Chapter 10

Consciousness and Metacognition

David Rosenthal

1. Metacognition and Conscious States

Metacognition has attracted considerable attention in recent psychological research. There has been extensive investigation, for example, of our metacognitive sense that information is present in memory even when that information is not consciously available, as well as investigation of the degrees of confidence people have about recall responses and about how well something has been learned. Moreover, the deficits that occur in such conditions as blindsight and visual agnosia and in certain forms of amnesia can usefully be seen as "[d]isruptions of metacognition" (Shimamura, 1994, p. 255). These and other investigations have highlighted the importance of studying the grasp we have of our own cognitive abilities and performance.[1]

Metacognitive functioning is plainly relevant to consciousness. Consider the so-called feeling-of-knowing (FOK) judgments that subjects make about what information is present in memory, even when that information is not currently accessible to consciousness, and also subjects' judgments of learning (JOLs) about how successfully something has been mastered. The metacognitive processes involved in both kinds of case result in subjects' conscious appraisals of their own cognitive condition. This has led one leading investigator to claim that metacognition can provide a useful model for studying the introspective access we have to our own mental states.[2]

Moreover, in blindsight, visual agnosia, and related disorders various indirect tests reveal that certain informational states occur of which subjects are wholly unaware, and which they routinely deny being in.[3] In the absence of the relevant neurological impairment, however, this visual and recognitional information would occur in the form of conscious mental states. So it is reasonable to hope that understanding these neural deficits will help us understand what it is that makes the difference between conscious and nonconscious mental states.[4]

Despite all this, current research into metacognition has not had all that much to say specifically about what it is in virtue of which conscious mental states are conscious. Consider, again, feeling-of-knowing judgments, which are representative of many of the phenomena on which investigation into metacognition has recently focused. A feeling-of-knowing judgment expresses a subject's sense "that a piece of information can be retrieved from memory even though that information currently cannot be recalled" (Miner & Reder, 1994, p. 47). In the terms used by Tulving & Pearlstone (1966, p. 115), the information is available to subjects but not accessible to them. Thus, in tip-of-the-tongue experiences,[5] conscious access to a word is blocked even though one consciously feels that the word is close to conscious recall. The relevant informational state in these cases is not conscious, but at best only potentially conscious. And in general, the mental states involved in feeling-of-knowing situations are not conscious mental states.[6]

How about research into blindsight and related neural deficits? Learning just what neurological processing is absent in these cases might well point to factors that, if present, would make those states conscious states.[7] But such research is still at a highly speculative stage.

Nonetheless, it is inviting to speculate that research into metacognitive functioning will be useful in understanding what it is for mental states to be conscious. What is the source of this theoretical intuition? At bottom, I think, it is due to the suggestion that for an informational or other mental state to be conscious is for one to be conscious *of* that state. More precisely, for a mental state to be conscious is for one to be conscious *that one is in that very state*.

It will be helpful, in this connection, to look again at the feeling-of-knowing situation. When one feels a word on the tip of one's tongue or has some other feeling-of-knowing experience, one has a conscious sense that the relevant informational state is there somehow, even though the state itself is not a conscious state. How can that be? As a first approximation, it is because one is conscious of being in some state or other that would fit the relevant informational bill, but one is not conscious of the particular informational state itself. So, despite one's conscious sense that the information is present, the particular informational state is not a conscious state. In feeling-of-knowing experiences one is conscious not of the state itself, but only that there is some relevant state. This lends provisional support to the idea that a mental state's being conscious consists in one's being conscious *of* that state.

As it stands, however, this is not quite right. On this account, the relevant informational states in feeling-of-knowing experiences fail to be conscious because one fails to be conscious *of* those states. But that is not what actually happens. Having a conscious sense that information is there somehow is one way of being conscious of the relevant informational state. One is conscious that one is in a state that bears the relevant information and, hence, conscious *of* that state as a state that bears that information. One is not, however, conscious of the state in respect of the information it bears.

We must therefore refine our first approximation. What matters is not just *that* one is conscious of the informational state, but *how* one is conscious of it. In feeling-of-knowing experiences, one is conscious of the target state as being a state that bears the relevant information, but not conscious of it in virtue of the actual information itself. Evidently what matters to a state's being conscious is not simply whether one is in some way conscious *of* the state, but whether one is conscious of the state in a way that gives one conscious access to its informational content. And for that to happen, one must be conscious that one is in a state *with that very information*. Informational states, after all, are individuated mainly in virtue of their content; so, if one is conscious of a state but not in respect of its content, one will have no subjective sense of that very state's being conscious. For a state to be conscious, one must be conscious of the state in virtue of its particular informational properties.

This fits well with the striking sense we have in feeling-of-knowing experiences that we are, somehow, both conscious and not conscious of the relevant informational state. As William James usefully notes, in tip-of-the-tongue experiences, the "gap" in our consciousness "is intensely active" (James, 1890/1950, p. 251). There is a vivid conscious difference between having one word on the tip of one's tongue and having another. We are conscious *of* the informational state somehow without being conscious of the information it bears.[8]

An analogy will help capture the oddness of such experiences. Somebody who knows that Mark Twain's real name is 'Samuel Clemens' could correctly be described as knowing *what* Twain's real name is; somebody who knows that Scott wrote *Waverly* could be truly said to know *who* wrote those novels. When we describe somebody's knowledge by way of a 'wh' complement – a clause governed by 'what', 'who', 'how', 'when', 'where', and the like – we abstract from the full content of that knowledge. We specify the knowledge only in terms of some question to which the knowledge would provide an answer.

In ordinary situations, we can truly describe somebody as knowing 'wh' – that is, as having knowledge specified with a 'wh' complement – only if the person has the relevant knowledge specified with a 'that' clause or its grammatical equivalent.[9] Feeling-of-knowing experiences occur precisely when the relevant knowing *that* isn't conscious. Suppose I have Twain's real name on the tip of my tongue; I have the feeling of knowing *what* that name is without, however, knowing consciously *that* his real name is 'Clemens'.

One sometimes has partial conscious access to the relevant information; one knows, say, that Twain's real name begins with 'c'.[10] One is conscious of the relevant informational state in respect of part of content, but not all. That suffices for the state to be conscious, though not conscious in respect of all its mental properties. And that corresponds to the way we experience such states. The way we are conscious of our conscious states in respect of of some, but not all, of their mental properties will be discussed at more length in section 5, below.

2. Metacognition and Higher-Order-Thoughts

We have seen that certain considerations pertaining to metacognitive phenomena make it inviting to hold that a mental state's being conscious is a matter of one's being conscious *of* that state. Close attention to feeling-of-knowing experiences, however, points to an even more fine-grained formulation. A mental state's being conscious consists not just in one's being conscious *of* that state, but in one's being conscious of it *under some relevant description*.

The requirement that one be conscious of the state under a relevant description has important consequences. There are two general ways in which we are conscious of things. One way is to see or hear the thing, or sense it in respect of some other sensory modality. Theorists who have recognized that a mental state's being conscious consists in one's being conscious of that state have almost invariably explained the particular way we are conscious of our conscious states in terms of our sensing or perceiving them.

There are several reasons why a perceptual model is attractive. Perhaps the most important has to do with the intuitive immediacy and spontaneity that characterizes the way we are conscious of our conscious mental states. When one has a conscious thought, perception, or feeling, one seems, from a first-person point of view, to be conscious of that state in a way that is entirely spontaneous and unmediated. There seems to be no reason for us to be conscious of the state, and no antecedent cause; it's simply the case that we are conscious of it. Moreover, nothing seems to mediate between those states and our awareness of them. So it may be tempting to hold that our awareness of our conscious states resembles in these respects the way we are aware of objects we perceive. From an pretheoretic, intuitive point of view, nothing seems to mediate between those objects and our perceptual awareness of them, and such awareness seems to arise spontaneously. Doubtless these analogies have done much to make an "inner sense" model of the way we are conscious of our conscious states seem inviting.

Sensing is not the only way of being conscious of things; we are also conscious of things when we think about them. Still, it may be tempting to think that sensing is the only way of being conscious of things that allows us to explain the intuitive spontaneity and immediacy characteristic of the way we are conscious of our conscious mental states.

Consider having a thought about some particular object, say, a particular house one saw yesterday. Having such a thought makes one conscious of that house. Suppose that the visually accessible properties of the house are insufficient to identify it uniquely; all that identifies it for one as a unique house is one's experience of seeing it. Perhaps, then, having thought about that very house requires that one have some relevant sensory image of it. One needn't, of course, actually sense the house to have a thought about it; but perhaps some relevant sensory content must figure in imagination for one's thought about that very house. And this may be the way it is with many cases of having thoughts about unique objects. If so, then at least in those cases, sensory content would always intervene between one's thoughts and the individual objects they are about. Since those thoughts would then be accompanied by suitable conscious sensory or imaginative states, they would presumably not seem to arise spontaneously.

This Aristotelian thesis about thinking and sensing[11] applies directly to the case at hand, since the conscious states we are conscious of are individuals. They are not individual objects, of course, but they are individual states. So if sensory content is often required for thoughts to be about unique individuals, we could not explain the intuitive immediacy and spontaneity of the way we are conscious of our conscious mental state by hypothesizing that we have thoughts about those states. We would instead need to adopt a sensory model of such consciousness.

But sensing is not in general necessary for thoughts to be about individuals. Individuation of the objects of thought takes place in many ways. In some cases we doubtless do pick out the objects of our thoughts by way of sensory content that pertains to those objects. But we also individuate objects of thought by describing them uniquely. Sometimes we describe them in terms of their unique relations to other individuals, and sometimes just as the unique individual satisfying some property at some particular time. Perhaps individuation of objects could not generally occur unless we individuated some objects by our sensing them. But even if that is so, it does not show that all individuals are picked out that way.

If objects of thought need not all be individuated by sensing them, it is hardly plausible that we need to individuate our own mental states that way. For one thing, it is natural to think of picking out mental

states uniquely by reference to the individuating mental properties they instantiate during a particular time span. This type of procedure will work more successfully with one's own mental states than with physical objects, because restricting attention to one's own mental states cuts down the range of competing items with similar identifying properties. And that method of individuation aside, we may also be able sometimes to individuate mental states by reference to nonmental individuals. Even if I need to individuate a house by sensing it, once the house is picked out I can then individuate the thought I have at a particular time that that very house is large.

There are, in any case, other compelling reasons to doubt that we individuate, or are conscious of, our conscious states by sensing them. Sensing characteristically proceeds by way of a sensory modality dedicated to discerning the presence of a range of sensible properties, say, color and visible shape in the case of vision. But there is no sense organ dedicated to discerning our own mental states and no distinguishing range of sensible properties that those states exhibit.[12] Moreover, sense organs characteristically respond to whatever relevant properties fall within their sensory range. So, if the operation of some such organ were responsible for our being conscious of our conscious states, why would that organ respond only to some of those states and not to others? Why wouldn't all our mental states be conscious, instead of only some? Since we have no independent reason to think that any such sense organ actually exists, any answer to this question would very likely be unconvincing and ad hoc.

If the way we are conscious of our conscious states is not by sensing them, the only alternative is that we are conscious of those states by having thoughts about them. Can such a model square with our intuitive sense that the way we are conscious of our conscious states is immediate and spontaneous? Plainly it can. Recall that our consciousness of our conscious states need not actually be unmediated and spontaneous; it need only seem that way, from a first-person point of view. And thoughts do occur in ways that seem, from a first-person point of view, to be spontaneous and unmediated. All that's necessary for that is that we not be conscious of anything that causes those thoughts or otherwise mediates between them and the things they are about. So there must be no conscious inferences that lead to such thoughts, that is, no inferences of which we are conscious. And inference aside, we must also be unaware of anything else as mediating or causing those thoughts. For example, if sensory mediation of the sort discussed earlier did occur, we must not be conscious of it.

This model of the way we are conscious of our conscious mental states is the *higher-order-thought hypothesis* that I have defended elsewhere.[13] One is conscious of one's conscious states because every such

state is accompanied by a higher-order-thought (HOT) to the effect that one is in that state.

Such HOTs cannot be dispositions to have the thought in question, since being disposed to have a thought about something doesn't make one conscious of that thing. And they must exhibit an assertoric mental attitude, since nonassertoric thoughts also don't make one conscious of things. These conditions rule out certain putative counterexamples, such as higher-order memories or the higher-order cognitive processing, which occur even when the lower-order mental states they are about aren't conscious states.[14] But memories are not stored as occurrent, assertoric intentional states, and higher-order processing will not exhibit an assertoric mental attitude.

HOTs need not themselves be conscious thoughts. Having a thought makes one conscious of the thing the thought is about even when we are not at all aware of having that thought – that is, even when the thought is not a conscious thought. This helps with our explanatory task. If we are conscious of a mental state by having a nonconscious thought about it, it's clear that our being conscious of that state will seem to be both unmediated and spontaneous. If we are unaware even of having the thought in question, how could it seem to us that anything causes it or mediates between it and the state it is about? More precisely, whenever one becomes conscious that one is conscious of a conscious state, the way one is conscious of that state will seem unmediated and spontaneous. Because one isn't ordinarily conscious of the HOTs in virtue of which one is conscious of those states, one won't normally think anything about how it is that one is conscious of them.

More important for present purposes, the HOT model helps with the explanatory task that arose in connection with certain metacognitive phenomena. In a feeling-of-knowing experience, one is conscious *of* a certain cognitive state even though that state is not a conscious state. The reason, I argued, is that one is conscious of the state in a way that fails to give one conscious access to its informational content. An informational state will not be a conscious state unless one is conscious of it in a way that gives one such access. And that means being conscious of the state under some relevant description.

But the mental state in virtue of which one is conscious of something under a description must have intentional content. One can be conscious of something under a description only if one is conscious of that thing in virtue of having a thought about it. So only the HOT model of the way we are conscious of our conscious states can do justice to the characteristic way in which we both are and are not conscious of informational states in feeling-of-knowing experiences. Consideration of these metacognitive phenomena helps us decide on the right model for explaining what it is for a mental state to be a conscious state.

3. The Transitivity Principle

I urged at the outset that metacognition is relevant to consciousness because what it is for a mental state to be conscious is at bottom a matter of one's being conscious of that state. This account is not a definition, but rather a hypothesis about what it is for mental states to be conscious. But even so, one might object, as Alvin Goldman has, that this idea is circular. We cannot explain a mental state's being conscious by reference to one's being conscious of that state, since in effect that is appealing to consciousness to explain consciousness (Goldman, 1993, p. 366).

But no circularity occurs here. A mental state's being conscious – what I have elsewhere called *state consciousness* – is distinct from the manifestly relational property of a person's being conscious *of* something or other – what we can call *transitive consciousness*. We are transitively conscious of things when we sense them or have thoughts about them. And we understand such transitive consciousness independently of understanding state consciousness; otherwise we couldn't even entertain the possibility of conscious creatures' having thoughts and other mental states that aren't conscious.

My claim, then, is that we can explain state consciousness in terms of one's being transitively conscious of that state. I shall refer to this claim as *the transitivity principle*. As intuitively obvious as this idea may seem, there have been several objections lodged against it in the recent literature.

One of the more interesting of these objections derives from a second threat of circularity to which Goldman has usefully called attention. Goldman considers the view that believing oneself to be in a particular mental state is "criterial" (Goldman, 1996, p. 8) for that state's being conscious. This is a version of the transitivity principle; a state is conscious if one is conscious of that state by believing that one is in that state. But, as Goldman points out, a state is not conscious if one thinks one is in it solely because one is taking somebody else's word for it. So we must distinguish cases in which the mental state one thinks one is in is conscious from cases in which it is not.

We can rule out the counterexample by appeal to the requirement that there be no conscious mediation between a conscious state and one's thought about that state, that is, no mediation of which one is conscious. If I believe that I am in some mental state only because you tell me that and I take your word for it, I am conscious of your statement as mediating between my mental state and my thought that I am in that state.

Goldman regards this way of avoiding circularity as falling into circularity at yet another point. The requirement that one's HOT be noninferential and nonobservational, he urges, amounts to stipulating that one's HOT be *introspective* (Goldman, 1996, p. 14). And introspective HOTs are simply HOTs that are about conscious mental states.

But a HOT's being noninferential in the relevant way is a matter only of that HOT's not being based on any inference of which we are conscious. And that condition is not circular, since it mentions only the transitive consciousness of such inferences. Nor is a thought's being noninferential in this way the same as its being introspective. A thought need not be about mental states at all to arise independently of any inference or observation of which one is conscious. Not all spontaneous thoughts are introspective thoughts.

Indeed, Goldman's suggestion here seems to get things reversed. We are introspectively conscious of our mental states when we are conscious of those states, and conscious that we are. But when I am conscious of some mental state I am in solely because I take somebody else's word for it, my HOT that I am in that state will very likely be a conscious thought. By contrast, when my HOT is independent of others' remarks and similar considerations, that HOT typically fails, itself, to be conscious. The HOTs in virtue of which our mental states are sometimes conscious states are seldom introspective HOTs.

Fred Dretske has developed an ingenious argument against the transitivity principle. Consider a scene consisting of 10 trees, and another just like it but with one tree missing. And suppose that you consciously see first one scene and then the other, and that when you do, you consciously see all the trees in each scene. But suppose that despite all this you notice no difference between the two scenes. This sort of thing happens all the time, for example, when one scene is a slightly later version of the other but altered in some small, unnoticed way.

We may assume, with Dretske, that in such a case you will have conscious experiences of both scenes, including all the trees in each. Moreover, there will be some part of the conscious experience of 10 trees that is not part of the conscious experience of 9 trees. That part is itself a conscious experience – it is a conscious experience of a tree. But, because you notice no difference between the scenes, you are not transitively conscious of the difference between them. Dretske concludes from this that you will not be transitively conscious of the experience of the extra tree. And that would undermine the transitivity principle; the experience of the extra tree would be a conscious experience of which you are not transitively conscious.[15]

But Dretske's argument is not sound. As we saw with feeling-of-knowing experiences, one can be conscious of a mental state in one respect and yet not conscious of it in another. One may, for example, be conscious of a visual experience as an experience of a blurry patch, but not as an experience of a particular kind of object. Similarly, one could be conscious of the experience of the extra tree as an experience of a tree, or even just as part of one's overall experience, without thereby being in any way conscious of it as the thing that makes the difference between

the experiences of the two scenes. This is presumably just what happens in the case Dretske constructs.[16] Dretske's case does not run counter to the transitivity principle.[17]

Ned Block has recently sought, in effect, to split the difference between the transitivity principle and its opponents. According to Block, the term 'conscious', as applied to mental states, is ambiguous as between two distinct properties. One is the property a mental state has when there is something it's like for one to be in that state; Block calls this property *phenomenal consciousness*. A state has the other property when, in Block's words, its content is "poised to be used as a premise in reasoning ... [and] for [the] *rational* control of action and ... speech" (Block, 1995, p. 231, emphasis in the original).[18] This second property he calls *access consciousness*. And he maintains that the two properties are, conceptually at least, independent. If so, there would be no single property of state consciousness.

Block's distinction has considerable intuitive appeal. The concept of access consciousness is meant to capture the intuitive idea that a mental state's being conscious is a matter of having some conscious access to that state. A metacognitive model will very likely be helpful in understanding that sort of consciousness.

But there is also a resilient intuition that consciousness has something specifically to do with the qualitative character of bodily and perceptual sensations. It is that property which Block's concept of phenomenal consciousness is meant to capture. And, because qualitative character is presumably intrinsic to sensory states, Block urges that phenomenal consciousness is not a matter of our having access to those states. If he is right, a metacognitive model cannot help here. Moreover, if the property of phenomenal consciousness is intrinsic to sensory states, the transitivity principle will fail for that kind of consciousness.

Many theorists maintain that the qualitative character of sensory states cannot occur without our having conscious access to it. Elsewhere I have argued against that doctrine (Rosenthal, 1986a, sec. 3; 1991, sec 1; 1993b, pp. 357–358; 1997b, pp. 732–733). Sensory states occur in subliminal perception, peripheral vision, and blindsight, and those sensations are not conscious in any intuitive way whatever. Moreover, mundane aches and pains that last all day may be felt only intermittently, and an ache or pain that isn't felt does not count intuitively as being conscious.

Block's notion of phenomenal consciousness is meant to capture the idea of a state's having some intrinsic qualitative character. But unless one has conscious access to one's sensory states, none of the properties of these states has any connection with consciousness, intuitively understood.[19] It is precisely because such access is absent for the sensory states in blindsight and subliminal perception that we refuse to count those states as conscious.

Block seeks to avoid this conclusion by defining phenomenal consciousness in terms of there being something it's like to be in our sensory states. After all, whenever there is something it's like for one to be in a state, that state is plainly a conscious state. Moreover, since the various sensations in blindsight and subliminal perception differ in qualitative character, won't they differ also in respect of what it 's like to have them?

Not in any sense of the phrase 'what it's like' that has any bearing on consciousness. When one lacks conscious access to a state, there is literally nothing it's like for one to be in that state. Without access to a state one has no first-person perspective on it, and so there is nothing it 's like to be in it. As Thomas Nagel has insisted, what matters for consciousness is that there be something it's like *"for* the organism" (Nagel, 1979, p. 166). And there will be something it's like for the organism only if the organism has conscious access to the relevant state. Block's phenomenal consciousness is not a kind of consciousness at all unless it involves one's having access to the sensory states in question.[20]

Block distinguishes a third concept of consciousness, which he calls reflective consciousness (review of Dennett, p. 182) or monitoring consciousness ("On a Confusion," [1995], p. 235). A state is conscious in this way, according to Block, if one has a HOT about it. But the states he counts as being conscious in this reflective or monitoring way are states we are *introspectively* conscious of: states that we are conscious that we are conscious of. Block is right, therefore, to classify this as a distinct kind of consciousness. But he is mistaken to define it simply in terms of the having of HOTs. For a state to have monitoring consciousness, in his sense, it must be accompanied not just by a HOT, but by a *conscious* HOT.

Block, Dretske, and Goldman all seek to explain why they find the transitivity principle unconvincing by casting doubt on the power of higher-order states of whatever sort to make mental states they are about conscious. How could being conscious of a mental state make that state conscious when being conscious of a stone does not make the stone conscious?[21] In Block's version, why should being conscious of a mental state make it conscious when being conscious of a state of the liver does not (Block, 1994)?

This very question, however, embodies a question-begging assumption. Being conscious of a mental state results in no change in that state's intrinsic properties, any more than being conscious of a rock or a state of one's liver changes anything intrinsic to the rock or the state of the liver. But a state's being conscious, on the transitivity principle, is not an intrinsic property of that state, but a relational property. Perhaps Goldman is right that "[o]ur ordinary understanding of awareness or consciousness seems to reside in features that conscious states have in themselves, not in relations they bear to other states" (1993, p. 367). But

it is well-known that common sense is often a highly unreliable guide about whether the properties of things are relational.[22] The rock objection is simply a vivid way of expressing the conviction shared by Dretske and Goldman that a mental state's being conscious is not relational.

Block's adaptation of the objection to states of the liver avoids the categorial disparity between states and objects. The question whether a rock is conscious is parallel not to whether a mental state is conscious, but to whether a person or other creature is conscious. It is plain that the property of a creature's being conscious – what I have elsewhere called *creature consciousness* (Rosenthal, 1993b, p. 355; 1997b, p. 729) – is distinct from the property of a mental state's being conscious, since a conscious creature can be in mental states that aren't conscious.[23] This is why the objection is more vivid when cast in terms of rocks. Whatever we say about state consciousness, creature consciousness is plainly not relational. It consists simply in a creature's being awake and responsive to sensory input.

Still, if state consciousness is relational, why wouldn't states of the liver be conscious if we were conscious of them in a way that seems unmediated? Since such seemingly unmediated access to states of our livers never occurs, it may not be entirely clear how our intuitions would go. But there is reason to think that we would not count such states as conscious. Suppose, only for the sake of illustration, that a particular mental-state token is identical with a particular brain-state token. And suppose that we follow the transitivity principle, and say that this mental state's being conscious consists in one's being conscious of it in a suitable way. Still, if one were conscious of that state solely by being conscious of being in a particular brain-state token, even in a seemingly unmediated way, we would not count that state as a conscious state. Conscious states are mental states we are conscious of *in respect of some mental properties*.[24]

4. Young Children and Metacognitive Development

The transitivity principle to one side, Dretske has appealed to certain metacognitive studies to argue specifically against the HOT model of state consciousness. Developmental work by John Flavell (1988), Alison Gopnik (1993), Josef Perner (1991), and Henry Wellman (1990) is sometimes taken to show that, before the age of about three years, children do not describe themselves as believing or experiencing things. Although they apply such words as 'think' and 'believe' to themselves and to others, there is reason to hold that these children do not mean what we mean by these words, since they tend not to distinguish what a person believes from what is actually the case.

In one well-known study (Perner, Leekam, & Wimmer, 1987), three-year-olds who saw a candy box opened to reveal pencils inside said that others who saw the box still closed would also believe it had pencils in it. These children apparently attribute beliefs when they take the beliefs to be true. Even more striking, this false-belief task yields the same sorts of result when the children apply it to themselves. Three-year-olds also say of themselves that, when they first saw the closed box, they believed that it contained pencils, although they had actually described themselves then as believing it contained candy (Gopnik & Astington, 1988; also Moore, Pure, & Furrow, 1990).

In many such studies, children three and younger elide the difference between what is actually the case and what they or others believe. So they ascribe to others and to their own past selves only beliefs whose content matches the way they currently take the world to be. They regard a person's believing something as always corresponding to that thing's actually being the case; saying how the world is believed to be does not, for them, differ significantly from just saying how the world is.[25] Since they ascribe no states whose content diverges from how they take things to be, the remarks these children make show little or no evidence of having the concept of a state that might so diverge.

Since content is all that matters to the way these children attribute beliefs to themselves and to others, it's tempting to conclude that they think of these states as involving only content; perhaps no mental attitude figures in their conception of these states. This divergence from the way adults and, indeed, slightly older children speak and think about these states encourages the idea that these children three and younger may simply not have the concepts of beliefs, thoughts, and experiences. Their use of such words as 'think', 'believe', and 'experience' does not express the ordinary folk-psychological concepts of these states.

Nonetheless, we can reasonably assume that these children are in many mental states that are conscious. And this, Dretske argues, causes difficulty for the HOT hypothesis, which holds that the mental states these children are in will be conscious only if they have HOTs about those states. And it is arguable that having such HOTs requires one to have concepts of the relevant types of state (Dretske, 1995, pp. 110–111).[26]

There is serious question about whether the data do show that these young children lack the relevant concepts, and question even about what concepts must figure in HOTs for these states to be conscious. But before getting to that, it's worth noting that children three and younger might in any case have HOTs about their mental states without ever expressing those HOTs in words. Even adults have thoughts they cannot express verbally, and doubtless children at various stages of cognitive development have many concepts that are not yet expressed in speech.

So these young children might have HOTs that draw the relevant distinctions even though their speech does not reveal any such HOTs.

One might simply reject the idea that these children have thoughts they never express in speech. The automatic, effortless way adults report their conscious states may make it seem that if one can speak at all, the only explanation for an inability to report one's conscious states must be that one lacks the requisite concepts. How could any language-using creature with the relevant concepts fail to be able to report its conscious states? [27]

We do readily express in words all the thoughts we are conscious of having, but not thoughts that are not conscious. And it is possible that some thoughts not only fail to be conscious but, given the psychological resources of these children, could not come to be conscious. These thoughts might well not be available for verbal expression. Indeed, this is the standard assumption for intentional states that are inaccessible to consciousness, for example, many states posited by cognitive and clinical psychologists. In any case, extrapolating from the adult situation is an unreliable way to determine what holds for young children. Much about early psychological development is still unclear, but it's plain that cognitive functioning in young children is in many ways strikingly different from that of adults. Since the inability of young children to talk about thoughts and experiences may well be due to some developmental factor that prevents them from expressing the relevant concepts in speech, we cannot conclude that concepts not expressed in the speech of these children also fail to figure in their mental lives.

These methodological caveats notwithstanding, the failure of children three and younger to use ordinary folk-psychological concepts of intentional states in speech does constitute some evidence that they may lack these concepts altogether. And that might create a problem for the HOT hypothesis. In feeling-of-knowing experiences, we saw, we are conscious *of* an informational state even though that state is not at that time conscious. The explanation was that we are conscious of these states in respect of the answers they would provide to specific questions, but not in respect of the specific informational content of the states. And that won't result in a state's being conscious. For a state to be conscious one must be conscious of it *as having* some particular informational content.

Content, however, is not the only mental property that characterizes intentional states; such states also exhibit some mental attitude that one has toward the content in question. So the question arises whether, in addition to having to be conscious of a state's specific intentional content for that state to be conscious, one must perhaps also be conscious of the state's mental attitude. If one did, a lack of any concepts for such attitudes would preclude one's forming HOTs of the requisite sort.

There is, moreover, good reason to think that this is so; a state will not be conscious unless one is conscious *of* that state in respect of its mental attitude as well as its content. If the way one is conscious of a state characterizes it only in terms of some content, that will not differ subjectively from one's being conscious of only of a particular state type, rather than any individual token. A state will not be conscious unless one is conscious of oneself as being in that state. When an intentional state is in question, one must be conscious of oneself as holding a particular mental attitude toward some intentional content.[28]

But one can be conscious of the mental attitude an intentional state exhibits without being conscious of that attitude in adult folk-psychological terms. And that would suffice for a state to be conscious, just so long as one is conscious of the intentional states as individual state tokens belonging to oneself. Even if three-year-olds have a concept of thoughts and beliefs on which the content of such states is always true, that weaker concept will suffice for their intentional states to be conscious. Even HOTs that represent target states as always having true content would enable these children to be conscious of themselves as being in the relevant state tokens. Their states would be conscious states.[29]

Indeed, there is compelling evidence that these children do conceive of the thoughts and beliefs they and others have as involving some mental attitude. They distinguish between one person's believing something and another person's believing the same thing. Beliefs are not for these children mere abstract contents; they are states of individual people. And that can be so only if they think of such states as involving some connection between an individual person and a content. Furthermore, even two-year-old children evidently distinguish believing from desiring, and understand that desires are not always fulfilled (Astington & Gopnik, 1991).[30] So even if the beliefs these children tend to attribute have contents they take to be true, they plainly conceive of believing as a type of mental attitude distinct from desiring.

I have assumed that the concepts children three and younger may have of intentional states may differ from the ordinary folk-psychological concepts of these states in order to show that even this would cause no difficulty for the HOT hypothesis. But the available evidence does not, in any case, establish that these children's concepts do differ from adult concepts. Even assuming that the children's speech accurately reflects the content of their thoughts, differences in how they attribute beliefs might be due not to their concepts of these states, but instead to the beliefs they have about these states. It might be that these children have just the concepts we do, but think different things about the intentional states they take themselves and others to be in.

Indeed, Jerry Fodor (1992) has developed a related hypothesis for explaining the relevant data, on which the children three and younger,

though having the same concepts, differ in the way they attribute beliefs because their computational capacity is more limited than ours. In effect, it's more difficult for them to figure things out and work out the consequences of their thoughts. Though their concept of beliefs allow them to recognize and understand the occurrence of false as well as true beliefs, limitations in their computational resources lead them to take shortcuts in what behavior they predict and what beliefs they ascribe.

When unique predictions of behavior result from relying solely on desire and ignoring beliefs, these children operate in that way. Their thus disregarding beliefs would be indistinguishable from their simply assuming that beliefs match the way things actually are. On Fodor's hypothesis, moreover, these children assume that a person's beliefs are true when that assumption results in unique predictions about how things appear to the person. These practices impair predictive accuracy somewhat, but they represent an acceptable tradeoff given the children's limited computational abilities. The children end up believing that most or all of the beliefs that they and others have match what is true, but they do not believe this because they have a different concept of belief. On this account the enhanced computational resources of four-year-olds results in their increasingly taking into account the possibility of believing falsely.[31]

If Fodor's hypothesis is correct, the HOTs in virtue of which three-year-olds are conscious of their conscious states would involve the very same concepts as the HOTs of adults, whereas on the more standard hypothesis actual change in concepts occurs.[32] But it is unlikely that it matters which hypothesis is correct insofar as we are concerned with the way three-year-olds are conscious of their conscious states. How one is conscious of one's conscious states is a function not only of the concepts that figure in the relevant HOTs but of how, independently of these concepts, one thinks about states of the relevant type. On both hypotheses, three-year-olds will think of the beliefs they and others have as having content that largely or always matches what is actually true. This way of thinking about beliefs will in either case dominate what it's like for these children to have conscious beliefs.

Older children, of roughly four to five, also differ from adults in how they describe themselves cognitively. Unlike children of two or three, the beliefs these preschoolers ascribe to themselves and others do diverge in content from the way the preschoolers take things to be. In this respect, these children exhibit adult concepts of these states and think about those states the same way. But work by John Flavell, Frances Green, and Eleanor Flavell reveals dramatic differences between the metacognitive abilities of these children and those of adults (e.g., Flavell, Green & Flavell, 1993; 1995a; 1995b; & Flavell, 1993).

For one thing, these children seem to lack any notion of a continuous stream of consciousness. They think, for example, that people while

awake may go for considerable stretches without thinking or feeling anything whatever. More striking, these children believe that people, while completely asleep, still think, hear, and feel things, and that they know at that time that they do. And they judge what a person is attending to and thinking about solely on the basis of immediate behavioral cues and environmental stimulation.

Because these preschoolers describe people while awake as going for periods of time without thinking or feeling anything at all, perhaps these children simply lack the ability to introspect their own ongoing stream of consciousness. If they could introspect, would they not know that thinking and feeling in the waking state is pretty much continuous?

But there is another explanation of these striking results. Children three and younger think about intentional states in somewhat different terms from adults, whether because they have different concepts of these states or different beliefs about them. Similarly, it may be that the way these older preschoolers describe thoughts and feelings reflects some different concepts they have of these states or at least a different rage of beliefs about them. The adult conceptions[33] of thoughts and feelings represent them as kinds of state that are often conscious, perhaps even, on some views, always conscious. Evidently these preschoolers do not think of thoughts and feelings in that way. Instead, their conception of these states is cast solely in terms of the states' content properties, the type of state in question, and the characteristic circumstances in which that state typically occurs.

Positing this difference in the conception these preschoolers have of thoughts and experiences helps explain the data. If these children do not think of thoughts and experiences as often or characteristically conscious states, it would be natural for them to regard people as being in such states even when they are asleep.[34] They would also see people as thinking or experiencing things only when some specific environmental or behavioral event points to the occurrence of a particular thought or experience. Presumably these children do think, feel, and experience things consciously; so there is something that it's like for them to think, feel, and experience. It is just that they do not yet think of themselves as being in conscious versions of these states, nor of the states themselves as possibly being conscious. They do not think of there being something it's like for them to be in conscious intentional states.[35]

Again, the HOT model fits well with this explanation. For a state to be conscious, one must be noninferentially conscious of being in that state. And one must be conscious of the state in respect of its content and mental attitude, though perhaps not in respect of a full conception of that content and attitude. But a state can be conscious without one's thinking of it as a conscious state, or even thinking of it as the type of state that could be conscious.

Indeed, one can readily see on the HOT model why one might not, at an early stage of development, think of one's conscious states as being conscious. In adults, HOTs are sometimes themselves conscious. When they are, one is conscious not only of one's thought or feeling, but also of having a HOT about that thought or feeling. Since a state's being conscious consists in having a HOT about it, when a HOT is conscious, one in effect is thinking about the mental state one's HOT is about *as a conscious state*. And, if one's HOTs are sometimes conscious, it will be evident to one that one's mental states are sometimes conscious. So one will conceive of mental states as being the sort of state that can, on occasion, be conscious.

By contrast, if one's HOTs are never themselves conscious thoughts, one will have no occasion to think of one's mental states, even those which are actually conscious, as being conscious states. One would think about and ascribe mental states, to oneself as well as to others, solely on the basis of behavioral cues and environmental stimulation. This suggests that the HOTs of four- and five-year-olds may well never be conscious, or hardly ever. If so, these children would have no basis on which to think of their thoughts and experiences as the sorts of state that might sometimes be conscious.[36]

Whenever we say anything, we express some intentional state that has the same content as what we say and a mental attitude that corresponds to our speech act's illocutionary force.[37] And almost without exception, whenever we say anything, the intentional state we express is conscious. Elsewhere I have argued that the HOT hypothesis can explain this striking regularity without following Descartes in invoking some unexplained connection between consciousness and speech (Rosenthal, 1990; 1998; see also Rosenthal, 1993a; 1986b).

Although ⌜p⌝ and ⌜I think that p⌝ have distinct truth conditions, adults recognize that the speech acts of saying these two things have the same performance conditions. Minor variation in conviction aside, whenever we say one we could equally well say the other. Moreover, this performance-conditional equivalence is, for adults, automatic and second nature. So whenever I say that ⌜p⌝, thereby expressing my thought that ⌜p⌝, I might equally well have said ⌜I think that p⌝. But saying that would have expressed my HOT to the effect ⌜I think that p⌝. Whenever I say anything, therefore, the automatic character of the performance-conditional equivalence between saying that thing and saying that I think it ensures that I will have a HOT to the effect that I do think that thing.

The developmental findings about children three and younger help us understand the performance-conditional equivalence invoked in this explanation. These children are already habituated to say ⌜p⌝ whenever they might say ⌜I think that p⌝ and conversely, though they would also say of others that they think that ⌜p⌝. So the automatic character of the adult performance-conditional equivalence between saying ⌜p⌝ and saying ⌜I think that p⌝ is a special case of the more general habit that children three and younger have of ascribing not only to themselves but to others as well only beliefs they take to be true. This habit is presumably central to young children's mastery of talking about believing and thinking. So the more general form of the performance-conditional equivalence will have become well-entrenched at a very early age.

The lack in four- and five-year-olds of HOTs that are conscious is presumably not due to any conceptual deficit, but to limits on the computational capacity needed to form the relevant HOTs. These third-order thoughts will, after all, be thoughts about thoughts about mental states. But even three-year-olds can say ⌜I think that p⌝, thereby verbally expressing a HOT to the effect that they think that ⌜p⌝. So why won't the HOTs of these children be conscious after all, given that verbally expressed thoughts are conscious?

Verbally expressed thoughts are conscious because of an automatic performance-conditional equivalence between saying ⌜p⌝ and saying ⌜I think that p⌝. But the embedding needed for the next level up arguably prevents that performance-conditional equivalence from being automatic and second nature. It is not at all natural, even for adults, to equate saying ⌜I think that p⌝ with saying ⌜I think that I think that p⌝ (see Rosenthal, 1990; 1998). So it is open for even the verbally expressed HOTs of four- and five-year-olds to be never or seldom conscious.

5. Nonveridical Metacognition and Consciousness

It is well-known that metacognitive judgments about subjects' cognitive states often fail to be fully accurate. Indeed, studies of feeling-of-knowing judgments, confidence judgments, and judgments of learning often, as noted earlier (p. 267; see esp. note 1), seek to assess their degree of accuracy, sometimes in comparison with one another or with other cognitive functions, such as ordinary recognition and recall.

It would not be surprising, therefore, if metacognition is never perfectly accurate, even in that form which pertains to our mental states' being conscious. Few today would endorse the traditional idea that our introspective grasp of what mental states we are in is invariably correct, to say nothing of the twin idea that introspection reveals all our mental states and does so in respect of all their mental features.[38]

As noted in connection with Block's concept of monitoring consciousness, introspective consciousness goes beyond the ordinary way in which mental states are conscious. When we introspect our mental states, we are conscious of them in a special way. We focus on them

attentively and deliberately. And we are conscious of doing so. At the very least, therefore, the HOTs we have about our mental states when we introspect must be conscious HOTs.

But if our conscious HOTs can represent our mental states in ways that are less than fully accurate, presumably our nonconscious HOTs can do so as well. Indeed, inaccuracy among nonconscious HOTs, since it would seldom if ever be detected, may well occur far more often than we know it does with their conscious counterparts. HOTs result in our being conscious of ourselves *as* being in certain states, in particular, as being in states whose nature is *as* those HOTs represent them. And, if the HOTs in question aren't conscious thoughts, we won't be in a position to evaluate their accuracy.

It is to be expected, in any case, that HOTs would sometimes be inaccurate. In section 1, we considered the odd way in which we seem in feeling-of-knowing experiences both to be and not to be conscious of a cognitive state. I argued that we can understand such experiences in terms of our being conscious of the state as a state that holds the answer to a certain question but not in respect of that state's specific informational content. And I urged that we could best understand this difference, in turn, on the hypothesis that we are conscious of these states by having HOTs about them.

On this explanation, the HOTs in virtue of which mental states are conscious represent those states more or less fully. Moreover, the way our HOTs represent the states they are about actually influences what those states are like from a first-person point of view. What it's like for one to recall something consciously is strikingly different from what it's like simply to have that thing on the tip of one's tongue. And, because conscious recall and a tip-of-the-tongue experience may involve the very same informational content, the two will differ only in the way one is conscious of the target informational state. Whether one consciously recalls or has the information remain on the tip of one's tongue will depend on whether or not one's HOT represents the target in respect of its informational content.

Moreover, that target will be a conscious state only if the HOT does represent the state's informational content. So, if one had a HOT that represented the target as having, instead, some different content, it should seem to one as though one were actually in a conscious state that has that different content. What it's like for one depends on how one's HOTs represent the states they are about.

There are examples other than feeling-of-knowing experiences of the way differences in how our HOTs represent their targets affect what it's like for us to be in those target states. For example, what it's like for one to have a particular gustatory sensation of wine arguably depends on how much detail and differentiation goes into the HOT in virtue of

which that sensation is conscious. Similarly for other sensory states; the degree of detail with which we are aware of a state makes a difference to what it's like for us to be in that state.[39] Indeed, if erroneous HOTs did not affect what it's like for us to be in target states, subjective errors could never occur, since what it's like to be in those target states would then always be accurate.

What HOT one has on any particular occasion, moreover, will depend not just on what target state one is in, but also on the size of one's repertoire of concepts, as well as on such transitory factors as one's current interests and how attentive one is. But if mental states do not by themselves determine what HOTs occur and how they represent their targets, there is no reason why those HOTs cannot sometimes misrepresent those targets. One would then be in a state of one type but have a HOT that represents one as being in a state of some different sort. And, since the content of one's HOT determines what it's like for one to be in a mental state, an erroneous HOT may well make it seem, from a first-person point of view, as though one were in a mental state that one is not in fact in.

There is reason to believe that this actually happens. Dental patients sometimes seem, from a first-person point of view, to experience pain even when nerve damage or local anesthetic makes it indisputable that no such pain can be occurring. The usual hypothesis is that the patient experiences fear along with vibration from the drill and consciously reacts as though in pain. Explaining this to the patient typically results in a corresponding change in what it's like for the patient when drilling resumes. But the patient's sense of what the earlier experience was like remains unaltered. The prior, nonveridical appearance of pain is indistinguishable, subjectively, from the real thing.

There is a also well-known tendency people have to confabulate being in various intentional states, often in ways that seem to make ex post facto sense of their behavior;[40] here it's plain that HOTs misrepresent the states that subjects are in. Similarly, it is reasonable to assume that repressed beliefs and desires often are actually conscious beliefs and desires whose content one radically misrepresents. Thus one might experience one's desire for some unacceptable thing as though it were a desire for something else, instead. The desire would not literally be unconscious; it would simply be a conscious desire whose character is distorted by inaccurate HOTs. And what it would be like to have that desire would fail accurately to reflect its actual content. Erroneous HOTs may well also figure in cases of so-called self-deception; there one's HOTs would misrepresent not what one desires but what one believes. These cases may even challenge our ability to distinguish in a nonarbitrary way between a HOT that misrepresents an actual target and a HOT whose target does not actually exist but is strictly speaking notional.

These two kinds of situation would presumably be indistinguishable from a first-person point of view.[41]

The variation in degree of detail with which consciousness represents our conscious states provides an important test for any explanatory model. Consider an ingenious proposal by Keith Lehrer (1991; 1997a, chap. 7; 1997b), which in a number of respects resembles the HOT hypothesis. A mental state is conscious, on Lehrer's model, in virtue of a mental process that leads from the state itself to a mental quotation of that state and then back to the state itself, this time considered as exemplifying the mental type to which it belongs. Lehrer argues that this process results in a metamental affirmation to the effect that the target state exemplifies the mental type in question. And, because the metamental affirmation that one is in a mental state of that type leads in turn to one's having easy, immediate knowledge about that target state, the affirmation makes the target a conscious state.

Lehrer recognizes that this metamental process is not immune from error (1997a, p. 170); it might go wrong, connecting a mental quotation of one token with some other token of a distinct type, treated as exemplar. As we have seen, such an occurrence would be undetectable from a first-person point of view. More important for present purposes, Lehrer notes that the metamental affirmation provides a thin characterization of the target, representing it only as being of that mental type which the target itself exemplifies. Lehrer urges that the information consciousness actually gives us about our conscious states is thin in this way, which doubtless is often so. But, as we see both from feeling-of-knowing experiences and from cases such as wine tasting, consciousness not only represents our conscious states more or less accurately, but also in strikingly greater and lesser detail. And it's unclear how such a model, on which our conscious states themselves determine how consciousness represents them, can make room for this variability in the way we are conscious of our conscious states.[42]

The foregoing considerations make it likely that inaccuracy affects not only our metacognitive judgments, but even the way our mental states are conscious. Moreover, there is considerable variation in the degree of detail that enters into both our metacognitive judgments and the way we are conscious of our conscious states. These parallels give us every reason to expect that research into metacognition and the errors that affect it will shed light on the nature of consciousness.

Acknowledgments

I am grateful to Alvin Goldman, Keith Lehrer, François Recanati, and Georges Rey for useful reactions to a slightly earlier version.

Notes

1 Excellent samples of such investigations can be found in Metcalfe & Shimamura (1994); Nelson (1992); Weinert and Kluwe (1987); Forrest-Presley, Mackinnon, & Waller (1985). On dissociative neural deficits, see Milner & Rugg (1992), and many of the essays in Marcel & Bisiach (1988).

The term 'metacognition' seems to have been introduced into the psychological literature by John H. Flavell and coworkers (e.g., in Flavell & Wellman, 1977).

2 Nelson (1996). Cf. Nelson & Narens (1994).

3 On blindsight, see Weiskrantz (1986; 1990; 1991; 1997), and Marcel (1993). On visual agnosia, see Farah (1990; 1994).

4 Some studies suggest that frontal-lobe function subserves feeling-of-knowing judgments, since deficits unique to the frontal lobe are correlated with impaired accuracy in such judgments, even when recall and recognition are normal. See Janowski, Shimamura, & Squire, 1989; Shimamura & Squire, 1986; and Nelson et al, 1990.

5 The tip-of-the-tongue (TOT) phenomenon that psychologists discuss involves having conscious access to partial information, perhaps the initial letters or phoneme of a name. In what follows I'll use the phrase 'tip of the tongue' in its commonsense usage to refer instead to cases in which we have a vivid sense, sometimes accurate, that the name or information could be accessed though we cannot get conscious access even to partial information.

6 That is, the state that carries the information one feels one knows. As already noted, the *judgment* that one knows is of course conscious.

7 Larry Weiskrantz has reported research by colleagues aimed at such results (personal communication).

8 And, even when the states in question are all intentional states, there is typically something it's like to have a tip-of-the-tongue or other feeling-of-knowing experience, as well as something different it's like when one retrieves the actual information.

9 For a particularly useful discussion, see Vendler (1972, chap. 5).

10 See note 5 above.

11 Though Aristotle holds that images and thoughts are distinct (1907, Γ8, 432a14), he also insists that all thinking requires images (e.g., 1907 A1, 403a9–10, Γ7, 431a16, Γ8, 432a9, 14; 1972, 1, 449b31) and, indeed, that one "thinks in images" (1907, Γ7, 431b2; cf. Γ8, 432a5).

12 That is, characteristic properties susceptible of being sensed, as against, the distinguishing *sensory* properties of sensory states. On that distinction, see Rosenthal (1999).

13 For example, in Rosenthal (1986a); (1991); (1993b), all to appear with other papers in Rosenthal (in preparation), and in Rosenthal (1997b).

14 The objection from higher-order memories was raised by Alvin I. Goldman, in discussion; the objection from higher-order processing in the nonconscious editing of speech and in the executing of certain motor intentions occurs in Marcel (1988, p. 140).

15 Dretske (1993), pp. 272–275; cf. Dretske (1995), pp. 112–113. Dretske holds that we are never transitively conscious of our conscious mental states, not

even in introspection. He argues that introspection resembles what he calls displaced perception. Just as we come to know how full the gas tank is by looking at the gauge, so we come to know what mental state we're in by noticing what physical object we're seeing. Although we come thereby to be conscious *that* we're in some particular mental state, we're not conscious *of* that state (Dretske, 1994/1995; 1995, chap. 2).

If Dretske is right about introspection, introspecting a mental state in effect means having a thought that one is that state. Dretske regards all thoughts as conscious; so this amounts to the claim that introspecting is having a *conscious* thought about one's mental states. This is exactly the account of introspection I have offered elsewhere (e.g., Rosenthal, 1986a, pp. 336–337; 1997, pp. 745–746). Dretske's view differs from mine only in his presupposition that all mental states are conscious states, and in denying that whenever one is conscious *that* something has a certain property one is thereby conscious *of* that thing.

16 Striking experimental results underscore how frequent cases occur in which one attentively sees two scenes that differ in a single respect, but without being conscious of the respect in which they differ. Grimes (1996) presented subjects with scenes that changed during saccades, as determined by eye trackers. Subjects were told to study the presentation material, and some were even told that a conspicuous change would occur. Most subjects fail on 7 of 10 trials to notice dramatic changes, such as a change in color of a large, salient object. Even people who informally view the presentation material without being fitted with eye trackers fail surprisingly often to notice such changes, presumably because of random saccades.

17 It is worth noting a complication in Dretske's discussion. Unlike being conscious of concrete objects and events, being conscious of a difference, according to Dretske, always amounts to being conscious "that such a difference exists" (Dretske, 1993, p. 275; cf. pp. 266–267). So he might insist that being conscious of a difference is always being conscious of it *as* a difference. But that cannot help. Even though the experience of the extra tree is that in virtue of which the two overall experiences differ, one can be conscious of the thing in virtue of which they happen to differ without being conscious that they do differ. As Dretske would put it, one can be conscious of that in virtue of which they differ but not of the difference between them. Indeed, he explicitly acknowledges that this very thing can happen: "[Those] who were only thing-aware of the difference between [the two arrays] were not fact-conscious of the difference between [them]" (Dretske, 1993, p. 275).

18 See also Block (1993, p. 184; 1992, pp. 205–206; 1990, pp. 596–598).

19 It won't help simply to assume that any state with sensory properties must be conscious in some way or other. If one has no conscious access to such a state, it won't be a conscious state in any intuitive sense.

20 Block's notion of access consciousness also has its troubles. That reconstruction trades on the idea that a state's playing various executive, inferential, and reporting roles involves one's having the kind of access to that state that is relevant to the state's being conscious. But that is often not the case. Many states play executive, inferential, and reporting roles without being conscious in any intuitive sense whatever. To reconstruct the kind of state consciousness that

involves having access to a mental state, we must provide that one actually be conscious of that state in an intuitively unmediated way. Given the argument in the text, we can conclude that, *pace* Block, phenomenal consciousness actually implies access consciousness. That fact is obscured by Block's reconstruction of access consciousness, since a state's playing various executive, inferential, and reporting roles does not intuitively seem to be what is needed for there to be something it's like to be in a sensory state. Being conscious of that state, by contrast, is exactly what is needed. See also Rosenthal (1997a).

Block's definition of access consciousness in terms of a state's being "poised" for various uses may be meant to capture the way conscious states seem more readily available for such use than nonconscious states. But even nonconscious states can be so poised. Defining access consciousness in such terms gives a dispositional mark of such consciousness. But that doesn't mean that access consciousness is a matter of one's simply being disposed to be conscious of one's mental states, as opposed to actually conscious of them. The states we are conscious of have many dispositional properties, among them being reportable and introspectible.

21 Goldman (1993, p. 366); see Dretske (1995, p. 109).

22 Dretske holds that a state's being conscious consists not in one's being conscious of the state, but in the circumstance that, in virtue of one's being in that state, one is conscious of something or conscious that something is the case. And that is an intrinsic property of every mental state. Because that property is intrinsic to all mental states, however, Dretske must hold that all mental states are conscious, which is highly implausible (1993, p. 271).

23 Cognitive and clinical theorists often posit nonconscious informational states specifically in connection with some mental subsystem. This encourages a tendency to describe all nonconscious mental states as states of subpersonal systems, in contrast with conscious states, conceived of as states of the whole creature. But there is good reason also to regard most nonconscious mental states as being states of the whole creature. Not only do nonconscious thoughts and desires, and the sensations that occur nonconsciously in peripheral vision and subliminal perception, have the very same distinguishing mental properties as conscious states; they sometimes come to be conscious. And it would be surprising if the shift from being nonconscious to being conscious involved a concomitant shift from being a state of a subpersonal system to being a state of the whole creature.

24 When I am conscious without apparent mediation of my veins' throbbing, I am conscious of two things: states of my veins, and a certain bodily sensation. Being conscious of the sensation in respect of its mental properties results in that sensation's being conscious. By contrast, being conscious of the veins, as such, may well result in no conscious state whatever.

Still, HOTs do not presuppose that one have a concept of mentality, since we can be conscious of a state in virtue of a mental property it has without being conscious that the property is a mental property.

25 In particular, the two are equivalent in truth conditions. The truth conditions for somebody's believing something of course also includes that person's existing and being in the relevant state, but the children presumably see these conditions as being obviously satisfied in both cases.

26 Dretske's argument affects the transitivity principle only indirectly, since it does, by itself, not preclude there being ways we might be conscious of our conscious states which do not require having any concepts at all of the relevant types of state. On the argument of section II, however, no other way will do, and elsewhere Dretske endorses those arguments (1993, p. 297).

27 This reasoning echoes Descartes's notoriously unconvincing argument that, if nonhuman animals had thoughts, they would surely express them in speech (letter to Newcastle, November 23, 1646, in Descartes 1984–1991, Vol. 3, p. 303), as though nothing more is needed for verbally expressing thoughts than having them.

28 This sort of consideration doubtless underlies the traditional view that consciousness attaches more directly to a state's mental attitude than to its content. Cf., e.g., Descartes's claim, in *Fourth Replies*, that it is with respect to "the operations of the mind," and not the contents of those operations, that "there can be nothing in our mind of which we are aware" (Descartes, 1984–1991, Vol. 2, p. 162).

29 Similar considerations apply to nonlinguistic animals, which presumably are also conscious of their conscious states in a way that fails to capture much about the specific attitudes those states exhibit. Indeed, nonlinguistic animals will also be conscious of the content of their conscious states in terms far simpler than those which figure in our HOTs, indeed, in terms that may be difficult to translate into our ways of describing things.

30 It's striking that some three-year-olds follow the candy-box pattern in judging others' likes and dislikes always to match their own (Flavell, Flavell, Green, & Moses, 1993).

31 Fodor (1992) also presents reasons for rejecting the standard hypothesis.

32 It is unclear that psychologists who take this second, more standard line have clearly distinguished between change in concept and change in the beliefs one has about the relevant things. If not, the difference between these hypotheses may be, at least in part, verbal rather than substantive.

33 I use 'conception' here as neutral between the concept one has of something and the central, characterizing beliefs one has about that thing.

34 An exception is seeing; these preschoolers report that people cannot see things when asleep. That fits well with this interpretation. Closed eyes constitute a behavioral cue that shows that a sleeping person is not seeing anything, but nothing parallel prevents one from regarding sleeping people as hearing and thinking.

As noted, when Flavell and his associates asked these preschoolers whether people, while sleeping, know that they think, hear, and feel things, they give mostly affirmative answers. But knowing itself may be merely tacit and hence not conscious, and these replies may have referred only to such nonconscious, tacit knowing. Indeed, this is to be expected if these preschoolers do not think of mental states at all as being conscious states.

35 Flavell finds this interpretation congenial (personal communication).

36 It would then turn out that these children do not, after all, introspect, since introspection consists in having HOTs that are themselves conscious.

The hypothesis that these children conceive of their mental states in terms of behavioral cues and environmental stimulation echoes, at the level of individual development, Wilfrid Sellars' idea that the characteristic way people have of reporting noninferentially about their own mental states might have been built on an earlier practice of reporting inferentially about their own mental states (Sellars, 1963, p. 189, §59).

37 An exception is insincere speech, but that is best understood as pretending to say things. See Rosenthal (1986b, sec. 4).

38 For a useful review, see Lyons (1986).

39 Compare Daniel Dennett's case of looking straight at a thimble but failing to see it as a thimble (1991, p. 336). One's sensation of a thimble is conscious, but conscious not *as* a sensation of a thimble but only, say, as a sensation of part of the clutter on a shelf. There are many cases in which the visual sensations one has of one's environment are conscious, but not conscious in respect of much of their significant detail. The best explanation of this is the coarse-grained character of the HOTs in virtue of which those sensations are conscious.

40 The classic study is Nisbett & Wilson (1977). That influential study focused both on cases in which subjects confabulate stories about the causes of their being in particular cognitive states and on cases in which they confabulate accounts about what states they are actually in.

41 Being inaccurately conscious of mental states is a possibility left open not only by the HOT model, but by any theory that upholds the transitivity principle. Some residual sense of the relative infallibility of consciousness may therefore underlie intuitive resistance to that principle. See, for example, Goldman (1996).

42 It is worth noting that the metamental affirmation posited by Lehrer's model cannot actually contain the target itself. Suppose the target is an intentional state. Its content can occur in the metamental affirmation, and that affirmation can represent the target's mental attitude. But the only mental attitude that can literally occur in the metamental affirmation itself is that which governs the entire content of the affirmation. This is evident from the fact that, whereas the metamental state must have an assertoric attitude, the target's mental attitude may be one of doubting, desiring, or denying. Since an intentional state can play the role of exemplar only if its mental attitude is suspended, the exemplar states in Lehrer's metamental affirmations cannot be the actual targets, but must be states derived from them. Similarly, targets cannot function as mental quotations of themselves, since mental quotations must also lack mental attitude.

This points to a difficulty. According to Lehrer, the metamental affirmation represents the target as being of the mental type exemplified by the target, and he explains our grasp of what that mental type is by appeal to our understanding of the mental token in question. But if the state that serves as exemplar is not the target state itself but some state derived from it, some additional account is needed of how we come to understand its content.

References

Aristotle (1907). *De anima.* Translation, introduction, and notes by R. D. Hicks. Cambridge: Cambridge University Press.

Aristotle (1972). *On memory.* Translated with commentary by Richard Sorabji. London: Duckworth.

Astington, Janet W., & Gopnik, Alison (1991). Developing understanding of desire and intention. In Andrew Whiten (Ed.), *Natural theories of mind: The evolution, development and simulation of second order representations* (pp. 39–50). Oxford and Cambridge, MA: Blackwell.

Block, Ned (1990). Consciousness and accessibility. *Behavioral and Brain Sciences 13* (4), 596–598.

Block, Ned (1992). Begging the question against phenomenal consciousness. *Behavioral and Brain Sciences 15* (2), 205–206.

Block, Ned (1993). Review of Daniel C. Dennett, *Consciousness explained. The Journal of Philosophy 90* (4), 181–193.

Block, Ned (1994). On a confusion about a function of consciousness. Unpublished manuscript version of Block (1995).

Block, Ned (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences 18* (2), 227–247.

Davies, Martin, & Humphreys, Glyn W., Eds. (1993). *Consciousness: Psychological and philosophical essays.* Oxford: Basil Blackwell.

Dennett, Daniel C. (1991). *Consciousness explained.* Boston: Little, Brown.

Descartes, René (1984–1991). *The philosophical writings of Descartes (3 vols.).* Eds. John Cottingham, Robert Stoothoff, and Dugald Murdoch. (Vol. 3 with Anthony Kenny.) Cambridge: Cambridge University Press.

Dretske, Fred (1993). Conscious experience. *Mind 102* (406): 263–283.

Dretske, Fred (1994–1995). Introspection. *Proceedings of the Aristotelian Society, 115,* 263–278.

Dretske, Fred (1995). *Naturalizing the mind.* Cambridge, MA: MIT Press/Bradford.

Farah, Martha J. (1990). *Visual agnosia: Disorders of object recognition and what they tell us about normal vision.* Cambridge, MA: MIT Press/Bradford.

Farah, Martha J. (1994). Visual perception and visual awareness after brain damage: A tutorial overview. In Carlo A. Umiltà & Morris Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing.* Cambridge, MA: MIT Press/ Bradford.

Flavell, John H. (1988). The development of children's knowledge about the mind: From cognitive connections to mental representations. In Janet W. Astington, Paul L. Harris, & David R. Olson (Eds.), *Developing theories of the mind.* Cambridge: Cambridge University Press.

Flavell, John H. (1993). Young children's understanding of thinking and consciousness. *Current Directions in Psychological Science 2* (2), 40–43.

Flavell, John H., & Wellman, Henry M. (1977). Metamemory. In Robert V. Kail, Jr. & John W. Hagen (Eds.), *Perspectives on the development of memory and cognition.* Hillsdale, N.J.: Erlbaum.

Flavell, John H., Flavell, Eleanor R., Green, Frances L., & Moses, Louis J. (1993). Young children's understanding of fact beliefs versus value beliefs. *Child Development 61* (4), 915–928.

Flavell, John H., Green, Frances L., & Flavell, Eleanor R. (1993). Children's understanding of the stream of consciousness. *Child Development* 64, 387–398.

Flavell, John H., Green, Frances L., & Flavell, Eleanor R. (1995a). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development* 60 (1) (Serial No. 243), 1–95.

Flavell, John H., Green, Frances L., & Flavell, Eleanor R. (1995b). The development of children's knowledge about attentional focus. *Developmental Psychology 31* (4), 706–712.

Fodor, Jerry A. (1992). A theory of the child's theory of mind. *Cognition 44* (3), 283–296.

Forrest-Presley, D. L., Mackinnon, G. E., & Waller, T. Gary, Eds. (1985). *Metacognition, cognition, and human performance.* Orlando, FL: Academic Press.

Goldman, Alvin I. (1993). Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition* 2 (4), 364–382.

Goldman, Alvin I. (1996). The science of consciousness and the publicity of science. Unpublished paper, Department of Philosophy, University of Arizona.

Gopnik, Alison (1993). How do we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* 16 (1), 1–14. With open peer commentary, 29–90, and author's response, Theories and illusion, 90–100.

Gopnik, Alison, & Astington, Janet W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development 59* (1), 26–37.

Grimes, John (1996). On the failure to detect changes in scenes across saccades. In Kathleen Akins (Ed.), *Perception* (pp. 89–110). New York: Oxford University Press.

James, William (1950). *The Principles of Psychology.* (New York: Dover Publications. (Original work published 1890.)

Janowski, J., Shimamura, Arthur P., & Squire, Larry R. (1989). Memory and metamemory: Comparisons between patients with frontal lobe lesions and amnesic patients. *Psychobiology* 17, 3–11.

Lehrer, Keith (1991). Metamind, autonomy and materialism. *Grazer Philosophische Studien* 40, 1–11.

Lehrer, Keith (1997a). *Self-trust: A study of reason, knowledge, and autonomy.* Oxford: Clarendon.

Lehrer, Keith (1997b). Meaning, exemplarization and metarepresentation. In Dan Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (this volume).

Lyons, William (1986). *The disappearance of introspection.* Cambridge, MA: MIT Press/ Bradford.

Marcel, Anthony J. (1993). Slippage in the unity of consciousness. In *Experimental and theoretical studies of consciousness* (pp. 168–186). Ciba Foundation Symposium No. 174. Chichester: John Wiley & Sons.

Marcel, Anthony J. (1988). Phenomenal experience and functionalism. In Anthony J. Marcel & Edoardo Bisiach (Eds.), *Consciousness in contemporary science.* Oxford: Clarendon.

Marcel, Anthony J., and Bisiach, Edoardo, Eds. (1988). *Consciousness in contemporary science.* Oxford: Clarendon.

Metcalfe, Janet, & Shimamura, Arthur P., Eds. (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press/Bradford.

Milner, A. D., & Rugg, Michael D., Eds. (1992). *The neuropsychology of consciousness*. New York: Academic.

Miner, Ann C., & Reder, Lynne M. (1994). A new look at feeling of knowing: Its metacognitive role in regulating question answering. In Janet Metcalfe & Arthur P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 47–70). Cambridge, MA: MIT Press/Bradford.

Moore, Chris, Pure, Kiran, & Furrow, David (1990). Children's understanding of the modal expression of speaker certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development, 61* (3), 722–730.

Nagel, Thomas (1974). What is it like to be a bat? In Thomas Nagel, *Mortal questions* (pp. 165–179). Cambridge: Cambridge University Press. Originally published in *The Philosophical Review 83* (4), 435–450.

Nelson, Thomas O., Ed. (1992). *Metacognition: Core readings*. Boston: Allyn and Bacon.

Nelson, Thomas O. (1996). Consciousness and metacognition. *American Psychologist* 51 (2), 102–116.

Nelson, Thomas O., Dunlosky, John, White, David M., Steinberg, Jude, Townes, Brenda D., & Anderson, Dennis (1990). Cognition and metacognition at extreme altitude on Mount Everest. *Journal of Experimental Psychology: General 119* (4) (December), 367–374.

Nelson, Thomas O., & Narens, Louis (1994). Why investigate metacognition? In Janet Metcalfe & Arthur P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25).

Nisbett, Richard E., & Wilson, Timothy DeCamp (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84* (3), 231–259.

Perner, Josef (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press/Bradford.

Perner, Josef, Leekam, Susan R., & Wimmer, Heinz (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology 5* (2), 125–137.

Rosenthal, David M. (1986a). Two concepts of consciousness. *Philosophical Studies 49* (3), 329–359.

Rosenthal, David M. (1986b). Intentionality. *Midwest Studies in Philosophy 10*, 151–184.

Rosenthal, David M. (1990). Why are verbally expressed thoughts conscious? Report No. 32/1990, Center for Interdisciplinary Research (ZiF), University of Bielefeld, Germany. To appear in Rosenthal (in press).

Rosenthal, David M. (1991). The independence of consciousness and sensory quality. In Enrique Villanueva (Ed.), *Consciousness: Philosophical issues, 1, 1991* (pp. 15–36). Atascadero, CA: Ridgeview Publishing Company.

Rosenthal, David M. (1993a) Thinking that one thinks. In Martin Davies & Glyn W. Humphreys (Eds.), *Consciousness: Psychological and philosophical essays* (pp. 197–223). Oxford: Basil Blackwell.

Rosenthal, David M. (1993b). State consciousness and transitive consciousness. *Consciousness and Cognition 2* (4), 355–363.

Rosenthal, David M. (1997a). Phenomenal consciousness and what it's like. *Behavioral and Brain Sciences* 20 (1), 64–65.

Rosenthal, David M. (1997b). A theory of consciousness. In Ned Block, Owen Flanagan, & Güven Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 729–753). Cambridge, MA: MIT Press.

Rosenthal, David M. (1998). Consciousness and its expression. *Midwest Studies in Philosophy 22*, 294–309.

Rosenthal, David M. (1999). *Consciousness and Mind*. Oxford: Clarendon.

Rosenthal, David M. (in preparation). The colors and shapes of visual experiences. In Denis Fisette, *Consciousness and intentionality: Models and modalities of attribution* (pp. 95–118). Dordrecht: Kluwer Academic Publishers.

Sellars, Wilfrid (1963). Empiricism and the philosophy of mind. In Wilfrid Sellars, *Science, perception and reality* (pp. 127–196). London: Routledge & Kegan Paul. Reprinted (1991) Atascadero, CA: Ridgeview Publishing Company. Republished (1997) as *Empiricism and the philosophy of mind*. Cambridge, MA: Harvard University Press.

Shimamura, Arthur P. (1994). The neuropsychology of metacognition. In Janet Metcalfe & Arthur P. Shimamura, (Eds.), *Metacognition: Knowing about knowing* (pp. 253–276). Cambridge, MA: MIT Press/Bradford.

Shimamura, Arthur P., & Squire, Larry R. (1986). Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition 12*, 452–460.

Tulving, Endel, & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior 5*, 381–391.

Vendler, Zeno (1972). *Res cogitans*. Ithaca, NY: Cornell University Press.

Weinert, Franz E., & Kluwe, Rainer H., Eds. (1987). *Metacognition, motivation, and understanding*. Hillsdale, NJ: Erlbaum.

Weiskrantz, Lawrence (1986). *Blindsight: A case study and implications*. Oxford: Oxford University Press.

Weiskrantz, Lawrence (1990). Outlooks for blindsight: Explicit methodologies for implicit processes. The 1989 Ferrier Lecture. *Proceedings of the Royal Society B 239*, 247–278.

Weiskrantz, Lawrence (1991). Dissociations and associates in neuropsychology. In Richard G. Lister and Herbert J. Weingartner (Eds.), *Perspectives on cognitive neuroscience* (pp. 157–164). New York: Oxford University Press.

Weiskrantz, Lawrence (1997). *Consciousness lost and found: A neuropsychological exploration*. Oxford: Oxford University Press.

Wellman, Henry W. (1990). *The child's theory of the mind*. Cambridge, MA: MIT Press/Bradford.

Metarepresentations

A Multidisciplinary Perspective

edited by Dan Sperber

OXFORD
UNIVERSITY PRESS
2000