Rev.Phil.Psych. (2010) 1:593–610
DOI 10.1007/s13164-010-0044-x

Self-ascriptions of Belief and Transparency

Pascal Engel

Published online: 20 November 2010

Abstract Among recent theories of the nature of self-knowledge, the rationalistic view, according to which self-knowledge is not a cognitive achievement—perceptual or inferential—has been prominent. Upon this kind of view, however, self-knowledge becomes a bit of a mystery. Although the rationalistic conception is defended in this article, it is argued that it has to be supplemented by an account of the transparency of belief: the question whether to believe that P is settled when one asks oneself whether P.

1 Introduction

The access that we have to the contents of our own minds, in contrast with the access that we have to the minds of others, has three main *prima facie* features. First, it is *authoritative*: we have a special authority upon what happens in our own minds, in the sense that if we think that we are in a certain mental state it seems that we cannot be challenged. We can indeed make mistakes: our mental states can fail to represent correctly our environment, but it seems that we cannot be wrong in thinking that we have them. Second, our self-access is *privileged*: it seems to us that we know the contents of our own minds always better than we know the contents of the minds of others. There is a characteristic asymmetry between self-knowledge and knowledge of other minds. Third, self-knowledge is also *transparent*, in the sense that we seem to have access to our own mental states and to their content when they occur: the very fact that we have them is inseparable from our being conscious of them in the first person.

These three features seem so specific that they have been taken as characteristic of the mental as such within a whole tradition in philosophy. Cartesianism, in its strongest form, is understood as the view that not only the knowledge that we have

P. Engel (✉)
University of Geneva, Geneva, Switzerland
e-mail: Pascal.Engel@unige.ch

of our own minds is authoritative, infallible and transparent, but also that these features *define* the mental. But this seems to fly in the face of common sense—and indeed in the face of cognitive science common sense—since it apparently excludes unconscious or dispositional mental states which are neither transparent nor authoritative nor privileged. The Cartesian theorist can bite the bullet and claim that unconscious thoughts and the like are just not mental states at all. But the price is high. Moreover we often go wrong on the contents of our own thoughts, and the traditional appeal to a mysterious faculty of introspection does not convince any more. Anti-Cartesians squarely deny authority, privileged access and transparency. Thus Ryle famously argues in *The Concept of Mind* that there is no special first-person authority, and that our access to the contents of our own minds has no privilege over our access to the contents of the minds of others, hence that it is no less fallible. This claim has recently been revived by Daniel Dennett (1991) in his attacks against the "Cartesian theater" of consciousness, and it seems to be in line with much contemporary work in cognitive psychology (*locus classicus*: Nisbett and Wilson 1977). In spite of these frontal attacks, the fact that we have a privileged access to our own minds seems hard to die. However self-deceived, cognitively dissonant and in the grip of countless unconscious influences we can be, it remains true that by and large we know ourselves better than others know us.

The resilience of this feature of the mind raises two questions. The first is: can we accept it without adopting the distinctive tenet of Cartesianism, i.e that it is *constitutive* of the mental? In the second place how can we account of it? If, unlike Ryle, we take seriously the idea that our self-knowledge of our own minds is indeed a kind of *knowledge* and not an illusory stance, there seems to be only two possible ways of explaining it. We can take self-knowledge to be a form of *inferential* knowledge, that is knowledge inferred from other knowledge or from other beliefs. But this view clashes with the apparent immediacy of self-knowledge. Alternatively we can take it to be a kind of *perceptual* knowledge, something like a perceptual capacity directed inwards, some kind of inner sense. But this capacity looks mysterious: why would an inner sense perception, in contrast to an outer sense perception be infallible? A third alternative consists in saying that self-knowledge is neither based on inference nor perception, for it is not a kind of knowledge at all, but something which is true of us in virtue of conceptual necessity. Just as a sentence like "I am here now" is necessarily true and does not need the exercise of a particular cognitive capacity to be recognised as such, our self-access yields thoughts which are, like Descartes' *cogito* constitutively true and needs no explanation at all (so we can call this kind of account *constitutive*). In Boghossian's (1989) terms: self-knowledge is explained "either by inference, or by observation, of by nothing".

In what follows, I first review some of the main reasons why neither the inferentialist nor the perceptual models of self-knowledge are correct. It follows that the only option left is the constitutive view. But not any version of it would do. I discuss two alternative conceptions of the constitutive view, one inspired by Peacocke (1999) and the other one inspired by Gareth Evans' reconstruction of the notion of transparency: the best way to discover whether one believes that p consists in asking oneself whether p. Our own beliefs are transparent to us in the sense that we do not need to self ascribe them, but only to look at whether their contents are true. How is this feature connected to the idea that we have an authoritative

knowledge of our own beliefs? How can it explain the kind of warrant in which self-knowledge consists and what is the source of this warrant?

Before trying to answer these questions, a word of caution is needed to indicate where I think that the present kind of approach stands with respect to contemporary work in cognitive psychology. The relationships between, on the one hand, our common sense conception of mind and knowledge and, on the other hand, our scientific conceptions of these, are very complex. The familiar options are: reduction of the former to the latter, elimination, and complete autonomy. It is not the place here to state my own view, but none of these seem to me satisfactory.[1] Both eliminativism and reductionism seem to me to suffer from incomplete analysis both in descriptive and conceptual terms. In the present case the very idea that we could get rid or explain in completely satisfactory scientific terms a feature of the mind like self-knowledge is preposterous. This does not mean that a large body of scientific literature is not relevant to explaining it: on the contrary, a lot of work in cognitive psychology bears upon it.[2] But before we can confront our folk psychological and epistemological concepts with empirical studies in cognitive psychology we need an accurate description of the folk concepts in question. We cannot directly bring to bear work on consciousness, agency or metacognition without having a kind of map of how we understand, within our ordinary scheme, these notions. A further problem has to do with the fact that in asking questions about self-knowledge we ask questions about how it is justified or warranted, which are, at least *prima facie*, normative questions, which cannot, at least on a number of views in epistemology, be settled in purely psychological or causal terms. Once we have a more refined description of our conceptual scheme, we can, at a later stage, establish where empirical evidence can confirm or infirm it (in other words this scheme is not fully *a priori*). This kind of top-down strategy is, no doubt, one which a number of eliminativists and of "experimentalists" in the philosophy of mind—such as Stephen Stich and his associates—will find uncongenial and question-begging. But it is their strategy which I find question begging: most of these are actually versions of the inferentialist or of the perceptual strategy.[3] The present approach, which can be characterised as an "armchair" one, is, however, perfectly compatible with a kind of reductionism.[4] Before we can hope to reduce, we need to describe. One does need to subscribe to Brentano's program and to a strictly first-personal phenomenological conception of the mind to grant its importance, even within a naturalistic framework.[5]

[1] I have developed it in Engel 1996

[2] Much work in the huge mind-reading literature, about the Theory of Mind, about mental simulation, pretense, the emotions, autism, psycho-pathology is relevant to it. The problem is: *how relevant*? It seems clear, for instance that children's abilities to attribute beliefs to others have a lot to do with the abilities that they have to attribute beliefs to themselves. But how are we to understand that kind of evidence in order to assess claims to knowledge in both cases?

[3] E.g Nichols and Stich 2003 classify cognitive psychological conceptions of our self awareness in two broad categories: either as a Theory-theory account (we have a theory of mind from which we infer beliefs about ourselves, or a detector-monitor account. The former is clearly a case of an inferentialist account, the second of the perceptual account. Indeed the way the psychological story in both cases is filled out in many more details than the rough sketch that I give of the inferential and perceptual analyses, but the essence of the view remains the same, however the details are filled.

[4] In the style of Kim's conception of functional reduction, or in the style of Jackson's.

[5] See Thomasson (2003)

2 Self-knowledge: Inferential or Perceptual?

Let us briefly review the reasons why the inferentialist conception of self-knowledge is implausible. A typical inferentialist is Ryle (1949) who holds that our access to our mental states is just as good—and just as bad—in the third person as in the first person. According to the inferentialist our access to our own thoughts has to be inferred from facts about our behaviour or about our other beliefs. The inferentialist denies the Cartesian claim that our access to our own thoughts is infallible and that we have a special capacity of introspection which would yield this privileged access, but he needs not deny that we have a better access to our own thoughts than to the thoughts of others; what he denies is that we owe this access to a special capacity of inner sense. According to him the reason why we have a privileged access is that we are better placed than others to infer our thoughts from our behaviour (Ryle 1949: 171). This may be correct for those of our beliefs which are typically associated to dispositions to act (*e.g* my belief that all spiders are dangerous), but it is utterly implausible for other beliefs (such as my belief that the *Well Tempered Clavier* is a masterpiece). This seems to fly in the face of the obvious fact that our access to our thoughts is immediate and direct, and not the product of an inference. On the common sense conception of self-knowledge—which is spontaneously Cartesian—if I want to know what I believe, I usually do not observe my behaviour. I just try to figure out what I think, and deliver the result. To paraphrase Robert Burns, we think that we have "some gifty grant to see ourselves as others can't". The inferentialist tells us that it would be nice to have such a gift. But he suggests that it is utterly implausible to suggest that we have it anyway.

The alternative conception of self-knowledge as based on a specific capacity of introspection, however, is no less implausible. But it makes more sense if one withdraws the claim, usually associated with the introspectionist story, that this faculty is infallible. The defenders of the perceptual model of self-knowledge hold that if we treat this capacity on the model of perception, we can understand both how it can be a specific capacity—a kind of sense—which gives us a privileged access, and how it can fail—as any sensory perception can. The perceptual view needs not even amount to the view that there is a particular sensory modality which delivers sensory information about our inner life, for we can conceive of perception as the acquisition of *beliefs*. Thus David Armstrong, one leading proponent of the perceptual view says:

> Eccentric cases apart, perception, considered as a mental event, is the acquiring of information, or misinformation about our environment. It is not an "acquaintance" with objects, or a "searchlight" that makes contact with them, but it is simply the getting of beliefs. Exactly the same must be said of introspection. It is the getting of information or misinformation about the current state of our mind. (Armstrong 1968, p. 326)

On this view, self-acquaintance or introspection is analogous to sense perception, because just as sense perception allows us to acquire beliefs about our external environment, introspection allows us to acquire beliefs about our internal environment. Armstrong also emphasizes the fact that the beliefs that we thus get about our mental happenings need not be about a special object, a *self*.

For him introspection is perfectly compatible with our being acquainted with a bundle of mental items, composing a Humean scattered self. He holds that the capacity of introspection is compatible with the denial of the existence of a mental substance.

The important point, on this view, is that there is a belief-producing mechanism, which produces beliefs "about oneself". A mechanism, by definition, has a causal nature, hence is contingent. In other terms, just as sense perception yields beliefs about external states of affairs though a causal regularity, inner-perception yields beliefs about internal states through a causal regularity. One can understand why, on this perceptual model of self-knowledge, self-knowledge is warranted. It is because the mechanism operates, in general, in a reliable way. "In general" because it is not without exceptions. In some circumstances, the mechanism can fail to produce true information about oneself; it can also produce *misinformation*, just as perception can lead us to mistaken beliefs. The causal mechanism of introspection is reliable, but it is not infallible.

To a certain extent, the perceptual model preserves the other features of self-knowledge, authority and privileged access, in so far as it admits that inner sense is a faculty which only its owner can have. But can it preserve the transparency feature associated with its necessarily first-personal character? Hardly. Think of ordinary perception. If we understand it, as Amstrong does, as an acquisition of beliefs, I can, through ordinary perception, acquire the belief, say, that this is a cat. But I can be mistaken, and wrongly judge what I see to be a cat, while it is in fact a Pekinese dog. Somebody can correct my mistake, or I can correct my mistake by attending to my first belief and by revising it in the light of contrary evidence. In order to see that my perceptual belief that this is a cat is false, I must attend to my belief either in a third-person way (when somebody points out to me that my belief is false), or in a second-person way, by reflecting on my previous beliefs. Hence access to the truth of my perception cannot be had in the first-person way, for if I only attend to the contents of my own thoughts, I cannot judge whether they are true or false. Now perception is not necessarily reflexive. Whether or not we take it to necessarily imply awareness, everyone agrees that our perceptual beliefs are not necessarily reflexive in the sense of having second-order beliefs. Now if self-knowledge rested upon a perceptual state, it would follow that I could entertain certain beliefs, and have a certain conception of the kind of states they are, without knowing that I have them, that is without being able to self-ascribe them to myself. In other words, I could say that I believe that P, without being able to say whether I believe that I believe that P. According to the perceptual model, a person could have an acquaintance with his own states, through a causal mechanism which is generally reliable, and therefore which yields knowledge in so far as reliability can be a necessary condition for knowledge, without conceiving of himself as having these states, that is without conceiving himself as the very subject of these states. Shoemaker (1996) calls a creature that would instantiate this possibility, a "self-blind" creature, and gives an argument to the effect that there cannot be any such creature. This argument purports to show that one cannot be a rational believer and be self-blind, hence that first-person knowledge cannot be a contingent feature of our mental constitution, but a necessary and conceptual one.

3 Self-blindness and the Rationality Account

"Moore's paradox" lies in the paradoxical sounding character of sentences of the form: "P, but I don't believe that P". The reason why it is not a paradox in the usual sense is that the first conjunct does not formally contradict the second, since both conjuncts might be true. It may well be true, for instance that the earth is round, but that I do not believe it. But the sentence is nevertheless contradictory, since the first conjunct implies that I believe that it is true, whereas the second denies this. As Wittgenstein (1980) said, the paradox shows something about the "logic of assertion": asserting that P is the usual way of expressing that one has the belief that P, and therefore denying that one has the corresponding belief seems to contradict the belief expressed by the first conjunct. But, according to Wittgenstein, it is not a *logical* contradiction, because such sentences as "I believe that P" are not descriptions of one's state of belief, but expressions (*Äusserungen*) of them. When ascriptions of beliefs are made in the second or in the third person, there is no corresponding oddity. For instance there is nothing paradoxical in saying "The earth is round, but he does not believe it", because the "logic", or the "grammar" of such third-person ascriptions of beliefs is such that they are actually *descriptions*, and not *expressions* of beliefs. The paradox is not a semantic, but a pragmatic one. We may put a similar point along Gricean lines. We could say that someone who would assert that P intends to convey to an audience that he believes that P, and has the higher-order intention of intending the audience to recognise his first-order intention. Assertion is an action done with the intention of producing the belief that one has the belief. The Moorean sentence defeats this purpose, and therefore does not successfully convey the intention conveyed by the first conjunct; indeed it cancels it.

All this is common wisdom about Moore's paradox. What is less often noted is that Moore's paradox is not only present at the level of language, or of the linguistic expression of thought, but at the level of thought or belief itself (Heal 1994). As remarked above, if we take the first sentence of the Moorean conjunction "P, but I do not believe that P" to express the belief that P, and the second sentence to express disbelief that P, there is no contradiction. There is no contradiction, because I may well believe something, and disbelieve it. For instance, I may at one time believe that de Gaulle was a great leader, and fail to believe that at a later time, or I may discover that I have both beliefs, without having noticed it until now. But of course there is a contradiction if I have both beliefs, and if I am aware that I have them, and if I go on believing them while being aware of this. In other terms, if the Moorean sentence is understood thus:

> P, and I don't believe that P, and I *believe* that I believe that P and that I don't believe that P

then there is a genuine contradiction. In other terms, the subject who entertains the beliefs expressed by the Moorean sentences can entertain such beliefs, but he cannot believe that he has these beliefs, unless he consciously contradicts himself. In this sense Moore's paradox is a paradox because there cannot be such beliefs as those expressed by the sentences, not because there is something wrong in their linguistic expression.

We have not yet explained why there is a constitutive connexion between believing and believing that one believes. Nor have we explained what it is to believe that one believes. There are two possible explanations.

One explanation involves the notion of conscious belief. We can say that asserting a sentence implies that one is conscious of the belief that it expresses. On one analysis of conscious belief, a conscious belief is simply a second-order belief, a belief that one has the first-order belief (Rosenthal 1993). The explanation of Moore's paradox would then be that the content of the paradoxical sentence cannot be *consciously* believed. We can conceive of conscious belief as a sort of mental counterpart of assertion, a mental assent to a given content which is presented, in some way, to our mind. Thus the explanation of the constitutive link between assenting that P and assenting to "I believe that P" would be that if one assent to the first content, one assent to the second.

But this first explanation ignores the fact that we can believe that P, while not assenting to our believing that P. Ordinary cases of self -deception or of Freudian unconscious beliefs provide numerous examples. Similarly we can accept the existence of tacit beliefs, to which we can assent only in certain conditions (Lycan 1986). The claim that assenting to P implies that one believes that P, and that one believes that one believes that P, comes down to the claim that *if a belief P is available for assent*, then the belief that one believes that P is available as well (Shoemaker 1996, pp. 79-81). The contrary claim just amounts to granting the possibility of self-blindness, the possibility of a being who would be able to have beliefs, but who would not believe, in a first-person way, that he has these beliefs, although he could believe that he has these beliefs in a third person way, for instance by gaining information about his behaviour. And, you will notice, this also amounts to the possibility of a creature who could assert such sentences as "P, but I do not believe that P", and who would not find any impropriety in holding the beliefs expressed by such sentences. For instance there could be a creature who believes that P, expresses this belief by asserting that P, but discovers, in a third-person way, that in fact he does not have this belief. I want to claim, with Shoemaker, that this is not possible.

Could the reason why it is not possible be that such a creature could not *utter* Moorean *sentences*? This could not be the reason, because there is nothing which prevents such a creature from having acquired the appropriate linguistic rule "Do not utter sentences of the form "P but I don't believe that P"". Such an individual could notice that, in his community, such sentences do not elicit successful communication. He could follow this rule, just in the sense of instantiating a certain regularity. He could notice, for instance, that assertion is the normal mean of expressing one's beliefs, and assert sentences with the intention of conveying to his audience that he believes that P. But given that, by hypothesis, he is self-blind, he would lack evidence for believing that he believes that P. But if this is so, the self-blind person would be unable to use the proposition that P in his reasoning. Here is why: normally a rational person who believes that P should be disposed to use this proposition as a premise in reasoning, and should know that, if the proposition is true, it is in his interest to act on the assumption that it is true. And—this is the important point—such a rational person should know that to act on the assumption that a proposition is true is to act as if one believes that proposition. He should also know that it is of his interest to manifest his beliefs through assertions, if he wants to communicate successfully with others. In other terms, even if we use a minimal

definition of belief as a disposition to act, and if we assume a minimal notion of rationality as satisfaction of one's interests, a rational believer is a person who would act as if he believes that P. But could it be that although he acts *as if* he believes that P, he actually has grounds *not* to believe that P, and hence frame Moorean thoughts? Such a person would, at least, find an inconsistency in his own actions, not simply in his own thoughts. He would be unable to plan his own actions in the future, and to ascribe these actions to himself. In other terms, he would lack a capacity for normal rational action, because he could not find continuity in his own actions, indeed not even to find that these actions are *his* actions. And if belief is a disposition to act, then he would be self-blind about his beliefs. But I have just claimed that this seems to be utterly implausible, given only minimal requirements on rationality. To the extent that a subject is rational, and possesses the concept of belief, believing that P brings with it the cognitive disposition to believe that P, one believes that P either explicitly or tacitly. In other terms, a self blind creature could not have genuinely the concept of belief.

Why does this argument cast doubt on the perceptual model of self-knowledge? Because this model implies that a being could lack the perceptual capacity of introspecting himself in order to see what beliefs he has, without being in any way cognitively impaired. But if the preceding argument is correct, this is impossible. As Shoemaker (1996: 31) makes clear, the claim defended by the self-blindness argument is not that all believers have the concept of belief—this would of course be false for children and animals who are indeed believers—but that it is impossible to conceive of a creature who has a conception of his mental states—such as desires, beliefs and intentions—and can entertain them but would be unable to become aware of the truth of his thoughts about his thoughts except in a third person way by inferring his thoughts from his behaviour or by using a quasi perceptual capacity. Indeed the argument presupposes that ordinary thinkers are capable of having the concept of belief and of being capable to acting rationally. But—animals, irrational thinkers and children aside—that does not seem to be a too unrealistic condition upon our everyday thinking.

4 The Constitutive View of Self-knowledge

If self-knowledge neither rests on an inferential nor or a perceptual ability, then on what is it based? Given that the two ordinary ways of justifying beliefs are perception and inference from other beliefs, it follows that self-knowledge is not based on any justified belief. The answer suggested by Boghossian's three terms alternative quoted above is: nothing. Self-knowledge, on this view, is not knowledge at all, in the sense of a cognitive achievement. Self-knowledge, on the constitutive view, is a necessary feature of our having mental states about ourselves. It is a conceptual necessity, which holds *a priori*, which we can formulate as follows:

> (CT) Given certain conditions C, S believes that P if and only if S believes that he believes that P.[6]

[6] The formulation adopted here comes from Byrne (2005) and Coliva (2009).

This can be decomposed, as any biconditional, into two parts:

(i) if S believes that P then he believes that he believes that P
(ii) if S believes that he believes that P, then he believes that P

The *if* part (i) of (CT) follows, according to the self-blindness argument, from the very fact that one has a belief and is a rational thinker. The very fact that a subject believes that P entails, by a conceptual and *a priori* that he has the belief that he has the belief, or his first-order believes entails that he has the second-order belief. This seems incorrect in the case of dispositional or tacit beliefs, which are, by definition, not necessarily conscious. But as we saw above with the anti-self-blindness condition, (i) needs not entail the overly intellectualistic thesis that all beliefs are reflexive: it is enough that a subject who has a belief has the capacity to have the corresponding second order belief.

The *only if* (ii) part of (C) also follows from the self-blindness argument. It says that having a first-order belief is actually entailed by the having of the reflexive second-order belief: if you believe that you believe that P, you can't fail to believe that P. This, in effect is the Cartesian condition from which we started. This condition too, seems to be incorrect, for instance in cases of self-deception and other forms of irrational belief: a subject who has a conscious second-order belief that P may well not believe that P, if he is self-deceived about his belief that he believes that P, and for instance believes that Q instead. But here again, (CT) is supposed to hold for a rational agent. It actually holds, as we saw about Shoemaker's version, for a *rational* believer. This is what the reference to conditions C in (CT) means: unless a subject is irrational, or in some sense deceived, it is constitutive of his believing that P that he can't be wrong about his believing that P. Hence the Cartesian condition holds by conceptual necessity.

The constitutive thesis entails immediately the Cartesian features of authority, privileged access and transparency: a subject who satisfies the C conditions is *by definition* capable of accessing his own thoughts in a privileged and transparent mode, and can't be wrong about them. This seems neat, but it also looks a bit like magic. For we would like to know *in virtue of what* the conceptual necessity holds. What explains it? But this question is misplaced, on the constitutive view. If (CT) is true a priori and as a matter of conceptual necessity, then there is no need to explain it further, not any more that one needs to explain more what it is to be bachelor than to say that it is true of unmarried individuals. But this still seems a bit too good to be true.

There are a number of versions of the constitutive thesis,[7] but we can, in order to see where the problem lies, consider Burge's version of it. Rather that talking about self-knowledge of our own mental states, Burge talks of our *entitlement* to have these thoughts. I am always *entitled* to have such thoughts or beliefs about my own thoughts and beliefs. What is the source of this entitlement? Burge's answer is very close to Shoemaker's:

> [our entitlement to self-knowledge] derives not from the reliability of some causal-perceptual relation between cognition an its object. It has two other

[7] Although it can be said to have a Kantian flavour, he constitutive thesis can be traced back, in contemporary philosophy, to Wittgenstein. See Wright (1998), Heal (1994). For a recent version see Coliva (2009).

> sources. One is the role of the relevant judgments in critical reasoning. The other is a constitutive relation between the judgments and their subject-matter, or between the judgments about one's own thoughts and the judgments being true. Understanding and making such judgments is constitutively associated both with being reasonable and with getting them right.... To be capable of critical reasoning, and to be subject to rational norms necessarily associated with such reasoning, some mental acts must be *knowledgeably* reviewable. The specific character of this knowledgeable reviewability requires that it be associated with an epistemic entitlement that is distinctive... there must be a non contingent, rational relation, of a sort to be explained, between relevant first-person judgments and their subject matter or truth. (Burge 1996, p. 98)

Burge explicitly rejects here the perceptual model of self-knowledge. His claim is that the reason why we are entitled to have beliefs about our own beliefs is that a being who would not have the possibility of framing such second-order beliefs would not be able to engage in "critical reasoning".

The carrying out of a proof, for instance, presupposes the ability to reasoning of this kind. A non-critical reasoner, Burge says, would reason blind, without appreciating reasons as reasons (p.99). It follows in particular that a critical reasoner needs to have the concepts of the attitudes that he has, and needs to *commit* himself towards the contents of his attitudes. If we could not be critical reasoners in this sense, "there could be no norms of reason governing how one ought to check, weigh, overturn, confirm reasons or reasoning". And there could be no such thing as epistemic responsibility, whereby we could be able to review, reflexively our reasons.

We need not enter into the details Burge characterisation of what he calls critical reasoning to understand his claim. According to him, our capacity to have second-order thoughts, reflexive beliefs, is *required* by the very possibility of engaging in such reasoning and is, therefore, its main source. Indeed the requirement goes the other way too: the capacity to frame second-order thoughts requires the capacity to engage in critical reasoning. There is a constitutive, intrinsic relation between the two.

Burge's analysis of the source of our entitlement to self-knowledge is thus very similar to Shoemaker's analysis. On both views, the perceptual model of self-knowledge is rejected, because it makes the source of our entitlement a merely causal and contingent source. but the source is not contingent or causal: it is indeed a necessary or a logical one. And in this sense it is *a priori*. It is an *a priori* requirement for self-knowledge that we can be critical reasoners, who are able to follow rational norms. This is why we may call this the *necessary entitlement* or the *constitutive thesis*.

5 The Redeployment Account

The constitutive thesis, although it is supposed to render self-knowledge obvious, is not itself obvious. One can raise three questions.

1) One condition, as we have seen, for our capacity to self-ascribe beliefs to ourselves is that such beliefs must be available or accessible, either through some conscious assent to their contents, or because they are tacit, and at least

subject to assent *in principle*. But what are the conditions of such an accessibility of availability? What are the "C" conditions mentioned in (CT)? The constitutive thesis is silent upon this, because it seems that asking this question would imply some causal account of the availability of contents for self-ascription, and this seems to be incompatible with its a *priori* or conceptual character.

2) This immediately raises a second question: if the conditions upon which we can assent to belief-contents matter for an account of self-knowledge, and if these conditions have a causal character, how can the necessary entitlement thesis avoid the introduction of such causal elements? And if such elements are present, does this not justify partly one of the suppositions of the perceptual model, namely that the connection between the belief-forming mechanism and the beliefs that it produces is a contingent, not a necessary or logical one?
3) This question in turn suggests a third one: given that, on the necessary entitlement thesis, a subject needs to have beliefs about its own beliefs in a reflexive way, that is to have or to possess the concept of belief, what is it to possess this concept? In particular is it really necessary to possess such a concept in order to engage in the activity of reasoning?

Unless we come back to the perceptual or the inferentialist model, two sorts of strategies are open to us if we want to preserve the essentials of the constitutive thesis. Each of these consists in introducing causal or psychological elements into the account. Let us analyse one version in this section, Peacocke's (1996, 1999) and a second one in the next.

To take up Burge's vocabulary, a number of self-ascriptions are "self-verifying". For instance, if I make such self-ascriptions as: "I judge, herewith, that there are physical objects", such ascriptions are such that one cannot doubt their truth. But some self-ascriptions are not self-verifying. One is the case of ascriptions made from the existence of a memory. For instance if I ask: "What is the city to which both Garibaldi and Mussolini marched to?" and if my memory presents to me the answer: "Rome", my self-ascription of the belief: "I believe that Garibaldi and Mussolini both marched to Rome" depends upon two things: the fact that I seem to remember that it is Rome to which they marched, and the fact that I take my memory as correct. But this is not self-verifying. For my memory may be wrong. The example rests on the hypothesis that the self-ascriber is not Italian, but, say, French. It would be different with an Italian schoolboy, who knows, so to say, automatically, without attending to his memory, that it is indeed Rome to which Garibaldi and Mussolini marched to. In such cases, there is neither need of a conscious memory nor of attending to it, just as when one is asked his phone number or his name. But here too I can misrepresent the information.

Now if we think of such cases, there is no reason to deny that the self-ascriptions are made in virtue of a causal element: it is *because* memory serves up the information that Garibaldi and Mussolini marched to Rome that he can self-ascribe this belief to himself. And it is also *because* there is some automatic access to the information in question in the Italian schoolboy that he can give the same answer. If the memory were not present, and if he were not willing to take his memory as correct, the thinker could not be entitled to the self-ascription of the belief in the first

place. Therefore it seems difficult to deny that this causal route to the availability of the belief is at least an important component of the entitlement. It follows that the necessary entitlement thesis cannot be simply an *a priori* claim about the constitutive relationship between believing and believing that one does believe, and must include this causal condition as well. Or perhaps we should formulate the necessary entitlement thesis as a thesis to the effect that a thinker who has *normal* access to his beliefs should also be able to self-ascribe them to himself. But this normality condition just is the causal condition that we have mentioned. Nevertheless, it does not follow that we are led back to the perceptual model of self-ascription, and that we should conclude that there is nothing necessary nor logical in the constitutive link between belief and belief about belief. For it is still the case that once the thinker is caused to assent to P by his memory, and is willing to take his memory as correct, he is warranted in his ascription. Why? Because the content of the second-order belief "I believe that P" must be the same as the content of the first-order belief "P" that was entertained by the thinker. We can reproduce here Burge's argument: if the first-order belief 'P' is mistaken, then the 'P' featuring in the second-order "I believe that P" is mistaken as well. It is the identity of the first-level contents, and of the first-level concepts contained in those contents, which ensures the security of the second-order beliefs, to the effect that one cannot fail to have them, once one has them. And this is quite different from ordinary perception, because ordinary perception can lead to error. A perceptual experience is never sufficient for the correctness of a perceptual belief, whereas a second-order belief is always sufficient for the correctness of the self-ascription of it (i.e (ii) above is true). So the recognition of the fact that there is a causal factor in our entitlement does not imply that we come back to the perceptual model. In fact the presence of this causal factor is compatible with the necessary entitlement thesis.

This answers our second question above. What about the third question raised about the necessary entitlement thesis, whether it is necessary that we have the concept of belief in order to have genuine beliefs and in order to be able to be genuine reasoners? Burge (1996) claims that it is necessary, because otherwise we would not be able to follow any norms of reasoning, and to recognise them as such. Shoemaker is less committed to the idea of there being such norms, but, as we saw, he claims that a creature who would not have such capacities would not qualify as a rational creature in his actions, in at least a minimal sense.

But this seems too strong, or too idealised. We can grant the view that to be able to reason is to be able to asses certain rational relations among one's beliefs, that is to revise them in the light of new evidence, and to act accordingly. But does it imply that the thinker has or possesses the relevant concepts of belief, or desire, or of other propositional attitudes, as the necessary entitlement thesis seems to imply? No, for there can be more primitive forms of reasoning which do not involve the possession of such concepts. Plenty of cases of belief revision involve beliefs which need not be reflexive. One need not have *fully* the concept of belief in order to engage into such reasonings. There is no reason why creatures more primitive than adults, say infants and some animals, would be incapable of engaging into such kinds of reasonings.

All this is by no means incompatible with the necessary entitlement thesis. We can say that the source of our entitlement is the result of the combination of the

capacity for such elementary reasonings together with the capacity, in Burge's sense, to be a critical reasoner. The answer to our third question, therefore, could be given along these lines: a belief content can be available when at least trains of primitive reasoning of the kind suggested can occur. This does not exclude the rational or normative requirements upon belief adduced by Shoemaker and Burge, for the beliefs may be tacit. But if they are tacit in our sense, they must be at least accessible through some causal route. And this is why the causal or psychological element matters.

This corrects the picture given by the necessary entitlement thesis. But it still does not explain why we are entitled to our self-ascriptions of beliefs. Peacocke's answer is this. In a self ascription, the thinker must entertain, in his second-order belief, the *very same* concepts as those that he entertains in his first-order belief. For such ascriptions to be possible, the following "redeployment claim" must hold:

> The concepts (senses, modes of presentation) that feature in first-level thoughts not involving propositional attitudes are the very same concepts which feature in thoughts about the intentional contents of someone's propositional attitudes. (Peacocke 1996, p. 131).

A thinker who self-ascribes beliefs to himself must redeploy, in his second-order beliefs, the very same beliefs contents as those that he deploys at the first level of his beliefs. The redeployment claim has a semantic motivation which is familiar from the literature on propositional attitudes ascriptions, but which I shall not consider here. Its main motivation, however, is that in order for self-ascriptions to be possible, the *very same* concepts as those of first-order thoughts must be redeployed at the level of second-order thoughts. Peacocke gives us examples dealing with demonstratives, such as: (1) *I believe that that man over there is French*. Suppose also that I believe, on the basis of evidence that I have, that (2) *that man over there does not like croissants*. Given that most French like croissants, it seems that there is a sort of inconsistency in my beliefs. But if the demonstrative "that man" does not have the same sense in (2) as in (1), the inconsistency would not go through. Indeed the inconsistency is very similar to the one that we discussed earlier with Moore's paradox, for we could conjoin (1) and (2) in (3) *That man over there is French, and I do not believe that he is French (for he does not like croissants)*. But in order to see the inconsistency, the demonstrative concepts must be the same.

Peacocke's redeployment claim seems plausible, and reinforces the conclusions reached earlier: it is indeed a requirement that when I ascribe beliefs to myself, I must employ the same concepts as those that I employed when I entertained these beliefs, so to say, unreflexively. But for the very same reason, it is hard to understand why Peacocke claims that this principle is "explanatory" (1996 p. 118) and "contributes to an explanation of the near infallibility of a thinker's knowledge of the contents of his conscious beliefs" (p.147). For it is one thing to say that such a principle *has* to hold if our self-ascriptions of beliefs are to be warranted. And it is another thing to say that because it has to hold, our self-ascriptions *are* warranted. For there is something which is left unexplained, which the nature and the source of the identity of concepts. We could ask: why do we have the capacity of redeploying our thoughts so that the very same concepts are entertained? And how do we know that they are the same?

As Peacocke mentions himself (p.150), a sceptic about all this could claim that when we redeploy our first-order contents in second-order beliefs, we do not need to rely on the latter's *identity* with the first, but only upon their *similarity* with them. After all many views about propositional attitudes ascriptions rely on the view that such ascriptions can be successful even when there is only an overall similarity between the concepts of the ascriber and the concepts of the ascribee. According to a well known view, we just project the contents of our own beliefs upon the other's beliefs, and this guarantees successful ascriptions in most cases. Why could it not be the same when we self-ascribe beliefs to ourselves?

But the considerations which precede seem to cast doubt upon this possibility. If the mechanism by which we ascribe beliefs to ourselves was such a mechanism of projection or of simulation analogous to the one that is said to operate in the ascriptions of beliefs to others, we would have to ascribe beliefs to ourselves in the same way as we ascribe beliefs to others, in the third person way. But this is impossible, for the mechanism of ascription by projection supposes that we have already an access to our own beliefs. So even according to the projective method of ascription, we need some sort of self-access. If this self-access were not warranted in some way, even our ascriptions to others could not be successful. So even on a similarity of content view, we would need privileged access.

6 Self Knowledge and Transparency

(CT) is only a schema and unless we make the conditions C more precise, it is either false—since it obviously clashes with the existence of dispositional or tacit beliefs—or empty, since it does not tell us what is the source of the entitlement to self knowledge which is supposed to flow from it.[8] There is, however a well known psychological fact which is obviously related to condition (i), although it is distinct from it, and which can serve as an *explanans* of this condition. It is the feature which has been noted long ago by Gareth Evans in the following well-known passage of *The Varieties of Reference*:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be a third world war?,' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans 1982: 225)

The idea that one can discover whether one believes that *p* simply by considering whether *p* is called the *transparency of belief*. In determining what one believes one 'looks through' the belief and focuses directly on the state of affairs that the belief concerns.[9]

[8] For similar criticism of Peacocke, see Coliva 2008. She argues that either Peacocke's account is explanatorily inadequate, or it fails to explain self-knowledge , being trivial.

[9] Evqns does not use the term "transparency", which is used Moran 2000 and it has been discussed widely since (Shah 2003; Engel 2006)

But transparency can be interpreted in several ways. It can be understood as expressing the fact that having a given belief entails a form of *commitment* to it. Thus Richard Moran says:

> ...as I conceive of myself as a rational agent, my awareness of my belief is awareness of my commitment to its truth, a commitment to something that transcends any description of my psychological state. And the expression of this commitment lies in the fact that my reports on my belief are obligated to conform to the condition of transparency: that I can report on my belief about X by considering (nothing but) X itself. (Moran 2000, p. 84).

This view is close to the constitutive thesis. But in this sense, the transparency of belief is hardly explanatory of our warrant or entitlement to self-knowledge. If it is only a stipulation, on the part of the believer, that, *as a rational agent*, he ought to commit himself to his belief by believing it to be true—by holding it true—then we are again confronted with the difficulties that we have encountered with Burge's and Peacocke's accounts. In order to give to transparency some bite, we need to treat it not as a stipulation, but as a psychological *fact*, and as *a method* to yield true beliefs. This is what Byrne (2005) proposes. According to him we can treat transparency as a method, encapsulated in the following rule:

> **Bel** If *p*, believe that you believe that *p*. (Byrne 2005: 95)

Note the similarity with condition (i) in (CT). What BEL adds is in the antecedent condition: the transparency of belief is used in the transition from the (first-order) awareness of the fact that P to the second-order belief. BEL is not a definition nor a stipulation as it is the case with the constitutive view. It does not hold *a priori*, but involves a contingent fact and a method. As Byrne proposes:

> [A]s a contingent matter, trying to follow BEL will usually produce knowledge of what one believes. Venturing out on a limb—of course the matter requires more discussion—we may tentatively conclude that privileged access is thereby explained. (*ibid*., 98)

This accounts, according to him for the authority as well as the privileged character of self ascriptions of beliefs

(1) 'Roughly: beliefs about one's mental states acquired through the usual route are more likely to amount to knowledge than beliefs about others' mental states (and, more generally, beliefs about one's environment).' (*ibid*., 80)
(2) '[K]nowledge of one's mental states is *peculiar* in comparison to one's knowledge of others' minds. One has a special method or way of knowing that one believes that the cat is indoors [etc.]...' (*ibid*., 81)

It would be wrong, though, and open to obvious counterexamples, to claim that BEL is purely contingent. Dispositional or irrational beliefs do not fit that mould. But what transparency thus understood captures are features of our possession of the concept of belief. Although it is not the point of this paper to argue for this (see Shah 2003; Engel 2008), one can attach two main features to the possession of the concept of belief: that it answers a certain kind of reasons, and the fact that it obeys the norm that *one ought to believe that P if an only if P*. Transparency is a fact about our

reasons to believe that P: that P is the best reason we can have for believing that P. Of course here "that P" is elliptic for "that P is true", and the transparency of belief is the direct counterpart, in the psychological mode, of the transparency of truth itself: to say that P is true is just to say that P.

There is also a direct connection between the transparency of belief and the norm of truth: if the fact that P is our best reason to believe that P, it is because belief is the only attitude whose correctness condition is truth. The very fact that belief is in this sense "transparent" seems to account for the way in which the norm of truth regulates belief: when, in the context of asking ourselves whether P is true, we determine the answer by thinking or asserting that P, we *implicitly* follow the norm. In doing so, we need not ascent to a second-order judgement "Do I believe that P?" and even less ask ourselves "What are my best reasons to believe that P?". Our recognition of this standard of correctness for belief is tacit, not explicit. There are indeed—as I remarked above about the self-blindness argument—thinkers who are so unreflective that they might even not have this tacit recognition. Perhaps those who are in the grip of wishful thinking, or self delusive subjects in the grip of Capgras's delusion, do not have this understanding of their own beliefs. But although I cannot argue for this here, even deeply delusive believers have at least a partial understanding of this condition.[10]

The fact that the norm of truth enters as a reason for our believing that P in the kind of conscious reasoning in which we engaged when we ask ourselves whether P is true constitutes the best way of understanding how this norm can regulate—or guide, or govern—our doxastic behaviour. Of course we cannot always—and indeed in most cases we don't—reach truth for our beliefs: sometimes we have only strong evidence, or perhaps only a certain degree of subjective probability for a given belief. For instance on asking myself whether it will rain tomorrow, I may not come with the answer "Yes", or "No", but only with a "maybe". But it does not show that the norm of truth does not operate here. For even if I cannot, in such cases, determine whether my belief that it will rain is true, I need to recognize the condition that it would be correct only if it *were* true.

Now what about the troublesome cases where we do not deliberate explicitly and consciously about whether to believe that P, such as wishful thinking, self-deception and other kinds of irrational beliefs? Should we say that transparency does not apply and that these are not regulated by the norm of truth? Certainly the wishful thinker, for instance the man who believes that he is going to pass his logic exam by reading the Koran, does not care for the norm of truth and does not consider it. Neither does the man who is under the delusion that his wife has been replaced by an impostor, or that he is dead. Certainly there can be exceptions to the norm. But does it mean that these people do not have the concept of belief and that they are unable to recognise the norm? Hardly so. Even though these people obviously do not reason consciously with and from their beliefs, and do not consider norms of evidence, it is less clear

[10] I thus agree with Tim Bayne: "Do delusion and self-deception involve departures from the operating norms of belief formation?

Self-deception—at least, everyday self-deception—need involve no departure from the operating norms of belief-formation. "Delusion and self deception: mapping the terrain, in (T. Bayne and J. Fernandez (eds) *Delusion and Self-Deception: Motivational and Affective Influences on Belief-Formation* (Psychology Press).

that they have no understanding at all of what a proper belief should be. The wishful thinker is wrong when he believes that reading the Koran will help in his logic exam. But he is at least conscious of the fact that he needs a reason to believe that he will pass his exam, and even if he is wrong on the reason, he has some dim idea of what it might be. There are degrees here, obviously. The self-deceived wife may forget, or pass under silence for herself the evidence that she has that her husband is trumping her. But the very fact that she reasons to the contrary shows that she is aware of some evidence that her husband is unfaithful, and that attending to evidence is relevant to her believing. So it is not clear that the norm of truth does not in such cases regulate thinking tacitly.

The view that I have tried to defend here lies midway between the perceptual model, which construes self-knowledge as a specific cognitive achievement, and the constitutive view, which denies that it is actually knowledge. I agree with the latter that self-knowledge is a conceptual feature intrinsic to being a thinker. But I disagree with the constitutivist if his view amounts to saying that self knowledge cannot be explained psychologically. I have suggested how an account of self-knowledge based on transparency can give us not only a normative condition on self-knowledge, but also a psychological one. In this sense it shares both features of the causal account and of the constitutive account. Indeed much more needs to be done to make it a genuine explanation. But it is promising one.[11]

References

Armstrong, D. 1968. *A materialist theory of the mind*. Cambidge: Cambridge University Press.
Boghossian, P. 1989. Content and self-knowledge. *Philosophical Topics* 17(1): 5–26.
Burge, T. 1996. Our entitlement to self knowledge. *Proceedings of the Aristotelian Society* XCVI: 91–116.
Byrne, A. 2005. Introspection. *Philosophical Topics* 33: 79–104.
Coliva, A. 2008. Peacocke's self knowledge. *Ratio (new series)* XXI 1 March 2008.
Coliva, A. 2009. Self-knowledge and commitments. *Synthese* 171(3): 365–375
Dennett, D. 1991. *Consciousness explained*. New York: Little Brown.
Engel, P. 1996. *Philosophie et psychologie*. Paris: Gallimard.
Engel, P. 2006. Making up one's mind. *Cahiers Parisiens*. University of Chicago, Press.
Engel, P. 2008. Belief and normativity. *Disputatio*, Lisboa, 2008, 153–177.
Evans, G. 1982. *The varieties of reference*. Oxford: Oxford University Press.
Fernandez, J. 2003. Privileged access naturalized. *The Philosophical Quarterly* 53(2003): 352–372.
Gertler, B. 2009. Self knowledge and the transparency of belief. In *Self-knowledge*, ed. Anthony Hatzimoysis. Oxford University Press.
Heal, J. 1994. Moore's paradox, a Wittgenstein solution. *Mind* 103, No. 409, (Jan., 1994), pp. 5–24.
Lycan, W. 1986. Tacit Belief. In *Belief*, ed. R. Bogdan, 61–82. Oxford: Oxford University Press.
Moran, R. 2000. *Authority and Estrangement*. Princeton: Princeton University Press.

[11] Other accounts using the transparency feature are Fernandez 2003 and Gertler 2009. The first argues that one can go all of the way down the empirical path to explain transparency. The other is critical and rejects the transparency account. I cannot deal wit the here, but they illustrate the tensions to which this view is subject;

An ancestor of this paper was written in 1996 at the invitation of Gloria Origgi for a seminar in Bologna, which did take place. Christophe Heintz proposed me to rewrite a new version for this occasion, and it turned to be quite different from the first. I am grateful to him, to Dario Taraborelli and to the referees for their excellent comments, for their generosity and their angelic patience. The paper was writen while I was visiting fellow in the Formal Epistemology Project, KUL, Leuven

Nichols, S., and S. Stich. 2003. How to Read your own mind: A cognitive theory of self-consciousness. In *Consciousness: New Philosophical Essays*, eds. Q. Smith and A. Jokic, 157–200. Oxford: Oxford University Press.http://ruccs.rutgers.edu/ArchiveFolder/Research%20Group/Publications/Room/room.html

Nisbett, R., and A. Wilson. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84: 231–259.

Peacocke, C. 1996. Our entitlement to self knowledge. *Proceedings of the Aristotelian Society*, reprinted, in Peacocke 1999.

Peacocke, C. 1999. *Being known*. Oxford: Oxford University Press.

Rosenthal, D. (1993). Thinking that one thinks. In *Consciousness*, Davies and Humphreys. Oxford: Blackwell.

Ryle, G. 1949. *The concept of mind*. London: Hutchinson.

Shah, N. 2003. How truth regulates belief. *Philosophical Review* 113.

Shoemaker, S. 1996. *The first person perspective and other essays*. Cambridge: Cambridge University Press.

Thomasson, A. 2003. Introspection and phenomenological method. *Phenomenology and the Cognitive Sciences* 2: 239–54.

Wittgenstein, L. 1980. *Remarks on the philosophy of psychology*, 2nd ed. Oxford: Blackwell.

Wright, C. 1998. Self-knowledge: The Wittgensteinian legacy. In *Knowing our own minds*, ed. C. Wright, B. Smith, and C. Macdonald, 13–45. Oxford: Clarendon.