XV*—UNITY OF CONSCIOUSNESS AND THE SELF

by David M. Rosenthal

ABSTRACT The so-called unity of consciousness consists in the compelling sense we have that all our conscious mental states belong to a single conscious subject. Elsewhere I have argued that a mental state's being conscious is a matter of our being conscious of that state by having a higher-order thought (HOT) about it. Contrary to what is sometimes argued, this HOT model affords a natural explanation of our sense that our conscious states all belong to a single conscious subject. HOTs often group states together, so that each HOT is about a cluster of target states; single HOTs represent qualitative states as spatially unified and intentional states as unified inferentially. More important, each HOT makes one conscious of oneself in a seemingly immediate way, encouraging a sense of unity across HOTs. And the same considerations that make us assume that our first-person thoughts all refer to the same self apply also to HOTs; becoming conscious of our HOTs in introspection thus leads to a sense that our conscious states are unified in a single self. I argue that neither essential-indexical reference to oneself nor the alleged immunity to error through misidentification conflicts with this account. I close by discussing the apparent connection of unity with free agency.

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

*Meeting of the Aristotelian Society, held in Senate House, University of London, on Monday, 23rd June, 2003 at 4.15 p.m.

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 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

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, 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, 55–67; and William G. Lycan, *Consciousness and Experience*, Cambridge, Massachusetts: MIT Press/Bradford Books, 1996, Ch. 2, 13–43, and 'The Superiority of *HOP* to *HOT*,' in *Higher-Order Theories of Consciousness*, ed. Rocco W. Gennaro, John Benjamins Publishers, forthcoming; and David M. Rosenthal, 'Varieties of Higher-Order Theory', also in Gennaro.

it's rare that HOTs are accompanied by third-order thoughts, and hence rare that HOTs are, themselves, conscious.[3]

The atomistic character of this model, however, may seem to prevent it from explaining 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?[4] Why wouldn't a conscious mind seem instead to consist, in Hume's famous words, of 'a mere heap or collection of different perceptions'?[5] 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.'[6] Because such empirical consciousness cannot explain unity, Kant posits a distinct, '*transcendental* unity of self-consciousness' (B132).[7] 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

3. For more on the HOT model, see David M. Rosenthal, *Consciousness and Mind*, Oxford: Clarendon Press, forthcoming 2004.

4. I am grateful to Sydney Shoemaker for pressing this question, in 'Consciousness and Co-consciousness,' presented at the Fourth Annual Meeting of the Association for the Scientific Study of Consciousness, Brussels, July 2000, and forthcoming in *The Unity of Consciousness: Binding, Integration, and Dissociation*, Axel Cleeremans, ed., Oxford: Clarendon Press.

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

6. Immanuel Kant, *Critique of Pure Reason*, trans. and ed. Paul Guyer and Allen W. Wood, Cambridge: Cambridge University Press, 1998, B133.

7. 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).

the mere subjective impression of unity. And Kant's observation about the dispersed character 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 transparency to itself, we cannot explain the appearance simply by appeal to the reality.[8] 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 mind and consciousness 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.

8. 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.

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.[9] 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.

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. Still, the HOT model suffices here; we can explain our consciousness of such inferential connections as resulting from HOTs' representing our intentional states as being thus connected.

9. For problems about the way we are conscious of qualitative states as spatially unified within sensory fields, see David M. Rosenthal, 'Color, Mental Location, and the Visual Field,' *Consciousness and Cognition*, 9, 4 (December 2000): 85–93, Section 4.

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 else would not make it a conscious state.[10]

By itself, however, such reference to a self 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 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 the 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.

10. Might there may 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.

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.[11] 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 the utterance; 'I' does not refer to the utterance itself.[12]

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 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.[13] 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

11. *Pace* Hans Reichenbach, 'Token-Reflexive Words', *Elements of Symbolic Logic*, New York: Macmillan, 1947, Section 50.

12. 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, 481–563, 505–507. Kaplan posit a character of 'I', which is a function whose value, for each context, is the speaker or agent of that context.

13. 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, 346), 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 [1993]: 137–154, 23, and 'An Examination of Four Objections to Self-Intimating States of Consciousness', *The Journal of Mind and Behaviour* X, 1 [Winter 1989]: 63–116, 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 David M. Rosenthal, 'Higher-Order Thoughts and the Appendage Theory of Consciousness', *Philosophical Psychology*, VI, 2 (1993): 155–167.

that very thought. Each first-person thought 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 any conscious inference, HOTs also 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; indeed, no HOT would ever be conscious unless one had 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.[14]

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

14. For more on introspective consciousness, see 'Introspection and Self-Interpretation', *Philosophical Topics* 28, 2 (Winter 2000): 201–233.

we 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.[15] 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.[16] The HOT model, by contrast, 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.

15. 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.

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

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.[17] 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

How We Identify the Self. 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 way our sense of the unity of consciousness.

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,

17. See, e.g., my 'Thinking that One Thinks', in *Consciousness: Psychological and Philosophical Essays*, ed. Martin Davies and Glyn W. Humphreys, Oxford: Basil Blackwell, 1993, 197–223, and 'Consciousness and Higher-Order Thought', *Macmillan Encyclopedia of Cognitive Science*, forthcoming. For Carruthers's view, see Peter Carruthers, *Phenomenal Consciousness: A Naturalistic Theory*, Cambridge: Cambridge University Press, 2000. For difficulties in Carruthers's defence of that view, See David M. Rosenthal, 'Explaining Consciousness', in *Philosophy of Mind: Contemporary and Classical Readings*, ed. David J. Chalmers, New York: Oxford University Press, 2002, 406–421, 410–11.

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 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.[18] The word 'I' and its mental analogue refer to whatever individual says or thinks something in first-person terms, but 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 subject, but exhibits instead a wholly unmediated conception of standing.[19] 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 co-reference obtains. The same holds for talking or thinking about oneself using different tokens of 'I' or its mental analogue.

18. Perhaps as in cases of so-called Multiple Personality or Dissociative Identity Disorder.

19. 'The First Person', in *Mind and Language: Wolfson College Lectures 1974*, ed. Samuel Guttenplan, Oxford: Clarendon Press, 1975, 45–65.

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 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.[20] 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,

20. 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, 6–18 (references below are to the reprinted version); David Lewis, 'Attitudes *De Dicto* and *De Se*', *Philosophical Review* LXXXVIII, 4 (October 1979): 513–543, reprinted in Lewis, *Philosophical Papers*, vol. I, New York: Oxford University Press, 1983, 133–59; and Roderick M. Chisholm, *Person and Object: A Metaphysical Study*, La Salle, Illinois: Open Court Publishing Company, 1976, Ch. 1, Section 5, and *The First Person*, Minneapolis: University of Minnesota Press, 1981, Ch. 3, esp. 29–32.

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.[21] 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 successive perceptions in our thought or consciousness' (*Treatise*, Appendix, 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.

21. 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 Section III.

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.[22] 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. 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 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

22. John Perry, 'The Problem of the Essential Indexical,' *Noûs* XIII, 1 (March 1979): 3–21. See also P. T. Geach, 'On Beliefs about Oneself', in Geach, *Logic Matters*, Oxford: Basil Blackwell, 1972, 128–129; G. E. M. Anscombe, 'The First Person', in *Mind and Language*, ed. Samuel Guttenplan, Oxford: Oxford University Press, 1975, 45–65; Steven E. Boër and William G. Lycan, 'Who, Me?', *Philosophical Review* LXXXIX, 3 (July 1980): 427–66; Hector-Neri Castañeda, 'On the Logic of Attributions of Self-Knowledge to Others', *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*'.

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, Section iii.

target as belonging to the individual who thinks that HOT. This representing is tacit, since as we saw two sections ago, it is this 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 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 HOT ascribes the spilling of sugar to the very individual who thinks that thought. 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 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 this 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.

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 depends on our applying some battery of descriptions to ourselves 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.

Reference to somebody, as such, occurs in cases other than the first person. I can describe others as having thoughts about

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.

25. See references in n. 20.

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 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.

When, however, an essentially indexical thought about myself is conscious, the HOT I have about that conscious thought in effect 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[26] 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

26. 'If I see, reflected in a window, the image of a man whose pants appear to be on fire, my behaviour 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', 533).

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.

Many of our 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.[27] And that desire would likely lead to my acting, not because 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.[28]

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

27. Affective states, such as happiness, sadness, anger, and the like, also have intentional contents cast in such evaluative terms. See David M. Rosenthal, 'Consciousness and its Expression', *Midwest Studies in Philosophy* XXII (1998): 294–309, Section IV.

28. 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 note 10).

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).

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.[29] 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 relation of acquaintance is identity. If self-ascribing follows that model, the content towards 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.

29. 'Postscripts to "Attitudes *De Dicto* and *De Se*"', *Philosophical Papers*, Vol. I, New York: Oxford University Press, 1983, 156–159, 156.

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. The essential indexical apart, there is another worry 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 from 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 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.[30]

Perhaps the phrase 'is conscious of' is factive; perhaps one's being conscious of something implies that that thing exists. But this wouldn't prevent us from being conscious of states we aren't

30. See, e.g., David M. Rosenthal, 'Explaining Consciousness', Section 5; '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, 265–295, Section 5; 'Consciousness, Content, and Metacognitive Judgments', *Consciousness and Cognition*, 9, 2, Part 1 (June 2000): 203–214, Section 5; and 'Metacognition and Higher-Order Thoughts', *Consciousness and Cognition*, 9, 2, Part 1 (June 2000): 231–242, Section 4.

in. Even the apparent factivity of consciousness allows for our being conscious of things in ways that aren't accurate; plainly I can be conscious of actual objects as being different from the way they actually are; I can be conscious, for example, of a red object as being green. So my having a HOT that describes me as being in a state that I'm not actually in will make me conscious of an existing object, namely, myself, as being in a state that that object is not actually in.[31]

Even if I can be in error about who I am and what conscious states I am in, it is sometimes held that there is a particular way in which some of one's first-person thoughts cannot be mistaken. Perhaps I can be mistaken about whether the conscious state I am in is pain, but it might still be impossible, if I think I am in pain, to be mistaken about who it is that I think is in pain. Similarly, if I think that I believe or desire something, perhaps I cannot be mistaken about who it is that I think has that belief or desire. These first-person thoughts would, in Sydney Shoemaker's now classic phrase, be 'immune to error through misidentification', specifically with respect to reference to oneself.[32]

But how can such immunity to error obtain if we identify the individual a first-person thought refers to by appeal to some heterogeneous battery of contingent properties? It could of course turn out that any or all of the properties in such a collection do not actually belong to one. So identifying oneself in that way seems to leave open the possibility that, when I take myself to

31. 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 meet that objection by construing our reference to conscious states as reference to notional states, states that are merely intentional items. Or we could instead construe the state that's conscious in these cases as being some relevant occurrent state that we're conscious of, but in an inaccurate way.

32. Sydney Shoemaker, 'Self-Reference and Self-Awareness', 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, 142–266, and James Pryor, 'Immunity to Error through Misidentification', *Philosophical Topics* 26, 1 and 2 (Spring and Fall 1999): 271–304. Shoemaker refers to Ludwig Wittgenstein, *The Blue and Brown Books*, Oxford: Basil Blackwell, 1958, 2nd edn. 1969, 66–7.

Pryor's useful distinction between *de re* misidentification and *wh*-misidentification hinges on the epistemic grounds for holding the relevant beliefs, and those grounds do not figure here.

I have profited from discussion of these issues with Roblin Meeks.

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.

When I have a conscious pain, I cannot erroneously think that the individual that has that pain is somebody that could be distinct from me, though I can of course be wrong about just who it is that I am. How can we capture this delicate distinction? What exactly is it 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 being in pain.

Wittgenstein contrasts statements such as 'I have a broken arm' and 'I am in pain'. Whereas one could plainly be wrong about whether it is one's own arm that's actually broken, Wittgenstein urges that 'to ask "are you sure that it's *you* who have pains?" would be nonsensical' (67, emphasis original). But this seductively brief discussion obscures the pivotal difference between error about who it is that is in pain and error about whether it's I who is in pain. It's only the second kind of error that arguably cannot occur.

The HOT model provides a natural explanation of such immunity from error as actually obtains. 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[33] represents its target state as belonging to the individual that thinks that very HOT. When a pain is conscious, the individual my HOT represents that pain as belonging to is the same as the individual that thinks the HOT itself. 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, since the HOT in virtue of which the pain is conscious in effect represents it as belonging to the individual that thinks that HOT. So there is no way to go wrong about whether it's I that I think is in pain. I am conscious of an individual both as being in pain and, in effect, as the one that's conscious of being in pain, and I use the mental analogue of 'I' to refer to that individual.

33. In effect, once again, because, although thoughts containing the mental analogue of 'I' do not refer explicitly to themselves, every first-person thought disposes us to have another thought that identifies the referent of that mental analogue as the thinker of the first-person thought. So we can think of every first-person thought as tacitly or dispositionally characterizing the self it is about as the thinker of that very thought.

But such immunity is strikingly thin, since 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's just a matter of one's being conscious of the pain as belonging to the individual that's conscious of it. The immunity from error through misidentification that results consists in the impossibility of one's being conscious of a state as belonging to an individual other than the individual that's conscious of being in it.

Because the error to which we are immune is so thin, it cannot 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 very many others.[34] And this thin immunity has no bearing on error in respect of the contingent properties in such a battery.

According to Shoemaker, we are immune to misidentification because there is no 'role for awareness of oneself as an object to play in explaining my introspective knowledge' of such things as my being in pain or believing something.[35] The foregoing considerations suggest that this is not so. Such immunity as does obtain results from one's awareness of the pain as belonging to the individual that's aware of the pain, and this involves being aware of oneself. My HOT that I am in pain makes me conscious of myself not only as the individual that has that pain, but also, in effect, as the individual that has the HOT itself. And, in the introspective case, that awareness will be conscious. The examples Shoemaker gives (210–1) make it likely that the awareness he means to deny is perceptual awareness, and doubtless no perceptual self-awareness does figure in these cases. Still, the

34. One might insist that essentially indexical self-reference does even better, in that it enables one to distinguish oneself from every other individual. But that is only relative to the thinking of the particular thought by means of which the essential indexical operates; I am distinct from everybody else because nobody else is in the relevant intentional-state token.

35. '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, 201–268, 211.

nonperceptual awareness that invoked in the foregoing explanation of immunity is enough to undermine the idea that such immunity results from there being no role in introspection for our being aware of ourselves.

The striking thinness of the immunity that does obtain emerges vividly when we consider Shoemaker's claim that not all first-person thoughts are immune to error through misidentification. I might, he suggests, see somebody's reflection in a mirror and wrongly take myself to be that person.[36] I might even in such a way wrongly take somebody I believe is Napoleon to be me, for example, by finding clues about that person that seem to lead to me. Why should error through misidentification be possible here but not when I think that I am in pain or that I believe something? Why aren't the two kinds of case parallel?

It turns out that they are. If I think that I see myself in the mirror, I can be wrong about whether the mirror 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 I see Napoleon. But there is also a thin way in which my identifying myself even here is, after all, immune from 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 individual.

Such immunity to error is wholly trivial. But the immunity that holds when I think that I am in pain or that I believe something is no more substantive. 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 somebody is. I cannot be wrong about whether the individual I take to be in pain is the individual who is conscious of the 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

36. 'Self-Reference and Self-Awareness', 7.

do not misidentify who it is 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 immunity that occurs in the self-ascribing of mental states would be special only if being conscious were intrinsic to mental states. A pain's being conscious is my being conscious of myself as being in pain, which is itself a way of taking myself to be in pain. So, if pains were intrinsically conscious, it would be intrinsic to my simply being in pain that it is my pain, and so intrinsic to my being in pain that I cannot be wrong about whether the individual I take to be in pain is the very individual who is conscious of the pain.

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

Suppose I see in a mirror somebody limping. 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 thinks that person is. The mirror case shows that this thin immunity extends to self-ascription not only of nonmental properties, but in the same way also to self-ascriptions 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.[37] It also helps explain

37. 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 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 failure 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

one other important source of intuitions about the unity of consciousness.

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 and the 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.[38] 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 all the mental antecedents of our conscious

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.

38. 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 (1999), '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, 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 David M. Rosenthal, '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: The MIT Press/A Bradford Book, 2002, and Daniel Wegner and and Thalia Wheatley, 'Apparent Mental Causation: Sources of the Experience of Will', *American Psychologist* 54, 7 (July 1999): 480–492.

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 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 withholding 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.[39] 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 behaviour. 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. This lends some support to the hypothesis that our sense of freedom is itself due to our typically not being conscious of the mental antecedents of our conscious desires and intentions, even when those desires and intentions do seem to be efficacious.

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

39. Is this kind of case the sort of thing Aristotle calls *akrasia* (*E.N.*, VII, 1–10)? *Akrasia*, as he describes it, occurs when one perceives some path as good and passion leads one instead to follow 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.

state's being 'poised to be used as a premise in reasoning, ... [and] for [the] *rational* control of action and ... speech'.[40]

But on the face of it, this idea should strike us as perplexing. The role these states can play in our lives is a function of their causal relations to one another and to behaviour. 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.[41] Why, then, 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 seems that they are free and uncaused that their being conscious seems to matter to our ability to reason and make rational choices.[42]

David M. Rosenthal
City University of New York Graduate Center
Philosophy and Cognitive Science
E-mail: dro@ruccs.rutgers.edu

40. 'On a Confusion about a Function of Consciousness', 231; emphasis Block's.

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

42. 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 the 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.