13

Unity of Consciousness and the Self

I. THE PROBLEM

One of the most central and important phenomena a theory of consciousness must explain is the sense of unity we have in respect of our conscious mental states. It seems that, for mental representations to be mine, they must, as Kant put it, "all belong to one self-consciousness" (*K.d.R.V.*, B132). Indeed, it was just such mental unity to which Descartes appealed in Meditation VI in arguing for the real distinction between mind and body. Whereas the geometrical essence of body guarantees its divisibility, the unity of consciousness ensures that mind is indivisible.

The unity of consciousness is the unity of an individual's conscious mental states. So understanding our sense of such unity requires knowing what it is for a mental state to be a conscious state. I've argued in a number of places that a state's being conscious consists in its being accompanied by what I've called a *higher-order thought* (HOT)—a thought to the effect that one is in the state in question. Let me briefly sketch the idea.

Suppose that one is in some mental state—one has, say, a thought or desire or emotion—but one is in no way whatever aware of being in that state. It will then subjectively seem to one as though one is not in any such state. But a state that one seems subjectively not to be in is plainly not a conscious state. So it's a necessary condition for a state to be conscious that one be aware, or conscious, of being in that state.[1]

In what way, then, are we aware of our conscious mental states? The traditional explanation appeals to inner sense; we are aware of our conscious states in something like the way we are aware of the things we see and hear.[2] It turns out that this idea is

[1] So there is no reason to suppose mental states, of whatever type, cannot occur without being conscious.

Ned Block's notion of phenomenal consciousness tacitly embodies the contrary assumption for qualitative states, since he holds that every qualitative state is phenomenally conscious. See "On a Confusion about a Function of Consciousness," *The Behavioral and Brain Sciences*, 18, 2 (June 1995): 227–247, and "Paradox and Cross Purposes in Recent Work on Consciousness," *Cognition*, 79, 1–2 (April 2001): 197–219.

[2] The phrase 'inner sense' is Kant's: *K.d.R.V.*, A22/B37. Locke uses the related 'internal Sense' (*An Essay Concerning Human Understanding*, edited from the fourth [1700] edition by Peter H. Nidditch, Oxford: Oxford University Press, 1975, II, i, 4, p. 105. For prominent modern exponents of the inner-sense model, see D. M. Armstrong, "What is Consciousness?", in Armstrong, *The Nature of Mind*, St Lucia, Queensland: University of Queensland Press, 1980, pp. 55–67; and William G. Lycan, *Consciousness and Experience*, Cambridge, Massachusetts: MIT Press/Bradford Books, 1996, ch. 2, pp. 13–43, and "The Superiority of HOP to HOT," in *Higher-Order Theories*

hard to sustain. Sensing occurs in various modalities, each with a characteristic range of mental qualities. But there is no distinctive range of mental qualities by way of which we are conscious of our conscious states.

The only other way we are conscious of things is by having thoughts about them as being present. So that must be how we are aware of our conscious states; a state is conscious if one has a HOT about that state. We seem to be conscious of our conscious states in a direct, unmediated way. We can capture that intuitive immediacy by stipulating that HOTs seem to one to rely on no inference of which one is conscious. We are seldom aware of any such HOTs. But we can explain that by supposing that it's rare that HOTs are accompanied by third-order thoughts, and hence rare that HOTs are, themselves, conscious.

The atomistic character of this model, however, may seem to prevent it from being able to explain our sense of the unity of consciousness. If each conscious state owes its consciousness to a distinct HOT, how could we come to have a sense of such unity? Why would all our conscious states seem to belong to a single, unifying self?[3] Why wouldn't a conscious mind seem instead to consist, in Hume's famous words, of "a mere heap or collection of different perceptions"?[4] It's this challenge that I want to address in what follows.

The challenge arguably poses a difficulty not just for an atomistic theory, such as one that appeals to HOTs, but for any account of the way we are actually conscious of our own conscious states. As Kant observed, "the empirical consciousness that accompanies different representations is by itself dispersed and without relation to the identity [that is, the unity] of the subject."[5] Because such empirical consciousness cannot explain unity, Kant posits a distinct, "*transcendental* unity of self-consciousness" (B132).[6] But it's unclear how any such transcendental posit could explain the appearance of conscious mental unity, since that appearance is itself an empirical occurrence.

In what follows, I consider whether the HOT model itself can explain the robust intuition we have that our conscious mental states constitute in some important way a unity, whether, that is, the model can explain why it seems, subjectively, that such unity obtains. One might counter that what matters is actual unity, not the mere subjective impression of unity. And Kant's observation about the dispersed character

of Consciousness, ed. Rocco W. Gennaro, Amsterdam and Philadelphia: John Benjamins Publishers, 2004, pp. 93–113. See also my "Varieties of Higher-Order Theory," in Gennaro, pp. 17–44, §§ II and III.

[3] I am grateful to Sydney Shoemaker for pressing this question, in "Consciousness and Co-consciousness," in *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. Axel Cleeremans, Oxford: Clarendon Press, 2003, pp. 59–71.

[4] David Hume, *A Treatise of Human Nature* [1739], ed. L. A. Selby-Bigge, Oxford: Clarendon Press, 1888, I, IV, ii, p. 207. Cf. Appendix, p. 634. For the famous "bundle" statement, see I, IV, vi, p. 252.

[5] If so, the way we are conscious of our conscious states cannot yield a sense of mental unity. Immanuel Kant, *Critique of Pure Reason*, tr. and ed. Paul Guyer and Allen W. Wood, Cambridge: Cambridge University Press, 1998, B133.

[6] And he warned against what he saw as the traditional rationalist error of relying on our subjective sense of unity to infer that the mind as it is in itself is a unity (First Paralogism, *K.d.R.V.*, B407–413).

of empirical consciousness suggests that no empirical account can help explain such actual unity.

But, whatever the reality, we must also explain the appearance of unity. And absent some implausible thesis about the mind's being transparent to itself, we cannot explain that appearance simply by appeal to the reality.[7] In any case, it is arguable that the appearance of conscious unity is, itself, all the reality that matters. The consciousness of our mental lives is a matter of how those mental lives appear to us. So the unity of consciousness simply is the unity of how our mental lives appear. We need not independently address the challenge to explain any supposed actual underlying unity of the self. Actual unity will seem important only on the unfounded Cartesian thesis that, for the mind, appearance and reality coincide.

II. CLUSTERS, FIELDS, AND INFERENCE

Our goal is to see whether the HOT model can explain the subjective impression we have of mental unity. One factor that helps some is that HOTs often operate not on single mental states, but on fairly large bunches. For evidence of this, consider the so-called cocktail-party effect, in which one suddenly becomes aware of hearing one's name in a conversation that one had until then consciously experienced only as part of a background din. For one's name to pop out from that seeming background noise, one must all along have been hearing the separate, articulated words of the conversation. But, since one was conscious of one's hearing of the words only as an undifferentiated auditory experience, the HOT in virtue of which one was conscious of one's hearing all those words must have represented the hearing of them *as* a single undifferentiated bunch, that is, as a background din. Doubtless this also happens with the other sensory modalities. That HOTs sometimes operate in this wholesale way helps explain our sense of mental unity; HOTs often unify into a single awareness a large bunch of experiences, on any of which we can focus more or less at will.

There is another, related kind of mental unity. When qualitative states are conscious, we typically are conscious of them not just individually, but also in respect of their apparent spatial relations to other states, of both the same sensory modality and others. We experience each conscious sensation in relation to every other, as being to the right or the left or above or below each of the others.[8] And, by calibrating such apparent locations across modalities, so that sights and sounds, for example, are coordinated in respect of place, we yoke the sensory fields of the various modalities together into what seems to us to be a single, modality-neutral field. Qualitative states are related in this way even when they are not conscious. But when we are conscious of the relevant mental qualities as being spatially related, this also contributes to our sense of having a unified consciousness.

[7] Indeed, the need to appeal to transparency makes any such explanation circular, since whatever plausibility such transparency may have rests in part on the apparent unity of consciousness.

[8] For problems about the way we are conscious of qualitative states as spatially unified within sensory fields, see my "Color, Mental Location, and the Visual Field," *Consciousness and Cognition*, 9, 4 (December 2000): 85–93, § IV. See also "Sensory Qualities, Consciousness, and Reception," ch. 7 in this volume, § VII.

A third factor that contributes to this sense of mental unity is conscious reasoning. When we reason consciously we are aware of our intentional states as going together to constitute larger rational units. We not only hold mental attitudes toward individual intentional contents; we also hold what we may call an *inferential attitude* towards various groups of contents. We hold, in effect, the attitude that we would never mentally deny some particular member of a group while mentally affirming the rest. This inferential attitude often fails to be conscious. But awareness of such rational unity not only results in an impression of causal connection among the relevant states; it also contributes to our sense of the unity of consciousness, since it makes one conscious in one mental breath of distinct contents and mental attitudes.

Indeed, it seems that most of our intentional states, perhaps all of them, fall into groups towards which we are disposed to hold such inferential attitudes. This encourages the idea that some special mental unity of the sort stressed by Descartes and Kant underlies all our intentional states. But the HOT model suffices to explain such unity; we can explain our consciousness of such inferential connections as resulting from HOTs' representing our intentional states as being thus connected.

III. THE SELF AS RAW BEARER

Wholesale operation of HOTs, of these sorts and others, doubtless helps to induce some conscious sense of unity among our mental states. But that will only go so far. Since no single HOT covers all our conscious states, the basic problem remains. How can we explain a sense of unity that encompasses states made conscious by distinct HOTs?

A HOT is a thought to the effect that one is in a particular mental state or cluster of states. So each HOT refers not only to such a state, but also to oneself as the individual that's in that state. This reference to oneself is unavoidable. Having a thought about something makes one conscious of it only when the thought represents that thing as being present. But being conscious of a state as present is being conscious of it as belonging to somebody. And being conscious of a state as belonging to somebody other than oneself would plainly not make it a conscious state.[9]

By itself, however, such reference to a bearer will not give rise to a sense of unity, since each HOT might, for all we know so far, refer to a distinct self. A sense of unity will result only if it seems, subjectively, that all our HOTs refer to one and the same self.

HOTs characterize their target states in terms of mental properties such as content, mental attitude, and sensory quality. But HOTs have far less to say about the self to

[9] Might there be types of creature for which the impersonal thought simply that a pain occurs would make that pain conscious, assuming no conscious inferential mediation? (I owe this suggestion to Jim Stone, personal communication.) Perhaps so, if there are creatures that literally don't distinguish themselves in thought from anything else. But all the nonlinguistic creatures we know of do seem to draw that distinction in a robust way, and few theorists now endorse the speculation that even human infants fail to do so.

One might question whether having a thought about something makes one conscious of that thing only if the thought represents it as being present. But independent of that, unless HOTs assign their target states to some specific individual, those HOTs will only be about mental-state types, as against individual tokens.

which they assign those states. A HOT has the content: I am in a certain state. So each HOT characterizes the self to which it assigns its target solely as the bearer of that target state and, by implication, as the individual that thinks that HOT itself. Just as we understand the word 'I' as referring to whatever individual performs a speech act in which the word occurs, so we understand the mental analogue of 'I' as referring to whatever individual thinks a thought in which that mental analogue occurs.

We must not construe HOTs as actually having the content that whoever thinks this very thought is also in the target state. The word 'I' does not literally mean *the individual performing this speech act*. Though each token of 'I' refers to the individual that uses it in performing a speech act, it does not do so by referring to the speech act itself.[10] We determine the reference of each token of 'I' by way of the containing speech act, but 'I' does not actually refer to that speech act. David Kaplan's well-known account suggests one way in which this may happen. The reference of 'I', he urges, 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.[11]

Similarly, every thought we could express by such a speech act refers to the individual that thinks that thought, but not because the thought literally refers to itself. What the mental analogue of 'I' refers to is determined by which individual thinks the thought, but not because that mental analogue actually refers to the containing thought. This is important because, if HOTs were about themselves, it would then be open to argue that each HOT makes one conscious of that very HOT, and hence that all HOTs are conscious. But as noted earlier, we are seldom aware of our HOTs.[12] Still, since we would identify what individual a token mental analogue of 'I' refers to as the individual that thinks the thought containing that token, we can regard the thought as in effect characterizing that referent as the individual who thinks that very thought. Each first-person thought thus disposes us to have another thought that identifies the self as the thinker of that first-person thought.

HOTs make us conscious not only of their target states, but also of the self to which they assign those targets. And, by seeming subjectively to be independent of

[10] *Pace* Hans Reichenbach, "Token-Reflexive Words," *Elements of Symbolic Logic*, New York: Macmillan, 1947, § 50.

[11] David Kaplan, "Demonstratives," in *Themes From Kaplan*, ed. Joseph Almog, John Perry, and Howard Wettstein, with the assistance of Ingrid Deiwiks and Edward N. Zalta, New York: Oxford University Press, 1989, pp. 481–563, pp. 505–507. Kaplan posits a character of 'I', which is a function whose value, for each context, is the speaker or agent of that context.

[12] In "Two Concepts of Consciousness," I wrongly suggested that we could so construe the content of HOTs (*Philosophical Studies*, 49, 3 [May 1986]: 329–359, § 11, pp. 344 and 346; reprinted as ch. 1 in this volume), and Thomas Natsoulas subsequently drew attention to the apparent consequence that all HOTs would be conscious ("What is Wrong with the Appendage Theory of Consciousness?", *Philosophical Psychology*, VI, 2 [June 1993]: 137–154, p. 23, and "An Examination of Four Objections to Self-Intimating States of Consciousness," *The Journal of Mind and Behavior*, X, 1 [Winter 1989]: 63–116, pp. 70–72). But a HOT need not explicitly be about itself to represent its target as belonging to the individual we can independently pick out as thinking that HOT.

It is also arguable that even if HOTs had the content that whoever has this thought is in the target state, HOTs still wouldn't refer to themselves in the way required to make one conscious of them. See my "Higher-Order Thoughts and the Appendage Theory of Consciousness," *Philosophical Psychology*, VI, 2 (June 1993): 155–167.

any conscious inference, HOTs make it seem that we are conscious of our conscious states in a direct, unmediated way. But that very independence HOTs have from conscious inference also makes it seem that we are directly conscious of the self to which each HOT assigns its target.

Every HOT characterizes the self it refers to solely as the bearer of target states and, in effect, as the thinker of the HOT itself. Nothing in that characterization implies that this bearer is the same from one HOT to the next. But there is also nothing to distinguish one such bearer from any other. And our seeming to be aware in a direct and unmediated way of the self each HOT refers to tilts things towards apparent unity. Since we seem to be directly aware of the self in each case, it seems subjectively as though there is a single self to which all one's HOTs refer, a single bearer for all our conscious states.

HOTs are not typically conscious thoughts; no HOT is ever conscious unless one has a third-order thought about it. So long as HOTs are not conscious, one will not be conscious of their seeming all to refer to a single self. But HOTs do sometimes come to be conscious; indeed, this is just what happens when we are introspectively conscious of our mental states. Introspective consciousness occurs when we are not only conscious of those states, but also conscious that we are.[13]

When HOTs do become conscious, we become aware both of the sparse characterization each HOT gives of the self and of the unmediated way we seem to be conscious of that self. So introspecting our mental states results in a conscious sense of unity among those states even when the states are conscious by way of distinct HOTs. This helps explain why our sense of unity seems to go hand in hand with our ability to engage in introspective consciousness. Indeed, being conscious of our HOTs when we do introspect leads even to our being conscious of the self those HOTs refer to as something that's conscious of various target states, and thus to the idea of the self as a conscious being, a being, that is, that's conscious of being aware of things.[14] Introspective consciousness results in a sense of one's conscious states as all unified in a single conscious subject.

It's worth noting in this connection that Hume's famous problem about the self results from his tacit adoption of a specifically perceptual model of introspecting; one cannot find a self when one seeks it perceptually.[15] The HOT model, by contrast,

[13] For more on introspective consciousness, see "Introspection and Self-interpretation," ch. 4 in this volume.

Occasionally we are even conscious of introspecting, which suggests the occurrence of fourth-order thoughts. But, though such thoughts would be about third-order thoughts, their explicit content would very likely be simply that one is introspecting. And if so, they would not have explicitly fourth-order content, which would arguably at best be difficult to process.

[14] This notion of a conscious being goes well beyond a creature's simply being conscious rather than, say, asleep or knocked out, what I have elsewhere called *creature consciousness*. A creature is conscious in this weaker way if it is awake and mentally responsive to sensory input. Creature consciousness thus implies that a creature will be conscious of some sensory input, but in principle that could happen without any of its mental states being conscious states.

[15] *Treatise*, I, IV, vi, p. 252. Similarly, various contemporary theorists seem to assume that introspective access to our mental states must be perceptual. See, e.g., Fred Dretske, "Introspection," *Proceedings of the Aristotelian Society*, CXV (1994/5): 263–278, and *Naturalizing the Mind*, ch. 2;

provides an informative explanation of the way we do seem to be introspectively conscious of the self.

Still, we have a sense of conscious unity even when we are not introspecting. We often become conscious of ourselves, in a way that seems direct, as being in particular mental states. And that leads us to expect that we could readily become conscious of all our mental states, more or less at will. We expect, moreover, that any such consciousness of our mental states will seem direct and unmediated. And that expectation amounts to a tacit sense that our conscious states form a unity even at moments when we are not actually conscious of any such unity. This tacit sense of mental unity arises in just the way our being disposed to see objects in particular places leads to a tacit, dispositional sense of where those objects are and how they fit together, even when we are not actually perceiving or thinking about them. We not only have an explicit sense of the unity of our conscious states, but a dispositional sense of unity as well.

The idea of being thus disposed to see our conscious states as unified may recall Peter Carruthers's view that a mental state's being conscious is a matter not of its being accompanied by an actual HOT, but rather of its being disposed to be so accompanied. This will not do, since being disposed to have a thought about something doesn't make one in any way conscious of that thing.[16] But we needn't adopt the dispositional HOT model to recognize that our sense of conscious unity can in part be dispositional; our sense of how things are is often a matter of how we are disposed to find them.

IV. THE BATTERY MODEL

The seemingly direct awareness each HOT gives us of the bearer of its target state leads to an initial sense that there is a single bearer to which all our conscious states belong. And the sparse way HOTs characterize that bearer bolsters that sense of unity. But this sparse characterization is not enough to identify ourselves; we do not, *pace* Descartes, identify ourselves simply as bearers of mental states. Still, it turns out that the way we do identify ourselves reinforces in an important respect our sense of the unity of consciousness.

John R. Searle, *The Rediscovery of the Mind*, Cambridge, Massachusetts: MIT Press/Bradford Books, 1992, pp. 96–7 and 144; Gilbert Harman, "Explaining Objective Color in terms of Subjective Reactions," *Philosophical Issues: Perception*, 7 (1996): 1–17, p. 8; reprinted in Alex Byrne and David Hilbert, eds., *Readings on Color, volume 1: The Philosophy of Color*, Cambridge, Massachusetts: MIT Press/Bradford Books, 1997, pp. 247–261; and Sydney Shoemaker, "Introspection and Phenomenal Character," *Philosophical Topics*, 28, 2 (Fall 2000): 247–273.

[16] See, e.g., my "Thinking that One Thinks," ch. 2 in this volume, and "Consciousness and Higher-Order Thought," *Macmillan Encyclopedia of Cognitive Science*, Macmillan Publishers, 2002, pp. 717–726. For Carruthers's view, see Peter Carruthers, *Phenomenal Consciousness: A Naturalistic Theory*, Cambridge: Cambridge University Press, 2000, and "HOP over FOR, HOT Theory," in *Higher-Order Theories of Consciousness*, ed. Gennaro, pp. 115–135. For difficulties in Carruthers's defense of that view, See my "Explaining Consciousness," in *Philosophy of Mind: Contemporary and Classical Readings*, ed. David J. Chalmers, New York: Oxford University Press, 2002, pp. 406–421, at pp. 410–11, and "Varieties of Higher-Order Theory," § IV.

We identify ourselves as individuals in a variety of ways that have little systematic connection, relying on considerations that range from personal history, bodily features, and psychological characteristics to current location and situation. There is no magic bullet by which we identify ourselves, only a vast and loose collection of considerations, each of which is by itself relatively unimpressive, but whose combination is enough for us to identify ourselves whenever the question arises.

Identifying oneself consists of saying who it is that one is talking or thinking about when one talks or thinks about oneself, that is, when one has first-person thoughts or makes the first-person remarks that express those thoughts. And one picks out the individual those first-person thoughts are about by reference to a diverse collection of contingent properties, such as those just mentioned. For any new first-person thought, the reference that thought makes to oneself is secured by appeal to what many other, prior first-person thoughts have referred to, and this process gradually enlarges the stock of self-identifying thoughts available to secure such reference. Just as we take distinct tokens of a proper name all to refer to the same individual unless something indicates otherwise, so each of us operates as though all tokens of the mental analogue of 'I' in one's first-person thoughts also refer to the same individual. It is not easy, moreover, to override this default assumption.[17] The word 'I' and its mental analogue refer to whatever individual says or thinks something in first-person terms, but absent some compelling countervailing reason we also take them to refer to one and the same individual from one thought or speech act to the next.

The analogy with proper names may recall G. E. M. Anscombe's well-known view that 'I' does not function at all like a proper name. According to Anscombe, the first-person thought that I am standing, for example, does not predicate the concept *standing* of any object, but exhibits instead a wholly unmediated conception of standing.[18] But this view cannot accommodate various fundamental logical relations, such as the incompatibility of my thought that I am standing with your thought that I am not. Even on the sparse characterization of the referent of 'I' described earlier, these logical relations demand that 'I' function as some type of referring expression.

Having a conscious sense of unity does not require having an explicit, conscious thought that all occurrences of the mental analogue of 'I' refer to a single thing. We typically have a sense that we are talking about one and the same individual when we use different tokens of a proper name even though we seldom have any actual thought to the effect that such coreference obtains. The same holds for talking or thinking about oneself using different tokens of 'I' or its mental analogue.

HOTs are first-person thoughts, and these considerations all apply to them. We appeal to a broad, heterogeneous collection of contingent properties to specify the individual each HOT represents its target as belonging to, and we take that battery of descriptions to pick out a single individual. Since this process extends to our HOTs, it enriches our description of the self to which our HOTs assign their target states, thereby reinforcing and consolidating the subjective sense each of us has that our

[17] Perhaps as in cases of so-called Multiple Personality or Dissociative Identity Disorder.

[18] "The First Person," in *Mind and Language: Wolfson College Lectures 1974*, ed. Samuel Guttenplan, Oxford: Clarendon Press, 1975, pp. 45–65.

conscious states all belong to a single individual. There is nothing special about the way we are conscious of our mental states or of the self they belong to that issues in this subjective sense. It results simply from an extension of our commonsense assumption that the heterogeneous collection of ways in which we identify ourselves combine to pick out one individual, that the 'I' in all our first-person thoughts and remarks refers to a single self.

It might be thought that the way we are conscious of ourselves must be special, since we identify ourselves, as such, by being conscious of ourselves, and identifying oneself, as such, is a precondition for identifying anything else.[19] But no informative identification of ourselves, as such, is needed to identify other things. Perceptually identifying objects other than oneself relies on some relationship that holds between oneself and those other objects, but the relevant relationship consists in the perceiving itself, and one needn't identify oneself to perceive something else. Perhaps in identifying an object relative to other things we often use as a fixed point the origin of one's coordinate system, and that may make it seem that identifying oneself is a precondition for perceptually identifying anything. But we do not ordinarily identify things perceptually relative to ourselves, but relative to a larger scheme of things that contains the target object. When appeal to that larger framework fails for whatever reason, nothing about the way we identify ourselves independently of that larger framework will come to our rescue.

Since this reinforced sense of unity results from our HOTs' functioning just as other first-person thoughts do to pick out a single individual, we are conscious of that reinforcement only when some of our HOTs are, themselves, conscious.[20] Introspective consciousness is once again pivotal for our conscious sense of mental unity.

Each HOT represents its target state as belonging to some individual. One secures reference to that individual by way of other first-person thoughts, each of which contributes to the heterogeneous collection of contingent properties by way of which we identify ourselves. We thereby identify the individual to which each HOT assigns its target as being the same from one HOT to the next. Since introspecting consists in being conscious of our HOTs, it results in our being conscious of those HOTs *as* seeming all to assign their targets to some single individual. One becomes conscious of oneself as a center of consciousness. Indeed, this provides an answer, which Hume despaired of giving, to his challenge "to explain the principles, that unite our

[19] On the idea that self-identification is a precondition for identifying anything else, see, e.g., Sydney Shoemaker, "Self-Reference and Self-Awareness," *The Journal of Philosophy*, LXV, 19 (October 3, 1968): 555–567, reprinted with slight revisions in Shoemaker, *Identity, Cause, and Mind: Philosophical Essays*, Cambridge: Cambridge University Press, 1984, expanded edition, Oxford: Clarendon Press, 2003, pp. 6–18 (references below are to the reprinted version); David Lewis, "Attitudes *De Dicto* and *De Se*," *The Philosophical Review*, LXXXVIII, 4 (October 1979): 513–543, reprinted in Lewis, *Philosophical Papers*, I, New York: Oxford University Press, 1983, pp. 133–159; and Roderick M. Chisholm, *Person and Object: A Metaphysical Study*, La Salle, Illinois: Open Court, 1976, ch. 1, §5, and *The First Person*, Minneapolis: University of Minnesota Press, 1981, ch. 3, esp. pp. 29–32.

[20] Simply operating as though 'I' has the same referent in all one's first-person thoughts is enough, however, to produce the tacit sense of unity mentioned at the end of § III.

successive perceptions in our thought or consciousness'' (*Treatise*, Appendix, p. 636). HOTs lead to our interpreting the states they are about as all belonging to a single conscious self.

It is important to stress that the single subject which we're conscious of our conscious states as belonging to may not actually exist. It may be, for one thing, that there is no subject of which we actually have direct, unmediated consciousness. Perhaps the subject one's HOTs refer to isn't even the same from one HOT to the next. Even though the mental analogue of 'I' refers in each first-person thought to whatever individual thinks that thought, perhaps the relevant individual is different, even for a particular person's HOTs, from one HOT to another. For present purposes, however, these possibilities don't matter. As noted at the outset, the aim here is not to sustain the idea that a single, unified self actually exists, but to explain our compelling intuition that it does.

V. THE ESSENTIAL INDEXICAL

There is, however, a well-known reason to question whether we do actually identify ourselves by way of a heterogeneous battery of contingent properties. The reason has to do with the special way in which we sometimes refer to ourselves when we speak using the first-person pronoun and frame thoughts using the mental analogue of that pronoun.

Consider John Perry's vivid example, in which I see a trail of sugar apparently spilling from somebody's grocery cart. Even if I am the one spilling it, my thinking that the person spilling sugar is making a mess does not imply that I think that I, myself, am making a mess.[21] Reference to oneself, as such, uses what Perry dubs the essential indexical, also called by traditional grammarians the indirect reflexive, because it plays in indirect discourse the role played in direct quotation by the first-person pronoun.[22] And such reference to oneself seems to operate independently of any contingent properties in terms of which one might describe and identify oneself.[23]

Every HOT refers to the self in this essentially indexical way. A HOT cannot represent its target as belonging to oneself under some inessential description; it must represent that target as belonging to oneself, as such. But a thought's being about

[21] John Perry, ''The Problem of the Essential Indexical,'' *Noûs* XIII, 1 (March 1979): 3–21. See also P. T. Geach, ''On Beliefs about Oneself,'' *Analysis*, 18, 1 (October 1957): 23–4, reprinted in Geach, *Logic Matters*, Oxford: Basil Blackwell, 1972, pp. 128–9; A. N. Prior, ''On Spurious Egocentricity,'' *Philosophy*, XLII, 162 (October 1967): 326–335; G. E. M. Anscombe, ''The First Person,'' in *Mind and Language*, ed. Samuel Guttenplan, Oxford: Oxford University Press, 1975, pp. 45–65; Steven E. Boër and William G. Lycan, ''Who, Me?'', *The Philosophical Review*, LXXXIX, 3 (July 1980): 427–66; Hector-Neri Castañeda, ''On the Logic of Attributions of Self-Knowledge to Others,'' *The Journal of Philosophy*, LXV, 15 (August 8, 1968): 439–56; Roderick M. Chisholm, *The First Person*, chs. 3 and 4; and David Lewis, ''Attitudes *De Dicto* and *De Se*.''

[22] And indirect discourse matters here in specifying intentional content.

[23] For an argument that this type of self-reference conflicts with the HOT model, see Dan Zahavi and Josef Parnas, ''Phenomenal Consciousness and Self-Awareness: A Phenomenological Critique of Representational Theory,'' *Journal of Consciousness Studies*, 5, 5–6 (1998): 687–705, § III.

oneself, as such, seems not to rely on any battery of contingent properties. How, then, does the idea that we identify ourselves in terms of such collections of contingent properties square with the requirement that one's HOTs refer to oneself, as such?

Mental states are conscious, when they are, in virtue of being accompanied by HOTs, and each HOT in effect represents its target as belonging to the individual who thinks that HOT. This representing is tacit, since as we saw two sections ago, it is not mediated by any actual reference to the thought itself.

Essentially indexical self-reference occurs not just with HOTs, but with all our first-person thoughts. Suppose I think that I, myself, have the property of being *F*. My thought that I, myself, am *F* in effect represents as being *F* the very individual who thinks that thought. In this way I refer to myself, as such. I refer to myself, as such, when I refer to something, in effect, as the individual that does the referring. No additional connection between first-person thoughts and the self is needed.

In Perry's case, I begin by thinking that somebody is spilling sugar and I come to realize that I, myself, am that person. What I discover when I make that realization is that the individual who is spilling sugar is the very same as the individual who thinks that somebody is spilling sugar; the person being said or thought to spill is the very person who is saying or thinking that somebody spills. By identifying, in effect, the individual a thought purports to be about with the individual who thinks that thought, the essential indexical tacitly links what the thought purports to be about to the very act of thinking that thought.

HOTs are just a special case of first-person thoughts, and all the same things apply to them. Each HOT tacitly represents its target as belonging to the individual that thinks that very HOT. In this way, every HOT represents its target as belonging to oneself, as such.

Reference to oneself, as such, seems to be independent of any particular way of describing or characterizing oneself. But we can now see that there is one type of characterization that is relevant to such reference. When I think, without any essentially indexical self-reference, that the person spilling sugar is making a mess, my thought is, as it happens, about the very individual who thinks that thought, though not about that individual, as such. By contrast, when I think that I, myself, am spilling sugar, my thought then does ascribe the spilling of sugar to the very individual who thinks that thought, as such. As I've stressed, that essentially indexical thought does not explicitly refer to itself. Reference to the thinker, as such, is secured not by descriptive content, but because it's that individual who holds a mental attitude toward the relevant content. The essential indexical ties intentional content to mental attitude.[24]

This connection between the individual that's thought to be spilling sugar and the individual doing the thinking obtains solely in virtue of the tie between content and mental attitude. So it's independent of any other contingent properties one may think of oneself as having. That connection, moreover, is all one needs to refer to oneself, as such. The mental analogue of the word 'I' refers to whatever individual thinks a thought in which that mental analogue occurs.

[24] Any account, such as Kaplan's, that relies on context to determine the referent of 'I' and its mental analogue, will appeal to the performing of the relevant speech act or mental act.

In the first person, the essential indexical in effect identifies the self it refers to as the individual who thinks a thought or performs a speech act. This thin way of identifying oneself provides almost no information. But, by the same token, there is no conflict between our referring to ourselves in this way and the battery model of how we identify ourselves. The essential indexical picks something out as the individual that thinks a particular thought; the battery model provides an informative way of saying just which individual that is. This is why we seem unable ever to pin down in any informative way what the essential indexical refers to. The essential indexical refers to the thinker of a thought; an informative characterization of the self depends on one's applying some battery of descriptions to oneself in an essentially indexical way.

A thought about oneself, as such, refers to the individual that thinks that thought, but its content does not explicitly describe one as the thinker of the thought. Since essentially indexical thoughts refer independently of any particular description that occurs in their content, it's tempting to see them as referring in an unmediated way, which might then even provide the foundation for all other referring.[25] But such reference is not unmediated and cannot provide any such foundation. Reference to the thinker, as such, is mediated not by descriptive content, but by the tie the essential indexical tacitly forges between a thought's content and its mental attitude.

There is a sense we sometimes have of ourselves that makes it hard to see how, as conscious selves, we could find ourselves located among the physical furniture of the universe.[26] It's sometimes urged that essentially indexical self-reference is responsible for that appearance of mystery about being the subject of conscious experience, since the essential indexical occurs ineliminably only in describing such subjects. The present account suggests an explanation. It may be that the self seems difficult to fit into our ordinary objective framework because essentially indexical reference to the self is secured not by a thought's descriptive content, but by a tie between that thought's content and its mental attitude.

Reference to somebody, as such, occurs in cases other than the first person. I can describe others as having thoughts about themselves, as such, and the same account applies. Thus I can describe you as thinking that you, yourself, are *F*, and your thought is about you, as such, just in case your thought, cast in the first person, refers to an individual in a way that invites identifying that individual as the thinker of that thought.

Thoughts need not be conscious, and essentially indexical reference to oneself can occur even when they are not. I realize that I, myself, am the one spilling sugar if I would identify the person I think is spilling sugar with the person that thinks that thought. If that thought fails to be conscious, my realization will fail to be as well.

Still, when a thought is about oneself in that essentially indexical way, there is a tendency for it to be conscious over and above whatever tendency exists for thoughts

[25] See references in n. 19.

[26] Consider the puzzled cognitive disorientation Wittgenstein writes of "when I, for example, turn my attention in a particular way on my own consciousness, and, astonished, say to myself: THIS is supposed to be produced by a process in the brain!—as it were clutching my forehead" (*Philosophical Investigations*, ed. G. E. M. Anscombe and R. Rhees, tr. G. E. M. Anscombe, Oxford: Basil Blackwell, 1953, § 412).

that aren't about oneself in that way.[27] When one essentially indexically thinks that one is *F*, one is disposed to identify the person one thinks is *F* with the person who thinks that thought. So one is disposed to have a thought about that essentially indexical thought, and so to be conscious of that thought.

When an essentially indexical thought about myself is conscious, the HOT I have about that conscious thought describes it as being about the individual that thinks the thought. That HOT also in effect describes its target state as belonging to the very individual that thinks the HOT, itself. So, when a conscious thought is about oneself, as such, one is in effect conscious of that thought as being about the individual that not only thinks the thought but is also conscious of thinking it.

Does essentially indexical self-reference make a difference to the way beliefs and desires issue in action? Kaplan's catchy example of my essentially indexical thought that my pants are on fire[28] may make it seem so, since I might behave differently if I thought only that some person's pants are on fire without also thinking that I am that person. Similarly, my thinking that I, myself, should do a certain thing might result in my doing it, whereas my merely thinking that DR should do it might not result in my doing it if I didn't also think that I was DR.

Such cases require care. My doing something when I think I should arguably results from that belief's interacting with my desire to do what I should. Since I very likely would not desire to do what DR should do if I didn't think that I was DR, I would then have no desire that would suitably interact with my belief that DR should do that thing. And if, still not recognizing that I am DR, I nonetheless had for some reason a desire to do what DR should do, my belief that DR should do something would then very likely result in my doing it. The need here for a belief to make essentially indexical self-reference is due solely to the essentially indexical self-reference made by the relevant desire.

The situation is similar with thinking that one's pants are on fire. Even disregarding perceptual asymmetries, the desires that would pertain to my belief that my pants are on fire will doubtless differ in relevant ways from desires that would pertain to my belief that your pants are on fire, and so to my belief that DR's pants are on fire if I don't know that I am DR.

Many of one's beliefs and desires, however, do not refer to oneself at all, as such or in any other way. I might want a beer and think that there is beer in the refrigerator. The content of that desire might refer to me; it could be a desire that I have a beer. Things might well then be different if I had instead a desire only that DR have a beer. But the desire need not refer to me at all; its content could instead be simply that having a beer would be nice.[29] And that desire would likely lead to my acting, not because

[27] On the tendency of thoughts in general to be conscious when they occur in creatures with suitable ability to think about their own thoughts, see "Why Are Verbally Expressed Thoughts Conscious?", ch. 10 in this volume.

[28] "If I see, reflected in a window, the image of a man whose pants appear to be on fire, my behavior is sensitive to whether I think, 'His pants are on fire' or 'My pants are on fire', though the object of thought may be the same" ("Demonstratives," p. 533).

[29] Affective states, such as happiness, sadness, anger, and the like, also have intentional contents cast in such evaluative terms. See my "Consciousness and its Expression," ch. 11 in this volume, § IV.

the content of the desire refers to me, but because I am the individual that holds the desiderative attitude towards that content. Essentially indexical self-reference is not needed for beliefs and desires to issue in action.[30]

According to David Lewis, the objects toward which we hold attitudes are best understood as properties. Holding an attitude, he urges, consists in self-ascribing a property. And he argues that this account not only handles attitudes toward essentially indexical contents, which he calls attitudes *de se*, but also provides a uniform treatment for the attitudes, whatever their content.

Lewis's main concern is to say what kind of thing the objects of the attitudes are. And he seems to take as primitive the notion of self-ascribing invoked in this account. But it's still worth examining just what would be needed to ascribe a property to oneself, as such. In particular, does one need explicitly to think about or to represent oneself, as such?

Lewis holds that all ascribing of properties to individuals takes place under a description, which in the relevant kind of case is a relation of acquaintance. Self-ascribing, then, is the special case in which one ascribes a property to oneself "under the relation of identity," which he characterizes as "a relation of acquaintance par excellence" (543/156; see n. 19).

Lewis goes on, then, to construe all ascribing of properties to individuals as the ascribing of some suitable property to oneself. The ascribing of a property, *P*, to some individual under a relation of acquaintance, *R*, is the ascribing to oneself of the property of bearing *R* uniquely to an individual that has that property, *P*.[31] The property one self-ascribes specifies the content of the attitude one thereby holds. And, since the relation of acquaintance, *R*, figures in the property one self-ascribes, it is part of the content toward which one holds an attitude. One explicitly represents the individual one thinks has property *P* as the individual with which one is acquainted in the relevant way.

Because regress would occur if one applied this account to the special case of self-ascribing, it is not obvious how the relation of acquaintance is secured in that case. Still, self-ascribing is the special case of the ascribing of properties in which the

[30] When action results from a desire whose content is simply that having a beer would be nice, an interaction between mental attitude and content is again operative: It's my holding that desiderative attitude that results in action. So it may well be that some such interaction between attitude and content is needed for belief-desire pairs to lead to action, whether or not that interaction issues in essentially indexical self-reference.

Philip Robbins has suggested (personal communication) that HOTs might operate, as do desires, without first-person content, in which case HOTs also would not need to be cast in essentially indexical terms. But explaining the consciousness of mental states makes heavier representational demands than explaining action. To explain an action we need a belief-desire pair that would plausibly cause that action; my holding a desiderative attitude toward the content that having a beer would be nice would plausibly do so. To explain a state's being conscious, however, we must explain an individual's being conscious of being in that state, and that means actually representing oneself as being in that state, at least for creatures that distinguish in thought between themselves and everything else (see n. 9).

[31] "Postscripts to 'Attitudes *De Dicto* and *De Se*'," *Philosophical Papers*, I, New York: Oxford University Press, 1983, pp. 156–159, at p. 156.

relation of acquaintance is identity. If self-ascribing follows that model, the content toward which one holds an attitude in self-ascribing will explicitly represent the individual to which one ascribes a property as being identical with oneself. And it is unclear how this might occur unless the attitude explicitly represents that individual as being identical with the individual doing the ascribing. And this, we saw in section III, would cause trouble for the HOT model, since if HOTs explicitly refer to themselves, each HOT would make one conscious of that very HOT, thereby making all HOTs conscious.

But, since Lewis seems to take the notion of self-ascribing as primitive, it needn't follow the general model of the ascribing of properties. And if it doesn't, we can construe the identity between the individual to which a property is ascribed and the individual doing the ascribing as built into the act of self-ascribing, rather than its content. It would then be the performing of that act, rather than some explicit representing, that secures the identity. And the potential difficulty for the HOT model would thereby be averted.

VI. IMMUNITY TO ERROR THROUGH MISIDENTIFICATION[32]

The essential indexical apart, there is another concern about whether we actually identify ourselves by way of a heterogeneous collection of contingent properties. Some of our first-person thoughts appear to be immune to a particular type of error, and it may not be obvious how such immunity is possible if we identify ourselves by way of a battery of contingent properties.

One can, of course, be mistaken in what one thinks about oneself and even about who one is; one might, for example, think that one is Napoleon. And I've argued elsewhere that one can also be mistaken about what mental states one is in, even when those states are conscious. One can be conscious of oneself *as being* in mental states that one is not actually in. The HOT model readily explains this as being due to the having of a HOT that is mistaken in the mental state it ascribes to one.[33]

It's tempting to think that the phrase 'is conscious' is factive, so that one's being conscious of something implies that that thing exists. But even if this is so, one could still be conscious of states that one isn't actually in. Plainly one can be conscious of an actual object as being different from the way it actually is; one can be conscious, for example, of a red object as being green. And one's having a HOT that describes one

[32] This section is significantly recast over the original; see also "Being Conscious of Ourselves," *The Monist*, 87, 2 (April 2004): 159–181, §IV.

[33] See, e.g., "Sensory Qualities, Consciousness, and Perception," ch. 7 in this volume, §V; "Explaining Consciousness," §V; "Consciousness and Metacognition," in *Metarepresentation: A Multidisciplinary Perspective*, Proceedings of the Tenth Vancouver Cognitive Science Conference, ed. Daniel Sperber, New York: Oxford University Press, 2000, pp. 265–295, §V; "Consciousness, Content, and Metacognitive Judgments," *Consciousness and Cognition*, 9, 2, Part 1 (June 2000): 203–214, §V; and "Metacognition and Higher-Order Thoughts," *Consciousness and Cognition*, 9, 2, Part 1 (June 2000): 231–242, §IV.

as being in a state that one is not actually in simply makes one conscious of an existing object, namely, oneself, as being in a state that that object is not actually in.[34]

Even if we can be in error about who we are and what conscious states we are in, perhaps there are other ways in which some of our first-person thoughts cannot be mistaken. Suppose that I consciously feel pain or see a canary. It may be that I can be wrong about whether the state I am in is one of feeling pain or seeing a canary. But if I do think that I feel pain or see a canary, perhaps it cannot then be that I am right that somebody feels pain or sees a canary but wrong that it is I who does so. Sydney Shoemaker has forcefully urged that a range of first-person thoughts cannot be in error in this specific way. And he describes those thoughts in a now classic phrase as being "immune to error through misidentification," specifically with respect to reference to oneself.[35]

Not all first-person thoughts are immune to error in this way. In particular, no first-person thought is thus immune if one comes to have it in the way we come to have thoughts about the mental states of others. Shoemaker recognizes this, noting that one might wrongly take a reflection one sees in a mirror to be a reflection of oneself, and thereby misidentify oneself as the person one sees in the mirror (7). And, if one thought that the person in the mirror is in a particular mental state, one might conclude that one is, oneself, in that state; one could then be right that somebody is in that state, but wrong that it is oneself that is in the state. So Shoemaker does not claim immunity to error whenever one has a thought that one has a particular property, but only when one's thought arises from the special way we have access to our own conscious states.

Even confining ourselves to these cases, however, such immunity to error would threaten the battery model. On that model, we identify the individual each first-person thought refers to by appeal to a heterogeneous collection of contingent properties. But it could turn out that any or, indeed, all of the properties in such a battery do not actually belong to one. So, if we do identify ourselves in that way, then

[34] One might further object, as Elizabeth Vlahos has ("Can Higher-Order Thoughts Explain Consciousness? A Dilemma," MS), that in such a case there is no state that's conscious. This, too, is not a problem. We can simply construe the conscious states our HOTs thus refer to as the notional objects of those HOTs. Or we could equally well say instead that we are conscious in these cases of some relevant occurrent state but in an inaccurate way. Either move effectively meets the objection. See "Metacognition and Higher-Order Thoughts," §I.

[35] Sydney Shoemaker, "Self-Reference and Self-Awareness," p. 8. Shoemaker thinks such immunity applies even when I think I'm performing some action. See also Gareth Evans, "Demonstrative Identification," in Evans, *Varieties of Reference*, ed. John McDowell, Oxford: Clarendon Press, 1982, pp. 142–266.

James Pryor's useful distinction between *de re* misidentification and *wh*-misidentification hinges on the epistemic grounds for holding the relevant beliefs, which are not relevant to what follows ("Immunity to Error through Misidentification," *Philosophical Topics*, 26, 1 and 2 [Spring and Fall 1999]: 271–304.

I have profited from discussion of these issues with Roblin Meeks, and with Michael Martin when an earlier version of this paper was presented at the June 23, 2003, meeting of the Aristotelian Society. See also Meeks, "Identifying the First Person," unpublished Ph.D. dissertation, The City University of New York Graduate Center, 2003, chs. II–IV.

whatever state I think I am in and whatever my basis for thinking that, I could be right that somebody is in that state but wrong that I am the person who's in that state. If immunity to error through misidentification holds, the battery model is wrong.

But there is reason to doubt that such immunity does actually obtain, even for the first-person thoughts under consideration. One might in certain circumstances have such strong empathy with another person that one becomes confused about a mental state of that person, and takes oneself to be in the state. So one might in this way go from seeing another person in some form of emotional elation or anguish to being conscious of oneself as being in that state.

Such overwrought empathy might in some cases lead one actually to be in a felt state of elation or anguish. But that need not happen. It might instead be that no such affective state occurs in one, but one comes nonetheless to be conscious of oneself as being in such a state. This is what would happen if one were conscious of oneself as feeling elation or anguish and yet displayed none of the other characteristic signs of the state one was conscious of oneself as being in.

Suppose I think that I am in pain in what seems to be the characteristic first-person way in which we have access to our own pains. And suppose that there is some pain my thought is about. Why does it seem that I can't be wrong about whether I'm the one who's in pain, as opposed to you? The temptation to think that must stem from the assumption that being conscious of a pain or other state in the relevant subjective way in some way guarantees that I can't be wrong about whether I'm the individual I'm conscious of as being in that state. Subjective access may not guarantee that I'm right about the character of the state I'm conscious of myself as being in, but on this view it does ensure that I'm the one who's in that state if anybody is. One might indeed wonder what else it could mean for one's access to a state to be subjective. If some individual is in a state to which I have subjective access, how could it be somebody other than me who is in that state?

One might simply stipulate that this guarantee is part of what it is for access to be subjective; access is subjective only if it is in this way access to oneself. But it is a substantive question whether the thoughts under consideration are immune to error through misidentification, and it begs that question to assume that all access that seems subjective is actually subjective in this stipulated sense.

Part of what makes access to something seem subjective is that it appears spontaneous and unmediated. And it may be tempting to think that access that seems thus spontaneous and unmediated can occur only in connection with the transparency of the mind to its own states. But the appearance of spontaneity and of lack of mediation is not due to any actual transparency. Access to things often seems spontaneous and unmediated without actually being so; perceiving typically seems spontaneous and unmediated, though we know that it isn't. So, even when we seem to have spontaneous, immediate access to our own mental states, the appearance of spontaneous immediacy may be due solely to the noninferential character of the way we're aware of those states. Our access to these states appears spontaneous and unmediated only because we are unaware of any mediation between the states and our awareness of them.

I have subjective access to a pain when I have a HOT that I am in pain, a HOT that does not appear to me to rely on any inference. And such a HOT can arise, as with all our thoughts, in ways that result in its being erroneous. I can have a seemingly noninferential HOT that I am in pain or some other mental state even though I am not. And, as the empathy case illustrates, I can have such a HOT even though it is you that is actually in that state, rather than I. Immunity to error through misidentification does not obtain. The battery model faces no difficulty from that quarter.

Shoemaker appeals to a passage in which Wittgenstein urges that the first-person pronoun is used differently in statements such as 'I have a broken arm' from the way it's used in statements such as 'I am in pain'.[36] Wittgenstein claims that, though one plainly could be wrong about whether it is one's own arm that's actually broken, "[t]o ask 'are you sure that it's *you* who have pains?' would be nonsensical" (67, emphasis original). Doubtless such a question is typically out of place, but remarks that are typically out of place need not on that account be nonsensical. And, as the example of extreme empathy shows, such a question will in exceptional cases be appropriate, even if rather surprising.

It's tempting to hold that immunity to error obtains because the seemingly spontaneous, unmediated character of the access we sometimes have to mental states seems to guarantee that those states cannot belong to anybody other than oneself. But, since we have no reason to think that this access is actually spontaneous and immediate, as against merely apparent, no such guarantee holds.

The temptation to think that this guarantee does hold rests on the assumption that being conscious of a mental state in the relevant subjective way ensures that I'm not wrong about whether I'm the individual I'm conscious of as being in that state. And this assumption suggests a weaker form of immunity that does hold, despite cases of extreme empathy. I may think in that seemingly immediate way that I am in pain and be right that somebody is in pain, and yet be wrong that I'm the one who's in pain. But I cannot in such a case be wrong about whether it is I who I think is in pain. Similarly, if I think in that seemingly immediate way that I believe or desire something, I cannot be mistaken about whether it is I who I think has that belief or desire. Because this constraint is weaker than the immunity to error Shoemaker describes, I'll refer to it as *thin* immunity to error through misidentification.

Such thin immunity seems unimpeachable; how could one be mistaken in such a case about whether the individual one thinks is in some particular state is oneself? But it may seem that even such thin immunity may threaten the battery model about how we identify ourselves. Suppose, as that model holds, that we do identify the individual each first-person thought refers to by appeal to a heterogeneous collection of contingent properties. Any or even all of the properties in such a battery might turn out not to belong to one. So identifying oneself in that way seems to leave open the possibility that, when I take myself to be in pain or to have some belief, I could be mistaken even about who the individual is that I think is in pain or has that belief.

[36] Shoemaker, "Self-Reference and Self-Awareness," p. 7; Ludwig Wittgenstein, *The Blue and Brown Books*, Oxford: Basil Blackwell, 1958, 2nd edn. 1969, pp. 66–7.

What exactly does this thin immunity guarantee? When I have a conscious pain, I cannot be wrong about whether it's I who I think is in pain, though I can of course be wrong about just who it is that I am; I may, for example, think that I'm Napoleon. How can we capture this delicate distinction? What exactly does thin immunity ensure that I cannot be wrong about?

When I have a conscious pain, I am conscious of being in pain. The error I cannot make is to think that the individual I think is in pain is distinct from the individual that's conscious of somebody's being in pain. The misidentification I cannot make is to take myself be to somebody distinct from the individual doing the identifying. Nonetheless, such thin immunity does leave it open for me to be wrong in some substantive way about who I am, for example, by thinking that I am Napoleon.

The HOT model provides a natural explanation of thin immunity. The mental analogue of the pronoun 'I' refers to whatever individual thinks a thought in which that mental analogue occurs. So each HOT in effect[37] represents its target state as belonging to the individual that thinks that very HOT. When a pain is conscious, the individual the relevant HOT represents that pain as belonging to is the same as the individual that thinks that HOT. So one cannot be wrong about whether the individual that seems to be in pain is the very same as the individual for whom that pain is conscious.

When I think I am in pain, there is no way to go wrong about whether it's I who I think is in pain. So there is no way to misidentify the individual I think is in pain as somebody else. I am conscious of a single individual both as being in pain and, in effect, as the individual that's conscious of being in pain. And I use the mental analogue of 'I' to refer to that one individual.

This thin immunity is no more than an echo of the immunity Shoemaker describes, and the error against which it protects is not substantive. I cannot represent my conscious pain as belonging to somebody distinct from me because a pain's being conscious consists in one's being conscious of oneself as being in pain. And that in turn is simply a matter of one's being conscious of the pain as belonging to the individual that's conscious of it. The thin immunity to error that results consists only in the impossibility of one's being conscious of being in a state and yet conscious of that state as belonging to an individual other than the individual who is conscious of being in it.

Thin immunity doesn't protect against substantive errors about who I am; I might still think I'm Napoleon. So there is no conflict with the battery model. One is trivially immune to error only about whether the individual one is noninferentially conscious of as being in pain is the individual that's conscious of the pain. The battery of contingent properties, by contrast, enables us to distinguish that individual from others, described in terms of various distinguishing properties.[38] Thin immunity has no bearing on error in respect of the contingent properties in such a battery.

[37] In effect, once again, because, although thoughts that contain the mental analogue of 'I' do not actually refer to themselves, still every first-person thought disposes one to have another thought that identifies the referent of that mental analogue as the thinker of the first-person thought. Every first-person thought in that way tacitly, i.e., dispositionally characterizes the self it is about as the thinker of that very thought.

[38] Essentially indexical self-reference also enables one to distinguish oneself from every other individual, but again only in a thin way. It allows me to distinguish myself from others only relative

According to Shoemaker, when one introspectively knows that one is in pain or that one believes something, there is no "role for awareness of oneself as an object to play in explaining my introspective knowledge" of those states. This, he suggests, is of a piece with the immunity to error of our first-person thoughts about our own mental states.[39] And, if awareness of oneself as an object does not figure in introspective access, there is no way for such access to go wrong in respect of which object is in the introspected state.

But this doesn't accurately reflect the way we're conscious of our conscious states. Being conscious of a state on its own, and not of the state as belonging to some individual, is being conscious only of a state type, and not of any particular token. States are perforce states of objects; only types of states are independent of particular objects. It's for that reason that HOTs must make one conscious of mental states as belonging to a particular individual, namely, oneself. And those HOTs will accordingly make one conscious of oneself as the object that is in those states.

As noted at the outset, the perceptual model of the access we have to our own mental states makes a mystery of how we could be aware of a self. But we no more perceive our mental states than we perceive the self to which they belong. Construing introspective awareness in terms of conscious HOTs explains how awareness of oneself occurs. One is introspectively aware of a state in virtue of having a conscious thought that one is in a particular state, and one is thereby aware of oneself as being in that conscious state. Introspective awareness involves conscious thoughts that ascribe mental states to oneself.

Shoemaker writes that "[m]y use of the word 'I' as the subject of [such] statement[s 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 predicate applies (9). 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 disposed, in this thin way at least, to identify as oneself the individual one takes to feel pain or see a canary.

As noted above, Shoemaker is clear that the strong immunity to error he describes would not hold for all first-person thoughts. I might see somebody's reflection in a mirror and wrongly take myself to be that person. Error through misidentification is plainly possible in that case. Only thoughts that result from transparent access to our mental states would be immune to error in Shoemaker's strong way; such immunity hinges on our being conscious of states in that special way.

Thin immunity, by contrast, does not depend on subjective access or on any special properties such access supposedly has. Rather, thin immunity is a function solely

to my thinking of a particular essentially indexical thought; I am distinct from everybody else in that no other individual is in that token intentional state.

[39] "Self-Knowledge and 'Inner Sense'," The Royce Lectures, *Philosophy and Phenomenological Research*, LIV, 2 (June 1994): 249–314; reprinted in Shoemaker, *The First-Person Perspective and Other Essays*, Cambridge: Cambridge University Press, 1996, pp. 201–268, at p. 211.

Shoemaker's claim here echoes Wittgenstein's idea that, whereas 'I have a broken arm' involves "the use [of 'I'] as object," 'I have a toothache' involves instead "the use [of 'I'] as subject" (*The Blue and Brown Books*, p. 66). Cf. also Anscombe's view, mentioned in §IV, that there is no object of which first-person thoughts predicate concepts.

of how one's awareness of oneself as being in some state represents the individual that one is thereby conscious of. So far as thin immunity goes, the mirror case is on a par with one's being aware of oneself by having first-person access to some conscious state.

If I think that I'm the person I see in a mirror, I can be wrong about whether that image is actually of me. I could even be wrong in a certain way about who it is that I think I see; I might think I'm Napoleon and so think that I see Napoleon. But there is a thin way in which my identifying myself even there is immune to error. If I think I see myself in a mirror, I cannot be wrong about who it is I think the individual in the mirror is. I identify the individual in the mirror as the very individual whom I could also pick out as doing the identifying. In that thin way, I in effect identify the person I am visually conscious of as the individual who is visually conscious of that person, and I use the mental analogue of 'I' to refer to that one individual.

Such thin immunity is plainly trivial. But the thin immunity that holds when I think that I am in pain or that I believe something is no less so. I can be wrong about whether the individual I think is in pain is DR or Napoleon; what I can't be wrong about is whether the individual I think is in pain is the very individual that thinks that somebody is. I cannot be wrong about whether the individual I take to be in pain is the individual who is conscious of somebody as being in pain.

Similarly, suppose I think I am Napoleon because I see somebody in a mirror suitably dressed and wrongly take myself to be the person I see. Though I misidentify myself as that person, I do not misidentify who it is that I think is Napoleon; it is I, myself, that I think is Napoleon. No error is possible about whether the individual I think is Napoleon is the very individual I could also identify as having that thought.

The substantive immunity to error Shoemaker describes does not obtain. But things would be different if a mental state's being conscious were an intrinsic property of that state. It would then be intrinsic simply to one's being in a conscious state that one is disposed to regard as being in that state the individual that takes 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, one's identifying oneself as the individual that is in the state would not be the result of any process. It would result simply from one's being in the state, since the state is intrinsically conscious. Since there would be no identifying process that might go wrong, there would be no way for one to be right in thinking that somebody is in a conscious state but wrong that it is oneself that is in the state.

So, if consciousness were an intrinsic property of our conscious states, Shoemaker's strong immunity would hold. But we cannot assume without argument that a state's being conscious is an intrinsic property of that state. And that view faces serious difficulties. For one thing, it requires that we can only individuate conscious states so that being conscious of the state is an intrinsic part or aspect of the state itself. But one need not individuate conscious states in that way; one gets an entirely satisfactory method of individuation by taking the property of being conscious to be extrinsic to every conscious state. And if we can individuate conscious states both in ways that make consciousness intrinsic and in ways that do not, consciousness is intrinsic only

nominally, relative to an optional method of individuation. And that is not enough to sustain strong immunity.[40]

Other considerations also tell against the idea that consciousness is intrinsic to conscious states. A single mental state may well have more than one content, but no single state can exhibit more than one mental attitude; no single intentional state could, for example, be both a case of doubting and an assertoric thought. Since an assertoric mental attitude figures in one's being conscious of one's mental states, the consciousness of any intentional state whose mental attitude is nonassertoric must be distinct from the state itself.[41] Since the Cartesian idea that consciousness is intrinsic to conscious states is untenable, it cannot sustain the strong immunity Shoemaker describes. Only thin immunity obtains.

Thoughts can make essentially indexical self-reference whether or not they are conscious thoughts. Does thin immunity to error occur only with conscious states? Or does such immunity affect nonconscious mental states as well?

Suppose I see in a mirror somebody limping. I am not in conscious pain. But I take that person to be me and, since I acknowledge the occurrence of pains that aren't conscious, I conclude that I am in a nonconscious state of pain. I can be wrong about whether the person I see is me. But here again I cannot be wrong about whether the individual I take to be in pain is the individual that's thought to be in pain. The mirror case shows that thin immunity extends to the self-ascription not only of nonmental properties, but in the same way to the self-ascription of mental states that are not conscious.

VII. UNITY AND FREEDOM

The present approach to unity also suggests natural ways to explain various failures of unity, such as the puzzling phenomenon of Multiple Personality, now more often known as Dissociative Identity Disorder.[42] It also helps explain one other important source of intuitions about the unity of consciousness.

[40] Shoemaker urges that the functional roles definitive of mental states sustain "a conceptual, constitutive connection between the existence of certain sort of mental entities [states] and their introspective accessibility" ("Self-Knowledge and 'Inner Sense'," Lecture II, "The Broad Perceptual Model," p. 225). Though he insists that this connection is weaker than the transparency of the mind to itself, it may be enough to support a taxonomy of conscious states on which their being conscious is an intrinsic property of those states. See Lecture II *passim*, and "On Knowing One's Own Mind," *Philosophical Perspectives: Epistemology*, 2 (1988): 183–209, reprinted in Shoemaker's *The First-Person Perspective and Other Essays*, pp. 25–49.

Still, since the connection with introspective accessibility is a matter of functional role, we need not regard that connection as a necessary part of taxonomizing mental states as such. Shoemaker's taxonomy may make consciousness an intrinsic property of conscious states, but that taxonomy is optional. These issues are also discussed in "Moore's Paradox and Consciousness," ch. 9 in this volume, §IV.

[41] See "Thinking that One Thinks," §IV. For more reasons to reject the idea that consciousness is intrinsic to mental states, see "Two Concepts of Consciousness," §II, and "Varieties of Higher-Order Theory," §V.

[42] The compelling appearance of distinct selves presumably results in part from there being disjoint sets of beliefs, desires, emotions, and other intentional states specific to the apparent selves, though

People have a compelling experience of many of their actions as being free, and that experience of seeming freedom encourages the idea of a unified, conscious self as the source of such actions. The HOT model provides a natural explanation of these Kantian ideas about freedom that doesn't posit any underlying unity of the self.

Even when we experience actions as free, we typically experience them as resulting from conscious desires and intentions. We do not experience the actions as being uncaused, but rather as being due to conscious desires and intentions that seem not, themselves, to be caused.[43] Actions appear to be free when they appear to result from spontaneous, uncaused desires and intentions.

Because our mental states are not all conscious, we are seldom if ever conscious of the mental antecedents of our conscious states. And conscious desires and intentions whose mental antecedents we are not conscious of seem to us to be spontaneous and uncaused. The sense we have of free agency results from our failure to be conscious

many general desires and background beliefs will be shared. But it's also very likely due to there being distinct sets of HOTs, each operating on a distinct group of intentional states. And, because each disjoint group of HOTs operates on a distinct set of first-person thoughts, that group of HOTs will assign its targets to an apparent self characterized by the battery that derives from that set of first-person thoughts. Such an individual will accordingly be conscious of itself in dramatically different terms, depending on which alter is active.

It is worth noting that such failures of unity are failures of apparent unity of consciousness, and do not by themselves speak to the issue raised at the outset about some underlying actual unity of consciousness. We can speculate that such apparent unity may also be diminished or even absent altogether in creatures whose mental lives are less elaborate in relevant ways. I am grateful to Josef Perner (personal communication) for pressing the question about absence or failure of unity.

[43] As always, it is crucial to distinguish the mental state one is conscious of from our being conscious of it, in this case, the event of desiring or deciding from our consciousness of that event. Indeed, robust experimental findings support this distinction, by establishing that our subjective awareness of decisions to perform basic actions occurs measurably later than the events of deciding of which we are conscious. See Benjamin Libet, Curtis A. Gleason, Elwood W. Wright, and Dennis K. Pearl, "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness Potential)," *Brain*, 106, Part III (September 1983): 623–642; and Benjamin Libet, "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action," *The Behavioral and Brain Sciences*, 8, 4 (December 1985): 529–539. This work has been replicated and extended by Patrick Haggard, Chris Newman, and Edna Magno, "On the Perceived Time of Voluntary Actions," *British Journal of Psychology*, 90, Part 2 (May 1999): 291–303; Patrick Haggard, "Perceived Timing of Self-Initiated Actions," in *Cognitive Contributions to the Perception of Spatial and Temporal Events*, ed. Gisa Aschersleben, Talis Bachmann, and Jochen Müsseler, Amsterdam: Elsevier, 1999, pp. 215–231; and Patrick Haggard and Martin Eimer, "On the Relation between Brain Potentials and Awareness of Voluntary Movements," *Experimental Brain Research*, 126, 1 (1999): 128–133.

For more on the connection between this research and intuitions about free will, see my "The Timing of Conscious States," *Consciousness and Cognition*, 11, 2 (June 2002): 215–220.

Related considerations have been advanced by Daniel Wegner, who presents experimental evidence that the experience of conscious will results from our interpreting our intentions as the causes of our actions. Wegner argues that such an interpretation arises when we are conscious of the intention as prior to and consistent with the action and we are conscious of no other cause of the action. See Daniel M. Wegner, *The Illusion of Conscious Will*, Cambridge, Massachusetts: MIT Press/Bradford Books, 2002, and Daniel Wegner and Thalia Wheatley, "Apparent Mental Causation: Sources of the Experience of Will," *American Psychologist*, 54, 7 (July 1999): 480–492.

of all our mental states. It does not point to any underlying metaphysical unity of the self.

This conclusion receives support from a certain type of weakness of will. Consider what happens when one is conscious of oneself as wanting to do something or withhold from doing it, but the desire one is conscious of oneself as having is not efficacious in producing or blocking that action. Doubtless in some cases one does not actually have the desire or intention one is conscious of oneself as having, or in any case not in the decisive way one is conscious of oneself as having it. In other cases the desire or intention may be present, but still not lead to action.[44] These cases all lead to a diminished subjective sense of freedom of the will, since one comes to see that causes one is unaware of sometimes play a decisive role in determining one's behavior. We become aware that the desires and intentions we are conscious of ourselves as having diverge somewhat from the actual mental determinants of our actions. Our diminished sense of freedom in these cases reinforces the hypothesis that our full sense of freedom on other occasions results simply from our being unaware of any causal determinants of the conscious desires and intentions that seem to lead to our actions.

These considerations also help explain the compelling sense we have that the consciousness of our thoughts, desires, and intentions makes a large and significant difference to the role those states are able to play in our lives. It's often held that our ability to reason, make rational choices, and exercise our critical capacities is enhanced by the relevant intentional states' being conscious. This inviting idea doubtless underlies Ned Block's explication of what he calls access consciousness in terms of a state's being "poised to be used as a premise in reasoning, . . . [and] for [the] *rational* control of action and . . . speech."[45]

But on the face of it, this idea should strike us as perplexing. The role that thoughts and desires can play in our lives is a function of their causal relations to one another and to behavior. And presumably those causal relations are due solely, or at least in great measure, to the intentional contents and mental attitudes that characterize the states. So it will not significantly matter to those causal interactions whether the states are conscious. Accompanying HOTs will of course add some causal relations of their own, but these will be minor in comparison to those of the target states.[46] Why, then,

[44] Is this the sort of thing Aristotle calls *akrasia* (*Nicomachean Ethics*, VII, 1–10)? *Akrasia*, as he describes it, occurs when one perceives some path as good and passion leads one to follow instead some other course. But perceiving the good, on his account, itself functions desideratively; the perception that a particular kind of thing is good together with the belief that something is of that kind leads to action. So, if passions can sometimes occur nonconsciously, the kind of case envisaged here will comfortably fall under Aristotle's notion of *akrasia*. I am grateful to Eric Brown for having raised this issue.

[45] "On a Confusion about a Function of Consciousness," p. 231; emphasis Block's.

It's arguable that Block's well-known distinction between phenomenal and access consciousness is best seen not as a distinction between two types of consciousness, but between two types of mental state, each of which can occur consciously. See my "How Many Kinds of Consciousness?", *Consciousness and Cognition*, 11, 4 (December 2002): 653–665.

[46] Nor, if content and mental attitude determine the interactions intentional states have with behavior and each other, should their being conscious matter much on any other explanation of what that consciousness consists in.

should consciousness seem, subjectively, to make such a difference to our ability to reason and make rational choices?

The answer lies in the connection consciousness has to the apparent freedom of our conscious thoughts, desires, and intentions. It's plausible that a state's arising freely would make a significant difference to the role it can play in our lives. And our conscious thoughts, desires, and intentions seem to us to arise freely because of the way we are conscious of them. So it seems, in turn, that our intentional states' being conscious must itself somehow make a significant difference to the role those states can play in our lives. It is because the way we are conscious of our intentional states often makes it seem that they are free and uncaused that their being conscious seems to matter to our ability to reason and make rational choices.[47]

[47] Earlier versions of this paper were presented as the Clark-Way Harrison Lecture at Washington University in St Louis, and at the Duke University meeting of the Association for the Scientific Study of Consciousness, the University of Salzburg, and Stanford University. Some work on this paper occurred during a semester's visit in the Program in Philosophy, Neuroscience, and Psychology at Washington University in St Louis; I am grateful for their support and hospitality.

CONSCIOUSNESS AND MIND

DAVID M. ROSENTHAL

CLARENDON PRESS • OXFORD

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2005

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Rosenthal, David M.
Consciousness and mind / David M. Rosenthal.
p. cm.
Includes bibliographical references and index.
1. Consciousness. 2. Thought and thinking. I. Title.
B808.9.R675 2006 126–dc22 2005020562

Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd, King's Lynn, Norfolk

ISBN 0-19-823697-2 978-0-19-823697-9
ISBN 0-19-823696-4 (Pbk.) 978-0-19-823696-2 (Pbk.)

1 3 5 7 9 10 8 6 4 2