Privileged Access to the Mind: What It Is and How It Can Fail

Johannes L. Brandl

1. Introduction

A basic fact of our mental life is – what the phenomenological tradition calls – our "inner consciousness": the fact that we do not just have experiences and thoughts, but that we are also aware of these phenomena. As the term indicates, it is meant to denote a type of consciousness that makes us aware of what is going on "inside" our own minds, in contrast to thoughts and experiences that make us aware of what is going on in the external world around us. Inner consciousness can therefore be described as a form of reflexive, self-directed awareness.

Although it is hardly disputable that we are creatures with an inner consciousness, there is little agreement about what this fact involves. Two basic questions need to be addressed here: "What is it?" – what are the cognitive mechanisms that make us conscious of our own mental states? – and "What does it do?" – what is the specific function of this type of consciousness? Answers to these questions will have to go hand in hand. One might first look at the function of inner consciousness and then ask what mechanisms could perform that function; or one might start with a mechanism that produces inner consciousness and then see what its function is.

The answer to the second question is traditionally given in epistemological terms. The function of inner consciousness, so it is said, is to provide subjects with a *direct* or *privileged access* to their own mind. This raises a number of further questions. What does it mean to know one's own mental states "directly", and what kind of epistemological privilege does this involve? If this means more than that it is "easy" for subjects to access their own minds, it must mean that their mental self-ascriptions have a special warrant or even certainty. That does not seem to be generally true however. The privileged access to one's mind – whatever it may be – is clearly limited. It fails in the case of unconscious mental states, e.g. anxieties or hopes that we are not aware of; and it fails when we suffer from self-deception, e.g. when we mistakenly take ourselves to be rational and consistent in our beliefs, decisions, and moral attitudes. These limits to our self-knowledge pose a special difficulty for a theory of inner consciousness. It has to explain how the access to our mind is restricted, without thereby denying that inner consciousness has a peculiar epistemological function. How is this challenge to be met?

In what follows I want to address this question from the point of view of the higher-order-thought theory of consciousness. According to the HOT-theory, as it is called, the inner awareness that we have of our own mental states arises from the fact that these states are accompanied by higher-order thoughts with the content that we are currently in those states. This is, first of all, a proposal about the *mechanism* that produces inner consciousness, hence an answer to the first question mentioned above. If correct, the

HOT-theory shows that inner consciousness does not require anything like an "inner sense" or "inner experiences", comparable to the perceptual experiences that our sense organs provide us with. This has been claimed to be a major advantage of the theory. But what about the second question? What answer does the HOT-theory offer concerning the function of inner consciousness? Does it accord with the view that its function consists in giving us a privileged access to our own mental states; and if so, how does it explain that privilege and the way in which our access may fail?

In section 2 I will give a brief summary of the main theses of the HOT-theory, as it has been proposed and defended by David Rosenthal. The main part of the paper will then deal with the question of whether the HOT-theory can be combined with a certain infallibility principle. In section 3 I introduce this question and consider Rosenthal's view on this matter. In section 4 I will argue – against him – that a qualified version of the infallibility principle can be defended by an argument based on Moore's paradox. In section 5 I will answer an objection to this result and indicate some further questions that need to be resolved when one uses the HOT-theory to explain the privileged access we have to our own mental states.

2. The HOT-theory as a theory of inner consciousness

In a series of papers, starting with "Two Concepts of Consciousness" (1986), David Rosenthal has developed what is now known as the HOT-theory of consciousness.[1] His approach is very broadly conceived. Rosenthal aims at a classification and, based on this classification, a systematic explanation of all the different kinds of consciousness there are. In what follows I will take a more limited perspective and focus on just that kind of consciousness that is traditionally called "inner consciousness". I thus set aside the recently much-discussed problems pertaining to phenomenal consciousness and the objection that the HOT-theory is merely a theory of access consciousness. It may well be that the notion of "state consciousness", as Rosenthal uses this term, does not cover as much as he intends it to cover. This is not my concern here. I want to see how far his analysis of state-consciousness can take us towards an understanding of the phenomenon of inner consciousness.[2]

The term "inner consciousness" plays no role in Rosenthal's own classification of the different types of consciousness. This is surprising because it would seem that an explanation of inner consciousness is the primary target of the HOT-theory. To start

---

[1] Several of these papers will be republished in a forthcoming book by Rosenthal entitled *Consciousness and Mind* (Oxford 2004). The core ideas of the theory, to which I restrict myself here, can be found in several places, e.g. in Rosenthal 1986, 1993, 1997 and 2002a. In the meantime, other types of higher-order representational theories of consciousness have been proposed that deviate in several ways from Rosenthal's theory. For comparisons see Carruthers 2001, Lycan 2001, and Rosenthal 2002b.

[2] The terminological differences here are very confusing. Rosenthal tries to square his own terminology with Ned Block's tripartite distinction between "phenomenality", "global access", and "reflexivity" (see Rosenthal 2002b). "Reflexivity" may be another term for "inner consciousness", not to be confused with what Rosenthal calls "introspective awareness" (see Rosenthal 1986, 336f.).

with, we can characterize "inner consciousness" in Rosenthal's terminology as a form of "transitive" consciousness (see Rosenthal 1993). It is definable, in terms of the two-place predicate "*S* is conscious of *y*", as follows:

Df. *S* has inner consciousness $=_{df}$ There is a mental state *M* of *S* such that *S* is conscious of *M*.

What we need, then, is an account of the transitive verb "being conscious of". According to Rosenthal the use of this verb is governed by the following principle:

(TC) *S* is conscious of *x* if and only if there is a mental state *M* (distinct from *x*) such that *M* makes *S* conscious of *x*.

The first main thesis of the HOT-theory says that principle (TC) characterizes not only cases of "external consciousness", like perception, but applies to our inner consciousness as well. Given the above definition, we can take *x* in principle (TC) also to refer to some mental state of *S*. We then get a three-term relation between (i) the subject *S*, (ii) a mental state M_1 of which *S* is conscious, and (iii) a mental state M_2 that makes S conscious of M_1. This application of principle (TC) can be disputed, but I will not go into this matter here.[3]

Principle (TC) is closely connected with another important assumption, namely that there are no intrinsically conscious mental states. Consciousness may be a highly important feature of the mind, as Rosenthal says, but it is "not necessary or even central to a state's being a mental state" (Rosenthal 1986, 330). This means that, for all mental states *M*, *M* might occur without the subject of *M* being conscious of it. Let me call this the "independence principle":

(IND) For all mental states *x*, it is possible that *x* occurs without there being a mental state *M* that makes *S* conscious of being in *x*.

This principle appears problematic in the case of experiences like smelling a rose or feeling a toothache, but I will not pursue this problem here either. Instead I turn to the more specific claims that the HOT-theory makes. Consider first the following – still very general – statement of the core of the theory:

> We are conscious of something, on this model, when we have a thought about it. So a mental state will be conscious if it is accompanied by a thought about that state. (Rosenthal 1997, 741)

This is meant to give us a rough idea of what the HOT-theory says. Several comments and qualifications need to be added to this simple statement of the theory. First of all, a

[3] Following Brentano, one might claim that inner consciousness involves a self-reflexive structure that goes against priniple (TC). See Brentano 1874, II, ch.2.

higher-order thought that leads to inner consciousness has to be a "non-dispositional, assertoric thought to the effect that one is in that very state." (Rosenthal 2002, 410) This tells us that inner consciousness is an *occurrent* phenomenon that cannot be generated by a mere ability or disposition to reflect about one's mental states.

Another important qualification concerns the particular form of the thoughts that have to accompany our mental states to make them conscious. These cannot be just any assertoric thoughts that one might have about those states. They must be in the first-person singular and in the present tense. And finally, the following important assumption is made by the HOT-theory:

> A state is conscious if whoever is in it is to some degree aware of being in it in a way that does not rely on inference, as that is ordinarily conceived, or on some sort of sensory input. (Rosenthal 1986, 334)

Rosenthal specifies here an important constraint on the mechanism at work in the formation of higher-order thoughts. A higher-order thought should *not* be the result of an inference – at least not an inference that is consciously made by the subject in question. For instance, I might come to believe that I am annoyed because somebody tells me that I look annoyed, but I may "still feel no conscious annoyance" (Rosenthal 2002, 409).[4] Furthermore, my higher-order thought should not result from a distinct sensory experience. If I consciously smell a rose, the only experiences at work here are those that make me conscious of the rose's smell; there is no *additional* sensory input that makes me aware of experiencing that smell. This separates the HOT-theory from the "inner sense"-model of inner consciousness. (see Rosenthal 2002, 409)

This leaves Rosenthal very little to say in positive terms about the mechanism that produces inner consciousness. His account of this mechanism consists of nothing more than the following two claims:

(HOT_1) A higher-order thought that makes a subject conscious of some mental state *M* must be contemporaneous with *M*, and
(HOT_2) it must be a thought in an assertive mode.

It may seem that this is not much of a mechanism, and that the HOT-theory therefore leaves it quite mysterious how we actually become aware of our own mental states. That would be a premature judgment, however. It is not so easy to have higher-order thoughts that satisfy all the requirements just specified. Usually we come to make a judgment or form a belief by drawing inferences from other beliefs or by relying on some new sensory input. Admittedly, we can *entertain* all sorts of thoughts quite freely, but we do not get so easily into the assertive mode without being driven by inferences or by experience. The insight of the HOT-theory might be that we can do this only in forming

[4] For more on this constraint see Rosenthal 2000.

higher-order thoughts about our own mental states. The mystery of inner consciousness may well be hidden in this fact.

There are other aspects of the HOT-theory that I have not mentioned yet. For instance, it is an open question how mental states and higher-order thoughts about them must be causally related if the latter are to make one conscious of the former. I will not pursue this matter any further here. I take the principles and definitions stated above to form the core of the HOT-theory and turn now to the question of what this theory can tell us about the privileged access that we have to our own minds.

3. The question of infallibility

There are many different things one can mean by the concept of "privileged access". In her introduction to a recent collection devoted to this topic, Brie Gertler lists five basic principles connected with this concept, which she calls "infallibility", "self-intimation", "epistemic asymmetry", "epistemic privilege", and "incorrigibility" (see Gertler 2003, xii). As Gertler mentions, this is already a great simplification in comparison to William Alston's analysis of the concept of "privileged access" that offers more than a dozen different readings of this term (see Alston 1971). I will cut down the space of possible interpretations here even further and consider only one aspect of this complex notion. My concern will be to see how far a principle of infallibility can be defended that may then be used to explain some of the other aspects mentioned above, like the epistemic asymmetry and the incorrigibility of mental self-ascriptions.[5]

There is undoubtedly a difference between the way in which we know about mental phenomena occurring in our own minds and mental phenomena occurring in other minds. It is perhaps less clear, but still plausible to assume, that what subjects know best are their own experiences and thoughts. Hence we have here an epistemic privilege that extends in two directions: we are privileged in accessing our own mental states relative to *other* subjects; and our self-ascriptions are privileged relative to beliefs that we hold about our bodily states and states in the external world.

Most theories of self-knowledge accept these asymmetries. At the same time, however, they emphasize that holding beliefs with a certain privilege to be true does not mean that one is *infallible* in forming these beliefs. This accords with the sceptical tradition in epistemology according to which there is always the possibility of being in error, even when we are as certain as possible in what we believe. As a general rule this fallibilist position should hold for beliefs about our own mental states as well. And it does seem to hold in this case too, when one thinks of cases of unconscious perceptions or of confabulation. For instance, I can be completely certain that I did not see a rabbit crossing the street, and yet I must have seen it – subliminally – since otherwise I would

[5] The only principle mentioned by Gertler that needs an independent treatment is the principle of "self-intimation" which says that whenever a subject is in some mental state, she believes of herself to be in that state. I will touch on this issue in the final section of this paper.

not have jumped on the brakes. Similarly, when people confabulate memories or dreams, they can be absolutely certain that they remember experiences they never had.

These examples do not necessarily show, however, that sceptical doubts are in order in each and every single case. There may still be some exceptional beliefs, like the Cartesian *cogito*, that cannot be rationally doubted. The reason may be that these are first-personal beliefs that concern only the present state of mind of the person ascribing this state to herself. So, perhaps, inner consciousness provides us not just with an access to the mind that is privileged, but even infallible?

The conjecture that first-person, present-tense, higher-order thoughts about one's own mental states are exceptional in this way, may be expressed in the following principle of infallibility:

(INF) It is psychologically impossible for a subject *S* to mentally assert "I am now in mental state *M*", unless *S* is in fact in mental state *M*.[6]

I assume here that in mentally asserting something a subject either manifests what she already believes or acquires the corresponding belief.[7] It thus follows from the above principle that if a subject mentally asserts that she is in certain mental state, she correctly believes that she is in that state. This principle is not refuted by the cases of subliminal vision and confabulation just mentioned. In the case of the rabbit that crossed the street I do not *assert*, but *deny* that I have had certain perceptual experiences. I might even deny that I am having such an experience right now, and yet see the rabbit at a subconscious level. A conflict with principle (INF) is avoided here, if we insist that higher-order thoughts must assert the *presence*, not the *absence* of a certain mental state or event. Neither do cases of confabulated memories or dreams tell against principle (INF). In this case subjects assert that they previously had certain experiences, and may later confess that they made them up. Whether it is possible to confabulate experiences that one presently has, is a more difficult question. Here the HOT-theorist may need a different strategy, as we shall see later. The usual cases of confabulation, however, concern mental states in one's past and therefore do not provide counter-examples to the infallibility principle as stated above.

Added to the HOT-theory of inner consciousness, this principle would provide a straightforward answer to the question of what the specific function of inner consciousness is. The answer would be that it provides subjects with a set of beliefs that are beyond any doubt for them. I should emphasize, though, that what is beyond doubt for a subject are propositions that they actually *believe* to be true. As long as one has not

---

[6] It will become clear later why I formulate this principle in terms of a "psychological" impossibility. The intention is not that principle (INF) might hold only for subjects with a certain psychological constitution. Rather, the principle should hold for *all* subjects capable of ascribing mental states to themselves by mentally asserting propositions of the specified form.

[7] This leaves open the possibility that people may hold beliefs for non-epistemic reasons without being willing to mentally assert what they believe. See Lehrer 1990, 11.

formed such a belief, one may be undecided whether one currently is in a certain mental state *M* and hence also doubt whether one is in state *M*. Principle (INF) does not deny this. It merely tells us that once a subject mentally asserts that she is in state *M*, no room for doubt is left. This is not just because one cannot believe and doubt something at the same time. There are sceptical doubts that one can have about one's beliefs even while holding onto them. For instance, one can be fully convinced that it is raining outside and yet admit that one might be mistaken in this belief. It is this kind of doubt that principle (INF) is meant to exclude. Subjects have the privilege of being *certain* that the higher-order thoughts they mentally assert are actually true.

Can we integrate this infallibility principle into the HOT-theory? At first, this does not seem likely since the theory has been developed in sharp opposition to the Cartesian conception of the mind (see Rosenthal 1986). And since it is one of the most prominent claims of Descartes that our self-knowledge provides an infallible foundation for all of our knowledge, one would expect that an anti-Cartesian theory distances itself from such claims as far as possible. Rosenthal takes it to be one of the great advantages of the HOT-theory over the perceptual model of inner consciousness that it can go a step further in this direction:

> Because being conscious of something is factive, using HOTs to explain the relevant transitive consciousness may seem less plausible than a perceptual model. After all, perceiving something is also arguably factive, whereas having thought is not. This should not lead one to adopt the perceptual model, however; […]; there is good reason to doubt that the way we are conscious of our conscious mental states guarantees truth; special views about privileged access notwithstanding, we can and do make mistakes about what conscious states we are in. (Rosenthal 1997, 741)

It is not clear which "special views about privileged access" Rosenthal has in mind here. But he clearly wants to endorse the view that higher-order thoughts can be just as mistaken as most of the other beliefs we have. This is shown, he thinks, by the confabulation cases that show that "we sometimes in effect invent the mental states that we take ourselves to be in" (ibid., 744).[8]

The official view of the HOT-theory, therefore, is that the Cartesian ideal of infallibility is an illusion. Rosenthal even offers two explanations of how this illusion might arise. First, he suggests that it may be prompted by the factive use of the phrase "being conscious of something". Strictly speaking, therefore, we should not say that a higher-order thought makes one conscious of a mental state, implying that one is actually in this state, but that it *seems* to make one conscious of a state that may or may not be there. Secondly, he suggests that the illusion of infallibility arises because the difference between appearance and reality may not be accessible from the first-person point of view: "[A] case in which one has a HOT along with the mental state it is about might well be subjectively indistinguishable from a case in which the HOT occurs but not the mental state." (ibid., 744) Hence, if higher-order thoughts provide us with a kind of

[8] He elaborates this point further in Rosenthal 2000.

privileged access, this privilege cannot be explained in terms of a Cartesian infallibility principle.

We need not accept this as the final word on the matter however. Maybe what we have here is a case of an author not fully appreciating the merits of his own theory. I will now try to show that a qualified version of principle (INF) is not only perfectly coherent with the HOT-theory, but can even be explained by it.

4. An argument based on Moore's paradox

G.E. Moore once made a famous observation that can be taken as a starting point for an argument to the effect that – under certain conditions – one cannot be mistaken in ascribing mental states to oneself. The argument is not simple, and there are several sidelines to it that I will have to omit. By focusing on the core points, I hope to make the case as convincing as possible.

A first difficulty here is how to interpret Moore's paradox itself. Several proposals have been made how to explain what seems so puzzling about it. I have nothing new to add here to this ongoing discussion.[9] For my purposes I will follow the analysis that David Rosenthal has suggested. [10] What interests me about it is not so much whether it does full justice to Moore's paradox, but why Rosenthal takes it to give no support to the view that higher-order thoughts might be infallible. This seems to be a mistake. Moore's paradox does have such implications, and I will try to show this by an argument that closely follows Rosenthal's own reasoning. There may be other ways to set up such an argument following some other interpretation of the underlying paradox, but I will not consider such alternatives here.[11]

Moore's observation was that it would be absurd to say: "I believe he has gone out, but he has not". [12] This would be something absurd to say, he suggests, because the second part of this statement – "He has not gone out" – implies, in a certain sense of "implies", that one does *not* believe that he has gone out. Yet this is asserted in the first part of the statement. The observation has been called a "paradox" because there is something puzzling about it: how can it be absurd to say something if that assertion could easily be true? After all, it could easily be true that my friend is at home, although I mistakenly believe that he has gone out. In this case, both parts of the above statement would be

---

[9] For a recent proposal about how to deal with Moore's paradox that also critically examines some of the older interpretations see Lee 2001.

[10] See Rosenthal 1995. A shorter version of this paper was published in *Philosophical Studies* 77 (1995), pp. 195–209, together with comments by Roger Albritton and Sydney Shoemaker, as well as with replies by Rosenthal. For an alleged counter-example and a defense of his analysis see Rosenthal 2002c.

[11] Peter Baumann convinced me of this point. I am grateful to him for a manuscript in which such a related argument can be found.

[12] See Moore 1944, p. 204f. It does not matter in the present context whether Moore is right about what he takes his observation to reveal about Russell's theory of definite descriptions.

true: it would be true that I believe he has gone out, and it would also be true that he has not gone out. Why, then, is it absurd for me to say what might plainly be true?

Rosenthal's answer to this question relies on the difference between "expressing" and "reporting" a belief in making an assertion. On his analysis, the absurdity of a Moore-paradoxical sentence results from the fact that a subject asserting such a sentence would express a belief and at the same time report that she holds exactly the opposite belief. Thus, when I say "I believe that he has gone out, but he has not" I am performing the following two speech-acts: I am *reporting* the belief that he has gone out, and at the same time *expressing* the belief that he has not gone out. This shows, according to Rosenthal, why such sentences cannot be unproblematically asserted despite having unproblematic truth conditions. Any situation in which one could assert "I believe that *p*" is *eo ipso* a situation in which one could assert "*p*", and conversely, any situation in which one could assert "*p*" is also a situation in which one could assert "I believe that *p*". Moreover, this is something that everyone knows when he makes an assertion. It is "second nature for us", as Rosenthal puts it, that statements of the form "I believe that *p*" and "*p*" have the same performance-conditions. (see Rosenthal 1995, 321) It is therefore as incoherent to say: "I believe that *p*, but not *p*", as it is to say: "*p*, but I do not believe it". Neither of these utterances makes a coherent assertion because it conjoins statements whose assertibility conditions exclude each other.

This is, in a nutshell, Rosenthal's analysis of Moore's paradox. So far it has been left open whether a speaker actually succeeds in making an assertion when uttering a Moore-paradoxical sentence. We may take it, however, that incoherent assertions are not successful assertions at all. They are merely *attempts* by a speaker to assert something without achieving his goal. A speaker cannot actually succeed in asserting both "I believe that *p*" and "*not-p*" in a single complex assertion. This leaves us with three options: he may succeed (i) in asserting that he believes that *p*, or (ii) in asserting that *not-p*, or (iii) he may not succeed in asserting anything at all. The only plausible option here is the third one. Why should we grant a speaker success in reporting his belief that *p*, but no success in expressing his belief that *not-p*? Or conversely, why should we grant him that he has successfully expressed the belief that *not-p*, but not that he has successfully reported his belief that *p*? I conclude, therefore, that uttering a Moore-paradoxical sentence means that *no* assertion has been successfully performed. Let me call this the "no-assertion-thesis":

(NAT) In uttering a Moore-paradoxical sentence a speaker does not succeed in asserting anything at all, and hence neither reports nor expresses any of his beliefs.

On the basis of this thesis I will now try to show how to get from Moore's paradox to a defense of – a qualified version of – the infallibility principle stated earlier. If this argument is sound, it shows that the consequences of that paradox for a theory of inner consciousness are greater than Rosenthal takes them to be.

The first step of the argument should be no problem. Moore's paradox does not only arise for *utterances* that make no coherent assertions, it also arises when we try to form *beliefs* that are not coherent. This suggests that we might establish the following instance of the infallibility principle formulated earlier:

(INF*) It is psychologically impossible for a subject *S* to mentally assert "I am now believing that *p*", unless *S* is in fact believing that *p*.

Support for this principle may be gained from the following self-experiment. Consider some state of affairs that is obviously not the case at the moment, e.g. that it is raining outside if you can see the cloudless sky. And now try to mentally assert "I believe it *is* raining outside". Will you succeed? Not as long as you do not change your mind about the current weather outside. At least you have first to "forget" what you just saw when you looked out the window, namely that it is not raining outside. Unless you withdraw this belief or ignore it, you will not be able to assert the higher-order thought "I believe that it is raining". You may *entertain* that thought, but you will not be able to think it in an assertive mode, i.e. to make it a judgment.

This experiment can be varied by starting with a question you are undecided about. Suppose you are on a trip to Salzburg and have no idea what the weather is like there at the moment. You will then not succeed in mentally asserting "I believe it is currently raining in Salzburg", unless you use the term "believe" to denote a conjecture or a guess, instead of a real belief. In order to mentally assert that you believe this, you have to bring in some new considerations. It might come to your mind that Salzburg is a rainy city at this time of the year, and thus you may convince yourself that you should believe that it is raining there at the moment. That means, of course, that you have changed your mind: you are no longer undecided about what the weather is like in Salzburg at the moment.

No one actually needs to perform this kind of experiment. We know in advance that things will turn out this way. Using Rosenthal's words: it is "second nature for us" that we are unable to form higher-order beliefs that are in conflict with the first-order beliefs that we would thereby ascribe to ourselves. The explanation for this inability is the same as before: we cannot form beliefs that performance-conditionally exclude each other.

It is more difficult to see this in the case of beliefs that have been tacitly formed or innate beliefs that subjects are not aware of. For instance, it is an allegedly widespread belief that women are less gifted in specific areas, e.g. in parking cars. Suppose a subject *S* holds this belief, yet at the same time vehemently denies it. *S* might instead claim to believe that there are no gender-specific differences of that kind. Apparently we have here a counter-example to (INF*) since *S* claims to believe that there are no such gender differences, but in fact *S* is tacitly prejudiced and believes that woman are not as good as men at parking cars.

That is not a convincing counter-example, however, for two reasons: first of all, if a subject merely claims to hold a certain belief that claim may be false. This is so when it is a *verbal* utterance or when subjects say to themselves "I believe that *p*." However, saying something to oneself is not the same as making a mental assertion. As I introduced this notion above, it is meant to denote an act of belief-manifestation or belief-formation, not a mentally performed speech act that may be insincere.[13] Taking the difference between linguistic and mental assertions into account, it follows that there are no insincere mental assertions, as Shoemaker has pointed out. (see Shoemaker 1995, 215) If *S* merely *claims* to believe that there are no gender differences of this kind, but is insincere in her claim, she cannot mentally assert what she claims to believe. But as long as no such assertion has been made, we have no counter-example to principle (INF*).

Suppose, however, that *S* is sincere and actually believes what she claims to believe. Then it seems that her second-order belief may be false if *S* belongs to that group of people who are tacitly prejudiced against women without admitting it (not even to themselves). However, there is also a different explanation available. What might actually happen in this case is this: *S* has a prejudice against women for whatever reason, but has never paid attention to it. At some point *S* becomes convinced that it is a mistake to be prejudiced in this way, still not knowing that she herself holds such mistaken beliefs. Instead of giving up her discriminatory beliefs – which she is not aware of – *S* adds to her belief-system further beliefs that make her entire system *inconsistent*. *S* has actually convinced herself that women are equally qualified in the relevant respects, and she can therefore mentally assert that she believes this. But she may still not have overcome her old prejudice and – inconsistently – also believe that women are *not* equally qualified drivers. In this way the principle that a mental assertion always implies that one actually holds the corresponding belief can be saved.[14]

As this example illustrates, it is much harder than it first appears to find a conclusive case against the infallibility principle (INF*). It would have to be a case in which all of the following conditions obtain: (i) *S* must either be undecided with respect to *p*, or if she believes *not-p* she must be completely unaware of this belief; otherwise it would be psychologically impossible for her to assert – contrary to what she actually believes – "I believe that *p*". (ii) If *S* tacitly believes that *not-p*, there have to be some new considerations that might lead her to change her mind about *p*. (iii) In fact, however, *S* must not change her mind either by giving up her tacit belief that *not-p* or by inconsistently believing in addition that *p*. Rather, *S* must consistently continue to believe that *not-p* and – without being aware of this belief – mentally assert "I believe that *p*."

Although it is hard to think of an example satisfying all these conditions, the possibility of such a case remains. We can take care of this possibility, however, by slightly

---

[13] This should also resolve the problem of whether *S* might mentally assert "*p*" without understanding what "*p*" means. In performing a speech act a speaker may use words that he does not understand, but not in manifesting or forming a belief. If *S* does not know what "*p*" means, he will not be able to mentally assert "*p*" and thereby manifest or form the belief that *p*.

[14] For a different view on this matter see Lee 2001, 363.

modifying our principle. Let us add as a further condition that in mentally asserting "I believe that *p*" *S* must not draw on considerations that might lead her to change her mind with respect to *p*. This gives us the following *qualified* infallibility principle with respect to one's own beliefs:

(INF**) It is psychologically impossible for a subject to mentally assert "I am now believing that *p*" while not relying on considerations that might lead her to change her mind about *p*, unless *S* is believing that *p*.

This principle, I claim, is not only consistent with the HOT-theory of inner consciousness, but is fully explained by this theory. The apparent counter-examples discussed above are also examples in which the constraints of the HOT-theory are not satisfied, either because a subject claims to mentally assert something, but merely entertains a higher-order thought without actually asserting it; or because *S* has formed her higher-order thought on the basis of considerations that violate the constraint that these thoughts should be formed non-inferentially and without relying on additional sensory evidence.

The third, and last, step in the argument is to generalize this result. So far we have only established a specific instance of the infallibility principle pertaining to beliefs. What are we to say about higher-order thoughts about other mental states, e.g. about what we desire, what we expect, etc? Can one believe that one desires *p* or expects that *p*, without actually desiring or expecting it? I can see no insurmountable obstacle here for generalizing the above findings. Here too, *S* can mentally assert that she has a certain desire or expectation without actually having it, only if she has some information from which she can infer this higher-order belief, where this information is such that it may change her desires and expectations. Once we rule out that such extra information comes into play, the infallibility principle will be correct in these cases too. More problematic are higher-order thoughts about what we currently experience. In this case the qualification added to the infallibility principle finds no application, since experiences are not mental attitudes that can be changed in the light of new considerations. Higher-order thoughts about experiences may therefore need a different treatment.

5. An objection and a look ahead

It is time to take stock. I have argued that by drawing the right lessons from Moore's paradox, a qualified version of the principle of infallibility can be defended for higher-order thoughts. These thoughts turn out to be infallible mental assertions precisely on the conditions specified by the HOT-theory. They will be infallible if they are formed independently of any conscious inferences and additional perceptual evidence. This shows why the ability to form such thoughts is an epistemological privilege. It is the privilege of having self-directed thoughts that, if formed in the right way, are guaranteed to be true. The qualification "if formed in the right way" carries all the weight here.

It is the crucial element in the theory that also explains how our privileged access may fail. In cases of self-deception and confabulation it always turns out that the self-ascriptions involved have not been formed by the subject independently of additional information, i.e. information that has not been used in forming the mental states that she ascribes to herself.

One might therefore object that it is not much of an infallibility-thesis at all. If subjects can be mistaken about whether their higher-order thoughts satisfy the requirements of the HOT-theory – and this possibility has not been ruled out – then they may also be mistaken about what their current mental states are. As Rosenthal observes, it may only *seem* to them as if these thoughts make them conscious of certain first-level thoughts, while these mental states may in fact not be present. (see Rosenthal 1997, 744)

The objection is well taken. We are indeed not infallible with respect to the *mechanism* by which our higher-order thoughts are produced, and this places severe limits on the privileged access we have to our own minds. Limited though it may be, it suffices to explain the epistemological asymmetry that distinguishes the access that we have to our own minds and the access that we have to the mental states of others. Only in our own case can we form higher-order thoughts in such a way that these thoughts – if formed in the right way – are guaranteed to be true. Nor can we say anything similar about the beliefs we have about our bodily states or the external world.

For this reason the HOT-theory is much more than a theory about the mechanism underlying our inner consciousness. It also has something substantial to say about the epistemological advantage that subjects have in accessing their own mental states, and about the way in which this access is privileged though limited. The final verdict about the epistemological merits of the HOT-theory, however, will depend on how this theory can handle the other issues arising in this context, in particular the question of transparency or "self-intimation". We have earlier seen one crack in the theory, namely the fact that the access we have to our own experiences requires a different treatment. This indicates that some changes in the HOT-theory are in order. Considering the question of transparency may show us what the required changes might be, but this is something for another occasion.[15]

References

Alston, William (1971), "Varieties of Privileged Access", in *American Philosophical Quarterly* 8, 223–241.

Baumann, Peter (2004), "BBp → Bp", Manuscript.

Brentano, Franz (1995), *Psychology From an Empirical Standpoint*. London: Routledge and Kegan Paul.

Carruthers, Peter (2001), "Higher-order Theories of Consciousness", in *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/sum2001/entries/consciousness-higher/

[15] For valuable comments on earlier drafts I am grateful to Marian David and Thomas Spitzley.

Gertler, Brie, ed. (2003), *Privileged Access. Philosophical Accounts of Self-Knowledge*. Ashgate: Ashgate Publishing Company.
Lee, Beyong D. (2001), "Moore's Paradox and Self-ascribed Beliefs", in *Erkenntnis* 55, 359–370.
Lehrer, Keith (1990), *Theory of Knowledge*. Boulder: Westview Press.
Lycan, W. Gregory (2001), "Representational Theories of Consciousness", in *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/spr2001/entries/consciousness-representational/.
Moore, G.E. (1944), "Russell's Theory of Descriptions", in *The Philosophy of Bertrand Russell*, edited by Schilpp, P. A., New York: Tudor Publishing Company, 175–226.
Rosenthal, David M. (1986), "Two Concepts of Consciousness", *Philosophical Studies* 49, 329–359.
— (1993), "State Consciousness and Transitive Consciousness", *Consciousness and Cognition* 2, 355–363.
— (1995), "Moore's Paradox and Consciousness", *Philosophical Perspectives* 9, 313–333.
— (1997), "A Theory of Consciousness", in *The Nature of Consciousness*, edited by N, Block, O. Flanagan und G. Güzeldere, Cambridge, Mass.: The MIT Press, 729–753.
— (2000), "Consciousness, Content, and Metacognitive Judgment", in *Consciousness and Cognition*, IX, 2, part 1.
— (2002a), "Explaining Consciousness", in *Philosophy of Mind. Classical and Contemporary Readings*, edited by D. Chalmers, Oxford: Oxford University Press, 406–421.
— (2002b), "How Many Kinds of Consciousness?", *Consciousness and Cognition* 11, 653–655.
— (2002c), "Moore's Paradox and Crimmins's Case", *Analysis* 62.2, 167–171.
— (2004), "Varieties of Higher-order Thought", in *Higher-Order Theories of Consciousness*, edited by R.J. Gennaro, John Benjamins (forthcoming).
Shoemaker, Sydney (1995), "Moore's Paradox and Self-Knowledge", in *Philosophical Studies* 77, 211–228.