself to say that I have this belief).[5] In other words, those judgements about our own mental states which *are* authoritative (namely, it seems, those which are about our *conscious* states), cannot be so in virtue of being judgements of a certain form, or judgements with a certain type of content, in the way that, say, judgements of the form 'I am hereby thinking that p' can count as authoritative simply in virtue of being of that form. Rather, our (non cogito-like) judgements about our own conscious states, if authoritative, must be so in virtue of being made on a certain *basis*, or reached in a certain *way* not available to others, that is, in virtue of our standing in a certain privileged position with respect to a certain class of our thoughts, which we do not stand in with respect to the thoughts of others, nor with respect to a wide range of our own thoughts (ie. our unconscious thoughts). To put things differently, given the possibility and indeed existence

[5] For this line of attack on the idea of the incorrigibility of a judgement considered *as such*, discussed more specifically in relation to perceptual judgements, see (Austin 1962, lecture 10).

of *non*-authoritative as well as authoritative judgements about one's own mind, there can be nothing about self-ascriptive judgements *in general* that makes them authoritative, but rather, something about the way in which some of them are reached.

The third characteristic feature of our introspective judgements is that, although they are not infallible (even when authoritative), they still seem to be immune to certain kinds of errors, in particular non-cognitive errors, or what Burge refers to as 'brute' errors, that is, errors not due to any cognitive deficiency (eg, irrationality, division of the mind, etc.) or conceptual deficiency (eg. misapplication, or incomplete grasp of a concept).[6] To illustrate this, if I am now consciously thinking to myself 'My keys are on the table', I cannot, it seems, fail to know that I am thinking this. That is, if I consider the matter, I can first of all not fail to know that a thought is occurring to me, nor can I mistake this occurrent thought that my keys are on the table, for, say, an occurrent thought that there is a book on the floor. In fact, if in these circumstances I were to assert 'My keys are on the table, but I do not believe that they are' or if I were to believe that I am thinking about a book being on the floor when in fact I am thinking about my keys being on the table, there would seem to be reason to question either my rationality, or my understanding of the terms I am using, or my sincerity. There would *not* however, necessarily be any reason to question my rationality or conceptual competence if I were to say 'My keys are on the table, but Jones does not believe that they are', or if I were to judge that Jones is thinking about a book being on the floor when she is in fact thinking about a set of keys being on the table (I may just be very bad at interpreting people). In other words, there seems to be a certain range of our thoughts and other attitudes, in

[6] See (Burge 1996)

particular our *occurrent conscious* attitudes, about which we cannot, without irrationality or misunderstanding, be mistaken about, whereas we *can* be so mistaken about the thoughts of others, and indeed even about a wide range of our own thoughts (ie. our unconscious thoughts), which in fact leads us to our next question:

2. What aspects of our own mind do we know immediately and authoritatively through introspection?

As the discussion so far has revealed, we only seem to know a certain very restricted class of our mental states in this special, immediate and authoritative way, namely our *occurrent conscious* states, or our *phenomenally* conscious states,[7] namely perceptual experiences (visual, auditory, etc.), sensations (pains, itches), but also, and more relevantly to our present concerns, phenomenologically occurrent thoughts, entertainings, processes of reasoning, acts of assent or judgement, pangs of desire, etc, that is, attitudes which are occurring in our phenomenal stream of consciousness, or which we are in some sense 'thinking to ourselves in words' or perhaps representing to ourselves in images, and which, to use a Nagelian turn of phrase, there is 'something it is like' for us be having them.[8]

Attitudes which are conscious in this sense need to be distinguished from attitudes which are not *phenomenally* conscious, but which might nonetheless still be

---

[7] This term is introduced by Block in (Block 1995) where he distinguishes 'phenomenal consciousness' from 'access consciousness'. However, for reasons which it is beyond the scope of this thesis to discuss, I will not be dividing mental states up in quite the same way as Block does, although I will, to some extent, be borrowing his terminology.

[8] See (Nagel 1974)

thought of as conscious in the Freudian sense, that is, attitudes which, we might say, are *non*-conscious but not *un*conscious. Such states are essentially dispositional states such as beliefs and desires which are not phenomenally conscious but which are nonetheless rationally integrated with the rest of our conscious attitudes (ie. they dispose us to behave in ways that make sense to us from our conscious point of view), and are such that they can *become* phenomenally conscious upon consideration of their subject matter.[9] An example of such a state might be that of my believing that it is not raining in this room. The proposition that it is not raining in this room is something to which I may not be currently attending, but it is nonetheless something that I believe, and something which I would consciously assent to, were I for some reason to consider the matter of whether or not it is raining in this room. It is also something which I would immediately be able to judge that I *believed*, were I to consider *this* question. But now, contrary to the claim made above, namely that the attitudes which we know immediately and authoritatively through introspection are only our *occurrent conscious* attitudes, we seem to have a case here where I would be able to immediately and authoritatively self-ascribe a *non*-conscious attitude simply upon considering the question of whether I have it. This is indeed true, but, my self-ascribing the belief that it is not raining in this room, in this case, would have to happen via my *attending* to the fact that it is not raining in this room, and so via my coming to *consciously* believe that it is not raining in this room, that is, via my coming to

[9] One might have doubts as to whether there can be such things as phenomenally conscious beliefs, given that beliefs seem to be essentially dispositional standing states, rather than mental *occurrences*. However, one might think of conscious acts of assent or conscious judgements as the conscious counterparts of non-conscious beliefs. That is, one might think of a conscious belief as one which is being consciously expressed. Following Peacocke's 'datum on conscious belief' (See Peacocke 1992, p.154), I will in fact be taking it that linguistically expressed beliefs can be thought of as *occurrent* or phenomenally *conscious* beliefs.

be in a phenomenally conscious state of assent to the proposition that it is not raining in this room, and not just in a *non*-conscious dispositional state. Our phenomenally conscious states thus still seem to remain the primary objects of our introspective self-awareness.

Finally, both conscious and non-conscious states need to be distinguished from states which are *un*conscious in the Freudian sense, that is, unconscious in the sense of being repressed, or for some other reason completely inaccessible to us, and not rationally integrated with the rest of our conscious thoughts, and which, even when known (ie. inferentially), still remain somehow alien to us in the way that other people's thoughts are. These are states which we are essentially unable to immediately control or influence through reasoning alone, unlike our conscious and non-conscious attitudes, which are immediately affected by rational reflection.

With these contrasts between conscious/ non-conscious/ and unconscious attitudes in mind, together with a clearer understanding of the various ways in which our introspective knowledge of our own minds is distinctive, we can now move on to raise, and outline an answer to, the basic question which will be guiding the rest of this investigation, namely:

3. How is immediate, authoritative knowledge of our own conscious mind possible?

Recent discussion of this special kind of self-knowledge divides the theoretical options as follows:[10] either we know the contents of our own minds inferentially, or we do so on the basis of observation through some form of 'inner sense'

[10] See (Boghossian 1989)

or perceptual 'self-scanning mechanism', or we do so not on the basis of any evidence, but rather, in virtue of the existence of some constitutive link between our first-order conscious states and our introspective judgements about them. In the chapters that follow, I will essentially argue that a close examination of these three lines of approach ultimately reveals, against all three of them, that our introspective judgements about our own conscious thoughts must be based on *reasons*, not however on inference or observation, but on our self-ascribed conscious thoughts themselves, considered as states of primitive self-awareness. More specifically, I will proceed as follows:

In chapter 2, I will examine the first two options, focusing more specifically on the second, and arguing that although perceptual models of self-knowledge have certain strong intuitive advantages, they ultimately cannot work because of their incompatibility with a number of essential features of introspective self-knowledge, in particular with its immunity to non-cognitive error, and with its being a way of knowing one's *own knowing mind*, that is, in a certain sense a way of knowing oneself as *subject* and not as object.

In chapter 3, having ruled out the standard reason-based options, I will be considering the third option, namely the possibility that our introspective knowledge of our own mind might be distinctive precisely by *lacking* reasons, the suggestion being that we know our own conscious thoughts immediately and authoritatively not on any *basis*, but rather in virtue of some *constitutive* feature either of our first-order conscious states, or of our second-order judgements. These views, I will divide into two categories, namely into the views according to which it is *ontologically* constitutive either of our first-order states or of our self-ascriptive judgements that we have the corresponding higher or

lower order state, and those according to which it is merely *conceptually* constitutive of having a lower-order state that one will normally tend to form the corresponding higher-order one. I will then argue first of all that strong constitutive views (ie. the former) cannot be right essentially because of their inability to adequately deal with the fallibility of our knowledge of our own minds. Secondly, I will argue that weak constitutive views (ie. the latter), when spelled out, either turn out to be no more plausible than purely reliabilist accounts (which will be discussed in the context of chapter 1), or they actually cease to be *non reason-based*, but turn into a kind of intermediate *reason-based* position between observational models and non reason-based accounts.

Chapter 4 will then be dedicated to examining this possible intermediate position, namely that according to which our lower-order conscious states *themselves* (and not any distinct experience of them) constitute our reasons for self-ascribing them. In doing so, I will suggest however, that adopting this position generates a new explanatory problem, namely that of explaining how a conscious state about the *world* can constitute an immediate reason for believing something about one's *mental states*, given that nothing about what mental states one is in follows from how things are in the world. Given this problem, I will then argue that none of the recent attempts made to uphold the intermediate position are satisfactory.[11] Nonetheless, I will show that a close examination of why these attempts are unsatisfactory will ultimately reveal how the problem should be answered, and indeed that the only way in which it *could* be answered, and thereby the possibility of self-knowledge accounted for, is by taking on the thesis that phenomenally conscious states, in appropriately conceptually equipped beings, are primitively *self-*

[11] I will be discussing in particular the views of Burge (1996) and Peacocke (1992; 1996; 1998)

conscious states, or states of *primitive self-awareness*. In other words, the inevitable conclusion of this investigation will be that we must assume that our second-order abilities are reflected in the very nature of our phenomenally conscious states, if the possibility introspective self-knowledge is to be accounted for.

In chapter 5, I will then lay out in outline how this view of our phenomenally conscious states as intrinsically self-conscious states might be developed in such a way as to avoid it just sending us off on an infinite regress of having to explain again, each time at a different level, how self-consciousness is possible. Finally, I will conclude with some remarks about the consequences of this conclusion, in particular for the way we think about phenomenal consciousness in ourselves, and for the way we think about the subjectivity of non-human animals.

Chapter 2: Looking Inside

Etymologically, the term 'introspection' suggests that we are aware of our conscious states through some form of perception, namely 'inner' perception. That is, to 'introspect' is in some sense to 'look inside'. But, one might ask, in *what* sense? The words 'perceive' or 'see' can be, and often are, pre-theoretically used in a variety of different ways, in particular to mean quite generally to 'know' or 'understand', as in 'seeing' the truth of a mathematical proposition, or 'seeing' what someone means. Similarly, 'look' may be used to mean 'consider' or 'think about', as in 'looking into some matter'. In its most usual modern theoretical sense, however, 'perception' refers fundamentally to *external sense perception*. Speaking of 'introspection' in the theoretical context of philosophy may thus easily seem to establish a (possibly false) analogy between so called 'inner perception' and external sense-perception, or between the knowledge we have of our own conscious mind on the one hand, and the knowledge we have of that which lies outside it, through sense perception, on the other. Asking whether this analogy is legitimate, is indeed the basic philosophical question of whether or not introspective self-knowledge can be said to be perceptual, and will, accordingly, be the guiding question of this chapter.

In other words, the question is: does our introspective knowledge of our own thoughts and attitudes have the features distinctive of the kind of knowledge we have

of the world, or even of ourselves, through sense-perception? In fact, what *are* the essential features of sense perception, and of the knowledge we thereby gain?

In what follows, I will address the issue of whether introspective self-knowledge can be properly conceived of as perceptual, in two stages: first, in section 2.2.1 I will consider the question purely from the point of view of the first-person phenomenology of the two kinds of knowledge (introspective and perceptual), and look at how, from this perspective, they seem to have certain important features in common, which give the analogy between them an initial appearance of overwhelming intuitive plausibility. I will then return, in section 2.2.2 to the question of what exactly is to be understood by 'perception' in this context, that is, to the question of what features of perceptual knowledge are essential to its qualifying as distinctively 'perceptual', and to which our introspective self-knowledge must therefore conform, if the analogy is to have any content. In so doing, I will argue that the analogy fails in a number of ways, some of which, I will suggest, are particularly damaging, and others indeed ultimately fatal to *any* perceptual account of self-knowledge, whatever its details. In other words, this chapter will be essentially guided by the three following questions: Why might one be initially drawn to thinking of our introspective self-knowledge as perceptual? What features must our knowledge of our own thoughts in fact have if it is to properly count as perceptual? And finally, *does* our knowledge of our occurrent conscious thoughts actually have these features? I will argue that it does not, and that introspective self-knowledge is in certain very distinctive ways *non*-perceptual.

But first, before starting to consider whether self-knowledge can be conceived of as based *directly* on observation, it is worth briefly considering the view that

it might, instead, be based *inferentially* on observation, not however of our mental states, but of our behaviour.

2.1 Self-Knowledge by Inference

As seen in chapter 1, we do sometimes know our own thoughts on the basis of inference from observation, in much the same way as we know the thoughts of others. This is so in particular of our knowledge of our unconscious or repressed thoughts. The important question here, however, is that of whether this is could be how we know our own thoughts in *all* cases, that is, even in those cases in which no inference *seems* to be involved, that is, in which we do not *seem* to consult any evidence regarding our self-ascribed mental states.

According to some views, Ryle's in particular,[12] this is indeed the way in which we know our own thoughts even in the so called 'introspective' cases. That is, we always know our own thoughts inferentially through observing our behaviour, in the same way as we know the thoughts of others, the only difference between our knowledge of our own thoughts and our knowledge of the thoughts of others being that we are far better and far quicker at interpreting our own behaviour than we are at interpreting that of others. In other words, the difference between the knowledge we have of our own minds and the knowledge we have of the minds of other people is, on this view, only a difference in *degree* and not a difference in *kind* of knowledge. The immediacy and authoritativeness of self-knowledge arises only from the fact that we are better placed to observe our own

[12] (Ryle 1966)

behaviour than we are the behaviour of others, since we have observed ourselves for far longer than anyone else, and indeed for so long that we are in a position to be able to immediately recognise patterns in what we do, what we say, etc.

Now, is spite of this view's appealingly straightforward simplicity, it is unsatisfactory for a number of reasons. To begin with, we seem to be able to know what we are currently thinking even when no behavioural evidence is available to us for interpretation, let alone actually consulted. I am for instance clearly able to know that I am now thinking that there are nine planets in the solar system, although I am merely sitting at my desk, displaying no behaviour from which I could possibly infer that this thought was occurring to me. In fact what kind of behaviour would that have to be? In the case of most of our thoughts, the only behaviour which could possibly be fine grained enough to allow us to know with precision what thoughts we were thinking would be verbal behaviour, and yet clearly, we are able to know what we are thinking even when we are silent.

A second, and equally decisive problem with this approach to introspective self-knowledge is that it seems to leave us without any explanation of why in some cases we are *not* authoritative about our own thoughts, while in other cases we *are* authoritative. How is it, for instance, that although I may behave consistently depreciatively towards a member of my family, I might nonetheless believe that I admire them, whereas someone else, who has observed me far less than I have myself, would quickly be able to tell that I do not admire this person at all? Or, how is it that in cases (like this one) where I have plenty of behavioural evidence from which I could infer that I have a certain attitude, I may nonetheless be completely unable to say that I have this

attitude, or even go so far as denying it, whereas in other cases, where I have *no* behavioural evidence regarding my attitudes, I am immediately and far more authoritatively able to say that I have them? Given these facts, the special status of the latter types of cases cannot possibly have to do with my being quicker at interpreting my behaviour, since, by hypothesis, I have far more behavioural evidence in the former kinds of cases than I do in the latter. In fact in the latter kinds of cases, I have no evidence at all from which I might infer what I am thinking. My knowledge of my own thoughts in these cases can therefore clearly not arise from inference.

Leaving therefore the inferential model aside, let us return to the more plausible suggestion that we might know our own thoughts not inferentially on the basis of observing our behaviour, but directly on the basis of observing our thoughts, through some form of 'inner sense' or perceptual 'self-scanning mechanism'.

2.2 Self-Knowledge Through 'Inner Perception'

2.2.1

To start from the purely phenomenological point of view, our introspective knowledge of our own conscious states seems to bear a number of striking similarities to perceptual knowledge. For one thing, just like our knowledge of the world through sense perception, our knowledge of our own thoughts is, as we have now established, immediate and non-inferential.

Secondly, our judgements about our occurrent conscious thoughts appear

to be somehow *rationally grounded* despite not resulting from inference.[13] That is, we do not seem to just find ourselves, in certain circumstances, with a sudden impulse to self-ascribe a belief or an experience. Rather, when we make judgements about our occurrent conscious states, these judgements seem to make rational sense to us from our own first-person self-ascribing perspective.

Moreover, not only do we seem to make our judgements about our first-order conscious states on the basis of *reasons*, but we indeed seem to do so on the basis of some kind of *awareness* we have of ourselves being in these states. Very often, in fact, we may be thinking to ourselves things like 'it is a nice day' or 'I should get some work done', or perhaps be engaged in some process of reasoning about how to resolve some practical problem, etc., when suddenly, it occurs to us that we are thinking these thoughts or engaged in this process of reasoning. When this happens, however, the information that we are thinking such and so does not usually strike us as a surprise. We feel that we were aware all along of ourselves having these thoughts or trying to resolve a certain problem, but were just not explicitly thinking about the fact that we were. In other words, we do not generally feel that we just come up with our beliefs about what we are thinking or doing from nowhere. Rather, these beliefs seem to be somehow based on our being aware of ourselves as having the attitudes in question. That is, we feel that we are always in some (as yet undetermined) sense aware of what is currently going through our phenomenal stream of consciousness, but do not always explicitly think about these occurrences. When we do, however, (either by suddenly 'catching' ourselves in the act,[14] or by considering

---

[13] See chapter 1 pp.7-8 above for the distinction between a belief's being *inferred* and its being rationally *grounded.*

[14] See (Boghossian 1989, p.11)

the matter), we generally seem to do so on the basis of this awareness we already had all along of what we were, or currently are, consciously thinking.

In fact, not only do we seem to base our judgements about what we are *currently* thinking, on some sort of awareness we have of ourselves doing so, but we also seem to base our judgements about what we were *just a moment ago* (or much longer ago) thinking or doing, on this same kind of awareness we then had of ourselves being in these conscious states. I may, for instance remember being engaged in a conversation with someone, or thinking that p, or wondering whether q, although at the time, I was not thinking about myself at all, but only about the topic of the conversation I was engaged in, or about p and q. In retrospect however, I feel that I can immediately and authoritatively state what I was thinking or doing at the time, because I *remember doing it*, just as I may later remember there being a green car parked outside my front door, although at the time, I was not thinking about the car being there. I was, however, perceptually aware of the car, and, similarly, at the time of conversing with my friend or thinking that p, I was also (or so it intuitively seems) in some sense *aware* of myself doing so.

In other words, it seems as though there may be more to the thought that our knowledge of our occurrent intentional states is perceptual, than a mere ambiguity in the terms 'look' and 'see'. Having said this however, appealing to a perceptual account of self-knowledge does not seem to be the only possible way of explaining the above phenomenological features. I will in fact suggest in chapter 5, below, a different way in which this might be done. In any case, it still remains that close attention to phenomenology, and to our first-person subjective standpoint as self-knowers, is essential

to fully understanding the phenomenon of introspection, and that therefore any account which is unable to explain, or somehow accommodate, the above considerations, ought, at best, to be regarded with some scepticism. To this extent, perceptual approaches to self-knowledge certainly seem to have at least one strong intuitive advantage. This is, however, by no means sufficient for drawing the conclusion that our introspective knowledge of our own thoughts is, in fact, perceptual. Other crucial criteria need to be met.

2.2.2

Amongst the many distinctive features of our knowledge of the world through sense perception, the following seem to be some of the most essential, and potentially most threatening to the analogy between perception and introspection:

(1) Firstly, it is sometimes claimed that perceptual beliefs are beliefs only about the intrinsic, non-relational properties of that which they are about, or at least that this is so when that which they bear a relation to is not also being perceived or otherwise known. To borrow an example from Boghossian, we can tell, by merely looking at a coin, that it is of a certain size, has a certain shape, is made out of a certain type of metal, has a certain colour, etc. but cannot, for instance, tell what monetary value it has.[15] This knowledge needs to be inferred from some further information about how the coin's intrinsic features relate to the possession of monetary value. Or, to borrow an example from Davidson, we cannot tell just by inspecting a burn on someone's skin whether or not it is a *sun*burn.[16]

---

[15] (Boghossian 1989, p.16)

[16] (Davidson 1986)

Following on this idea, a recent objection against perceptual accounts of self-knowledge has it that we cannot possibly know the contents of our own thoughts perceptually, since, given a plausible externalism about mental content, the contents of our thoughts are not intrinsic, non-relational properties of these states[17]. This objection, however, seems to assume both a very narrow conception of what an attitude is, and a very narrow conception of what 'inner sense' might be. That is, first of all it assumes that a propositional attitude is an internal state of a person which happens to also have certain relational 'content' properties, rather than something which is *intrinsically* a relation to the world, and therefore whose content properties could, conceivably, be perceived. Secondly, it assumes that 'inner sense' is necessarily to be thought of as 'inner' in the sense of being directed inward towards our *heads*, rather than inward towards our *perspective*, a perspective which may also be thought of as "reaching out into the world".[18] When looked at this way, the relational nature of our attitudes no longer seems to pose any imminent threat to the conception of self-knowledge as perceptual.[19] A different

---

[17] See for instance (Shoemaker 1996), and (Boghossian 1989).

[18] Shoemaker considers, in a footnote (1996, p.212), the idea that 'instead of thinking of a belief as something internal to the person, and its contents as constituted by its relations to other things, one could think of it as "reaching out into the world"'. However, he then dismisses the thought that this might make inner sense models more plausible, because, it seems, he continues to assume that 'inner sense' is 'inner' in the sense of being directed inward towards our heads, that is, that 'inner sense' must be a way of seeing internal states of a person. In fact he writes: '...how could inner sense reach out into the environment?'

[19] The more serious threat thought to be posed by externalism is not one that applies restrictedly to perceptual accounts, but to the authoritativeness of self-knowledge in general - thereby motivating Rylean types of positions. I will not, however, be going into this issue, essentially because, in agreement with (Burge 1988; 1996), I take the issue of externalism/internalism about mental content to be altogether orthogonal to the debate about a certain kind of self-knowledge. Externalism, it seems to me, may be relevant to the question of whether we are authoritative in our knowledge of what the meanings of our words or the contents of our thoughts *consist in*, but not to that of whether we are authoritative in our knowledge of we are thinking or what we believe *understood in our own terms*. In the former sense, I do not think we actually are authoritative, and in the latter sense, although we are authoritative, the truth or falsity of externalism is, I believe, irrelevant to this being the case, and it is only this kind of self-knowledge which is our present concern.

feature of perception, however, which does seem to make the analogy with introspection somewhat more problematic, is the following:

(2) In perceptual knowledge, there is always an intermediate perceptual state, or sense-impression (visual, auditory, proprioceptive, etc.) which is distinct from the perceptual belief or knowledge it gives rise to, as well as distinct from the object of perception that causes it.[20] My perceptual belief that there is a cup of coffee in front of me, for instance, is based on my being perceptually aware (through sight, and perhaps also smell) of there being a cup of coffee in front of me. My awareness of the coffee, is in turn somehow caused by the presence of the object. Similarly, my knowledge, through proprioception, that I am leaning forward rather than standing upright, is based on a distinctive sensation (a sense of a particular kind of imbalance; an impression of leaning forward), which is itself somehow caused by my body's being in a certain position. In other words, in both proprioception and more straightforward cases of external sense-perception, our perceptual beliefs seem to always be based on some intermediate informational state of awareness of that which they are about, distinct both from our belief and from its object. In fact, it seems that there *must* be such a distinction if such things as the possibility of misperception and the possibility of disbelief in one's senses are to be accounted for. In other words, insofar as it is possible to disbelieve one's own senses, and hence to see something without believing it, one's beliefs and one's sense experiences cannot possibly be one and the same state. They must be distinct. Similarly, given the possibility of misperception (eg. the brute misidentification of an object, or the misperception of some of its properties), one's perceptual experiences and the objects

[20] See (Shoemaker 1996, p.205)

thereby perceived must also be distinct. No mismatch between them would otherwise be possible. But now, if this is an essential feature of perceptual knowledge, we seem to have here an important dis-analogy with self-knowledge. In fact in self-knowledge, most would agree, and this in spite of some of the phenomenological features mentioned earlier, that there is no separate informational state (an experience of our first-order conscious states) in between our first-order states and our second-order judgements about them, which is caused by the former and which rationally grounds the latter.[21]

There are however, it seems, two ways in which one might try to bypass this dis-analogy: (a) one might deny that there actually is a dis-analogy with introspection here, and argue that our introspective beliefs *are* based on a kind of experience of our conscious states, namely on the 'phenomenal feels' or 'what it is like' properties associated with having them, by which we might be able to 'sense' that we have them. Or, (b) one might deny that perception actually involves the existence of any perceptual states distinct from our perceptual beliefs.

To start with (a), this idea would have the great virtue of providing us with an account of what it is about an attitude of ours being *conscious*, as opposed to *un*conscious, that enables us to know that we have it in a special way in which we do not know our unconscious attitudes (these not being accompanied by any phenomenal feel by which we could sense them). However, on closer examination, this view is unsatisfactory for the simple reason that it is not clear at all how there could be a recognizable, and sufficiently fine grained, phenomenal feel associated with each occurrent belief or other conscious attitude, which would allow us to know that we had it. In fact if this view were

---

[21] See for instance (Shoemaker 1996, p.207) and (Burge 1996, p.105).

correct, it would have to entail that my consciously thinking, say, the thought that space is not Euclidean, would always be associated with some specific kind of phenomenal feel which would be recognizably the same on each occasion in which this thought occurred to me, and which I would recognize even on the first occasion on which I thought it. It is not clear at all, however, that any purely phenomenal feel associated with such a complex thought could possibly be focused enough, or fine grained enough, to ground a belief that one has exactly that thought and no other.

Turning therefore to option (b), on some views of perception, in particular Armstrong's,[22] what is central to a belief's being perceptual is not that there be some intermediate experience between this belief and that which is believed, but simply that there exist a reliable, contingent, causal mechanism linking the objects perceived to one's beliefs about them. In fact, on this view, perception amounts to nothing more than the acquiring of beliefs through a reliable causal mechanism. If so, then holding a perceptual view of self-knowledge does not require being committed to the existence of separate phenomenal experiences of our conscious states on which our introspective beliefs could be rationally based. Armstrong himself, in fact, holds such a view of self-knowledge.[23] The central claim of this kind of model is indeed that there exists a reliable, but contingent, causal mechanism in our brain, a 'self-scanning process', which can be thought of as analogous to many of our other perceptual mechanisms (sight, hearing, etc.), and whereby our first-order conscious states directly cause us to have second-order beliefs about them. If one accepts, therefore, a purely reliabilist approach to perception, one can hold on to

[22] (Armstrong 1968)

[23] (Ibid)

a perceptual account of self-knowledge in spite of its non conformity to feature (2) above. The question however is this: should we accept a purely reliabilist approach to perception? And, indeed, if we do (or even if we do not), what are the advantages of thinking of our knowledge of our own minds on this model?

To start with the first question, a general problem with this whole approach to perception seems to be that it does not allow for the possibility of disbelief in one's senses. This is particularly problematic when thinking about ordinary cases of *object* perception, where disbelief in one's senses clearly does seem possible. I may in fact in certain circumstances mistrust what I see (perhaps because I believe I have been given some drug) and therefore not form any beliefs corresponding to my perceptions. Seeing, in this case, would therefore not involve *believing*. Secondly, a purely reliabilist approach to perception seems unable to make room for the fact that, from the phenomenological point of view, when we make a perceptual judgement about something, we do not seem to just find ourselves with an impulse to make it, in the way that we do just 'find ourselves' having perceptual experiences through no rational choice of our own. That is, our perceptual judgements are judgements which, we generally feel, make rational sense to us. They are judgements which we feel we could *withhold* were we to have reason to mistrust our senses, in a way that we could not, out of any rational motive, withhold a perceptual experience if our eyes are open. In other words, a purely reliabilist account of perception seems unable to make room for the important phenomenological difference between seeing and believing, and is therefore perhaps not the right way of thinking about perception. Of course, even if this is granted, pure reliabilism might still be the right approach to *intro*spection, although in this case, it would turn out not to be a *perceptual*

approach in at least one important respect. In fact, not only would this not be a truly perceptual model, but it is unclear what the advantages of thinking of our knowledge of our own minds on this model would be, even if it *could* be thought of as a genuinely perceptual model. In fact, as seen in section 2.2.1 above, the main appeal of thinking of self-knowledge as perceptual was that it made it come out as *reason-based* (in the sense of making rational sense to us from our own point of view)[24] in accordance with the phenomenology of many common cases of introspective belief. But now, if to adopt a perceptual model of self-knowledge is to adopt a purely reliabilist, non reason-based account, we are left without any clear reason for preferring a perceptual model of self-knowledge to a non-perceptual one - other than, perhaps, the fact that perceptual accounts, whether reason-based or purely reliabilist, seem to be better able to accommodate the fallibility of self-knowledge than certain non-perceptual accounts. This can of course only constitute a reason for preferring a perceptual account, if it turns out that the kind of fallibility to which our knowledge of our own minds is subject, can indeed be properly likened to the fallibility of our senses, which leads us indeed to consider the analogy between perception and introspection on the question of fallibility.

(3) In perception, the objects of perception and perceptual knowledge (eg. a table) bear no constitutive, conceptual or rational relation to anyone's awareness of them or perceptual beliefs about them (eg. one's experiences or beliefs about the table). That is, in perception, the relation between awareness and object of awareness (although perhaps not between awareness and perceptual belief), is a purely causal relation between

[24] One might hold a view about reasons according to which reasons are just to be reduced to probabilistic links between beliefs formed and facts the beliefs concern. I am not sure, however, how this could possibly capture the intuitive idea behind our concept of a reason. Whether or not it can, however, this is not the sense in which I am here using the expression 'reason-based'.

two *distinct* and entirely *independent* existences. Brute error or misidentification, due to no cognitive failure of ours, is therefore always possible.

Now, as already seen earlier however,[25] brute errors of the kind possible in perception do not seem to be possible in introspection. That is, it seems impossible that someone could, for instance, mistake one occurrent conscious belief for another, or a belief that p for a desire that q, in the way that one might, due to no cognitive failure or perceptual malfunction, perceptually mistake one object for another, say, some clothes hanging on a chair in the dark, for a seated person. Similarly, it seems impossible that one might (again due to no cognitive deficiency) fail to be able to say what, or whether anything at all, is currently going through one's mind, in the way that one might fail altogether to be able to say what, or whether anything at all, is directly in front of one, due to some very think fog, for instance.

Of course, a perceptual theorist of self-knowledge might just reply that our inner scanning mechanism is far more accurate and reliable than any other of our perceptual mechanisms, and that, moreover, there just happen to be no external factors (the equivalent of lighting conditions, etc. in the case of visual perception) in our heads, which might interfere with our perception of our thoughts. An immediate response to this, however, is that all errors in self-knowledge seem to be due to some cognitive failure or other: irrationality, division of the mind, etc., and never occur in the absence of such failure. In fact consider the following examples. People with split personality disorders might fail, in one of their personalities, to be aware of a conscious thought had by them when in another personality, and yet we would not diagnose these failures as cases of

---

[25] See chapter 1, pp.10-11 above

benign misperception, but rather as cases of division of the mind, that is, as something cognitively pathological. Other extreme examples of failures of self-scanning might be found in schizophrenics who complain of 'thought insertion', and who deny that they are responsible for the thoughts they are having, that is, who deny that it is *they* who are the authors of the thoughts they are introspectively aware of. Again, although we would seem to have here a case of downright failure to identify oneself as the *subject* of one's thought, we would not classify such a failure as a mere error of identification made by a perfectly rational subject who, say, was not looking closely enough. Rather, we would attribute such a failure to a serious failure in one's rational thought processes. More common cases of error in self-knowledge might be cases of self-deception, or other cases of the kind already considered in chapter 1 above. In fact, leaving aside cases which we clearly know to be pathological, if someone were to come up to us and sincerely say things like 'It is 12 o'clock but I do not believe that it is', or 'someone is thinking that it is now 12 o'clock but I do not know whether it is I who is thinking this' or 'I am not sure whether it is now occurring to me that it is raining or whether I am wondering about the nature of self-knowledge', we would immediately assume that there was something wrong either with their rational thinking or with their understanding of what they were saying. We would not take such errors to be cases of misperception without irrationality or conceptual deficiency. In other words, in self-knowledge, we do not seem to have any cases which we would classify as simply failures of some inner scanning mechanism without irrationality or other cognitive breakdown, whereas in sense perception, almost all cases of failure in perceptual mechanism which lead to erroneous belief, we would *not* classify as involving any irrationality, conceptual incompetence, or other cognitive deficiency. If

there is an inner scanner, therefore, its proper operation is far more closely tied to the ascription of rationality than in the case of sense perception. This in fact would suggest that there must be, in introspective self-knowledge, a *rational* relation between our conscious thoughts and our self-ascriptive judgements about them, in addition to whatever underlying causal mechanism may or may not also be involved at the sub-personal level. That is, having a first-order conscious state must somehow constitute a *reason* for believing that one has it, as only this could seemingly explain why a failure to believe that one has it should count as a *rational* failure.

(4) A fourth, and not unrelated point about perception, is that speaking of perception or observation (understood as external *sense*-perception) generally implies the idea of a perspective or point of view *on* something, on something which lies *outside* the observing perspective, something which is *external* to it. If so, however, the following question immediately arises: does it actually make sense at all to think of our knowledge of the contents of our own minds, that is, of the contents of our *own knowing perspective*, on the model of a kind of knowledge which is, of its very essence, knowledge only of that which lies *outside* our knowing perspective?

This in fact relates back to a familiar worry about how a subject could, *qua subject*, become an object to itself,[26] and the apparent incoherence of this idea that we could know ourselves as subjects through perception, given that in perception, by its very nature, things only present themselves to us as *objects*.

It is not part of the aim of this investigation to go into the question of how there could be perception of the self or of the 'I' which thinks. However, the above

[26] See in particular (Hume 1888). See also Shoemaker's discussion in (Shoemaker 1986).

problem arises even when restricting ourselves to the issue our knowledge of our own thoughts, the problematic question being the following: how could we perceive a thought of ours *from the inside* so to speak, that is, from the very same point of view from which we are thinking it? If perception involves having a point of view *on* something, looking into our own mind or into our own point of view, would have to involve having a point of view on our own minds, thereby creating an immediate distance between our observing perspective and the perspective of the thoughts observed. And, if we then tried to look into our *observing* perspective, another dissociation would occur between *this* perspective and the new observing perspective, and so on *ad infinitum.*[27] This is a problem which most people would feel the pull of, and yet it is somewhat obscure what we should make of it. In fact, one might ask, what would it be to know one's own thoughts as the subject of these thoughts, or from 'the same point of view' as these thoughts, or 'from the inside'? And, indeed, why exactly could this knowledge not be perceptual? One way of getting at this idea is by first considering the notion of a 'mind' or a 'perspective'.

A 'mind' or a 'point of view' might, in this context, be taken to consist roughly in a coherent system of rationally related intentional states, that is, a system of states which together form a single unified picture of the world, and which not only fill the same logical space (in the way that two different people's attitudes may also do), but are also immediately causally and/or explanatorily related. That is, beliefs and desires within this system, which have appropriately related contents, will immediately explain, affect, or give rise to actions or other attitudes with relevantly similar contents, within this system. Now given this understanding of a mind or a perspective, knowing one's own

[27] For a discussion of this phenomenon, although not specifically in relation to perceptual accounts of self-knowledge, see (Sartre 1969, chapter 2, section III).

attitudes from the inside, or having direct knowledge of one's *own knowing perspective*, might be understood as a case of having a belief about a certain attitude, where one's belief and the attitude it is about form part of the same rational system, that is, part of the same point of view, and bear, therefore, a direct rational relation to each other.

But now, if this is what introspective self-knowledge amounts to, then it cannot, it seems, be *perceptual*, since, as seen in the last section, the relation between a perceptual experience and its object must be a purely causal, *non*-rational relation, if brute non-cognitive error of the kind which clearly *is* possible in perception, is indeed to be possible. To put things differently, if the objects of perceptual awareness formed part of the same point of view as our perceptual experiences or perceptual beliefs about them, they would have to bear a direct *reason-giving* relation to these perceptual beliefs, characteristic of what it is to occupy the same point of view. If so however, a failure to perceive these objects, or to form a belief corresponding to them, would constitute a failure to take account of one's reasons, that is, a failure of rationality. From this it would follow that error without irrationality would not be possible, and yet being subject to such error is, we have seen, of the very essence of perceptual knowledge. Knowledge had from and about the same rational perspective can therefore not be perceptual.

But perhaps we never do have knowledge of ourselves from the inside in this very strict sense, and, if so, what we call *self*-knowledge or *introspective* knowledge, may well just turn out to be a kind of perceptual knowledge. In fact, a perceptual theorist could argue that our knowledge of our own thoughts is a kind of knowledge which is somehow *distanced* from the perspective of our first-order thoughts, that is, a kind of knowledge where our thoughts are objectivised as if they were someone else's thoughts.

One could in fact argue, following Armstrong, that '... it is only an empirical fact that our direct awareness of mental states is confined to our own mind. We could conceive of a power of acquiring non-verbal non-inferential knowledge of current states of minds of others. This would be direct awareness, or perception, of the minds of others. Indeed, when people speak of 'telepathy' it often seems to be this they have in mind'.[28] In other words, the idea is that there is nothing inconceivable about the possibility of knowing the thoughts occurring in other people's minds through some contingent, non-rational, brutely causal telepathic perceptual mechanism, so why should there be a problem with supposing that our knowledge of our own thoughts is perceptual in this way? In fact wouldn't this make more sense than to think otherwise? Why should we think that our knowledge of our own thoughts is a special kind of knowledge of these thoughts had *from the inside*, that is, had in virtue of some *direct rational* relation holding between these conscious thoughts and our judgements about them?

Well, there does not seem to be any problem with the supposition that we might be able to know *some* of our thoughts perceptually through some form of extra-sensory perception, such as some of our unconscious attitudes, for instance. However the crucial question here is whether the knowledge we *actually* have when we introspect, given its peculiar features (eg. the fact that only certain types of errors are possible, and not others), and given the uses to which we put it (eg. in reflective reasoning) could be of this kind. That is, do we in actual fact have knowledge of some of our own thoughts as the *subjects* of these thoughts, or is all self-knowledge had from a different perspective from that which it is knowledge about, and so a kind of knowledge which we could

---

[28] (Armstrong 1968, p.325). See also Churchland's discussion of telepathic knowledge in (Churchland 1991, pp. 610-611).

conceivably also have of the minds of other people?

There are, it seems, at least three reasons for believing that introspective self-knowledge is knowledge of our own attitudes from the inside. Firstly, if it were not, it would be difficult to explain why, in self-knowledge, error is so closely tied to irrationality. Secondly, again if self-knowledge were not truly 'inner' in this way, it would be difficult to explain why when we non-inferentially self-ascribe a conscious attitude, say, a belief that p, we generally seem to do so with a certain commitment to the view that indeed p, such as when we say things like 'I believe it is time to go.' In such cases, we do not seem to just be reporting a belief in the uncommitted way in which someone other than ourselves might do, by saying 'She believes it is time to go'. Thirdly, an actual example of knowledge of our own thoughts had from the same perspective from which we are thinking them can be found in our practices of critical reasoning.[29] In fact, to illustrate this, let us digress and briefly consider in more detail what it is to reason critically.

Following Burge, critical reasoning is essentially reasoning where a thinker:

(1) recognizes her attitudes, reasons and reasoning *as* attitudes reasons and reasoning,

(2) reasonably evaluates these attitudes, reasons and reasoning by reference to rational norms, and

(3) where these reasonable evaluations constitute immediate reasons, and immediately rationally result in, explicit confirmation, review, or supplementation of the attitudes,

---

[29] See especially (Burge 1996). See also Shoemaker's discussion of reflective reasoning in (Shoemaker 1988).

reasons and reasoning reasoned about.[30]

Now, if I start thinking about the various beliefs I hold about the nature of introspective self-knowledge (stage 1), and begin evaluating them and considering whether they are reasonable beliefs to hold, thereby reaching the conclusion that one of my beliefs is unreasonable (stage 2), this evaluation, that is, my conclusion that one of my beliefs about self-knowledge is overall unreasonable, will constitute an immediate reason for my dropping this belief (stage 3), in a way that my coming to this conclusion about someone's else's beliefs would not itself alone, constitute an immediate *prima facie* reason for them to drop *their* belief. In fact consider the following situation: I am listening to a philosopher expressing her views on self-knowledge, and I thereby start reasoning about her views as a result of which I come to the conclusion that one of her views is unreasonable. My merely coming to this conclusion is clearly not in itself enough to make it the case that there will be immediate reason for her to change her belief. In order for my evaluation to result in a change in her views, I would first have to convince her of the truth of my evaluation. My coming to think it alone would not immediately rationally result in her changing her mind, in the way that my coming to this conclusion would itself have an immediate effect on what views *I* hold. My reasoning about her views would therefore not constitute a process of genuine critical reasoning in the sense defined above. In other words, the point to take from this is that genuine critical reasoning seems to involve there being a certain *rational integration* of our first and second-order attitudes. They must immediately rationally influence each other. That is, the first and second-order thoughts

[30] As Peacocke points out (1996), we do not always reason fully reflectively in this manner, and other animals most likely never do. The only relevant point here however, (not that Peacocke denies it) is that *we do* sometimes engage in this kind of reasoning, or are at least able to.

involved in critical reasoning must form part of the same point of view, since the two levels (first and second-order) in such reasoning, if such reasoning is to be possible, must stand in immediate reason giving relations to each other.[31] If this is right, then it looks as though we do sometimes have knowledge of our own thoughts from the inside, that is, as the *subject* of these thoughts. Moreover, given that the judgements we make when we introspect are the very judgements we use in critical reasoning, our introspective self-knowledge must be of this very kind, and so cannot be perceptual.

* * *

To conclude, we have seen in this chapter that the terms 'introspection' and 'inner sense',[32] in suggesting an analogy between our knowledge of our own minds and our perceptual knowledge of the world, capture certain important aspects of the phenomenology of much of our introspective self-knowledge, and yet are ultimately misleading for the following reasons: first of all, any perceptual account of self-knowledge which is not committed to the existence of inner sense experiences, distinct from both our first-order thoughts and our second-order judgements about them, ends up turning into an intuitively implausible, purely reliabilist account, which it is not even clear whether it can be properly thought of as perceptual. Secondly, given, on the one hand, the nature of

[31] I will return to discuss the relevance of critical reasoning in more detail in chapter 4 below in the context of discussing Burge's views on self-knowledge.

[32] 'Inner sense' is not always used to refer to a form of inner perception in the sense discussed in this chapter. In fact, Kant for instance, in speaking of 'inner sense' in the first Critique, does not seem to have anything like a perceptual model of self-knowledge in mind. In spite of the misleading term 'sense', 'inner sense' in Kant just seems to mean self-consciousness, or that primitive self-awareness that comes with all other conscious states including perceptual states; that 'sense', so to speak, which accompanies all others, but which is not itself one of them.

perception as essentially involving a dissociation between the observing and the observed perspectives, and given, on the other hand, the notion of 'inner' knowledge as a way of knowing the contents of our *own* observing perspective, we end up having to choose between taking introspective self-knowledge to be truly 'inner', and taking it to be truly perceptual, that is, between taking it to be knowledge had *from the inside*, or knowledge had *by looking*. It cannot be both. Since its immunity to non-cognitive error together with the actual existence of practices involving a rational integration of our first and second-order attitudes show it to be 'inner', it follows that it cannot be perceptual.

Having seen that our knowledge of our own minds, given its distinctive features, can neither be based on inference from observation, nor on direct observation of our thoughts, a natural next option to consider is that it might be distinctive precisely by not being based on anything at all, that is, by *lacking* reasons. I now turn to consider this option.

Chapter 3: 'No-Reasons' Accounts [33]

The view that self-knowledge is not reason-based, is shared by a wide variety of positions ranging from strong constitutive views, according to which it is somehow constitutive of believing that one believes that p that one actually does believe that p, all the way to purely reliabilist views, according to which our first and second-order beliefs are both ontologically and conceptually independent, although somehow causally linked at the sub-personal level. These very different positions share however a common commitment to the view that our immediate introspective attitudinal avowals are not based on reasons, that is, that they are not rationally grounded in any way: they are neither based on other beliefs, nor on observation (whether of our behaviour or directly of our conscious states), nor even on our conscious states themselves.[34] Rather, on each of these accounts, there is something else (if anything at all) in virtue of which our mental self-ascriptions have their distinctive features of immediacy, authoritativeness, immunity to certain types of error, etc. Having already discussed pure reliabilism in chapter 2, in this chapter I will be examining the various lines of 'constitutive' non-reason-based approaches to mental self-ascriptions available, dividing them into two categories: (1) *artefact of grammar*

---

[33] This expression is borrowed from (Peacocke 1998).

[34] In the case of reliabilism, this is of course only true of *certain types* of reliabilist positions, in particular *not* of those which hold reliabilism across the board for all knowledge, and therefore according to which to be reason-based is just to be produced by a reliable purely causal mechanism.

views or *strong constitutive views*, according to which it is in one way or other ontologically constitutive of having a second-order attitude that one actually does have the corresponding first-order attitude or vice versa, and (2) *weak constitutive views*, according to which it is conceptually constitutive of having a first-order conscious attitude that one will generally tend to form a correct second-order belief about it. In each case, I will ultimately argue that the position offered is unsatisfactory, and that an epistemological approach to mental self-ascriptions must therefore be returned to, although neither an inferential nor a perceptual one.

3.1 Artefact of Grammar Views

In essence, the fundamental claim of these views is that the immediacy, authoritativeness and otherwise specialness of our knowledge of our own minds is just an artefact of a grammatical misconstrual, a misconstrual either of *expressions* of beliefs as truth-evaluable *assertions* about them, or of a mere *language game* as the reflection of a language-independent reality and special way of knowing it upon which this language game is consequential, or simply the misconstrual of self-verifying judgements as reflecting a special way of gaining knowledge.

To be more specific, according to the first kind of artefact of grammar view, which could be called the 'expressivist view', we sometimes, although not always, say things like 'I am in pain' or 'I believe that p' simply as an alternative way of saying

'ouch' or 'p'.[35] It is in *these* cases, according to this view, that our mental self-ascriptions are immediate, non-inferential, authoritative, and immune to certain kinds of error. When our mental self-ascriptions are used as actual *assertions*, on the other hand, they do not, on this view, have any of the special features we generally associate with first-person attitudinal avowals, but are just as indirect, and based on exactly the same kind of evidence as our judgements about other people's attitudes. The problem of how it is that we can have special, immediate, and authoritative knowledge of some of our own mental states is, on this view, just an illusion which arises from mistaking uses of 'I believe that p' as *expressions*, for uses of them as *assertions*.[36]

According to the second kind of artefact of grammar view alluded to above, there is actually nothing there to be explained about why our first-person avowals have the distinctive features they have, or nothing there to be said about that in virtue of which our avowals have these features; they just do. That is, it is just part of our practices with the words 'believe', 'desire', 'intend', 'pain', etc. that a person's immediate claims about her own states are taken as correct and authoritative, in all cases in which there are no strong overriding reasons for rejecting them.[37] That is, on this view, what someone

---

[35] This is sometimes taken to be Wittgenstein's position (Wittgenstein 1953), and is defended amongst others by Heal (1994). Wright (1998) however denies that this is actually Wittgenstein's view. Whether or not Wittgenstein actually held this position, however, is not important for the purposes of the present discussion. As far as this discussion goes, it only matters that this is one possible 'no-reasons' view, and one which is not without certain advantages.

[36] There are many ways in which this approach might be made to look more plausible, such as by saying, following Heal (1994), that these expressions of belief are not *only* expressions, but are at the same time to be taken as self-descriptions of oneself as satisfying certain behavioural criteria. See (Heal 1998, p. 21). However, whatever the details might be of any particular account along these lines, what I am interested in here is only the particular strategy that such accounts appeal to in order to explain the distinctive features exhibited by our non-inferential utterances or thoughts of the form 'I believe that p', and whether this strategy works.

[37] For a discussion of this position, which Wright calls the 'default view', see Wright (1998).

believes, desires, intends, feels, etc., is not to be *inferred* from her avowals (as one might infer from someone's screaming that they are in pain), but indeed in part to be *identified* by what this person (when sincere) claims to believe, desire, etc.

On the third type of view, held in particular by Burge, although solely with respect to strict cogito-like judgements, our judgements of the form 'I am hereby thinking that p' are immediate, non-inferential, first-person authoritative and immune to error simply in virtue of their self-verifying form. I cannot indeed be thinking to myself that I am hereby thinking that there are physical objects, without in fact thereby thinking to myself that there are physical objects.[38]

In other words, on one kind of strong constitutive view, it is constitutive of someone asserting non-inferentially that they believe that p, that they do actually believe that p, because asserting this is just another way of expressing their belief. On another such view, it is a basic unanalysable fact about our practices with the word 'believe' that if someone non-inferentially, sincerely, and with understanding asserts that they believe that p, then they do, in virtue of that very fact, count as believing that p. That is, uttering or being disposed to utter 'I believe that p' is constitutive of believing that p. On yet another strong constitutive approach, thinking a higher-order thought involves quite literally thinking the corresponding lower-order thought. These being the basic claims underlying the various types of strong constitutive accounts of self-knowledge, let us now turn to consider what some of the advantages might be of adopting one or other of these positions.

---

[38] See (Burge 1988)

3.1.1

(1) To begin with, strong constitutive views have the advantage of providing a straightforward account of why it is that our mental self-ascriptions (or at least some of them) exhibit the features of non-inferentiality, authoritativeness, a kind of transparency, etc. In fact if to assert that one believes that p is also in some sense to assert that p, or if to believe that one believes that p constitutes in one at the same time the belief that p, then obviously no inference from first to second-order belief is needed, nor can one ever come out as being wrong or ignorant about one's first-order attitudes, nor can any third-person judgement about the same states equal the authoritativeness of first-person self-ascriptions of them.

(2) Concerning the expressivist proposal more specifically, this view has the virtue of providing an appealing solution to one of Moore's paradoxes which other strategies might seem unable to deliver. That is, it has the virtue of providing an explanation of why one seems to contradict oneself when asserting things like 'I believe that p, but not p' although it is perfectly possible that one may believe that p and yet for it not to be the case that p, and moreover for there to be nothing wrong or contradictory about someone *else*'s judging this to be the case.[39] If the expressivist proposal is right in suggesting that judging 'I believe that p' is in some cases just an alternative way of

[39] See (Heal, 1994). Moore's other paradox concerns statements of the form 'p, but I do not believe that p'. This paradox, Heal grants, could be dealt with by appealing to the consciousness of our self-ascribed thoughts (where a thought's being 'conscious' is taken to consist in, or just to somehow involve, one's being aware of oneself having it). In this way, the utterance 'p', expressing a conscious belief that p, can be expanded into 'I believe that p', thereby generating the contradiction 'I believe that p, but I do not believe that p' which is of the basic form 'p, but not p'. Using this strategy to explain the second paradox 'I believe that p, but not p' however does not work. It only generates 'I believe that p, but I believe that not p' which is not itself a contradictory statement, but only an acknowledgement of the fact that one has contradictory beliefs.

asserting 'p', it becomes immediately clear why, in these cases, these Moorean utterances are contradictory: they amount to asserting 'p, but not p'.

(3) A closely related advantage of this kind of no-reasons view, is that taking our immediate attitudinal avowals of the form 'I believe that p' to be mere substitutes for assertions of the form 'p', fits well with the datum pointed out by Evans, drawing on a remark by Wittgenstein, that when asked whether we believe that p, what we do is not look at ourselves and consider the evidence regarding our beliefs, but rather, we look out at the world and consider whether or not p.[40] That is, if I am asked whether I believe that it is raining, I will not look at myself but out the window and consider whether it is or is not raining. And indeed, if, following the expressivist, to say 'I believe that it is raining' is roughly to say 'It is raining', nothing should seem more obvious than that the evidence appealed to in order to make this avowal should be evidence regarding the weather.

(4) One final virtue of artefact of grammar views (although there may well be others which I am overlooking) is that they fit well with the fact that when we sincerely and non-inferentially say things like 'I believe that this is the right thing to do', we generally seem to do so with a certain conviction and commitment to the view that this *is* indeed the right thing to do, which we do not do when saying things like 'Jones believes that this is the right thing to do'. On a strong constitutive approach, according to which asserting 'I believe that p' either constitutes in one the belief that p, or is just an expression of this belief, there is no difficulty in explaining this. In fact this point links up with the discussion in chapter 2 above, about our first-order conscious attitudes and our self-

[40] See (Evans 1982, p.225)

ascriptive judgements being held from the same cognitive perspective. Put in these terms, adopting a strong constitutive approach to self-knowledge whereby our first and second-order attitudes are not truly distinct attitudes, would again provide us with a simple account of how both attitudes can be held from the same point of view.

In brief, artefact of grammar approaches to avowals seem to have much to recommend themselves. However, having now listed a number of their virtues, it is time to re-examine these points with a more critical eye.

3.1.2

Concerning the first advantage of these views, it should be pointed out that the mere fact that the strong constitutive approach is able to accommodate the distinctive marks of first-person avowals is not enough to tip the balance in its favour. It only puts it on a par with all other approaches which are *also* able to provide an explanation of these distinctive marks.

Concerning the second advantage of this approach, namely that of being able to provide an explanation of why Moorean utterances of the form 'I believe that p, but not p' seem to be contradictory, we have here again only a negative advantage if it turns out that the contradictoriness of such Moorean utterances can also be generated *without* appealing to some constitutive link between second and first-order thoughts. Heal claims that it cannot, and is indeed lead to embracing an expressivist account of avowals essentially as a result of her attempt to solve this Moorean paradox.[41] It seems to me however, that there is another way in which this paradox could be dealt with, indeed a way

[41] See (Heal 1994)

which Heal herself briefly mentions but does not pursue in her paper.[42] The idea is that, given the datum that the evidence we appeal to in order to self-ascribe our conscious beliefs is not evidence about our beliefs but essentially evidence about the world, insofar as we self-ascribe a belief that p on the basis of evidence we have for p, then to say 'I believe that p, but not p' is in effect to be asserting 'not p' in spite of the fact that we are in possession of evidence for p, and have therefore immediate reason to assert 'p'. A certain contradiction would indeed be involved if one were to sincerely assert 'it is not raining' while looking out the window and clearly seeing that it *is* raining (assuming one has no reason to mistrust what one sees).

But now, one might ask, how exactly is this point supposed to count against the artefact of grammar approach to avowals? In fact another virtue of such accounts (ie. point (3) above) was precisely that they fitted well with Evans's datum about what evidence we consult when considering what we currently believe. It was in fact suggested earlier that this datum seemed to support, rather than count against, the view that our first-person attitudinal avowals of the form 'I believe that p' are just a different way of asserting 'p'. Now although this is true, this fact cannot itself decide things one way or another regarding whether artefact of grammar views are right or not, since, it is not clear that this is the only approach to avowals which is supported by Evans's datum. In fact, there seems to be a possible intermediate position between a perceptual account and a no-reasons view,[43] according to which our self-ascriptions of our occurrent conscious attitudes are ontologically distinct from the thoughts self-ascribed, and yet

[42] (Ibid, p.19)

[43] I will be discussing this position in some detail in chapter 4 below.

according to which the former are rationally *based* on the latter, and therefore according to which looking at the world will also come out as being the right way to go about making a correct self-ascription, since by looking at the world one will come to form a conscious belief about the world, which will in turn constitute an immediate reason for self-ascribing it.

In other words, it seems that neither the fact that artefact of grammar views can explain the contradictoriness of Moorean assertions of the form 'I believe that p, but not p', nor the fact that they also fit the datum about mental self-ascriptions discussed by Evans, can decide the issue between an artefact of grammar strong constitutive non-reason based approach, and an intermediate *reason-based* one.

Finally, concerning point (4), the fact that asserting non-inferentially that one believes that p tends to involve a certain commitment to the belief self-ascribed, does not need to be explained be reference to any constitutive principle. In fact, if the belief that p, which we are self-ascribing, is indeed *our* belief, then of course we will be committed to the view that p when we assert that this is what we believe, without this commitment having to be constitutive of our self-ascriptive judgement. To sum up, none of the virtues of artefact of grammar accounts of avowals seem to be exclusive to them. We must therefore look elsewhere in order to decide between this strong constitutive approach and its alternatives. In particular we need to consider the fundamental constitutive claims themselves, and look at whether they can plausibly be maintained.

3.1.3

To start with the expressivist thesis, one simple consideration seems to

count decisively against it: we are able to know immediately what thoughts are going through our mind even if we are silent, in a dark room, and, say, tied up and thus unable to move. This is a consideration which, we have seen, also counts against Rylean views of self-knowledge according to which we know our own thoughts in no different way than we know the thoughts of others.[44] The expressivist proposal, however, might have seemed to be an improvement on the Rylean position, in that, as we have seen, it is actually able to accommodate the distinctiveness of our mental self-ascriptions in certain cases, namely in those cases in which these self-ascriptions are supposedly being used as mere alternative ways of making statements about the world. But if self-knowledge only has the distinctive features of immediacy, non-inferentiality and special authority in cases where this so called 'self-knowledge' is not actually *knowledge*, but a mere *expression* of our occurrent thoughts, the awareness we are able to have of thoughts of ours which we are not actually *expressing* still remains entirely unexplained.

So much, therefore, for the expressivist proposal. But what about this whole general idea that there is an ontologically constitutive link between making an attitudinal avowal and having the corresponding first-order attitude? The problem with this view more generally, is that it does not seem to leave any room for the phenomena discussed above of error and self-deception about what we believe, desire, etc., that is, for the fact that first and second-order attitudes can, and often do, come apart: we can sometimes have second-order beliefs without having the corresponding first-order beliefs.[45]

---

[44] See chapter 2 above, pp.19-20

[45] This objection is raised amongst others by Boghossian (1989) and Martin (1998).

One could of course reply to this objection by arguing, following Bilgrami for instance, that cases of error and self-deception, are not actually cases where we believe that we have a certain attitude which we do not in fact have, but rather, these are cases where we happen to have both the attitude self-ascribed *and* an attitude which is inconsistent with it.[46] This line of defence of the constitutive thesis however seems hopeless. To say that the self-deceived son in Bilgrami's example, who behaves *consistently* and *only* contemptuously towards his father (the only evidence of his admiring him being his asserting that he does), believes both that his father *is* a fine person and that his father is *not* a fine person, is just, it seems, either to deny the obvious (ie. the possibility of actual self-deception), or simply to reiterate in different terms that he believes that he believes his father is a fine person, but in fact does not believe that his father is a fine person, and hence is self-deceived.

Another possible line of defence of the strong constitutive thesis is that of arguing that it is only constitutive of our self-ascriptions of our *conscious* attitudes, that we actually have the attitudes self-ascribed, thereby leaving room for the possibility of being mistaken about a wide range of other beliefs, desires and other attitudes of ours. The problem with this approach however, is that it seems to force us to say that it is ontologically constitutive of only *some* second-order conscious beliefs of ours that we have the first-order beliefs they are about, whereas it is not ontologically constitutive of other such beliefs with the same content.[47] That is, for example, we end up having to say that in some instances, it is constitutive of my believing that I believe that it is raining that

---

[46] See (Bilgrami 1998, in particular p.218)

[47] See (Martin 1998)

I actually do believe that it is raining, and in other instances it is not. But now given that there seems to be no difference in the nature of the *second*-order beliefs in these two kinds of cases (other than the ad hoc difference that some are infallibly correct in virtue of some constitutive principle which does not happen apply to the others), and given that the whole difference was supposed to hang on the (conscious or unconscious) nature of the first-order states self-ascribed, a more sensible approach at this point would be to say that first and second-order states are actually *distinct* states, and to try to see what it might be about a first-order state's being *conscious* that connects it particularly intimately to second-order beliefs about it. This, however, is almost by definition not a line that defenders of a strong constitutive theory of avowals would want to take. Not doing so however, at this point, just seems to be to insist without argument that the constitutive thesis is right (perhaps because no better account seems to be at hand)[48], and in effect just to say that our second-order judgements about our mental states are infallibly correct in all cases (ie. those in which our second-order states constitute in us at the same time the states self-ascribed) except those in which they are *not* infallibly correct (ie. those in which having a second-order state does *not* constitute in us having the states self-ascribed).

This whole problem in the end links back to an issue raised in chapter 1 above about the sense in which our self-ascriptive judgements can be taken to be authoritative. It is, it seems, a fundamental mistake of artefact of grammar approaches in general, to take the problem of introspective self-knowledge to be a problem about a

[48] In fact, if it turns out that none of the more explanatory theoretical options work, we may have to just admit defeat and resort to a quietist position of just describing how things are, and admitting that there is nothing more illuminating to be said. As Wright points out however, this view, which he calls the 'default view', is indeed a *default* view, that is, a view we can only be satisfied with if it is shown that no better or more illuminating account of avowals can be given. See (Wright 1998, especially pp.44-45)

certain kind of *statements*, namely 'avowals', or about knowledge with a certain *subject matter*, ie. 'mental knowledge'.[49] The basic idea behind these views is that judgements about mental states are problematic because we take them to be authoritative, incorrigible, we do not ask people to defend their avowals, etc. However, it is unclear that statements about our mental states taken *as such*, actually *are* problematic, since in many cases (eg. when these statements are about unconscious attitudes which we have come to know inferentially on the basis of behavioural evidence) there is no difficulty in explaining how they are possible. It is only in certain *specific cases* that our attitudinal avowals are problematic in that in *these cases* they are non-inferred, authoritative, and it does not make sense to ask one to defend them. The distinctiveness of these judgements in certain cases, given that in other cases these same judgements are *not* distinctive, must therefore be a matter of *how*, in some cases and not in others, we are able to make them non-inferentially and authoritatively. Insofar then as the question of *competence* arises, the possibility of authoritative self-knowledge raises traditional epistemological issues which cannot be dealt with through a purely metaphysical account.

At best, a strong constitutive approach to attitudinal avowals might work in the case of strict cogito-like judgements. Indeed, insofar as these judgements are self-verifying, their being immediate, non-inferred and authoritative can be taken to be a feature of these judgements *as such*, rather than of these judgements as reached in a certain way. However, following Boghossian,[50] it seems clear that such an account is not sufficiently general. It takes us nowhere nearer understanding the immediacy,

---

[49] See in particular the way in which the problem of self-knowledge is set up by Wright (1998), as a problem about 'avowals', rather than as a problem about the way in which some of them are reached.

[50] (Boghossian 1989)

authoritativeness and immunity to non-cognitive error of our knowledge of other attitudes of ours which are either not strictly contained in our self-ascriptions of them, or not strictly simultaneous with these self-asrciptions. If I was just a moment ago thinking to myself 'It is cold in here' I am immediately able to say that just a moment ago I was thinking this. Or, if I non-inferentially judge 'I fancy a cup of coffee', my desire for a cup of coffee is in no way contained in this judgement. I could be lying, or indeed I could be wrong. It may in fact turn out that once I get the coffee, I realize that what I actually wanted was a beer, but was deceiving myself about it.

In other words, adopting a strong constitutive approach might be the right way of dealing with *some* cases of self-knowledge, namely strict cogito-like cases, just as adopting an inferential account of self-knowledge may well be the right way of dealing with other cases, namely cases of knowledge of our unconscious thoughts, and just as adopting a perceptual model of self-knowledge might be the right way of dealing with yet other cases, namely cases, were they to exist, of telepathic knowledge of some of our thoughts. The problem however, is that none of these approaches seems able to deal with the most common and most problematic cases of self-knowledge, namely those where we are able to know what we are currently consciously thinking authoritatively, without having to infer this knowledge from anything, and yet in a way which is not infallible, but nonetheless not subject to certain kinds of error. The problem of introspective self-knowledge therefore still remains. Let us therefore consider a slightly different strategy, namely the weak constitutive no-reasons approach.

3.2 The Weak Constitutive View

On this view, which may also be referred to as the 'weak special access functionalist theory',[51] it is thought to be conceptually (although not ontologically) constitutive of having first-order states that one will generally tend to form second-order beliefs about them when one considers the matter. In other words, the idea is that, following a functionalist theory of mind, according to which mental states are to be individuated by way of their functional role, that is, by reference to their relations to other mental states and to behaviour, self-ascriptive higher-order beliefs are amongst the mental states a lower-order state's relations to which are constitutive of it. For a state to be a belief, for instance, it must first of all tend to give rise to other beliefs, and tend to combine with desires and other attitudes to give rise to yet other attitudes and actions. The claim about self-knowledge is then that first-order attitudes have not only conceptually constitutive dispositional links to other first-order attitudes and behaviour, but also to second-order beliefs about themselves. That is, for one to qualify as having, say, a certain first-order belief, one must not only be disposed to form other appropriately related first-order beliefs and other attitudes, but also, in certain circumstances, to non-inferentially self-ascribe this belief itself, that is, to form a *second-order* belief about it.

Now at first, this idea might seem to make a lot of sense, given that we would not generally be prepared to ascribe to someone, without thorough consideration of any overriding evidence, a mental state which they themselves did not believe they had. In addition, this kind of constitutive view has the advantage over strong constitutive views of leaving plenty of room for the possibility of error and self-deception, given its

[51] See (Fricker 1998)

commitment to the ontological distinctness of first and second-order states, combined with the thought that it is constitutive of first-order states only that they tend *normally* (in circumstances of full rationality, reflectiveness, etc.) to give rise to second-order beliefs about themselves. There therefore seems to be something very appealing about this approach to self-knowledge. However, two features of it seem worthy of notice:

Firstly, the supposedly constitutive feature of mental states that they will tend to give rise to second-order beliefs about themselves, needs to be relativised to *conscious* mental states, since many psychological states of ours *never* give rise to non-inferential self-ascriptions of them (eg. our unconscious states), and it can therefore surely not be conceptually constitutive of *these* states, namely the unconscious ones, that they be dispositionally linked to second-order beliefs about them, although we would certainly still want to count them as genuine *beliefs* and *desires*. In other words, the weak constitutive functionalist thesis can only apply to *conscious* states, that is, it can only be conceptually constitutive of having a *conscious* state that one will tend to form a second-order belief about it. The basic idea of this view therefore has to be that it is not conceptually constitutive of having mental states that they be dispositionally linked to second-order beliefs simply in virtue of their being *beliefs* or *desires*, but essentially in virtue of their being *conscious* beliefs and desires. But now this might tempt one to ask what it is about a state's being conscious (as opposed to *un*conscious) that makes it the case that if one has it, one will tend not only to form other first-order states with appropriately related contents, but also to self-ascribe it? The importance of this question will become clearer as we proceed.

A second, and not unrelated point is that when we consider the nature of

the first-order states to which a certain state's relations are conceptually constitutive of it, it is striking to notice that these are all states which have relevantly similar contents to the one being individuated. That is, on most functionalist theories, the conceptually constitutive functional role of, say, a belief that p, is it's tending to combine with other beliefs and desires with relevantly similar contents, to give rise to yet other beliefs and actions with relevantly similar or appropriately related contents. It is, for instance, part of the functional role constitutive of my believing that there is ice-cream in the refrigerator, that this belief will tend to combine with my desire for ice-cream, to give rise to an action of going to the refrigerator and getting the ice-cream. In other words, a functionalist individuation of mental states, on a theory according to which it is *conceptually* constitutive of mental states *qua* mental (and not, say, *qua* physical) that they be related to other mental states and behaviour, seems to be in effect an individuation of them by reference to what states are good reasons for having them, and what states they themselves are good reasons for having.[52] Thus, the idea of mental states having *constitutive* dispositional links to other mental states and to behaviour, is the idea that mental states are to be individuated in terms of what are typically good *reasons* for having them, or what attitudes they themselves are typically good reasons for having, and, these reason-giving relations can be made sense of, it seems, by reference to certain appropriate

[52] I will discuss the notion of a reason-giving transition more fully in chapter 4. For present purposes, however, the idea is, briefly, that the notion of a reason-giving transition between two states in these cases can be understood roughly by reference to logical relations, or to some appropriate overlap, holding between their contents. For example, my believing that there is ice-cream in the refrigerator combined with my desire for ice-cream will *not* constitute an immediate reason for my starting to read a book on self-knowledge, although it *will* constitute a reason for my going to the refrigerator and getting the ice-cream. Similarly, my believing that p, combined with my believing that q, will *not* constitute an immediate reason for my believing that z, although it may well constitute an immediate reason for my believing that (p and q).

relations holding between their contents.

But now the problem with the weak constitutive functionalist view of self-knowledge is the following: if the above is what it is for it to be conceptually constitutive of a mental state to be dispositionally related to other mental states and behaviour, and given the unrelatedness of first and second-order contents,[53] how are we to make sense of the thesis that it is conceptually constitutive of having a first-order attitude that it will tend to give rise to a second-order belief about it, or at least how are we to make sense of the idea, lying at the very basis of the weak constitutive theory of self-knowledge, that this constitutive link is on a par with the constitutive link our first-order states bear to other first-order states and behaviour? There are, it seems, two ways in which one could go from here:

(1) One could try to explain how a first-order conscious state can stand in the same kind of relation to a second-order state as it does to other first-order states in relation to which it is conceptually individuated (ie. a rational relation), and this could be done perhaps by explaining what it is about a state's being *conscious* that allows it to stand in such a relation. That is, one could try to explain how a first-order conscious state can stand in a *rational* or *reason-giving* relation to a second-order state, in spite of the unrelatedness of first and second-order contents. To do this however would be to drift away from a no-reasons view.

(2) One can hold onto a no-reasons view, and argue that, although the conceptually constitutive connection a first-order state bares to other first-order states happens to be associated with *reason-giving* relations holding between them, this just

---

[53] First-order states are about the *world*, while second-order state's are about *mental states*.

simply happens not to be the case of the constitutive connection between a first-order state and a self-ascription of it. If so, however, that is, if our first and second-order states are both ontologically and rationally unrelated, then it is not clear how the relation between them can be *conceptually* constitutive of having them. That is, it cannot be constitutive of a first-order state *qua mental state* that it will tend to give rise to a second-order belief about it, but only perhaps constitutive of its physical realization in the brain.

But now if this is what the view amounts to, it ceases to be clear how it is any better than an entirely *non-constitutive* purely reliabilist view of self-knowledge, given that it will end up being just as incapable as non-constitutive theories of explaining, amongst other things, why we intuitively feel that it is a matter of *conceptual* necessity that one cannot have a conscious state without being disposed to self-ascribe it, or of explaining why Moorean utterances of the form 'p, but I do not believe that p' are *conceptually* odd, and indeed *contradictory*. In other words, if this (ie. (2)) is what the weak constitutive view amounts to, it is no more plausible than a purely reliabilist account of self-knowledge. If on the other hand it amounts to more than this (ie. (1)), then it is no longer a no-reasons view.

* * *

At the end of chapter 2, we came to the conclusion that introspective self-knowledge cannot not be based on any of the normal ways we have of gaining knowledge, namely by inference or through direct observation, and yet in this chapter, we have seen that the issue of the distinctiveness of introspective self-knowledge cannot be a purely

metaphysical issue regarding a certain class of judgements, as suggested by strong constitutive accounts. A certain version of the weak constitutive approach, however, appears to be somewhat more promising, in that it seems less subject to the problems faced by its stronger counterparts. However, this turns out to be so essentially in virtue of its not actually being a *non reason-based* view after all, but a view implicit in which is the idea that our mental self-ascriptions are based on reasons, not however on inference or observation, but directly on the mental states self-ascribed. In other words, in spite of the failure of the three standard options of inference, observation or nothing, we seem to have here a possible intermediate position between the second and the third. This position, however, seems to immediately gives rise to a further problem, which any satisfactory account along these lines is going to have to resolve, namely that of explaining how having a first-order conscious belief or other attitude can in itself constitute a reason for self-ascribing it. To put things differently, we are now faced with the following question: how can having a belief about the *world*, constitute an immediate reason for believing something about one's *beliefs*, given that nothing about what one believes, follows from how things are in the world?

I now turn to examine this question more closely, and consider how successful recent attempts made to uphold this intermediate reason-based approach to self-knowledge, are at generating an answer it.

Chapter 4: Our Grounds for Self-Knowledge

There are, it seems, two possible questions of justification that one could ask: (1) the question of what makes a *proposition* justified, and (2) the question of what makes someone's *belief* in a proposition justified. The two are not unrelated.

A *proposition* that p can be said to be justified, roughly, when there is (impersonally speaking) reason to believe it. Someone's *belief* that p, on the other hand, can be said to be justified when *this person* has reasons or grounds for believing it. More precisely, answering a question of the form 'why is the proposition that p justified?' involves providing an *argument* for p, that is, providing a story which either entails the truth of p, or makes it probable that p. For example, an answer to the question of why the proposition that Mary will be in college tomorrow is justified, would involve mentioning, for instance, (1) that Mary goes to college almost every day, (2) that she said she was going to be in college tomorrow, (3) that she usually does what she says, etc.

On the other hand, answering a question of the form 'why is S's *belief* that p justified?' would involve considering why the proposition that p is justified (ie. answering the first question), and then restricting one's answer only to those facts to which the believer has access. To apply this to our example, an answer to the question 'Why is John's belief that Mary will be in college tomorrow justified', would involve mentioning, for instance, that John heard her say that she would be in college tomorrow

and that he believes that she usually does what she says. If, however, John was neither aware of Mary's daily habits, nor heard her speak of her plans for the following day, or *did* hear her but did not take her to be trustworthy, he would not count as justified in his belief, although in this case the *proposition* that Mary will be in college tomorrow would still be justified.

In other words, on a plausible internalist conception of justification, a belief or a perceptual experience will count as a reason or a rational ground for believing that p, if the content of this belief or experience figures as a relevant premise in a possible argument which either entails or makes probable that p, that is, an argument from which it follows either deductively or inductively that p. [54]

Taking this general understanding of the notion of what it is for one belief to be a reason for another as a starting point,[55] the explanatory problem faced by the approach to self-knowledge according to which there is a direct reason-giving relation between our first and second-order conscious states becomes apparent: given that from facts about how the world is nothing follows about what one is going to believe, how can

---

[54] I am taking it here that perceptual experiences, and not just beliefs, can constitute rational grounds for beliefs, although of course they do not constitute *inferential* grounds. It is sometimes held, however, that beliefs can only be justified by other beliefs, because perceptual experiences are not subject to revision; they are not something we are responsible for; and so it would not make sense to say that one *ought* to believe that p, if one perceives that p. This, however, does not seem to me to be entirely right, given that we would actually judge someone to be *irrational* if they saw that it was raining but did not believe that it was (assuming of course they had no reason to mistrust their senses). Moreover, given the possibility of disbelief in one's senses, a perception can surely be taken as an evidential ground, or as reason amongst others, for believing something. In any case, this would fit with the above model in that perceptions are a kind of 'access' that a believer might have to a justifying fact.

[55] There are of course other possible accounts of what constitutes a rational relation, in particular reliabilist or other externalist accounts of justification. These, however, seem to me to go clearly against the phenomenology of belief formation. Generally, in fact, our conscious beliefs make rational sense to us from our own *personal level* point of view; we do not just find ourselves, as Evans would put it 'with a yen to apply some concept' (1982, p.229).

having a conscious belief about the world constitute an immediate rational ground for believing something about one's beliefs? To illustrate this, we have the following:

John heard that Mary said that she will be in college tomorrow

John believes that Mary usually does what she says

John believes that Mary will be in college tomorrow

Here, John's reasons for his belief that Mary will be in college tomorrow, correspond to the premises of the argument on the right hand side, which make it probable that p, or from which it follows (although not strictly) that p.

Contrast this with the case of the self-ascription of a belief:

John believes that Mary usually does what she says

John believes that John believes that Mary usually does what she says

Now, clearly, the conclusion of the argument on the right hand side in no way follows from the premise, so how can John's belief in the premise constitute an immediate rational ground for his belief in the conclusion?

There are two ways in which one could respond to this problem: (a) one could take it to constitute an actual *problem* for the 'intermediate' reason-based position,[56] and hence a reason to reject it, or (b) one could take it as an explanatory *challenge* which must and can be answered.

[56] 'Intermediate' as in intermediate between a perceptual model and a no-reasons account; ie the view that our first and second-order states stand in a direct reason-giving relation to each other.

In what follows, I will be examining two intermediate reason-based accounts of self-knowledge and of our entitlement to it, namely Burge's and Peacocke's.[57] Neither, I will argue, is able to provide a satisfactory answer to this question, and yet, a close examination of their shortcomings will, I will suggest, first of all show that this question *must* be answered, and secondly, reveal the only way in which it *could* be answered.

More specifically, I will first consider Burge's account of our entitlement to self-knowledge and suggest that, although it is ultimately unsatisfactory given the present explanatory aim, it can be used to generate a *reductio ad absurdum* of the view that our introspective higher-order states might not be directly and rationally related to our first-order conscious states. The intermediate reason-based position, I will suggest, therefore *must* be adopted, thereby making approach (a) no longer viable. Approach (b) will therefore have to be taken, and the question of how a conscious first-order attitude can constitute an immediate reason for self-ascribing it will have to be addressed. Secondly, I will consider Peacocke's account, and argue that it ultimately faces the same predicament as Burge's, although its limitations end up narrowing down the options to the point of revealing the only possible way in which our question *could* be answered, and thereby the only way in which the possibility of immediate authoritative introspective self-knowledge can be accounted for. Let us begin with Burge's account of our entitlement to self-knowledge.

---

[57] Unless otherwise stated, all references to Burge in this chapter will be to (Burge 1996). For Peacocke see (Peacocke 1992; 1996; 1998)

4.1 Burge

According to Burge, the following claims hold:

(1) We have, and are entitled to, a distinctive kind of non-perceptual, epistemically special self-knowledge (by which he means, in the end, knowledge had from and about the same cognitive perspective - understood in the sense defined in chapter 2 - and therefore knowledge which stands in an immediate rational relation to that which it is about) .

(2) This entitlement to this special kind of self-knowledge arises essentially from the role of this knowledge in critical reasoning.

In support of the first claim, Burge puts forward a transcendental argument to the effect that critical reasoning requires being entitled to a special kind of self-knowledge, and, since we are critical reasoners, it follows that we must be so entitled. More specifically, the argument runs as follows:

(1) We are able to reason critically. That is, we are able to reason in the fully reflective way spelt out in chapter 2 above.[58]

(2) Critical reasoning requires the following:

i. that we be epistemically entitled to certain judgements about our attitudes and reasons

ii. that these judgements constitute knowledge (ie. that they be normally true and not just accidentally so)

iii. that this knowledge be distinctive in being directly rationally related to the attitudes it is about.

---

[58] See chapter 2, pp.37-38

(3) Therefore, by (1), we must be entitled to this distinctive kind of self-knowledge.

Concerning (i) it is clear that we could not possibly reason critically if we were not entitled to our judgements about our attitudes, that is, if we were not being reasonable in making them. In fact, if we were not reasonable in our reflective judgements about our attitudes, we could not be reasonable in our conclusions derived from reasoning based on these judgements, nor would we be reasonable in our reviews of our attitudes based on these conclusions. Critical reasoning would therefore not be possible, given its nature as reasoning which involves *reasonable* confirmation, change, or supplementation of our attitudes based on our reflection on these attitudes. If we were not entitled to our second-order beliefs, this would mean that reviewing or confirming our attitudes on the basis of rational reflection on these attitudes and on our reasons for holding them, would not actually be a reasonable enterprise to engage in, and so, insofar as we are rational, we would not engage in it.

Concerning (ii), the basic idea is, once again, that if these judgements to which we are entitled did not constitute knowledge, then genuine critical reasoning would not be possible. In fact, we engage in critical reasoning essentially for the purpose of arriving, through reflection on our attitudes, at more reasonable beliefs and at a rationally more coherent set of attitudes. That is, the point of reasoning critically is to *control* or *guide* our beliefs and other attitudes in such a way as to further their reasonability and rational coherence. If, however, our second-order judgements were either never true or only accidentally so, then changing our attitudes on the basis of conclusions derived from reasoning based on these judgements, would not be a case of rationally *controlling* or *guiding* our attitudes, even if it somehow accidentally resulted in the promotion of their

reasonability and rational coherence.

Finally, concerning point (iii), as already suggested in chapter 2 by way of an example, critical reasoning seems to require that the knowledge we use in reasoning critically be had from and about the same cognitive perspective, and must therefore be *directly* and *rationally* related to the attitudes it is about. In other words, the idea is that our first-order conscious thoughts must constitute *immediate reasons* for judging that we have them, and similarly, our second-order evaluative judgements about these first-order thoughts (eg. a judgement that a first-order thought of ours is unreasonable), must constitute *immediate reasons* for confirming, reviewing or supplementing them.

At this point, however, given the problem that this view of the relation between our first and second-order thoughts has given rise to, one might be sceptical about just going along with this argument. In fact, the only kind of argument which could give us a real reason for accepting, rather than rejecting, this intermediate reason-based approach to self-knowledge in light of the clear problem it gives rise to, would be one which could show us that critical reasoning would be *impossible* if it were done from and about a different cognitive perspective, and, insofar as we clearly *are* able to reason critically,[59] our conscious attitudes and our introspective knowledge of them *must* be held from and about the same cognitive perspective, and so *must* stand in immediate reason-giving relations to each other.

It seems to me that such an argument can indeed be provided, namely the following *reductio ad absurdum* of the view that the immediate self-ascriptive judgements

---

[59] I am, for instance, critically reasoning right now in considering and evaluating my beliefs on self-knowledge, with the aim of coming, through this process of reasoning, at more reasonable beliefs on this topic.

we use in critical reasoning could be held from a different cognitive perspective from that of the attitudes they are about.

Let us start by assuming, for the sake of argument, that the self-ascriptive judgements we use in critical reasoning are made from a *different* perspective from the perspective of the attitudes reasoned about, and that they are therefore entirely independent from these attitudes, to which they are related only by a reliable, but purely contingent, non-rational, causal mechanism. Let us then ask whether on these assumptions genuine critical reasoning, as defined in chapter 2 above, is possible.

For the sake of clarity, let us consider a case in which the relevant dissociation of perspectives, is a dissociation between the perspectives of two different people. In fact, let us imagine that my beliefs about your attitudes and your attitudes are linked by some reliable causal mechanism, and, moreover, that I decide to engage in critical reasoning about your attitudes, and begin to reflect on them and on their reasonability, and that I thereby come to the conclusion that you are not being reasonable in one of your beliefs. What would happen next? The process of critical reasoning would seem to be blocked at this stage. The problem is that in critical reasoning, the conclusions arrived at from reasoning based on reflection on one's attitudes, should *immediately* and *rationally* result in an explicit reasonable change or confirmation of the attitudes reasoned about. So, if I were genuinely reasoning critically, then my coming to the conclusion that you were being unreasonable in one of your beliefs, should make it immediately rational for you to change your beliefs, which it does not seem to do. In order for my process of reflection on your attitudes to result in your explicitly and reasonably changing your beliefs, this process would somehow first have to result in your coming *yourself* to believe

that a belief of yours is unreasonable. This could happen indirectly, for instance, if I were to somehow succeed in convincing you of the unreasonableness of your belief by making you go yourself through the reflection I just went through about your attitudes, in which case, however, the reasonable review of your attitudes would end up ultimately resulting from *your* reflection on your own attitudes, and not directly from mine. Alternatively, a *direct* way in which you could come to hold the belief that a certain belief of yours was unreasonable, as a result of my reflection on your attitudes, would be via some direct (contingent, non-rational) causal mechanism linking the conclusions of my reflection to your beliefs. The problem with this, however, would be that this causal mechanism could not be the same kind of mechanism by which my second-order beliefs are related to your first-order beliefs since such a mechanism would only entitle you to the belief that *I* believe that you are being unreasonable in one of your beliefs, but not to the belief that you actually *are* being unreasonable. It would therefore again not immediately follow that there would be reason for you to change your belief. The relevant kind of mechanism would therefore have to be one whereby whatever conclusion I arrived at from reasoning based on my judgements about your attitudes, would immediately, and non-rationally, cause you to hold this *same* belief. Of course now the problem would be that critical reasoning would *still* be impossible, given that such reasoning involves review or confirmation of attitudes on the basis of conclusions derived *rationally* from reasoning about one's attitudes, whereas in this case, although *my* conclusion that a certain belief of yours is unreasonable would be directly and rationally derived from reasoning about your attitudes, *your* belief that a belief of yours is unreasonable would *not* be directly rationally derived from any such reasoning. For critical reasoning to be possible, in other words, all the reflective

beliefs about your attitudes as well as the reasoning about them and the conclusions thereby reached would have to be ultimately held by *you*, that is, by the same person whose attitudes are being reasoned about. And, to return to the case of single individuals, this means that in order to reason critically, all our reflective second-order beliefs about our attitudes must be held from the same rational, cognitive perspective as our first order attitudes.

Our nature as critical reasoners thus requires that our introspective knowledge of our own occurrent thoughts be had from and about the same cognitive perspective, and, by definition of a cognitive perspective,[60] that there must be a *direct rational* relation between our first and second order states, in addition to whatever underlying causal relation may or may not also hold between them. In other words, if our introspective self-knowledge were not of this kind, we could *know* our own thoughts and *reason* about them, but could not thereby immediately rationally influence them, just as, if we had telepathic knowledge of someone else's thoughts, this would not give us any kind of immediate rational control over them. But, since in reasoning critically our reflective judgements *do* immediately rationally influence our first order thoughts, these judgements must be of this distinctive kind. Any judgements about our attitudes which are not based in this direct rational way (such as, for instance, judgements about our unconscious attitudes), could therefore not be part of a genuine process of critical reasoning. Since, however, the immediate introspective, authoritative judgements we ordinarily make about our occurrent conscious attitudes are the very judgements we use in critical reasoning, these judgements must be of this distinctive kind, which however, just

[60] See chapter 2, p.34 above.

leads us straight back to our initial problem about how this can be, the only difference now being that the problem is even more pressing, since we can no longer respond to it by rejecting the approach to self-knowledge which generates it. The only way forward is thus to try to provide a head-on answer to it. Can such an answer be extracted from Burge's account of our entitlement to self-knowledge?

According to Burge, our entitlement to this special kind of self-knowledge is to be found essentially in our nature as critical reasoners, that is, the role of this knowledge in critical reasoning is supposed to be the very *source* of our entitlement. However, in reading Burge's (1996) paper, one might find that nothing beyond the earlier mentioned transcendental argument, is put forward in support of the claim that the role of our introspective judgements in critical reasoning is the *source* of our entitlement to them, and yet, all this argument seems to show is that critical reasoning *presupposes* that we are entitled to a special kind of self-knowledge, but not *why* we are so entitled, or how it is that we can be so entitled. One might therefore argue, following Peacocke,[61] that Burge's account gets things the wrong way around. In fact, if critical reasoning requires being entitled to self-knowledge, then one must first be entitled to self-knowledge if one is to be able to reason critically. Critical reasoning can therefore not possibly be the *source* of our entitlement but only a consequence of it.

Now, in a sense, it seems that Peacocke is right in his objection to Burge. To be fair to Burge, however, it is not clear that he is speaking of 'entitlement' or 'source' of entitlement in the same sense as Peacocke. In fact, by way of an example, we can distinguish the following two notions of entitlement:

---

[61] (Peacocke 1996)

(1) The sense in which I may, for instance, be entitled to use the concept of a cause (as opposed to just that of constant conjunction) because without it, following Kant's second analogy,[62] objective experience would be impossible. The very existence of experience itself therefore warrants me in my use of it. This, however, says nothing about

(2) what *grounds* I have on a particular occasion for saying that A caused B. What was the basis for my judgement? Did I infer if from some information I already had? Was my judgement based on what I saw? Was it purely a guess? etc. In this sense, I am entitled to a judgement if it is based on, or grounded in, a good reason.

In the case of our entitlement to self-knowledge, Burge seems to have in mind sense (1), whereas Peacocke, in his objection to Burge, has in mind sense (2). That is, for Burge, our nature as critical reasoners warrants or justifies us in our immediate, non-inferential, authoritative, directly reason-based self-ascriptions, in that without them critical reasoning would not be possible. In this sense, the role of our immediate introspective judgements in critical reasoning can indeed be thought of as the *source* our entitlement to them. Burge is not concerned with what entitles us to self-knowledge in any other sense than this. But, one might ask, why is he not? *Should* he be concerned with explaining what our immediate judgements about our occurrent conscious attitudes are based on? Perhaps his thought is that there is nothing illuminating there to be said about what entitles us to self-knowledge in this sense. Our judgements are just immediate; they are not based on any evidence.

As pointed out earlier, however,[63] a judgement can be immediate without

---

[62] (Kant 1929)

[63] See, chapter 1, p.7 above

being rationally ungrounded. Moreover, given Burge's account of critical reasoning, it looks like our second-order judgements *are* based on something, namely our first-order thoughts, that is, it would seem, on evidence about the world. In fact, as seen above, for critical reasoning to be possible, our first-order thoughts must constitute *immediate reasons* for judging that we have them, and likewise our second-order evaluative judgements must constitute *immediate reasons* for reviewing or confirming our first-order attitudes. In other words, Burge's account does seem to suggest that our introspective second-order beliefs are based on reasons, that these reasons are the very first-order attitudes they are about, and, moreover, that these first-order attitudes must count as *good* reasons for our second-order judgements, if critical reasoning is to be a reasonable activity to engage in. Burge's account thus implies that we have a special kind of entitlement to our mental self-ascriptions in sense (2) and not just in sense (1), but it does not *explain* this entitlement. In fact, his account just leads us more forcefully back to the question set out at the beginning of this chapter, namely that of how our conscious *object* -oriented thoughts can possibly constitute immediate reasons for our *thought*-oriented thoughts. That is, how can it be directly rational to move from a conscious thought about the world, to a thought about oneself and one's mental states?

Burge does not address this question, nor does he go into the issue of the rational grounding of our immediate introspective judgements at all, because, perhaps, he takes it that nothing illuminating can be said about how there can be a direct reason-giving relation between our first and second-order conscious thoughts; all that can be said is that it must be so, because if it were not, we would not be able to reason critically, and it so happens that we *are* able to reason critically.

I, however, believe that more can, and indeed *needs* to be said, or at least that any option in this direction must be fully explored before resorting to the not obviously coherent idea that the relation between our first and second-order thoughts must necessarily be thought of as *rational*, and yet in a way which bares no resemblance to the way in which other transitions between intentional states are rational. Perhaps Peacocke has the answer. His account certainly attempts to take things further in the suggested direction than Burge's. Let us therefore turn to examine it in more detail.

4.2 Peacocke

According to Peacocke, it is inscribed in the very possession conditions for the concept of belief that anyone who possesses this concept will find it primitively compelling to judge that they believe that p, whenever they have an occurrent conscious belief that p (and consider the matter), and they will find judging so primitively compelling *because* they have this belief, that is, for the very *reason* that they consciously believe that p. In fact, on this view, one will not count as possessing the concept of belief *unless* one has, in appropriate circumstances, a tendency to make such 'consciously-based self-ascriptions'.[64] In other words, on this view, someone who possesses the concept of belief will make immediate judgements about their occurrent conscious thoughts *directly* and *rationally* on the basis of these very thoughts themselves. Moreover, these judgements will constitute *knowledge*, that is, *warranted* true beliefs (and not just true beliefs), because, when made for the very reason that one does actually currently have the self-ascribed

[64] See (Peacocke 1992; 1998)

attitude, they will be bound to be true. This warrant, or entitlement to self-knowledge, has its source in the very conditions of what it is to possess the concept of belief, according to Peacocke, since these conditions are such that anyone who possesses the concept of belief will normally make self-ascriptions in the direct knowledge-yielding way outlined above, that is, on the basis of the very conscious beliefs self-ascribed, ie. only in circumstances in which they actually do have these beliefs. Given this account, the next question is: how does it compare to Burge's?

In the last section we saw how Burge's transcendental argument from critical reasoning shows that we must be entitled to immediate judgements about our conscious thoughts based directly and rationally on these conscious thoughts themselves, but it does not actually explain *why* moving from a first-order conscious attitude to a self-ascription of it is a rational or warranted transition; it only suggests that it *must* be. Peacocke's account, on the other hand, does, in a sense, seem to explain this: we are reasonable in making judgements about our occurrent conscious thoughts on the basis of these occurrent thoughts themselves, that is, the reasons on the basis of which we self-ascribe our beliefs are *good* reasons, because they are reasons which guarantee the truth of the self-ascriptions. In this sense therefore, Peacocke's account might seem to be an improvement on Burge's. However, the truly fundamental question to which Burge's account led us but did not in fact provide an answer to, was not that of how consciously-based self-ascriptions, given that such self-ascriptions are possible, can generally be veridical, but rather that of how it is possible at all for a first-order conscious state to rationally issue in higher-order knowledge of itself, given the general understanding laid out at the beginning of this chapter, of what it is for a judgement to be reason-based, or

what it is for a transition between two states to be, at the personal level, immediately rational.

To put things differently, what Peacocke's account seems to do, is explain how it is that the immediate judgements we make about our occurrent first-order states can generally be veridical, and his answer is that these judgements are generally veridical because, given the possession conditions for the concept of belief, these judgements will normally be made only in circumstances which guarantee them to be true. This, however, does not address the issue of how a judgement about our own thoughts can be reason-based in this way at all. That is, it does not answer the question of how a state with one content can rationalize a state with a completely unrelated content.[65]

Having said this, it is not entirely true to say that Peacocke does not address this question. There in fact seems to be at least the beginning of an answer to it in his discussion of the 'conscious' character of the thoughts which we are able to self-ascribe non-inferentially and authoritatively.[66]

For Peacocke, a 'conscious' thought is essentially a thought which is conscious in the sense defined in chapter 1 above, that is, a *phenomenally* conscious thought, or a phenomenologically *occurrent* one, or, in Peacocke's terms, a thought which is currently 'occupying our attention', and which 'contributes to what, subjectively, it is like for the person who enjoys it'.[67] Now the relevance that the occupation of attention is supposed to have to our ability to move directly from a conscious state to a self-ascription

---

[65] Martin raises a similar problem for Peacocke in (Martin 1998).

[66] (Peacocke 1998)

[67] (1998, p.64)

of it, is that in the case of such self-ascriptions, these judgements are not rationalized by the attitude self-ascribed primarily in virtue of this attitude's having a certain *content* (in fact, we have seen, how could it?), but essentially in virtue of the fact that this attitude is one which is currently *occupying our attention*, and somehow contributing to what things are like for us subjectively. Does this resolve our problem? It seems not, until we get an answer to the question of how it is that a first-order state's being *conscious* or being such that it contributes to what things are like for us subjectively, might enable it to stand as an immediate reason for self-ascribing it.

One possible answer would be to say that having a phenomenally conscious attitude in this sense involves being somehow implicitly aware of oneself having it. Peacocke, however, for various reasons seems to want to resist this option. In particular, he seems to want to allow for the possibility that non-human animals might have conscious states of the same kind we do. That is, he wants to leave room for the possibility that animals are, roughly, the same as us, except that we possess the concept of belief but they do not.[68] In fact, he writes: '...what is involved in a belief's being conscious can be fulfilled by a creature who does not even possess the concept of belief. What is true is that if a thinker does have the concept of belief and has a certain conscious belief, then he will be willing to judge that he has the belief'.[69] But now if having a conscious state involves being aware only of the *world*, we are back to our original question: how can a belief about the world rationalize a belief about itself? Or, what else might it be about a first-order state's being 'conscious' that might enable it to stand as an

---

[68] See especially (1992, pp.151-154) and (1998, p. 96)

[69] (1992, p.153)

immediate reason for self-ascribing it?

Another option might be to say that it is in virtue of there being something specific *it is like* to have an occurrent belief that p, (as opposed to, say, an entertaining that q), that such a belief can immediately rationalize a self-ascription of itself. But how does this help? If the proposal is that we base our self-ascriptions of our conscious attitudes on some kind of 'phenomenal feel' that comes with having them, and by which we can somehow 'sense' that we have them, then this just leads us back into a perceptual model of self-knowledge, and thereby to all the problems that go with it.[70] This is therefore not a viable option, nor indeed one which Peacocke would want to accept, given that his aim is precisely to put forward a view of self-knowledge which is reason-based *without* being perceptual.

In other words, on Peacocke's account, the relevance that a state's being 'conscious' has to this state's ability to constitute an immediate reason for self-ascribing it, can have nothing to do with one's being in any way aware of having it (whether in virtue of this state's being accompanied by a higher-belief about it,[71] or in virtue of its being accompanied by some phenomenal feel by which we might be able to sense it, nor in virtue of its being somehow intrinsically a state of self-awareness). Peacocke's answer, in the end, arises from his account of what it is to possess the concept of belief, together with his account, derived from his theory of concepts, of what it is for a transition between two mental states to be *rational.*

Essentially, Peacocke's view is that the transition between a conscious

---

[70] See chapter 2, pp.27-28 above.

[71] Peacocke spends some time arguing against higher-order thought theories of consciousness both in (1992, chapter 6) and in (1998).

belief and a self-ascription of it is a *rational* transition because it is a transition the making of which is inscribed in the very possession condition for the concept of belief, the relevant clause of which is: 'A relational concept R is the concept of belief only if [...] the thinker finds the first-person content that he stands in R to p primitively compelling whenever he has the conscious belief that p, and he finds it compelling because he has that conscious belief'.[72]

But now, one might raise a similar objection to Peacocke as Peacocke does to Burge, and say that this being the right possession condition for the concept of belief just *presupposes* that we make mental self-ascriptions directly and rationally on the basis of the self-ascribed conscious beliefs, since indeed we would not count someone as possessing the concept of belief unless they found it primitively compelling to self-ascribe a belief whenever they had a conscious belief (and considered the matter), and to do so for that very reason. This however does not explain *why* having a conscious belief makes it immediately rational for us to self-ascribe it, nor does it explain what it is about the self-ascribed belief's being *conscious* that enables it to stand in such a direct reason-giving relation to our self-ascriptions of it.

Peacocke can of course just reply that such a transition is rational *precisely because* it is inscribed in the possession condition for the concept of belief, and it is only transitions between *conscious* beliefs and self-ascriptions of them that are inscribed in the relevant clause of the possession condition. In fact, Peacocke seems to have a slightly different view of what it is for a transition to be *rational* than the one I laid out at the outset of this chapter. In brief, he takes justification to have to do with

---

[72] (Peacocke 1992, p.163)

transitions between states of mind, some of which count as rational and others of which do not, the former being those transitions which are inscribed in the possession conditions for a concept. For example, a transition from a perceptual experience as of red, to a judgement that there is something red in the near vicinity, is a *rational* transition on this view, because it is one which is written in the possession condition for the perceptual concept of red, namely, in essence, the condition that someone who possesses the perceptual concept of red will find the content that there is something red in the near vicinity primitively compelling whenever they have a conscious experience as of red, and they will find it primitively compelling *because* they have this experience. Similarly, a transition from believing that p and believing that p entails q, to believing that q, is a *rational* transition, because it is amongst the transitions the making of which or the finding compelling to make which, is written in the possession condition for the logical concept of entailment. That is, in brief, for someone to count as possessing the concept of entailment, they must find, amongst other things, contents of the form q primitively compelling, whenever they believe that p and believe that p entails q, and to do so *because* they believe that p and believe that p entails q.

But now, one might still complain that in the case of *these* transitions, between states which are inscribed in the possession conditions for perceptual or logical concepts, we are actually able to make sense of *why* someone who possesses the relevant concept, and who is in the former state of the transition, should find the content of the second state of the transition primitively compelling. We can make sense of this by reference to the model of justification set out at the beginning of this chapter, that is, by reference to the idea that they find the latter content primitively compelling in such

circumstances because they have *evidence* for it, that is, because they have a certain *access* to facts that *justify* this content. In fact, one could represent these rational moves as follows:

(1)

S perceives that ____p____

S believes that p

(2)

S believes that p

S believes that __p entails q__

S believes that q

(3)

S believes that ____p_____

S believes that S believes that p

In cases (1) and (2), we can make sense of why moving from the first (or first two) mental states of the transition, to the second (or third) state, might be a *rational* transition, or, to put it in Peacocke's terms, why anyone who possesses the perceptual concept of p, or the concept of entailment, and who is the first (or first two) states, might find the content of the second (or the third) state primitively compelling, by reference to the fact that if the proposition that p (or the propositions that p, and that p entails q) to which they have (loosely speaking) access are true, then the content of the second (or

third) state of the transition would *follow*, that is, it would also be true.

In the case of the move between having a conscious belief and self-ascribing it, on the other hand, it remains utterly unclear why, someone in the first state of the transition, would find the content that they *are* in this state primitively compelling, unless, in having the first-order conscious state, either they have at the same time some sort of access to the fact that they are in this state, or they have some sub-personal causal mechanism in their brain which makes them find themselves with a sudden compulsion to self-ascribe a belief whenever they have a conscious belief. Peacocke, however, we have seen does not want to accept any version of the former option, nor does he want to accept the latter, since, he insists, his account is supposed to be an account of the transition between first-order conscious beliefs and self-ascriptions of them as a *personal* level transition, one which, to use his words, 'makes sense to the subject himself given [the subject's] point of view'[73] which is why he insists that the first-order state in such a transition must be a *conscious* state.

However, it seems that, until we are given a story about what it is about a state's being conscious, other than that a transition between conscious (and not unconscious) beliefs and self-ascriptions of them are written into the possession condition of the concept of belief, it is not clear what sense can be made of this idea that moving from a conscious state to a self-ascription of it is a transition which makes rational sense to us from our own point of view. In fact, what sense are we to make of why, from Joe's point of view, his consciously believing, say, that there is a cat on the mat, should make him find the content that he *believes* that there is a cat on the mat primitively compelling,

[73] (1998, p.96)

if his consciously believing that there is a cat on the mat does not in any way involve him being aware of believing this. To put things differently, what is about the cat's being on the mat, from Joe's point of view, that it follows that he *believes* that there is a cat on the mat?

But perhaps to be asking these questions, is just to be caught in the assumption that there is more to be said about personal level justification than a certain transition's being inscribed in the possession condition for a concept. It is difficult however not to ask them, especially as, first of all, it seems that more *can* be said in the case of the transitions involved in the possession conditions for all concepts other than that of belief, and secondly, because Peacocke himself stresses the importance of the self-ascribed state's being *conscious* if it is to be able to constitute an immediate reason for self-ascribing it. No real story, however, of what it is about a state's being conscious that makes it able to constitute an immediate reason for self-ascribing it, is in the end given to us.

To sum up, one of Peacocke's main motivations for his account was to be able to maintain, contra what he calls the 'no-reasons' view, that the transition between a first-order conscious state and a self-ascription of it is a rational *personal*-level transition, that is, a transition which somehow makes sense to us from our own self-ascribing perspective. However, the only way in which such a transition can seemingly be a *personal*-level transition, is if that which fixes the content of our self-ascription is something to which we are sensitive to, or have *access* to at the personal level, namely something we are *aware* of. This sensitivity could then rationally ground our self-ascriptions. Peacocke of course does not deny that making a self-ascription involves being sensitive to the psychological nature of our conscious states, since it is in virtue of their

psychological nature that the content our self-ascription is fixed. However, we have seen that if this sensitivity is just a 'phenomenal' sensitivity, then we end up with a perceptual model of self-knowledge and all the problems associated with it. On the other hand, if this sensitivity is a kind of implicit awareness, intrinsic to having the thought itself, of ourselves as thinking the relevant thought, then we can no longer hold on to the view (which Peacocke *does* want to hold on to) that having a conscious state involves the same thing across species. In resisting this latter option though, his position ends up collapsing either into a perceptual model or into a non-reason based account whereby the nature of the psychological state self-ascribed itself (and not any personal-level sensitivity to it) directly causes a self-ascription of it, at the sub-personal level. Taking either of these two lines however goes against Peacocke's initial intentions to find an intermediate line between them, and so he would not want to take them, yet given his commitments about the nature of our conscious states, it is difficult to see what other options are left open to him, other than the anti-explanatory one of just insisting that there is no further sense to be made of the idea that the transition between first and second-order states is rational than mentioning that it is inscribed in the very fact of what it is to possess the concept of belief.

In other words, in the end, Peacocke's account seems to face a similar limitation to Burge's, namely that of failing to provide a satisfactory explanation of how it is that it is rational for us to move from having a conscious belief about an object to believing that we have a belief about that object, given that nothing seems to follow from how things are with that object, about what anyone believes. Peaocke's account only seems to point to the fact that we *must* be able to make such self-ascriptions directly on the basis of our object-oriented conscious thoughts, given that doing so is presupposed

by the very fact that we possess the concept of belief, but does not explain *how* this is possible. Nonetheless, his account does seem to move a step further than Burge's, first of all in its actually addressing the question, and secondly in its suggesting that it must be something about a self-ascribed state's being *conscious* that enables it to stand as an immediate reason for a second-order belief about it. He fails however to provide a satisfactory story of what this distinctive feature of conscious thought might be.

* * *

To conclude, the examination of Burge's transcendental argument from critical reasoning, first of all, showed us that our nature as critical reasoners presupposes a 'rational integration'[74] of our first-order conscious thoughts and our second-order judgements about them, thereby lending strong support to the view that the 'intermediate' approach must be right. Secondly, Peacocke's plausible account of what it is to possess the concept of belief also strongly supports the view that our immediate introspective judgements about our occurrent conscious beliefs are based directly and rationally on these conscious beliefs themselves, since indeed we not would not tend to regard someone as possessing the concept of belief unless they found the content that they believed that p primitively compelling, whenever they did consciously believe that p, and for that very reason. Thirdly, the very phenomenology of mental self-ascription seems to support the view that we do in fact self-ascribe our own thoughts on the basis of these conscious thoughts themselves. In fact, following Evans's datum, when we consider what our beliefs

[74] (Burge 1996, p.103)

are, we do not look for evidence concerning our beliefs, but rather, we search for evidence about the world.[75] If I am asked whether I believe that it is raining, I do not look at myself but out the window, and then judge on the basis of what I see, and on the basis of what I thereby come to believe about the weather, that I either do, or that I do not believe that it is raining. Moreover, if the relation between first and second-order conscious states is immediate and rational, we also have an explanation of why error in self-knowledge is so closely tied to the ascription of rationality, and why in fact introspective self-knowledge seems to be altogether immune to non-cognitive error.

In other words, there are a number of positive reasons (in addition to the negative ones seen in chapters 2 and 3) for adopting this intermediate line of approach to self-knowledge. Burge's account fails however to be satisfactory given a certain explanatory aim, and Peacocke's attempt to fill this explanatory gap by reference to the possession conditions of the concept of belief ultimately faces the same predicament. Nonetheless, the problems faced by Peacocke's specific proposal, were not sufficiently general to undermine such a reason-based line of approach altogether. Rather, the problems encountered merely revealed that the following three claims, which Peacocke wants to hold together, do not in fact seem to be compatible:

(1) Our immediate authoritative attitudinal self-ascriptions are based on *reasons.*

(2) Our reasons for our self-ascriptions are our first-order conscious attitudes.

(3) Our conscious attitudes are not *self-conscious* attitudes; they are strictly about the *world.*

Given the arguments of chapter 3, we have to accept (1), and adopt an

[75] See (Evans 1982, Chapter 7, in particular p.225)

epistemological approach to self-knowledge. Given the arguments of chapter 2, we have to accept (2) , and adopt the intermediate reason-based position, since our self-ascriptions cannot be based either on inference, nor on a direct perceptual experience of our first-order attitudes. The only option left open to us is therefore to reject (3), and to maintain that our special, immediate, authoritative knowledge of our own occurrent conscious thoughts, beliefs, and other attitudes is based on *reasons*, that these reasons are the very conscious states thereby known, and that such a relation between first and second-order states is possible essentially in virtue of the intrinsically *self-conscious* nature of our first-order conscious states. In other words, we must ultimately assume that our conscious states are themselves *self-conscious* states, if we are to account for the possibility of introspective self-knowledge.

Chapter 5: Self-Conscious Thoughts

What does it mean to say that our first-order conscious thoughts are *self-conscious* thoughts or to say that they are *intrinsically self-intimating*, or states of *primitive self-awareness*? What is it for a world-oriented state to involve not only awareness of the world, but also of itself? Our examination of the various available theoretical lines of approach to introspective self-knowledge revealed that some such claim must be true, that is, that our first-order world-oriented states *must* somehow be at the same time states of implicit self-awareness, if knowledge of our own thoughts from the inside, looking without, is to be possible. This however, one might feel, still leaves us without an entirely clear sense of what the claim is supposed to amount to, and in particular without a clear sense of how exactly an account of introspective self-knowledge as based on the nature of our conscious states as *self-conscious* states, might differ from the accounts of self-knowledge already on offer. Spelling out this thesis more clearly will thus be the first task of this final chapter. The second task will then be to briefly consider some of the consequences of this view, in particular for the way we think about phenomenal consciousness in ourselves, and for the way we think about the subjectivity of non-human animals.

In the course of this investigation, we arrived at the thesis that our first-order conscious thoughts are self-intimating essentially through seeing that they *must* be so if self-knowledge is to be possible. The best way therefore, of beginning to understand what exactly this view is supposed to amount to, is by looking at what it *must* amount to if it is indeed to constitute the view it is supposed to constitute, namely one which allows us to account for the possibility of authoritative introspective self-knowledge in a way that avoids all the problems faced by the other available accounts. Bearing this in mind then, I will now turn to consider and spell out this thesis, as much as it is actually possible to do so, by first looking at what it does *not* amount to, then by going over what it *has to* involve, and finally by considering what further positive sense might be made of it.

To begin with, the implicit pre-reflective self-awareness suggested to be involved in consciously thinking about the world should not be taken to be a kind of *perceptual* awareness. That is, in particular, the suggestion that in appropriately conceptually equipped beings, having a conscious mental episode involves at a the same time being implicitly (and somehow pre-reflectively) aware of oneself as having it, should not be taken to mean that our mental states have some kind of phenomenal buzz associated with them by which we can sense them, as this would just lead us straight back into the already discussed problems with perceptual models of self-knowledge. This is of course not to deny that our occurrent conscious thoughts are states which there is something it is like for us to be in them. The only point here is that 'what-it-is like' properties are not sufficient to ground mental self-ascriptions. The sense in which the first-order thoughts which occur in our phenomenal stream of consciousness must be intrinsically self-intimating, can thus not amount to the sense in which we are sensitive to

them via their phenomenal character.

Secondly, the thesis that our phenomenally conscious states are self-conscious states should neither be taken to mean that our conscious states are states which are always accompanied by (possibly non-conscious) second-order beliefs about them, as although this may be true, merely stating this does not explain why it might be so, or how it *can* be so. To say this would in fact just lead us straight back to the beginning, that is, to the problem of having to explain *why* our conscious states are such that they are always accompanied by non-conscious second-order beliefs about them, assuming that this is indeed the case. One might of course argue, following higher-order thought theories of consciousness such as Rosenthal's,[76] that this is just a primitive fact about our conscious states, which cannot be explained any further. In brief, on this view a state is conscious if it is accompanied by a non-conscious second-order belief about it, and its being conscious indeed *consists* in its being so accompanied. There are however, I believe, a number of compelling arguments against this approach to consciousness, such as the simple consideration that we can have second-order thoughts about attitudes which are *not* conscious, thereby revealing that a state's being accompanied by a higher-order thought about itself is clearly not sufficient for it to be a conscious state, let alone the very fact that makes it conscious. I may for instance discover through psychoanalysis that I have feelings of resentment towards a member of my family, and thereby come to hold the second-order belief that I have these feelings. This however, will not necessarily make my feelings become conscious. They could in fact perfectly well remain completely repressed, to the point that I may even start doubting whether what my analyst led me to believe is true.

---

[76] See (Rosenthal 1991)

Moreover, without having to appeal to repressed attitudes, one could imagine having both a first and a second-order belief, both of them playing an active role in affecting one's actions and one's thoughts, without either of these beliefs actually occurring to one, that is, without either of them actually being phenomenally conscious. In other words, this type of higher-order thought theory of consciousness seems to be neither plausible as it stands, nor *a fortiori* one which could be taken as the fundamental primitive fact underlying the possibility of introspective self-knowledge.

Finally, the thesis being put forward here should not be taken to be a version of the strong constitutive no-reasons view of self-knowledge. In saying that the possibility of introspective self-knowledge presupposes that our conscious thoughts are intrinsically self-conscious thoughts, I am not suggesting that it is somehow ontologically constitutive of having a first-order conscious thought, that one also has the corresponding higher-order thought. That is, I am not suggesting that our first and second-order states are one and the same state. In fact, to do so would just be to end up back with the problem associated with the strong constitutive approach, of how to accommodate the fallibility of self-knowledge within this picture.

In other words, our first-order conscious thoughts and our second-order reflective judgements about them must be *distinct* states, the latter being *rationally based* on the former in virtue of the intrinsically self-conscious nature of the former; a self-conscious nature which can consist neither in their being associated with a phenomenal feel by which we can sense them, nor in their being accompanied by a non-conscious higher-order belief about them, nor in their not being distinct states from our reflective judgements about them after all. But if this is what being intrinsically self-conscious does

not amount to, one might ask, what *does* it amount to? What else is there for it to possibly be? Let us look at this in context.

In chapter 2, in discussing perceptual models of self-knowledge, we saw that to know one's own thoughts introspectively, is not to know them by looking inside, but rather to know them by looking *from* the inside outward at the world, that is, to know them as the *subject* of these thoughts thinking about the world. We then saw in chapter 3, however, that this could not be explained by adopting a purely metaphysical account according to which our conscious thoughts about the world and our reflective judgements about them are not actually distinct states. However, having to take an epistemological approach to this special kind of self-knowledge, gave rise to the problem of having to explain how having thoughts about the world, or being in possession of evidence regarding the world, could possibly directly and non-inferentially ground knowledge about one's mental states. More concretely, the question was, what is it about, say, a cat's being on the mat from our point of view, that it immediately follows that we *believe* that there is a cat on the mat? The conclusion reached in chapter 4, was that it must be precisely something about the cat's being on the mat *for us*, that is, *from our own conscious point of view*, that it indeed follows, from our point of view, that we believe it. That is, it must be something about the very nature of our way of experiencing or consciously thinking about the world that makes our doing so constitute an immediate reason for judging that we are doing so. In fact, the idea was that conscious thought, at least of the kind had by beings with second-order abilities, must be somehow primitively *self-conscious* thought. Or, to put things differently, the idea is that for fully reflective self-consciousness to be possible, this self-consciousness must already be present in a pre-reflective form in

conscious thought itself.

Now the question this obviously gives rise to is that of what it is for a thought to be *pre-reflectively self-conscious*? But of course, when understood in these terms, the problem with spelling out this idea, is that in order to articulate it we would have to introduce a reflective dissociation between the awareness and that which it is awareness of (ie. by saying that it is a kind of awareness of ourselves thinking), thereby going against the very idea that it is *pre-reflective* self-awareness, that is, something which comes *prior* to, or is more basic than, fully articulated reflective self-consciousness. Trying to articulate the pre-reflective self-conscious nature of conscious thought can in fact only give rise to the question all over again of how this fully articulated, and hence reflective, kind of self-consciousness is possible, thereby sending us off on an infinite regress of having to keep explaining how self-consciousness is possible, and endlessly having to appeal to the pre-reflective self-conscious nature of conscious thought itself, which, when articulated, immediately raises the same question of how reflective self-consciousness is possible all over again. The intrinsically self-conscious nature of phenomenally conscious thought, can therefore not be articulated beyond a certain point; it must ultimately be taken as primitive, and yet as something which *must* be present in conscious thought itself, as the ground for its possible reflective articulation. Making reflective judgements about what thoughts or experiences we are having, on the basis of experiencing or thinking only about the *world*, we have seen, would be impossible if the latter were not at the same time a way of thinking or experiencing the experienced part of the world, as being the world *as experienced* from our point of view, or *as thought about* by us.

This is, I believe, very close to what Sartre might have had in mind when

speaking of the 'pre-reflective cogito', as a kind of self-awareness implicit already in conscious thought itself, and made explicit in the fully reflective Cartesian cogito, for which it stands as its pre-cognitive basis.[77] A similar thought can also be found implicit in Kant's point about it having to be at least possible for the 'I think' to accompany all our representations,[78] thereby suggesting that the grounds for the possibility of reflective self-consciousness must already be present in experience itself. Now, ultimately, Kant wants to derive from this the conclusion that experience must be experience of an objective unified world, on the grounds that only experience of a world conceived of as objective can make room for the possibility of the self-ascription of experiences. In relation to our present concerns though, this can be taken to suggest that conscious experience as of an objective world has the required dual aspect of being both awareness of the world, and in some primitive sense, awareness of itself. By appealing to this idea in fact, which is discussed in particular by Strawson, about how experience of the world conceived of as objective makes room for thought about itself, we might be able to extract a less metaphorical sense of the notion of pre-reflective self-consciousness, which I now turn to consider.[79]

Strawson's suggestion is essentially that experience of a world conceived

---

[77] See (Sartre 1969, especially pp.xxvi-xxvii, and chapter 2, section III)

[78] See (Kant 1929, p.153 or B 132)

[79] In appealing to Kant's point here, I am not following the actual dialectic of his argument. In fact Kant's aim is to show that self-consciousness presupposes experience as of an objective world, and he does this, at least on Strawson's interpretation, roughly by arguing that if self-consciousness is to be possible, experience must be such as to provide room for thought about itself, and it so happens that experience being as of an objective world provides room for this thought; experience must therefore be of a world conceived of as objective. I am starting instead from the fact that our thoughts and experiences *are* of a world conceived of as objective, and taking it that the way in which such experiences make room for thought about themselves, might help elucidate the idea that our conscious thoughts about the world are primitively *self-conscious* thoughts.

of as objective can be said to have a dual aspect, in that taking the order and content of a series of such experiences together (say, as articulated in a series of judgements), would give us 'on the one hand a (partial) description of an objective world and on the other a chart of a single subjective experience of that world. Not only the series as a whole, but each member of the series, has a double aspect'.[80] In relation to our present concerns then, the idea is that experiencing the world as objective, involves implicitly making the above distinction between how things are, and how things are experienced as being from our point of view. That is, to experience and think about the world as an *objective* world is to implicitly conceptualize the present content of one's experiences as not *exhausting* the world, and to conceptualize the order in which the world presents itself, as not necessarily being the order in which things *exist*; which is in effect to conceptualize the content of one's thoughts or experiences, as being the world only *as thought about by us*, or *as it appears to us to be*, or *as experienced from our point of view.* If this is right, then it in fact turns out not only that conscious thought is pre-reflectively self-conscious, but that it cannot *but* be, given that it is indeed of a world conceived of as an objective world. In other words, in this Kantian account of objective experience as having a dual aspect, we may have the beginning of a positive account of how our second-order abilities might be reflected in the very nature of our way of experiencing and thinking about the world; an account, moreover, which does not just replace the problem of how self-knowledge is possible, with the problem of how self-consciousness is possible at the level of first-order thought.

One question may still remain though, namely that of how exactly this

---

[80] (Strawson 1966, pp. 105-106)

solution to the problem of immediate first-person authoritative self-knowledge should be taken. In particular, should it just be taken as a kind of 'default' view which we have reason to accept only because no better solution to the problem of introspective self-knowledge (or to the question of how to account for the world's being objective in our experience) seems to be available, or is it actually a plausible account on independent intuitive grounds? I believe that it is the latter. In fact, although this account was arrived at here essentially through an examination of the problems faced by its alternatives, upon reflection, it seems to ultimately be the one which fits best with the actual phenomenology of introspection discussed at the very beginning of chapter 2.[81] In fact to begin with, if our phenomenally conscious states are intrinsically *self-conscious* states, we can immediately see why self-ascribing them might make immediate rational sense to us from our own self-ascribing point of view. Secondly, if this thesis is correct, we also have an explanation of why when it suddenly occurs to us that we are thinking about something, this information does not usually strike us as a surprise, but rather, we feel as though we were aware of what we were doing all along, but were just not explicitly thinking about it. Finally, if it is true that in consciously thinking about the world we are also pre-reflectively aware of ourselves doing so, we also have an explanation of why we are able to remember thinking thoughts which we were not, at the time, reflecting on. This would, for instance, explain why I may actually be able to remember *thinking* to myself, as a child, that Santa Claus does not exist, thereby now putting me in a position to be able to immediately self-ascribe this thought, although at the time I was only thinking about the existence or not of Santa Claus and not about my mental states. In many ways, therefore, this solution to the

[81] See chapter 2, pp.21-23 above

problem of introspective self-knowledge seems to be a very intuitive one. However, there remains one important respect in which it might be taken to be, quite on the contrary, *counter*-intuitive, namely the respect in which it seems to rule out non-human animals and very young children, whom we do not take to be self-conscious, from qualifying as having conscious states of the same kind we do. This question must be addressed, since it tends to constitute a primary reason for wanting to avoid theories which entail that consciousness presupposes self-consciousness. I will therefore conclude this investigation by considering how threatening this point really is to the present thesis.

Upon reflection, this issue does not, I believe, turn out to be as damaging to the intuitive appeal of the present proposal as it might have seemed. In fact, for one thing, most people would agree that neither non-human animals nor very young children have conscious *thoughts*. That is, we would not intuitively say that animals and young children think to themselves in words, or wonder about things, entertain possibilities, etc. In fact, they do not intuitively seem to do so even in the form of images. They may well, however, have non-occurrent, or non-conscious dispositional intentional states such as beliefs and desires, which guide their behaviour. Their having such states though, is in no way ruled out by the present account of self-knowledge. All that is ruled out is that they make judgements, or assent to propositions, which seems to be a perfectly intuitive thought, except perhaps in the case of some primates; primates which, however, in such cases, we would most likely take to be to some extent self-conscious.

The only real clash between the conclusion of this thesis and our intuitions about young children and non-human animals, therefore only arises when thinking about conscious *experience*. However, even in these cases, upon reflection, it is not clear that

we would want to say that the conscious experiences of non-human animals are of the same kind as our own. In particular, it is not clear that we would want to say that they experience the world as an objective unified world, in the way that we do, rather than, as McDowell suggests, as a series of obstacles, opportunities, problems and other pressures from the environment, not conceptualized as such, but merely dealt with as they come.[82] For example, would we want to say that bats experience the world as an objective world? Would we want to say that dogs do, or primates? In the case of bats we would probably want to say that they do not, while in the case of dogs we might be more hesitant, and finally, in the case of primates we might even be tempted to say that some types of primates possibly do. In each case however, the degree to which we would be prepared to say that the creature's conscious experience of the world is similar to our own, seems to correspond roughly to the degree to which we would be prepared to attribute to it some level of self-consciousness. In other words, it is not clear that there actually is anything deeply counterintuitive about the present thesis in this respect. In fact, to say that non-self-conscious creatures do not experience the world in the same way that we do, is not to suggest that they do not have phenomenal states or a subjectivity of any kind at all. Rather it is only to say that their subjectivities are to varying degrees very different in kind from our own, which, when taking the example of bats, for instance, certainly does not seem to be a counterintuitive claim to make. In fact, would we really want to say that what it is like to be a bat must be roughly like what it is like to be ourselves, except with echolocatory experiences? Upon consideration, it would actually be highly surprising from an intuitive point of view, to find out that both linguistic and non linguistic creatures alike,

---

[82] See (McDowell 1994, chapter 6)

humans as well as bats, all experienced the world in the same way, the only real difference between species lying in an additional faculty for self-knowledge possessed by some but lacked by others. Indeed if we *did* discover that all animals experienced the world in exactly the same way as we do, and that therefore, for instance, the way in which fish subjectively experienced their environment was just like the way we would experience it, if we were in a fish's body swimming under water, would we not then be inclined to say that these animals were self-conscious? If we would, this would suggest that our reluctance to attribute self-consciousness to animals, actually reflects an underlying intuitive reluctance to think of them as subjectively experiencing the world in the same way that we do.

In the end, therefore, the thesis that our second-order abilities are reflected in the very nature of our phenomenally conscious states, seems to be not only necessitated by the shortcomings of all other possible theoretical approaches to introspective self-knowledge, but arguably also the most intuitively plausible.

Bibliography

Armstrong, D.M. (1968), *A Materialist Theory of Mind.* London: Routledge and Kegan Paul

Austin, J.L. (1962), *Sense and Sensibilia.* Oxford: Oxford University Press

Bilgrami, Akeel (1998), 'Self-Knowledge and Resentment', in Wright, Smith and Macdonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Block, Ned (1995), 'On a Confusion About a Function of Consciousness', in *Brain and Behavioural Sciences* (1995)

Boghossian, Paul (1989), 'Content and Self-Knowledge', in *Philosophical Topics 1989*

Burge, Tyler (1988), 'Individualism and Self-Knowledge', in the *Journal of Philosophy 1988*

Burge, Tyler (1996), 'Our Entitlement to Self-Knowledge', in *Aristotelian Society* Supplementary Volume 1996

Burge, Tyler (1998), 'Reason and the First Person', in Wright, Smith and Macdonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Churchland, P.M (1991), 'Eliminative Materialism and the Propositional Attitudes', in D. Rosenthal (ed) *The Nature of Mind.* Oxford: Oxford University Press

Davidson, Donald (1987), 'Knowing One's Own Mind', in *The Proceedings and Addresses of The American Philosophical Association,* 60 (1987) 441-58

Descartes, Rene (1912), *A Discourse on Method, Meditations and Principles.* Toronto: Dent

Evans, Gareth, (1982), *The Varieties of Reference.* Oxford: Clarendon Press

Fricker, Elizabeth (1998), 'Self-Knowledge: Special Access versus Artefact of Grammar - A Dichotomy Rejected', in Wright, Smith and Macdonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Heal, Jane, (1994), 'Moore's Paradox: A Wittgensteinian Approach', in *Mind* (January 1994)

Hume, David (1888), *Treatise of Human Nature.* L.A. Selby-Bigge (ed), Oxford: Clarendon Press

Kant, Immanuel (1929), *Critique of Pure Reason.* London: Macmillan Press

Lormand, Eric (1996), 'Nonphenomenal Consciousness', in *Noûs* (1996)

Martin, M.G.F (1998), 'An Eye Directed Outward', Wright, Smith and Macdonald (eds.) *Knowing Our Own Mind.* Oxford: Clarendon Press

McDowell, John (1994), *Mind and World.* Cambridge, Massachusetts: Harvard University Press

McDowell, John (1998), 'Response to Crispin Wright', in Wright, Smith and Macdonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Nagel, Thomas (1991), 'What Is it Like to Be a Bat?', in D. Rosenthal (ed) *The Nature of Mind.* Oxford: Oxford University Press

Peacocke, Christopher (1992), *Study of Concepts.* Cambridge, Massachusetts: The MIT Press

Peacocke, Christopher (1996), 'Entitlement, Self-Knowledge and Conceptual Redeployment', in *Aristotelian Society* Supplementary Volume (1996)

Peacocke, Christopher (1998), 'Conscious Attitudes, Attention and Self-Knowledge', in

Wright, Smith and Macdonald *Knowing Our Own Mind.* Oxford: Clarendon Press
Rosenthal, David (1991), 'Two Concepts of Consciousness', in Rosenthal (ed) *The Nature of Mind.* Oxford: Oxford University Press
Ryle, Gilbert (1966), *The Concept of Mind.* Harmondsworth: Penguin
Sartre, Jean-Paul (1969), *Being and Nothingness.* London: Routledge
Shoemaker, Sydney (1986), 'Introspection and the Self', in French, Vehling and Wettstein (eds.) *Studies in the Philosophy of Mind.* (Midwest Studies in Philosophy, Minneapolis 1996)
Shoemaker, Sydney (1988), 'On Knowing One's Own Mind', in *Philosophical Perspectives, 2, Epistemology* (1988)
Shoemaker, Sydney (1996), 'Self-Knowledge and 'Inner-Sense'', in Shoemaker, S. *The First Person Perspective and Other Essays.* Cambridge: Cambridge University Press
Strawson, P.F. (1966), *The Bounds of Sense.* London: Routledge
Williams, Bernard (1978), *Descartes: The Project of Pure Enquiry.* London: Penguin
Wittgenstein, Ludwig (1953), *Philosophical Investigations.* Oxford: Blackwell
Wright, Crispin (1998), 'Self-Knowledge: The Wittgensteinian Legacy', in Wright, Smith and Macdonald (eds.) *Knowing Our Own Mind.* Oxford: Clarendon Press