Annals of the Academy of Romanian Scientists
Series on Philosophy, Psychology and Theology
ISSN 2067-113X Volume **1**, Number **1**/**2009**

BEING CONSCIOUS OF OURSELVES

David M. ROSENTHAL*

Rezumat. *Studiul explică gândirea, care se presupune că ar avea la bază o atitudine mentală asertivă, şi totuşi, a pune la îndoială şi a interoga ceva în legătură cu un obiect nu face o persoană să fie şi conştientă de acel obiect. Simpla dispoziţie de a gândi despre ceva nu face pe cineva conştient de acel obiect; gândul trebuie să fie ocurent. dar un gând ocurent şi asertiv despre un obiect considerat prezent face o persoană conştientă de acel obiect la modul intuitiv. Aceste aspecte sunt investigate pe tot parcursul celor patru părţi ale acestui studiu: Conştiinţa de sine; Conştiinţa şi HOT-urile; Conştiinţa de sine şi indexarea esenţială şi Conştiinţa de sine şi imunitatea în faţa erorii.*

Abstract. *The study explains the thought that must presumably have an assertoric mental attitude, yet, doubting and wondering something about an object do not make one conscious of the object. Simply being disposed to have a thought about something does not make one conscious of it; the thought must be occurrent. But having an occurrent, assertoric thought about an object as being present does intuitively make one conscious of that object. These aspects are investigated throughout the four parts of this study: Consciousness of the Self, Consciousness and HOTs, Self-Consciousness and the Essential Indexical and Self-Consciousness and Immunity to Error.*

Keywords: consciousness, self, HOTs

1. Consciousness of the Self

What is it that we are conscious of when we are conscious of ourselves? Hume famously despaired of finding self, as against simply finding various impressions and ideas, when, as he put it, "I enter most intimately into what I call *myself."*[1] "When I turn my reflexion on *myself*, I never can perceive this *self* without some one or more perceptions; nor can I ever perceive any thing but the perceptions."[2] It is arguable that the way Hume attempted to become conscious of the self seriously stacked the deck against success. Hume assumed that being conscious of a self would have to consist in perceiving that self. Perceiving things does make one conscious of them, but perceiving something is not the only way we can be conscious of it. We are also conscious of something when we have a thought about that thing as being present. I may be conscious of an object in front of me by seeing it or hearing it; but, if my eyes are closed and the object makes no sound, I may be conscious of it instead by having a thought that it is there in front of me.

---

*Mind philosophy Professor at City University of New York, in the cognitive sciences and linguistic program. *Study taken from "Mind" magazine, issue 91/2008.*

[1]David Hume, *A Treatise of Human Nature* [1739], ed. L. A. Selby-Bigge, (Oxford Clarendon press, 1888, 2nd ed'n., revised by P. H. Nidditch, 1978), Book I, Part IV, sec. vi, p. 252.

[2]D. Hume, *A Treatise of Human Nature,* Appendix, p. 634.

Not all thoughts one can have about an object result in one's being conscious of that object. We resist the idea that having thoughts about objects we take to be distant in place or time, such as Saturn or Caesar, makes one conscious of those objects; so the thought must be about the object as being present to one. And the thought must presumably have an assertoric mental attitude; doubting and wondering something about an object do not make one conscious of the object. Nor does simply being disposed to have a thought about something make one conscious of it; the thought must be occurrent. But having an occurrent, assertoric thought about an object as being present does intuitively make one conscious of that object. Hume would presumably have argued that this alternative way of being conscious of things has no advantage here, since he maintained that thinking consists simply of pale versions of qualitative perceptual states. "All ideas," he insisted, "are borrowed from preceding perceptions."[1] And his problem was about finding anything other than mental qualities.

But there is good reason to reject this view about the mental nature of having thoughts. For one thing, it is difficult to see how perceptions could be combined to yield thoughts with complex syntactic structure. For another, though qualitative mental states arguably do represent things,[2] they do so in a way that is strikingly different from the way the intentional content of thoughts represents things.[3] Nor is there anything in qualitative mentality that corresponds to the mental attitudes exhibited by intentional states. Rejecting Hume's perceptual model of thoughts makes room for a more promising way to understand how it is that we are conscious of ourselves. We are conscious of ourselves by having suitable thoughts about ourselves.

The contrast between Hume's sensory approach and the alternative that relies on the thoughts we have about ourselves mirrors a contrast between two views about what it is for a mental state to be a conscious state. On the traditional inner-sense model, a mental state is conscious if one senses or perceives that state;

---

[1] *Ibidem*.

[2] One can capture the way qualitative states represent things by seeing each mental quality as representing the perceptible physical property that occupies a corresponding place in the quality space of the relevant perceptual modality. I have defended this view in 'The Colors and Shapes of Visual Experiences,' in *Consciousness and Intentionality Models and Modalities of Attribution,* ed. Denis Fisette (Dordrecht, The Netherlands Kluwer Academic Publishers, 1999), pp. 95-118, and "Sensory Quality and the Relocation Story," *Philosophical Topics*, 26, 1, 2 (1999), 311-50.

[3] *Pace* the representationalist or intentionalist views of writers such as D. M Armstrong, *The Nature of Mind* (St. Lucia, Queensland, Australia: University of Queens land Press, 1980), ch. 9; William G. Lycan, *Consciousness* (Cambridge, MA: M.I.T. Press, 1987), ch. 8; Gilbert Harman, "Explaining Objective Color in Terms of Subjective Reactions," *Philosophical Issues: Perception,* 7 (1996), 1-17; and Alex Byrne, "Intentionalism Defended,' *The Philosophical Review,* 110, 2 (2001), 199-240. 1 discuss representationalism in "Introspection and Self-Interpretation," *Philosophical Topics* 28, 2 (2000), 201-33.

this is doubtless the most widely held view about the consciousness of mental states.[1] The higher-order-thought (HOT) model, by contrast, holds instead that a mental state's being conscious consists in its being accompanied by a suitable thought that one is, oneself, in that state. On the version of the view that I have developed and defended, the thought must be assertoric and non-dispositional. And, because the thought has the content that one is, oneself, in that state, the thought automatically represents the target mental state as being present.[2]

The difference between the inner-sense and HOT views about what it is for a mental state to be conscious sheds light on the two models of consciousness of the self. Suppose that one's mental states are conscious in virtue of one's sensing those states. Sensing a state consists in having a higher-order sensation that represents the sensed states. But nothing in one's sensing a mental state would make reference to or in any other way represent any self to which the target state belongs. So nothing in one's sensing a mental state would make one conscious of the self. Things are different if one is, instead, conscious of one's conscious states by having thoughts about those states. One will then have a thought that one is, oneself, in the target state. And that HOT will thereby make one conscious not only of the target state, but also of a self to which the HOT represents the target state as belonging. The HOT model explains not only how we are conscious of our conscious mental states, but how we are conscious of ourselves as well.

But can the HOT model of how mental states are conscious do justice to the particular way we are conscious of ourselves? There are two main reasons to doubt that it does. The way we are conscious of ourselves seems, intuitively, to be special in a way that simply having a thought about something cannot capture. For one thing, there is a difference between having a thought about somebody that happens to be oneself and having a thought about oneself, as such; HOTs presumably must be about oneself, as such. But having a thought about oneself, as such, may seem to require some special awareness of the self that is antecedent to

[1]Kant first used the term 'inner sense' (K.d.r.V., A22/B37); Locke used the similar 'internal Sense' (Essay, II, i, 4). The view is currently championed by D. M. Armstrong, "What is Consciousness?", in *The Nature of Mind* and by William G. Lycan, *Consciousness and Experience* (Cambridge, MA: M.I.T. Press, 1996), ch. 2, pp. 13-43, and "The Superiority of HOP to HOT," forthcoming in *Higher-Order Theories of Consciousness*, ed. R. W. Gennaro (John Benjamins Publishers, 2004).

[2]See, e.g., Rosenthal, "Two Concepts of Consciousness," *Philosophical Studies* 49, 3 (1986), 329-59; "Thinking that One Thinks," in *Consciousness: Psychological and Philosophical Essays*, ed. M. Davies and G. W. Humphreys (Oxford: Basil Blackwell, 1993), 197-223; "A Theory of Consciousness," in *The Nature of Consciousness*, eds. N. Block, O. Flanagan, and G. Guzeldere (Cambridge, MA; M.I.T. Press, 1997), 729-53; and "Explaining Consciousness," in *Philosophy of Mind: Contemporary and Classical Readings*, ed. David J. Chalmers (New York: Oxford University Press, 2002), 406-21. The first two will appear, along with other papers that develop the HOT model, in *Consciousness and Mind* (Oxford: Clarendon Press, forthcoming 2004).

and independent of the thought. In addition, it seems too many that we are conscious of ourselves in a way that affords certain immunity from error. The special epistemological access to the self which these phenomena seem to suggest have even been thought to provide a foundation for other identifications of objects How could such special self-awareness arise if we are conscious of ourselves simply by having thoughts about ourselves as being bearers of particular mental states?

There is a second challenge to a view about consciousness of the sell that relies simply on such HOTs. Although it seems that we are conscious of ourselves in a way that is special in the ways just sketched, our consciousness of ourselves also fits with our ordinary, everyday ways of identifying and locating ourselves in the world. Each of us is a being with many conscious mental states. But each of us is also a creature that interact; with other objects in the world. And we are conscious of ourselves in both respects. How can simply having HOTs about our mental states explain the way we are conscious of ourselves as located within physical reality? How could having HOTs ground our identifying ourselves as creatures interacting with many other physical things?

The two challenges seem to pull in opposing directions. It may be unclear at first sight how we could be conscious of ourselves in a way that underwrites some kind of immunity from error and yet also captures our contingent location and identity in the physical world. In what follows, I argue that a model of self-consciousness based on HOTs can meet both these challenges. In the next section I briefly sketch the reasons why the HOT model is preferable to the inner-sense model in explaining what it is for a mental state to be conscious. In Section 3, then, I take up the first challenge, to explain how self-consciousness based on HOTs can capture the way our consciousness of ourselves seems special. And in Section 4 I conclude by showing how the account fits with the way we identify and locate ourselves as creatures in the world.

2. Consciousness and HOTs

There is extensive evidence from both everyday life and experimental findings that mental states occur without being conscious. Such evidence relies on situations in which there is convincing reason to believe that an individual is in some particular mental state even thought that individual sincerely denies being in the state. Such sincere denials indicate that the individual is not conscious of being in the state in question. We take it as decisive that a mental state is not conscious if an individual is in that state but is in no way conscious of being in it. It follows that whenever a mental state is conscious, the individual in that state is, in some suitable way, conscious of being in it. Both the inner-sense and HOT models agree thus far. They differ in what that suitable way is of being conscious

of a state, in virtue of which that state is conscious. It has often been emphasized that, for a mental state to be conscious, one must be conscious of it in a way that carries some kind of immediacy. But the way we are conscious of our conscious states need only be *subjectively* immediate. It need not be that nothing actually mediates between the state and one's awareness of that state, but only that nothing *seems* to mediate.

Ordinary perceiving seems to operate in just this way. Nothing seems, subjectively, to mediate between the things we see and hear and our seeing and hearing of them. From a first-person point of view, perceiving seems to be direct. This feature of perceiving makes the inner-sense model appealing; since higher-order perceiving can do justice to the way one's consciousness of one's conscious mental states seems to one to be unmediated. But the HOT model is no less successful in capturing this aspect of the way we are conscious of our conscious states. Some of the thoughts we have about things seem, subjectively, to rely on inference, while others do not seem to do so. When a thought seems subjectively to occur independently of any inference, I shall refer to it as a *non-inferential* thought. Since subjective impression is all that matters here, a non-inferential thought may actually arise as a result of some inference, as long as one is wholly unaware of the way the inference figures in that thought's occurring So, if one is conscious of being in a state by having a non-inferential HOT that one is in that state, it will seem subjectively that nothing mediate between that state and one's being conscious of it. HOTs can explain the intuitive immediacy that our awareness of our conscious states exhibits. Thus far, HOTs and inner sense seem equally good at explaining what it is for a mental state to be conscious. But inner sense faces a difficult that disqualifies it from serious consideration. Sensing and perceiving things take place by way of qualitative mental states. And, when sensing or perceiving is conscious, there is something it's like for one to be in these states, something it's like in respect of the mental quality that these states exhibit. So, if we are conscious of our conscious states by way o inner sense, there are higher-order qualitative states in virtue of which we are conscious of our conscious mental states. But it is clear that no such higher-order qualitative states actually occur. One way to see this is theoretical. Every mental quality belongs to some distinctive perceptual modality; but what modality might the mental qualities of such higher-order qualitative states belong to? It could not be the modality of the target conscious state, since that perceptual modality corresponds to a particular range of perceptible properties, and the first-order target state does not exhibit those perceptible properties. Visual states, for example, exhibit mental qualities that reflect the similarities and differences among the common-sense physical properties perceptible by sight; but the visual states, themselves, do not exhibit properties perceptible by sight. So the mental qualities of higher-order qualitative states could not simply reduplicate the first-order mental qualities. And

there is no other perceptual modality to which such higher-order qualities could belong. Subjective considerations point to the same conclusion. When our mental states are conscious in the ordinary, everyday way, we are not conscious of the higher-order states in virtue of which we are conscious of those first-order conscious states. But very occasionally we are actually conscious of being conscious of those first-order states; when we introspect, we are conscious of being aware of the introspected states. But even when we introspect, we are never conscious of any mental qualities that characterize the states in virtue of which we are conscious of those introspected mental states. The higher-order states in virtue of which we are conscious of our own mental states are not qualitative states.

The only alternative is that those higher-order states are not qualitative, but intentional. We have already seen that being in an assertoric, non-dispositional intentional state that represents the thing it is about as being present makes one conscious of that thing. And, if one is not conscious of any inference on the basis of which one holds that assertoric attitude, so that the awareness it results in seems spontaneous and unmediated, one will be conscious of the target state in the subjectively unmediated way characteristic of our conscious states. We are conscious of our conscious states by having HOTs to the effect that we are in those states. As noted at the outset, HOTs have the advantage over inner sense that, unlike higher-order sensations, a HOT makes one conscious of its target state as belonging to a self. So each HOT makes one conscious of that self. And just as a HOT, by being non-inferential, makes one conscious of its target state in a way that is subjectively unmediated, so that HOT will also make one conscious of the relevant self in a way that is subjectively unmediated. HOTs do justice to our intuitive sense that we have special, unmediated access to ourselves.

A proponent for the inner-sense model might argue that, whatever one thinks about the foregoing considerations, inner sense has a decisive advantage over the HOT model. When qualitative states are conscious, there is something it's like for the subject to be in those states, and this is absent when qualitative states are not conscious. It seems, however, that HOTs could not be responsible for this difference, since HOTs have no qualitative mental properties. We can explain why there is something it's like for one to be in conscious qualitative states, the argument goes, only if the higher-order states in virtue of which we are conscious of the first-order states are themselves qualitative states. But this argument misconceives the situation. The higher-order states are typically not themselves conscious; they are conscious only when we are introspectively aware of our conscious states.

The reason to invoke higher-order states in virtue of which some mental states are conscious is not because we are normally conscious of such higher-order states, but because invoking such higher-order states is theoretically well-

founded. The higher-order states, whether sensations or thoughts, are theoretical posits, which we only occasionally become subjectively aware of.

But, since the higher-order states typically are not conscious, their being qualitative in character could not help explain their being something it's like for one to be in conscious qualitative target states. There will be something it's like for one to be in a conscious qualitative state if one is conscious of oneself, in a way that seems subjectively to be unmediated as being in a state of that qualitative type. And HOTs plainly make us conscious of ourselves in that way.

There is some indirect evidence that HOTs actually do result in then being something it's like for one to be in conscious qualitative states. We sometimes become conscious of more fine-grained difference among our qualitative states by learning new words for the relevant mental qualities Consider the way new mental qualities seem consciously to emerge when we learn new words for the gustatory mental qualities that result from tasting similar wines or the auditory mental qualities that arise when we hear similar musical instruments.

We can best explain how the learning o words for mental qualities can have that effect by supposing that we come to deploy new concepts corresponding to those words, which enable us to have new HOTs about our qualitative sates. HOTs with more fine-grained content result in our qualitative states being conscious in respect of more fine-grained qualities. And, if the intentional content of HOTs makes a difference to what mental qualities we are conscious of, we can infer that HOTs also make the difference between there being *something* it's like for one to be in those states and there being nothing at all that it's like. HOT do result in our qualitative states "lighting up."

3. Self-Consciousness and the Essential Indexical

A HOT makes one conscious of oneself as being in a particular mental state because it has the content that one is, oneself, in that state. So a HOT must somehow refer to oneself. But as already noted, not any way of referring to oneself will do.

There are many descriptions that uniquely pick me out even though I am unaware that they do so; I might believe that some other individual satisfies the description, or simply have no idea who if anybody does. Consider John Perry's now-classic example of my seeing in a grocery store that somebody is spilling sugar from a grocery cart and not realizing that the person spilling sugar is me. My thought that the person spilling sugar is making a mess refers to me, though not to me, as such. Perry refers to this special way of referring to oneself as the essential indexical; classical grammarians know it as the indirect reflexive, since it captures in indirect discourse the role played in direct quotation by the first-person

pronoun.[1]

For a state to be conscious, it is not enough that the individual one is conscious of as being in that state simply happens to be oneself. Suppose that I am the unique *F* and I have a thought that the unique *F* is in pain. That would not make me conscious of myself as being in pain unless I am also aware that I am the unique *F*. Suppose I thought you were the unique *F*. My thought that the unique *F* is in pain would then hardly result in my pain's being conscious; it would not in any relevant way make me conscious of myself as being in pain.

Essentially indexical self-reference is one way in which our consciousness of ourselves is special. And it is sometimes argued that essentially indexical self-reference is required for identifying everything other than oneself.[2] But we rarely do identify other objects by reference to ourselves. We almost always use some local frame of reference in which we figure but which we identify independently of ourselves, by the presence of various objects we perceive and know about. Such local frames of reference occasionally fail, but when they do, referring to ourselves seldom helps. Essentially indexical self-reference cannot sustain such foundationalist leanings. What exactly, then, does such essentially indexical self-reference consist in? How is it that we are able to refer to ourselves, as such? An essentially indexical thought or speech act about myself will have the content that I am *F*. So we must consider how the word 'I' functions in our speech acts and the mental analogue of T functions in the thoughts those speech acts express.

The word 'I' plainly refers to the individual who performs a speech act in which that word occurs. Similarly, the mental analogue of that word refers to the thinker of the containing thought, the individual that holds a mental attitude toward the relevant intentional content. The word 'I' does not have as its meaning *the individual performing this speech act;* nor does the mental analogue of 'I'

---

[1]John Perry, "The Problem of the Essential Indexical," *Nous* **XIII,** 1 (1979), 3-21. For reference to oneself as such, see P. T. Geach, "On Beliefs About Oneself," *Analysis* 18, 1 (1957), 23-24; A. N. Prior, "On Spurious Egocentricity," *Philosophy,* XLII, 162 (1967), 326-35; Hector-Neri Castaneda, "On the Logic of Attributions of Self-Knowledge to Others," *The Journal of Philosophy,* LXV, 15 (1968), 439-56; G. E. M. Anscombe, "The First Person," in *Mind and Language,* ed. Samuel Guttenplan (Oxford: Oxford University press, 1975), pp. 45-65; David Lewis, "Attitudes De Dicto and De Se," *The Philosophical Review* LXXXVIII, 4 (1979), 513-43; and Roderick M. Chisholm, *The First Person* (Minneapolis, MN: University of Minnesota Press, 1981), chs. 3 and 4.

[2]See, e.g., Sydney Shoemaker, "Self-Reference and Self-Awareness," *The Journal of Philosophy* LXV, 19 (1968): 555-67, reprinted with slight revisions in S. Shoemaker, *Identity, Cause, and Mind: Philosophical Essays* (Cambridge: Cambridge University Press, 1984) pp. 6-18 (pages references are to the printed version); Roderick M. Chisholm, *Person and Object; A Metaphysical Study* (La Salle, IL: Open Court, 1976) ch. 1, §5, and *The First Person,* ch. 3, esp. pp. 29-32; and David Lewis, "Attitudes De Dicto and De Se."

express the concept *the individual holding a mental attitude toward this content.* One can refer to oneself using 'I' and its mental analogue without explicitly referring to any speech act or intentional state. On David Kaplan's well-known account, the reference of 'I' is determined by a function from the context of utterance to the individual that produces that utterance; 'I' does not refer to the utterance itself.[1] The connection between the words uttered and the act of uttering them is pivotal. 'I' refers to whatever individual produces a containing utterance, but not by explicitly referring to the utterance itself. Similarly, the mental analogue of 'I' refers to whatever individual holds a mental attitude towards a content in which that mental analogue figures, though again not by explicitly referring to that intentional state, as such.

Suppose, then, that I have the essentially indexical thought that I air *F*. My thought in effect describes as being *F* the individual that thinks that very thought, "in effect" because, although the thought does not describe the individual in that way, it still does pick out just that individual. It does not pick out that individual because the intentional content of my thought so describes the individual. But whenever I do have a first-person thought that I am *F*, my having that thought disposes me to have another though that identifies the individual that thought is about as the thinker of that thought. In that way, every first-person thought thus tacitly or dispositionally characterizes the self it is about as the thinker of that thought. Nothing more is needed for essentially indexical self-reference.

HOTs are simply first-person thoughts, and function semantically jus as other first-person thoughts do. So, when I have a HOT that I am in a particular state, my thought describes as being in that state the individual who thinks that thought. Though the thought itself does not describe that individual as thinking that thought, the individual that thinks the thought I disposed to pick that individual out in that way, by being disposed to have another thought that does so identify the individual the first thought is about. Because HOTs function semantically as other first-person thought do, the HOT hypothesis explains why, when a mental state is conscious, one is conscious of oneself in an essentially indexical way as being in that state. It is important for the HOT model that when a thought refers to oneself in this essentially indexical way, its content does not describe to individual it refers to as the thinker of the thought. If an essentially indexical first-person thought did describe the individual it is about as the thinker of that thought, simply having that thought would make one conscious of having it. And, since HOTs are essentially indexical first person thoughts, one could not have a HOT without being conscious o oneself as having it. But we are wholly

[1]David Kaplan, "Demonstratives," in *Themes From Kaplan*, eds. J. Almog, J. Perry, and H. Wettstein, with the assistance of I. Deiwiks and E. N. Zalta (New York: Oxford University Press, 1989), pp. 481-563, pp. 505-07.

unaware of most of our HOTs. It is sometimes objected to the HOT model that non-linguistic beings, including human infants, could not have HOTs. But this is far from obvious. Non-linguistic beings presumably do have some thoughts, and the conceptual resources HOTs use to describe their target states might, for these beings, be fairly minimal. These beings would not be conscious of their conscious states in the rich way distinctive of adult humans, but that is not implausible.

Still, it might be thought that the essentially indexical self-reference HOTs make preclude their occurring in beings without language. It is natural to suppose that such creatures have, in any case no HOTs about their thoughts. And a thought can make essentially indexical self-reference only if one is disposed to identify the individual one's thought refers to as the thinker of that thought. But it is natural to suppose that such non-linguistic beings would indeed so identify the individual their HOTs refer to if they had suitable conceptual resources. And that should be enough for a HOT to result in the creature's being conscious of itself as being in the state in question.

As noted above, the phenomenon of essentially indexical self-reference encourages the idea that a certain kind of reference to oneself occurs which provides an epistemic foundation for the identification of all other objects. And if so, it might be tempting to urge that we must have some special access to the self that is independent of the thoughts we have about it. Some other form of self-consciousness, antecedent to those thoughts, might then be needed for the essentially indexical self-reference that occurs in our HOTs.

The foregoing explanation of the essential indexical helps dispel that illusion. Reference to oneself as such is simply reference to an individual one is disposed to pick out as the very individual doing the referring. This disposition is independent of the thought that refers to oneself in an essentially indexical way, and that may encourage the idea that essentially indexical self-reference requires independent, antecedent access to the self. But the disposition to have another thought that identifies the individual the first thought is about as the thinker of that thought does not rest on or constitute independent access to the self. It is simply a disposition to have another thought. Essentially indexical self-reference raises no difficulty for an account of self-consciousness in terms of HOTs.

4. Self-Consciousness and Immunity to Error

There is, however, another way in which our consciousness of our own conscious states appears to raise problems for such an account. On a well-known traditional view, our awareness of our conscious states is both infallible and exhaustive. When a mental state is conscious, on this view, there is no feature we are conscious of the state as having which it fails to have, and no mental feature the state has of which we fail to be conscious. There is thus no distinction between

the reality of mental states and their appearance in consciousness.[1] Few today would endorse such a strong form of privileged access. There is doubtless much about the mental natures of our conscious states that we are unaware of, and much that we are wrong about. Our access to our mental states often falls short of exhaustiveness; we are often unclear about what we actually think about things. Nor is that access infallible; there is robust experimental evidence, for example, that we are sometimes wrong about what intentional states issue in our choices and other actions. People often confabulate being in intentional states to explain their choices in situations in which the reported intentional states could not have been operative.[2] In these cases, we are conscious of ourselves as being in states that we are not actually in. Errors of both types occur not only with intentional states, but also in connection with conscious qualitative states. When we consciously see red, we are often conscious of the conscious sensation in respect of a relatively generic shade. Though the sensation exhibits a fairly specific shade of red, as subsequent attention reveals. And it may even happen that we have one type of bodily sensation or emotion but we are conscious of ourselves as having a type different from that. When local anaesthetics blocks any actual pain, a dental patient may still react to the fear and vibration caused by drilling by seeming to be in pain; in such a case, one is conscious of oneself as being in pain, though no pain actually occurs.

The idea that our access to our conscious states is privileged often goes hand in hand with the view that a state's being conscious is an intrinsic property of that state. If a state's being conscious were intrinsic to that state, that would explain our subjective sense that nothing mediates between the states we are conscious of and our consciousness of them. We have that subjective sense because nothing actually does mediate. And it may be tempting to hold that, if nothing mediates between our consciousness of a state and the state itself, consciousness could not be erroneous; there would be no room for error to enter. But that picture is unfounded. Even if one's consciousness of a state were intrinsic to that state, it could still go wrong.

---

[1]See, e.g., Thomas Nagel: "The idea of moving from appearance to reality seems to make no sense" in the case of conscious experiences. In "What Is It Like to Be a Bat?" *The Philosophical Review* LXXXIII, 4 (1974), 435-50; reprinted in *Mortal Questions* (Cambridge: Cambridge University Press, 1979), pp. 165-79, p. 174.
If every mental state is identical with some physical state, then every mental state has both mental and physical properties. This thesis, that we have exhaustive access to our own mental states, therefore applies only to the mental properties.

[2]The classic study is Richard E. Nisbett and Timothy DeCamp Wilson, "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* LXXXIV, 3 (1977), 231-59. A useful review of the exlcnsive literature that follows that study occurs in Peter A. White, "Knowing More than We Can Tell: 'Introspective Access' and Causal Report Accuracy 10 Years Later," *British Journal of Psychology, 79,* 1 (1988).

But there is an echo of such privilege which persists in a view about the way we are conscious of ourselves. This echo pertains not to the mental nature of the states we are conscious of ourselves as being in, but to the self we are conscious of as being in those states. Suppose I consciously feel pain or see a canary. Perhaps I can be wrong about whether the state I am in is one of feeling pain or seeing a canary. But it may well seem that, if I think I feel pain or see a canary, it cannot be that I am right in thinking that somebody feels pain or sees a canary, but wrong in thinking that it is I who does those things. Such first-person thoughts would, in Sydney Shoemaker's now classic phrase, be "immune to error through misidentification," specifically with respect to reference to oneself.[1]

Shoemaker recognizes that such immunity to error fails if one comes to have such thoughts in the way we come to have thoughts about the mental states of others. As he notes, I can wrongly take a reflection I see in a mirror to be a reflection of myself; I thereby misidentify myself as the person I see in the mirror.[2] I might thereby think that I have some property, being right that somebody has that property but wrong that it is I who has the property. So such immunity to error through misidentification does not occur whenever one has a thought that one has a particular property. It must be that one's thought that one has that property arises from the special way we seem to have access to our being in conscious states. There is reason to doubt, however, that such immunity to error actually obtains. The way we have access to being in conscious states is a matter simply of our having non-inferential HOTs that we are in those states. We have a subjective impression that this access is special, since it appears to arise spontaneously and without mediation. But that subjective impression arguably results simply from the relevant HOTs being based on no conscious inference, and indeed from their typically not themselves being conscious in the first place.

As with other thoughts, we come to have these HOTs in a variety of ways, and the process by which HOTs arise can, like any other process, go wrong. So, however unlikely it may be that one is ever right in thinking that somebody is in a particular state but wrong that the individual in that state is oneself, such error is not impossible. One might, perhaps, have such strongly empathetic access to another's state that one becomes confused and thinks that it is oneself that is in

[1]Sydney Shoemaker, "Self-Reference and Self-Awareness," p. 8. Shoemaker urges that such immunity applies even when I take myself to be performing some action. See also Ludwig Wittgenstein, *The Blue and Brown Books* (Oxford: Basil Blackwell, 2nd edition. 1969) pp. 66-67; Gareth Evans, "Demonstrative Identification," in Evans, *Varieties of Reference*, ed. John McDowell (Oxford: Clarendon Press, 1982) 142-266; Jose Luis Hermudez, *The Paradox of Self-Consciousness* (Cambridge, MA: M.I.T. Press, 1998) chs. I and 6; and Robin Meeks, *Identifying the First person,* (Ph.D. dissertation, The City University of New York Graduate Center, 2003) chs. II-IV.

[2]S. Shoemaker, "Self-Reference and Self- Awareness," p. 7.

that state. Such strong immunity to error through misidentification does not obtain. Still, something like this immunity to error does hold. I can be mistaken about whether the conscious state I am in is pain, for example, and perhaps even about whether I am the individual that is actually in pain. But, if I think I am in pain, it seems that I cannot be wrong about whether it is I that I think is in pain. Similarly, if I think that I believe or desire something, perhaps I cannot be mistaken about whether it is I that I think has that belief or desire.

This differs from the immunity to error that Shoemaker and others have described. On that stronger sort of immunity, if I think I am in pain and am right that somebody is, I cannot go wrong about whether it is I who is in pain. On the weaker type of immunity considered here, all I am immune to error about in such a case is who it is I think is in pain. I shall refer to this as *thin immunity.*

Plainly there are ways in which we can misidentify ourselves. Not only might I misidentify myself by wrongly taking a reflection I see in a mirror to be a reflection of myself; I might wrongly take myself to be Napoleon, perhaps because of delusions of grandeur, perhaps because of evidence about Napoleon that seems to lead to me. And, if I do misidentify myself as the person in the mirror or as Napoleon, I will also in that way misidentify the person who has the pains, thoughts, desires, and feelings that I am conscious of myself as having.

How can we capture the specific kind of misidentification that thin immunity rules out? What distinguishes thin immunity to error through misidentification from the ways in which we plainly can and sometimes do misidentify ourselves? The error I cannot make is to think, when I have a conscious pain, for example, that the individual that has that pain is somebody distinct from me, but I can be mistaken about just who it is that I am. How can we capture this distinction? And how can we explain the thin immunity we actually have?

When I have a conscious pain, I am conscious of myself as being in pain. If I think I am Napoleon, I will think that Napoleon is in pain. What I cannot go wrong about is simply whether it is I that I think in pain, that is, whether it is I whom I am conscious of as being in pain. The question is what this amounts to.

The earlier discussion of essentially indexical self-reference gives us a clue. When I refer to myself as such, I refer to the individual I could also describe as doing the referring. Similarly, the error of misidentification I cannot make when I am conscious of myself as being in pain is to think that the individual I think is in pain is distinct from the individual who is conscious of some individual's being in pain. We can readily explain this in terms of the HOT model. The mental analogue of the word “I” refers to whatever individual thinks a thought in which that mental analogue occurs. So every HOT tacitly represents its

target state as belonging to the individual that thinks that very HOT.[1]

Suppose, then, that I have a conscious pain. Since the pain is conscious, I also have a HOT to the effect that I am in pain, and that HOT tacitly represents the pain as belonging to the very individual that thinks that HOT, itself. The HOT in virtue of which my pain is conscious in effect represents the pain as belonging to the very individual who thinks the HOT. But the individual who has the HOT is thereby the individual for whom that pain is conscious; so one cannot in that respect misidentify the individual that seems to be in pain. I am conscious of a single individual as being in pain and also, in effect, as the individual who is conscious of being in pain. The reason I cannot misidentify the individual I take to be in pain as being anybody other than me is simply that my being conscious of myself as being in pain involves my identifying the individual I take to be in pain as the very individual who takes somebody to be in pain. These considerations clarify the connection between such immunity to error through misidentification and essentially indexical self-reference. The word 'I' and its mental analogue refer to the speaker or thinker, thereby forging a connection between an intentional content in which 'I' or its mental analogue figures and the mental attitude held toward that content or the illocutionary act that verbally expresses it. My essentially indexical use of 'I' or its mental analogue to refer to myself relies on that connection; my thought or assertion that I am *F* in effect represents as being *F* the very individual that thinks that thought or makes that assertion. Similarly, because I identify the individual I take to be in pain as the individual who takes somebody to be in pain, no error of misidentification is in that respect possible.

This explanation leaves open all manner of mundane misidentification, such as my taking myself to be Napoleon or the person in the mirror. All that I am immune from error about is whether the individual I take to be in pain is me, that is, whether it is the very individual that takes somebody to be in pain. My immunity is simply a reflection of the way the first-person pronoun and its mental analogue operate. But however 'I' operates, I can mistakenly think that I am Napoleon or the individual in the mirror.

The stronger immunity to error that Shoemaker describes trades on the special way we have access to being in conscious states. Since the access is a matter of non-inferential HOTs, which like any other thoughts can be mistaken, such strong immunity fails. Thin immunity, by contrast is wholly independent of the processes by which HOTs arise. No matter how one comes to have a HOT, one is disposed to identify the individual represents as being in a particular state as the individual that thinks the HOT. And this amounts to representing the

[1]Tacitly, once again, because the content of a HOT never explicitly describes the individual as the thinkers of the HOT.

individual that is in the targe slate as being oneself. One cannot go wrong about its being oneself on represents as being in the state.

Shoemaker writes that "[m]y use of the word 'I' as the subject of [such statements as that I feel pain or see a canary] is not due to my having identified as myself something" to which I think the relevant predicat applies.[1] But one is disposed to identify the individual one takes to do these things as the individual who takes somebody to do them. So one is after all, disposed at least in this thin way to identify as oneself the individual one takes to feel pain or see a canary. Shoemaker offers the mirror case as an example of a thought above oneself that is not immune to error through misidentification; I see somebody's reflection in a mirror and mistakenly think that I am that person. But so far as thin immunity goes, this case is completely parallel to that of conscious pain. If I take the person in the mirror to be me, I can be wrong about whether the reflection is actually of me. But even here I cannot be wrong about who it is that I take the reflection to be of; I take the reflection to be of the very individual who is doing the taking. In just that way, I could be wrong about whether the person I take to be in pain is Napoleon but I cannot be wrong about whether the person I take to be in pain is the individual doing the taking. The contrast Shoemaker sees between cases in which immunity does and does not occur echoes Wittgenstein's idea that, though I could be mistake] about whether a particular broken arm is mine, I cannot be mistaken about whether a particular pain is mine. He writes: "To ask, 'are you sure that it' *you* who have pains?' would be non-sensical".[2]

But the cases do not differ in any significant way. The error at issue for the strong immunity Shoemaker and Wittgenstein see may be less likely for cases of conscious pain than for broken arms, but it is not impossible. And the cases are parallel in respect of thin immunity. I can be wrong about who the individual is whose arm is broken or who is in pain. But just as I cannot be wrong about whether the individual who takes somebody to be in pain is the individual taken to be in pain, so I cannot be wrong about whether the person who takes somebody's arm to be broken is the person taken to have a broken arm. Thin immunity results simply from the way "I" and its mental analogue function in our first-person thoughts and speech acts.

As noted earlier, claims of privileged access to conscious states tend to rely on the view that a state's being conscious is an intrinsic property of the state itself. But the way one is conscious of a mental state could misrepresent that state even if it were intrinsic to that state. Misrepresentation need not be external to the thing being represented. The idea that being conscious of our mental states is

[1]S. Shoemaker, "Self-Reference and Self-Awareness," p. 9.
[2]Ludwig Wittgenstein, *The Blue and Brown Books,* p. 67. Emphasis original.

intrinsic to those states does shed light on why, even in respect of thin immunity, the mirror and broken-arm cases seem to be different from the pain case.

Suppose I am in pain and the pain is conscious. Its being conscious consists in my being conscious of myself as being in pain. And suppose that the pain's being conscious is intrinsic to the pain itself. It follows that my being conscious of myself as being in pain will then be intrinsic to the pain itself. But my being conscious of myself as being in pain means that the individual I am conscious of as being in pain is the very individual who is conscious of somebody as being in pain. So it will then be intrinsic to my being in pain that I cannot, in that respect, be mistaken about the individual I am conscious of as being in pain.

When I take myself to be Napoleon or to be reflected in a mirror or to have a broken arm, the individual I take to have these properties is again the individual doing the taking. But now an apparent difference from the pain case emerges. Even if one is conscious of oneself as being Napoleon or having a broken arm, one's being thus conscious plainly is not intrinsic to those conditions. So, if a pain's being conscious were intrinsic to the pain, the Napoleon and broken-arm cases would indeed differ from the case of a mental state.

The idea that a mental state's being conscious is intrinsic to that stat even helps explain the initial plausibility of the stronger immunity that Shoemaker describes. If being conscious of a mental state were intrinsic to that state, it would be intrinsic simply to being in a conscious state that one is disposed to regard as being in that state the individual that take somebody to be in that state. Since it would be intrinsic to one's being in a conscious state that it is oneself that one takes to be in that state, then would be no process that leads to one's identifying oneself as the individual that is in the state in question. So there would be no identifying process that could go wrong, and so no way for one to be right in thinking that somebody is in a conscious state but wrong that it is oneself who is in the state.

It is subjectively tempting to see consciousness as an intrinsic feature of our mental states precisely because we are seldom aware, from a first person point of view, of anything that mediates between conscious state and our consciousness of them. To sustain this subjective impression, however one would need some way of individuating mental states on which our awareness of a conscious state is not distinct from the state itself. It is hard to see what means of individuation would have this result which would be independent from the subjective impression under consideration.

Indeed, the way we actually individuate intentional states seems to deliver the opposite result. No intentional state can have two distinct types of mental attitude, such as the attitudes of mental affirmation and doubt. And having a doubt

about something does not result in one's being conscious of that thing. So when a case of doubting is conscious, our consciousness of that doubting must exhibit an assertoric mental attitude. And that means that the consciousness of the doubting is distinct from the doubting itself.

There are other more general reasons to reject the idea that being conscious of a mental state is intrinsic to that state. States may be conscious; one time but not another, as with minor aches or pains that last all day but are not always conscious. If a state's being conscious were intrinsic to the state, it would be puzzling how a particular state could at one moment to be conscious but not at another.

And, if consciousness of mental states is not intrinsic to those states, there is no reason to hold that the stronger immunity Shoemaker describes obtains, nor that the thin immunity that holds for conscious states differs from that which holds for any other self-ascription.

What about an account, then, of the way we are conscious of ourselves that appeals simply to HOTs?

It seemed possible at the outset that the phenomena of essentially indexical self-reference and immunity to error through misidentification might undermine a HOT account of our consciousness of the self. That was because both immunity and the essential indexical seemed to presuppose our having some special access to the self independent of whatever thoughts we have about the self.

Essentially indexical self-reference, we saw, presents no such problem. Essentially indexical first-person thoughts refer to oneself in effect as the individual doing the referring; they refer to oneself as an individual one is disposed to pick out as the individual that thinks the essentially indexical thought. So having such thoughts requires no access to the self beyond that which we have by having thoughts about the self.

A similar conclusion holds for immunity to error through misidentification. It might seem that such immunity requires a privileged type of access to the self; how, otherwise, could we be immune to error in referring to the self? But the error to which we are immune is trivial. It is the error, when I take myself to have some property, *F*, of supposing the individual taken to be *F* as distinct from the individual that takes somebody to be *F*.

We are immune to error through misidentification of the self. But that immunity presupposes no special access we have to the self. It is simply that one cannot, when one thinks that one is, oneself, *F*, *be* wrong about whether it is the individual doing the thinking that one takes to be *F*. No antecedent access to the self figures here, only a particular form of self-reference. So nothing about

immunity to error through misidentification blocks a HOT account of the way we are conscious of ourselves.[1]

5. Identifying Oneself and Self-Consciousness

The idea that immunity to error through misidentification occurs in connection only with the self-ascribing of conscious states, but not with broken arms and being Napoleon, may encourage the Cartesian view that we identify ourselves in the first instance as mental beings. Why else would misidentifying oneself be impossible only in connection with mental self-ascriptions?

This picture is unfounded. For one thing, the error we are immune to does not, in any substantive way, involve the identifying of anything. It is simply the error, when I take myself to have the property, *F*, of thinking that the individual taken to be *F* is distinct from the individual that takes somebody to be *F*. It is perhaps even a bit misleading to describe the error we are immune to as that of misidentification.

Such immunity fails to support the Cartesian conclusion for other reason us well. Since the immunity applies not simply to the ascribing to oneself of conscious states, but to the self-ascribing of non-mental properties a well, it cannot sustain the idea that we identify ourselves primarily in mental terms. How, then, do we identify ourselves? And how does our identifying ourselves fit with the way our thoughts about ourselves involve essentially indexical self-reference and are immune to error through misidentification.

There is no single way we identify ourselves. We rely on a large and heterogeneous collection of factors, ranging from considerations that are highly individual to others that are fleeting and mundane. We appeal to location in time and place, current situation, bodily features, the current and past contents of our mental lives, and various psychological characteristics and propensities, indeed, to all the properties we believe ourselves to have. Contrary to pictures conjured up by essentially indexical self-reference and immunity to error, the factors that figure in our identifying ourselves are theoretically uninteresting and have relatively little systematic connection among themselves.

Each of these factors reflects some belief we have about oneself, such as what one's name is, where one lives, what one's physical dimension and location are, and what the current contents are of one's consciousness And all these beliefs

[1]I discuss essentially indexical self-reference and immunity to error through misidentification in some detail in the context of the apparent unity of our conscious states in "Unity of Consciousness and the Self," *Proceedings of the Aristotelian Society* 103, 3 (2003), 325-52. Here, unlike in the earlier discussion, I stress the difference between thin immunity and the kind of immunity described by Shoemaker.

self-ascribe properties by making essentially indexical self-reference, and they are all immune from error through misidentification in the thin way described above. Such immunity and self-reference figure in the way we identify ourselves not because they provide or presuppose any special access to the self, but only because the first-person beliefs on which all self-identification relies exhibit such immunity and self-reference.

We can be in error abut any of these beliefs about ourselves; indeed we could be in error about most of them. One could be wrong about all one's personal history, background, and current circumstances. One migh even be mistaken about one's location relative to other objects if, for example, one lacked relevant sensory input[1] or the input one had was suitably distorted.

One can be wrong even about what conscious states one is currently in. One may take oneself, in a distinctively first-person way, to have belief and preferences that one does not actually have and to lack those one has, and one may in this way be wrong even about the sensations or affective states one is conscious of oneself as being in.

We identify ourselves by reference to batteries of descriptions which our first-person thoughts and beliefs ascribe to ourselves. And we can successfully distinguish ourselves from others even if many of those descriptions are inaccurate. What, then, if all identifying thoughts and beliefs of the sorts just described are erroneous? Can one identify oneself even then?

Arguably not. We distinguish ourselves from other beings, just as we distinguish among all other individuals, on the basis of various properties. So, if one's beliefs about what properties one has are all incorrect, one has nothing accurate to go on. Our incorrect self-ascriptions would still make essentially indexical self-reference and would still exhibit the thin immunity to error described above. But these features of one's self-ascriptions would not help in identifying oneself, since they tell us only that the individual thought to satisfy a particular description is the individual doing the thinking. Essentially indexical self-reference and immunity to error through misidentification cannot short circuit the need to appeal to properties in identifying oneself.

Even if I am conscious of myself as having some thought or desire or as being in pain, I may nonetheless lack that thought, desire, or pain. Consciousness of our mental states is not infallible. But can I also, in such a case, be wrong even about whether it *seems* that I have that thought, desire, or pain? Perhaps there is, after all, a kind of privileged access we have not in connection with whether our

[1]As G. E. M. Anscombe imagines, "The First Person," in *Mind and Language*, ed. S. Guttenplan (Oxford: Oxford University Press, 1975), pp. 45-65, p. 58.

consciousness is correct, but in connection with what our consciousness is. Perhaps appearance and reality coincide at least in that respect; perhaps it makes no sense to talk about the way things seem to seem to one, as against simply the way things seem.[1] If so, perhaps our conscious states do provide an unimpeachable basis for identifying ourselves, not because we are infallible about the states we are conscious of ourselves as being in, but because we are infallible about whether we are conscious of being in them.

But infallibility does not arise here either. One may be wrong about any mental state one takes oneself to be in. But being conscious of oneself as being in some mental state is, itself, just another higher-order mental state; on the HOT hypothesis, it is a thought one has that one is, oneself, in that state. So one could be wrong even about whether one is in that higher-order state, about whether one has the HOT in question. Such higher-order infallibility fares no better than infallibility about first-order states, and can provide no certain foundation for identifying oneself.

A HOT account of the way we are conscious of ourselves relies on a subset of the essentially indexical first-person thoughts we have about ourselves, namely, our HOTs. But the HOTs an individual has are about the same individual as all the other essentially indexical first-person thoughts of that individual. In this respect, if not in others, the pronoun 'I' and its mental analogue function somewhat as proper names do. When we use a proper name, we take each token to refer to the same individual as other tokens do unless countervailing information overrides that default. Similarly with 'I' and its mental analogue; we assume each token refers to the same thing unless something interferes with that default assumption.[2]

The upshot is that we take all our essentially indexical first-person thoughts and beliefs to refer to one and the same individual. The way we are conscious of ourselves is therefore but one aspect of the way we identify ourselves as individuals. We are in the first instance conscious of ourselves by way of our HOTs, but we identify ourselves by way of all our essentially indexical first-person thoughts and beliefs.

---

[1]See Daniel C. Dennett's contention that it makes no sense to talk "about the way things actually, objectively seem to you even if they don't seem to seem that way to you." In *Consciousness Explained*, (Boston, MA: Little, Brown, 1991), p. 132.

[2]As presumably happens with so-called "Multiple Personality Disorder" (now more often known as "Dissociative Identity Disorder").