Chapter 1  Awareness and identification of self

David Rosenthal

I. SELF-AWARENESS AND THE SELF

It is a crucial aspect of everyday mental functioning that we are in some way aware of ourselves. But it is far from clear at first sight just what this self-awareness consists in, and indeed just what the self is that we are aware of.

It is possible to give an answer to the second question that is mundane and unproblematic. The self one is aware of in everyday life is simply the individual that walks, talks, sleeps, and eats; the self one is aware of is, to echo Aristotle’s provisional definition of the soul, the living, functioning creature.

I will argue in the end that something in that spirit is correct and defensible. But many have held that this bland account of what the self is fails to capture what is important about the self, at least insofar as we are aware of that self. Thus Descartes argued that the self is that thing the awareness of which makes its existence indubitable, even if we doubt everything about the functioning and existence of physical reality. And the self we are aware of is also typically taken to provide a kind of mental unity that binds together all of one’s contemporaneous mental states. As Kant put it, for mental representations to be mine at all, they must “all belong to one self-consciousness” (Kant [1787] 1998, B132; cf. Shoemaker 2003, 59–71). It is something like that idea of the self that presumably underlies Ned Block’s observation that there is a kind of “me-ishness” about at least many of one’s conscious mental states (1995, 235). And the self that binds one’s mental functioning into a unity may even thereby underwrite personal identity through time.
Hume notoriously challenged this notion of a self. “When I turn my reflexion on myself,” he wrote, “I never can perceive this self without some one or more perceptions; nor can I ever perceive any thing but the perceptions.” It is, he concluded, solely “the composition of these” perceptions that “forms the self” ([1739] 2000, Appendix, 399; emphasis Hume’s). What metaphysicians call the self is “nothing but a bundle or collection of different perceptions” ([1739] 2000, i, iv, vi, 165).

It is worth noting that Hume does not just contest the existence of a self that independently binds our mental states into a unity; in denying that he perceives anything that could count as such a self, he denies that we have any reason to think that there is any such self. We lack, Hume urges, any subjective appearance that we could properly regard as the appearance of a self, at least a self that is anything distinct from “a mere heap or collection of different perceptions” ([1739] 2000, i, iv, ii, 137–38).

One may be tempted to respond to Hume’s concerns by appealing to Kant’s “transcendental unity of self-consciousness” (B132). But it is not clear how that addresses the issue at hand. Kant was well aware that “the empirical consciousness that accompanies different representations is by itself dispersed and without relation to the identity [that is, the unity] of the subject” (B133). In effect, he thereby followed Hume in holding that the way we are aware of our conscious states cannot, by itself at least, yield a subjective sense of mental unity or of the self. Even if we must, as Kant argued, posit a transcendental self to be able to think coherently about mental functioning, perhaps Hume is nonetheless correct to hold that there is no subjective appearance of any such self.

But we can undercut Hume’s concerns without invoking a transcendental self that lies beyond any possible experience. Hume’s way of assessing whether we are aware of a self negatively stacked the deck in a way that is seldom noticed. He assumed that the only way one could be aware of a self would be to perceive that self. But that assumption is unfounded. Perceiving things does make one aware of them, but perceiving is not the only way we are aware of things. We are also aware of something when we have a thought about that thing as being present to us. I am aware of an object in front of me if I see it or hear it; but if my eyes are closed and the object
makes no sound, I can also be aware of it simply by having a thought that it is there in front of me.

Not all of our thoughts, of course, result in our being aware of the things those thoughts are about. Having thoughts about objects we take to be distant in place or time, such as Saturn or Caesar, does not intuitively make one aware of those objects. So the thought must be about the object as being present to one. And the thought must presumably have an assertoric mental attitude; having doubts or wondering about an object also does not make one aware of it. And the thought must be occurring; simply being disposed to have a thought about something does not make one aware of it. But having an occurring, assertoric thought about an object as being present does make one aware of that object.

Hume would have contested that the appeal to thoughts is a real alternative to perceptual awareness, since he maintained that thinking itself consists simply of pale versions of qualitative perceptual states. “All ideas,” he insisted, “are borrow’d from preceding perceptions” (Appendix, 399). His problem about the self therefore rested on his view that all awareness is perceptual awareness.

But there is good reason to reject Hume’s view about the perceptual nature of ideas and thoughts. For one thing, mental qualities cannot be combined to form thoughts with syntactic structure. In addition, intentional states have characteristic mental attitudes, such as mental assertion, doubt, wondering, and so forth, and there is nothing in perceiving that corresponds to mental attitude or even holds any hope for explaining it. And though mental qualities do arguably represent the perceptible properties that correspond to them (e.g. Berger manuscript; Lewis 1972, 257; Rosenthal 2005, chs. 5–7, 2011), the way they represent is unlike that of intentional states. Intentional states represent truth-evaluable units, such as states of affairs; by contrast, the impressions and ideas that Hume’s view accommodates simply represent perceptible properties, which are not truth evaluable.  

Once we reject Hume’s perceptual model of what thoughts consist in, a more promising way is available to understand just how we are aware of our mental selves. We are aware of ourselves by having suitable thoughts about ourselves, and aware of ourselves as mental beings by having
thoughts that describe ourselves in mental terms. And because these thoughts are occurrent, empirical thoughts, we also avoid Kant’s appeal to the transcendental.

The contrast between Hume’s sensory approach and the alternative that relies on the thoughts we have about ourselves in mental terms reflects a crucial contrast between two views about what it is for a mental state to be conscious. The most widely held traditional view about the consciousness of mental states appeals to inner sense, and holds that a mental state is conscious if one senses or perceives that state. The higher-order thought (HOT) theory, by contrast, holds that a state’s being conscious consists in one’s having a suitable thought that one is, oneself, in that state. On the version of the view that I have defended, the thought must be assertoric and non-dispositional. And because the thought has the content that one is, oneself, in that state, the thought automatically represents the target mental state as being present (Rosenthal 2002, 2005).

The difference between the inner-sense and HOT theories sheds light on the two models of self-awareness. If one’s mental states were conscious in virtue of one’s sensing those states, there would be for each conscious state a higher-order sensation that made one aware of that state. But sensations represent at best only corresponding perceptible properties; a sensation of red represents not the red physical object, but only its redness. So a higher-order sensation would represent only the target mental state, and not also a self to which that state belongs. So nothing in one’s sensing a mental state could make one aware of such a self.

Things are different if one is, instead, aware of one’s conscious states by having thoughts about those states. One will then have a thought that one is in the state in question. And that HOT will thereby make one aware not only of that state, but also of a self that the HOT represents the state as belonging to. The HOT theory explains not only how we are aware of our conscious mental states, but also how it is that we are thereby aware of ourselves.

A sensation of a mental state would make one aware just of the state by itself, and not of the state as belonging to a self. That is what gives rise to Hume’s professed inability to find a self “[w]hen,” as he put it, “I turn my reflexion on myself.” One might wonder whether a thought about a mental
state could also make one aware of that state, as a sensation would, without making one aware of it as belonging to a self. But a mental state is in its very nature the sort of item that is a state of something. A perception or feeling or thought is always somebody’s perception, feeling, or thought. So having a thought about a mental state is always to have a thought about that state as belonging to some individual. One might have a thought about a type of mental state, for example, a thought that you have the same feeling as I do; but a thought about a token mental state always ascribes that state to somebody.

Being aware of our mental states by having HOTs about those states results in our being aware of ourselves in mental terms. And being thus aware of ourselves explains why it appears to us that there is a self that is the subject of each of our mental states. The mental appearance of a self that eluded Hume is due to the HOTs we have about our various mental states.

There is, however, a stronger appearance of self than simply the subjective impression that each conscious state belongs to some self or other. In addition, it seems subjectively to us that each of our conscious states belongs to the same self as every other. This stronger idea of a unity of consciousness does not follow simply from the weaker appearance of a self for each conscious state. Nonetheless, it is tempting to hold that all one’s conscious states do belong to a single unified consciousness. And we need to explain that appearance, which goes beyond the sense that each conscious state belongs to some self or other. Once again, Kant’s transcendental unity of consciousness cannot help here, since that is only a unity we must posit, which cannot explain the robust subjective impression we have of a unity of consciousness.

Still, one might also have doubts about whether a HOT theory of how mental states are conscious can do justice to the particular way we are aware of ourselves. There are two main concerns, both reflecting the inviting idea that the way we are aware of ourselves is special in a way that simply having a thought about something cannot capture.

One worry stems from the well-known difference between a thought about somebody that happens to be oneself and a thought that is about oneself, as such, a thought that involves the so-called essential indexical. If
I have a thought that DR is in pain but do not know that I am DR, that thought would not result in my being aware of myself, as such. And that thought would not result in my pain’s being conscious; HOTs must ascribe their target mental states to oneself, as such. But it may not seem obvious how to capture this essentially indexical aspect of HOTs. And that may lead one to suspect that having a thought about oneself, as such, requires some special awareness of the self that is antecedent to and independent of the thought.

In addition to that worry, many have followed Sydney Shoemaker in holding that we are aware of ourselves in a way that precludes any error through misidentification of the self. And even if we can explain how HOTs can be about oneself in an essentially indexical way, it is unclear why having a thought that one is, oneself, in a particular mental state could be immune to error through misidentification of oneself. There would have to be something that goes beyond mere HOTs, it seems, to explain such immunity.

There is a third source of doubt about whether a view of self-awareness that relies on HOTs could be satisfactory. The way we are aware of ourselves as being in various conscious states may well be special in the respects just sketched. But our awareness of ourselves as being in conscious states must also somehow fit with our awareness of ourselves in ordinary, non-mental respects. Each of us is an individual that is in many conscious states. But each of us is also a creature that interacts with other objects in the world; we walk, talk, eat, and do many other non-mental things as well. And we are aware of ourselves not only as being in various conscious states, but also as engaging in these various physical activities. And it is unclear how the awareness of self that HOTs confer fits with awareness of ourselves as physical creatures.

It is not obvious, moreover, how to address this last concern about a HOT view of self-awareness without making things worse for the challenges that stem from the essential indexical and immunity to error through misidentification. Some special self-awareness seems immune to error through misidentification. But it may seem difficult to square any such special awareness with the way we are aware of ourselves as physically functioning creatures.
The sense of lack of fit between two types of self-awareness is especially vivid in connection with issues about unity. The unity that is relevant to our awareness of ourselves as bodily creatures is along the lines of the unity of enduring physical objects generally. There must be coherence of bodily properties and composition, and spatio-temporal continuity; the subjective unity that figures in connection with one’s conscious states, by contrast, seems to rely on altogether different considerations.

In what follows, I argue that a model of self-awareness based on HOTs can meet all three of these challenges. In §II I show how HOTs can accommodate essentially indexical awareness of oneself without invoking any special, antecedent self-awareness. In §III, then, I argue that a crucial assumption that underlies the claim of immunity to error through misidentification is unfounded, namely, the assumption that no self-identification figures in our awareness of our own conscious states. And in §IV I discuss the particular kind of identification of self that figures in our higher-order awareness of our conscious states and how that relates to the self-identification that underlies our first-person thoughts generally.

II. ESSENTIALLY INDEXICAL SELF-AWARENESS

The way we are aware of our own conscious states requires essentially indexical self-reference. So if we are aware of those states by having HOTs about them, HOTs themselves must somehow incorporate the essential indexical. Is such self-reference, in thought and speech, basic and unanalyzable? Or can it be discharged by appeal to more ordinary kinds of reference?

Consider John Perry’s well-known example, in which I see a trail of sugar apparently spilling from somebody’s grocery cart and, not realizing that it is spilling from my cart, think that the person spilling sugar, whoever it is, is making a mess. Though I am that person, having that thought does not imply that I think that I myself am making a mess. Reference to oneself, as such, uses what Perry dubs the essential indexical, called by traditional grammarians the indirect reflexive because it plays a role in indirect discourse played in direct quotation by the first-person pronoun.
For a mental state to be conscious, it will not do simply for one to be aware that somebody that happens to be oneself is in that state; one must be aware of oneself, as such, as being in that state. As noted earlier, if I am in pain but aware just that DR is in pain, that would result in my pain’s being conscious only if I were also aware that I am DR. If I thought instead that you were DR, my awareness that DR is in pain would not result in any pain’s being conscious; it would not make me aware in any relevant way of myself as being in pain.

Essentially indexical self-reference is one way in which our awareness of ourselves and our conscious states is special. So if we are aware of ourselves as being in particular mental states by having HOTs that we are, those HOTs must refer to oneself in this essentially indexical way. HOTs cannot represent the states they are about as belonging to oneself described in some inessential way; they must represent those states as belonging to oneself, as such.

It is sometimes argued that this requirement undermines any HOT account of what it is for a mental state to be conscious (e.g. Zahavi and Parnas 1998, §III). So it is crucial to see just what such essentially indexical self-reference consists in, and whether HOTs can refer to oneself in that special way. An essentially indexical thought or speech act about myself has the content that I myself am $F$. But what is it that results in its referring to oneself in an essentially indexical way?

Essentially indexical self-reference occurs not just with HOTs, but also with all our first-person thoughts. It does not matter whether the property the first-person thought ascribes to oneself is that of being in a mental state or something else. So let us consider the general case. 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.

It is this tacit identification of oneself as the thinker of the first-person thought that underlies one’s reference to oneself, as such. I refer to myself, as such, when I in effect refer to something as the individual that does the referring. That is why essentially indexical self-reference and self-awareness forge a referential connection that seems intuitively to be independent of any intentional content.
In Perry’s case, I begin by thinking that somebody is spilling sugar and I come to realize that I, myself, am that very person. What I discover, in effect, is that the individual who is spilling sugar is the very same as the individual who thinks that somebody is spilling sugar; the person that is being said or thought to spill is the very person who says or thinks that somebody is spilling. By tacitly identifying the individual that a thought purports to be about with the individual who thinks that thought, the essential indexical links what the thought purports to be about to that very act of thinking.

HOTs make essentially indexical self-reference in just this way. Suppose I have a pain, and that the pain is conscious; I am aware of myself as having that pain. On the HOT theory, I have a HOT to the effect that I have that pain. My HOT succeeds in referring to me in an essentially indexical way by in effect ascribing the pain to the very individual who has that HOT.

HOTs are just a special case of first-person thoughts. But there is a complication that affects HOTs alone. We are seldom aware of our HOTs; HOTs typically are not conscious thoughts. Indeed, that is just what the HOT theory predicts; a HOT would be conscious only if one were aware of having that HOT, and that would happen only when one has a third-order thought about it. And we evidently seldom have such third-order thoughts.

This is where the complication with HOTs comes in. Consider the case of conscious pain. If a HOT had the explicit content that the individual who thinks this thought is in pain, that would make one aware not just of the pain, but of the HOT itself. And since HOTs are seldom conscious, they cannot have explicit content of that sort. So it cannot be that the HOT explicitly identifies the individual that is in pain as the thinker of that HOT; rather the HOT identifies the individual in that way tacitly, only in effect. But what exactly does that amount to?

A HOT must represent tacitly that a pain belongs to the individual who has that HOT, but the HOT cannot have the explicit content that the pain belongs to the individual that has that HOT. Instead, the HOT’s content describes the pain as belonging to some particular individual, and the individual that has that HOT is, in addition, disposed to identify the individual the HOT refers to as the individual that has that HOT. The
content of each HOT refers to the individual it asserts to be in pain by the mental analogue of ‘I.’ Though the HOT does not describe that individual as the thinker of the HOT, the individual is disposed to do so should the question ever arise. Needless to say, the question seldom if ever does arise; so the individual that has the HOT may never actually perform that identification. But the disposition to do so constitutes a tacit identification of the self that the HOT ascribes pain to. And that constitutes the essentially indexical self-reference.¹¹

Indeed, that is exactly how things work in Perry’s sugar case. When I discover that I am the one who is spilling sugar, the thought that embodies that discovery refers to myself in thought in an essentially indexical way. But I do not on that account have a thought with the explicit content that the person spilling sugar is the person that has this very thought. Rather, I come to be disposed to identify the individual that is spilling sugar in that way, that is, as the individual who has the thought I can express by saying that I am spilling sugar. In ordinary circumstances, I would never have a thought that explicitly identifies the individual who is spilling sugar as the individual who has that thought; but my disposition to perform that identification suffices for my essentially indexical self-reference. So it is with the essentially indexical reference that occurs in HOTs.

This disposition to identify an individual as the thinker of a thought explains why essentially indexical reference seems to operate independently of any particular way of describing or characterizing oneself. It is not that no description of oneself is relevant to secure such reference; the relevant description is oneself as the thinker of a particular thought. But because essentially indexical self-reference relies on a disposition to identify the relevant individual in that way, and not on any explicit identifying content, such self-reference seems to be independent of self-description altogether. And this helps explain the intuitive elusiveness both of essentially indexical reference and what such reference picks out.

A thought about oneself as such refers, by way of a disposition to identify oneself, to the individual that thinks that thought, though the content of such a thought does not explicitly describe one in that way. Since essentially indexical thoughts refer independently of any particular description that occurs in their content, it is tempting to see them as
referring in an unmediated way, a way that might thereby provide the foundation for all other referring. If so, perhaps we need some special awareness of the self, antecedent to any thoughts we might have about it, to underlie any essentially indexical self-reference.

But such reference is mediated by a disposition to identify oneself as the thinker of a particular thought; it does not rest on or constitute some independent access to the self, but simply on a disposition to have another thought. So essentially indexical self-reference cannot provide any foundation for the identifying of anything else.

And it is independently implausible that essentially indexical self-reference is required for identifying anything other than oneself. We rarely identify other objects by reference to ourselves; we almost always identify other things by some local frame of reference that we fix in turn by appeal to various objects we perceive and know about, independently of ourselves. Such local frames of reference occasionally fail, but when they do the solution is seldom to refer back to ourselves. Essentially indexical self-reference cannot underwrite foundationalist epistemological leanings.

There is a sense we sometimes have of ourselves that can make it seem difficult to see ourselves, insofar as we are conscious selves, as located among the physical furniture of the universe. And it might be thought that essentially indexical self-reference is responsible for this appearance of mystery about the subjects of conscious experience, since the essential indexical occurs ineliminably only in describing such subjects. (See, e.g., Anscombe 1975 and Nagel 1965, §V.) The present account suggests an explanation. It may be that the self seems to elude our objective framework because essentially indexical self-reference is secured not by explicit descriptive content, but only by a disposition to identify the individual a thought is about with the thinker of that thought.

Reference to an individual, 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. 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 that you are disposed to identify 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 who thinks that thought; in that case, my thought makes essentially indexical reference to myself. If the thought fails to be conscious, my realization will fail to be as well.

Does essentially indexical self-reference make a difference to the way beliefs and desires issue in action? David Kaplan’s catchy example of my essentially indexical thought that my pants are on fire 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 did not 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 a relevant occurrent desire I have to do what I should. Since I likely would not desire to do what DR should do if I did not think I was DR, I would in that case 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, all without any essentially indexical self-reference. The belief must refer to oneself in an essentially indexical way only if the relevant desire does as well.

The situation is the same 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. Similarly with my belief that DR’s pants are on fire if I do not know that I am DR.

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. And the content of my desire might refer to me; it could be a desire that I have a beer. Things might then be different if instead I simply had a desire just that DR have a beer.
But the relevant desire need not refer to me at all; its content could be simply that having a beer would be nice. And that desire would likely lead to my going to get a beer not because the content of the desire refers to me, but because I am the individual that holds the desiderative attitude toward that content. Essentially indexical self-reference is not needed for beliefs and desires to issue in action.

It is sometimes objected to the HOT theory that non-linguistic beings, including human infants, could not have HOTs. But this is far from obvious. Many non-linguistic beings likely do have some thoughts, and the conceptual resources that HOTs use to describe their mental states could well, in the case of infants and non-human animals, be fairly minimal. Such beings would not, for example, have to conceptualize their mental states as being mental to be aware of those states in a way that results in their being conscious states. Perhaps language is required for HOTs about purely intentional states, such as thoughts. (See Rosenthal 2005, ch. 10, §v.) But it is also far from obvious that our pretheoretic view about non-linguistic creatures requires ascribing to them conscious intentional states. Non-linguistic creatures and human infants might well be in intentional states, but it may also be that only their qualitative mental states are conscious.\(^\text{14}\)

Essentially indexical self-reference involves a disposition to identify oneself as the individual that thinks a particular thought. So if one cannot be aware of oneself as being in intentional states unless one has language, essentially indexical self-reference will be available only for individuals that have language. And if HOTs require the essential indexical, creatures without language would not, on the HOT theory, be able to be in mental states that are conscious.

But weaker requirements arguably figure for the HOTs of non-linguistic creatures. The requirement that HOTs must refer to oneself in an essentially indexical way is to exclude irrelevant, inessential ways of referring to oneself. If I have a thought that DR is in pain without also thinking that I am DR, that will not result in my pain’s being conscious. I will not in that case be aware of myself, as such, as being in pain.

But human infants and non-linguistic animals have no irrelevant, inessential ways of referring to themselves in thought. They do distinguish
themselves from everything else, and can thereby refer to themselves in thought. But their HOTs do not require the essential indexical, since distinguishing themselves from everything else provides the only way they have to refer to themselves.

HOTs must make essentially indexical reference to the self to which they ascribe mental states. So if essentially indexical self-reference required special access to the self, a form of self-awareness would be needed that is more immediate and direct than the awareness of self that HOTs can deliver. But since we can explain essentially indexical self-reference by appeal to a disposition to identify the self that each first-person thought refers to with the individual that thinks that thought, we need not appeal to any such special, unmediated self-awareness. The reference to a self that occurs in each HOT, together with a disposition to identify that self with the thinker of that HOT, will suffice.

III. SELF-AWARENESS AND IMMUNITY TO ERROR

It has often been held that our awareness of our conscious states is both infallible and exhaustive, that no conscious state fails to have any feature that we are aware of it as having, and it has no mental feature of which we are unaware.\(^{15}\)

Few today explicitly assert that our access to our own mental states is privileged in such a strong way. Still, there is a tension that affects some thinking about this issue. Some authors are tempted to hold that the reality of conscious mental states automatically coincides with the way they appear to us in consciousness, and even that it is groundless to distinguish appearance from reality in the case of conscious states.\(^{16}\)

But the appearance of conscious states simply is the way they are from a first-person point of view, what it is like for one to be in those states. So it is not clear how the appearance of conscious states could coincide with their reality unless our access to our conscious states is, after all, exhaustive and infallible. And if the way conscious states are from a first-person point of view invariably coincides with the reality of those states, the way they appear to us will be both accurate and exhaustive.
But it is plain that it is not. There is doubtless much about the mental natures of our conscious states that we are unaware of, and much that we are, occasionally at least, wrong about. Often, for example, we are aware of a color experience in a relatively generic way, perhaps just as a perception of some red or other, though we have good reason to think that our visual sensations register exact shades.

Nor is our access to our conscious states infallible. There is robust experimental evidence that we are sometimes wrong about what thoughts and desires have led to our choices and actions. People often confabulate having thoughts or desires that explain their choices in situations in which we have independent evidence that the reported states could not have been operative. And expectations can affect subjective qualitative experience, even as regards qualitative experiences such as pain (e.g. Koyama et al. 2005). In this kind of case, we become aware of ourselves as being in states that differ in their mental properties from the states we actually are in.

If awareness of our conscious states were infallible or exhaustive, that would be surprising and would call for some special explanation. And it is unclear what explanation would be credible. If a mental state’s being conscious is due to a distinct higher-order awareness of that state, whether a HOT or any other kind of awareness, perhaps that higher-order awareness is typically accurate. But there is no guarantee that it will invariably be so.

Some theorists have urged that this higher-order awareness is intrinsic to the states it represents one as being in. And it might seem tempting to hold that a higher-order awareness of a state that is intrinsic to that state would have to represent it correctly. If a higher-order awareness were intrinsic to the state, perhaps nothing could mediate between the state and awareness of it, and perhaps nothing could then interfere with the accuracy of the awareness.

But misrepresentation remains an open possibility whether the awareness is intrinsic to or distinct from the state it is about. If they are distinct, it is clear that the awareness can misrepresent, since extraneous factors could interfere with the awareness. But even if it is intrinsic, the awareness would be only one of the state’s mental properties, and the awareness might still misrepresent the state’s other mental properties. Since the awareness, even
if intrinsic to the state, is distinct from the mental properties it represents the state as having, there can still be no guarantee that the awareness represents those other properties accurately or fully. Whatever our account of consciousness, awareness of one’s own mental states is neither infallible nor exhaustive.

But accurately and fully representing a state’s mental properties is not the only type of privilege that is been proposed. Even if I may be wrong about the intentional content and qualitative character of my mental states, there is something else, Shoemaker has argued, that I cannot be in error about in connection with my mental states. “[I]t cannot happen,” he urges, “that I am mistaken in saying ‘I feel pain’ because, although I do know of someone that feels pain, I am mistaken in thinking that person to be myself.”

Perhaps I can be wrong in such a case about whether it is pain that I feel. But I cannot be wrong in thinking that I consciously feel pain because, and only because, though I am right that somebody consciously does feel pain, I misidentify that person as being me. Such first-person thoughts would, in Shoemaker’s classic phrase, be “immune to error through misidentification,” specifically with respect to reference to oneself.

Shoemaker appeals to a well-known passage in *The Blue Book* in which Wittgenstein distinguishes two ways we use the first-person pronoun. On one use, which Wittgenstein labels “the use as object” ([1958] 1969, 66), I might believe that my arm is broken, and I could in that case be mistaken about whether the arm that is broken is actually mine; it might be the arm of another person. But Wittgenstein maintains when I say that I have a pain, this is a distinct use, on which “there is no question of recognizing [which] person” it is that is in pain. “To ask ‘are you sure it is you who have pains?’ would be nonsensical” ([1958] 1969, 67; emphasis Wittgenstein’s).

Since no recognition figures in using ‘I’ in the second way according to Wittgenstein, he evidently holds also that when ‘I’ is used in that way, its use does not involve one’s identifying oneself, as such or in any other way. And since no identification of oneself occurs in connection with one’s using ‘I’ in that way, misidentification is not a possibility. There can be no failure of identification if there is no identification to begin with.

Shoemaker evidently sees things this way as well. Sometimes we refer to ourselves in a first-person way that does require identification of oneself. I
may see somebody reflected in a mirror and identify the person thus reflected as myself. But sometimes the first-person reference one makes to oneself is, Shoemaker insists, “reference without identification.” In such cases, he writes, “[m]y use of the word ‘I’ as the subject of my statement is not due to my having identified as myself something of which I know, or believe, or wish to say, that the predicate of my statement applies to it” (Shoemaker 1968, 9). And since “identification necessarily goes together with the possibility of misidentification” (13), when no identifying figures in first-person reference to oneself, no misidentification is possible either.

One might suppose that introspection enables us to identify ourselves in a way that precludes possible misidentification. It seems difficult to see how one could misidentify oneself in introspection. But Shoemaker argues that no identification occurs even in connection with introspective self-awareness. Suppose I am aware of a pain I have. For self-identification to figure in introspective awareness, Shoemaker urges, such awareness would have to be parallel to perception. And “[a]n essential part of the explanation of my perceptual awareness that John has a beard is the fact that the observed properties of the man I perceive, together with other things I know, are sufficient to identify him for me as John” (13).

But that model cannot, Shoemaker maintains, apply to my awareness of my pain. If there were a property that served to identify myself as myself, I would have not only to be aware of it, but also to be aware that it is I who is aware of it. “[I]n order to identify a self as myself by its possession of this property, I would have to know that I observe it by inner sense, and this self-knowledge, being the ground of my identification of the self as myself, could not itself be grounded on that identification” (14). No properties that one is aware of could, Shoemaker concludes, identify the possessor of those properties with the self that is aware of them.

I urged earlier that Hume’s denial of awareness of a self resulted from his adoption of an exclusively perceptual model of awareness. Shoemaker construes Hume as rather having had in mind this concern about identification. He sees Hume’s insistence that we are never aware of a self as amounting only to a denial that when one is aware of one’s mental states, the resulting awareness of oneself rests in any way on some identification of oneself. It is unlikely that Hume meant no more than that, given his
repeated explicit denial of awareness of self, though that denial does carry with it a denial that our awareness of our conscious states involves any identifying of a self.

Apart from how we understand Hume, the issue about identification is independent of whether our higher-order awareness of our conscious states is perceptual or due to HOTs. On a HOT theory, one is aware of each conscious state by having a HOT with the content that one is in that state. So Shoemaker’s insistence that no identification underlies our awareness of ourselves raises a challenge that a HOT theory must address. The theory must explain what ensures that each HOT refers to the right self, that is, to the same self as the self that has that HOT. And it may seem that any such explanation would have to appeal to some prior awareness of self that does not itself rest on any kind of identification of the self. If so, HOTs cannot explain consciousness without appealing to the very awareness without identification that Shoemaker and Wittgenstein describe.

Our awareness of ourselves seems to differ in important ways from our awareness of everything other than ourselves. And it may be tempting to explain that intuitive difference by appeal to the idea that self-awareness and self-reference depend on no identification of self. But explaining the difference in that way does not square with the way we actually operate when we assert that we are in some mental state or we are aware of ourselves as being in that state.

When I assert that you are in pain, I must identify the self I ascribe pain to in some way independent of your simply being in pain. But there is no similar need, according to Shoemaker, to identify the individual I take to be in pain when I assert that I am in pain or am aware of myself as being in pain. But it is far from clear that this is so. For one thing, when I am aware of myself as being in pain, I identify the individual I take to be in pain as the individual that is aware, in a distinctively first-person way, of being in pain. Such identification is very thin, and does not take one far. But if I am aware of myself as being in pain, it does distinguish the individual I take to be in pain from you, as well as any other person.

Such thin identification is seldom explicit. As noted in §II, if my higher-order awareness of a pain had the explicit content that the individual I am aware of as being in pain is identical with the individual that is thus aware,
that higher-order awareness would itself be conscious. Since such higher-order awareness is seldom itself conscious, it cannot have that content. So the tie between the self one is aware of being in pain and the self that is thus aware must be due to a disposition to identify them, a disposition that is sometimes exhibited in an occurrent thought, though not all that often.

But this disposition is nonetheless sufficient to provide a way of identifying the individual that one is aware of as being in pain. When I am aware of being in pain or assert that I am in pain, I am disposed, if the question arises, to identify the individual I say is in pain as the individual that asserts or is thus aware. Needless to say, the question hardly ever arises, perhaps occurring only in theoretical discussions about self-identity and the like. But that hardly shows that awareness of oneself as being in a conscious state is not accompanied by a disposition to reply in that thin way, and perhaps with some impatience, to a question about who it is one is aware of as being in pain.

This thin self-identification operates along the same lines as essentially indexical self-reference. When I realize in an essentially indexical way that I, myself, am the one who is spilling sugar, I am disposed to identify the person that I think is spilling sugar with the person that thinks that thought. Similarly, when I am aware of myself as being in pain, I am disposed to identify the individual I am aware of as the very individual that is doing the identifying. That disposition to identify, thin as it is, distinguishes that individual I am aware of as being in pain from other candidates.

Because such identifying is so thin and because it is almost always a mere disposition to identify, it is natural to overlook it altogether and assume that no identification underlies one’s awareness of oneself as being in a conscious state. But a total lack of any underlying identifying runs the risk of reducing the self in question to nothing. That indeed is the picture of self that figures in Wittgenstein’s *Tractatus Logico-Philosophicus*: “The self does not belong to the world: rather, it is a limit of the world” (Wittgenstein 1974, 5.632: 69). And it is natural to assume that this picture remains operative in *The Blue Book* in connection with Wittgenstein’s insistence that when I say that I have a pain, “there is no question of recognizing [which] person” it is that is in pain.22
Is a mere disposition to identify enough to undermine Shoemaker’s claim that there is no need to identify, and Wittgenstein’s that there is no question of recognition? Compare again Perry’s sugar case. When I realize that I am the person spilling sugar, I have no thought to the effect that the person spilling is the person who thought somebody was spilling; it is just that I am now disposed to have that thought.

And compare a case of my identifying you as the person who is in pain; perhaps somebody cried out in pain and after looking around I see that it was you. I might have a thought with the explicit content that you, grasping your foot, are the person who cried out, but more likely I just have a thought that you are in pain, with an underlying disposition explicitly to identify you with the person that cried out. Most identifying, whether first or third person, takes place by way of dispositions to think identifying thoughts.

There is another reason to hold that awareness of oneself as being in a conscious state is sustained by a disposition to identify in some way the individual one is aware of as being in that state. When I am aware of, or refer to, myself as being in pain, I am aware of or refer to something that you can be aware of and refer to as well. The self I am aware of as being in pain is a self that you too can be aware of as being in pain. When you are aware of me as being in pain, it will not be the seemingly unmediated awareness characteristic of one’s awareness of one’s own pain, but we are nonetheless aware of the very same thing as being in pain. And when I say of myself that I am in pain, I refer to the very same self that you speak of when you say of me that I am in pain. Otherwise my statement that I am in pain would not say the same thing as your statement, ‘You’re in pain,’ addressed to me, nor would that statement of yours contradict my statement that I am not in pain.

Immunity to error through misidentification occurs if I cannot be wrong in thinking that I consciously feel pain because, and only because, though I am right that somebody does consciously feel pain, I misidentify that person as being me. And Shoemaker argues no such misidentification can occur in such a case because no identification figures in the first place.
But a disposition to identify oneself does underlie one’s awareness as being in pain or some other conscious state. One is disposed in any such case to identify the self that one is aware of as being in the relevant conscious state as the self that is thus aware. Since that disposition to identify does underlie one’s awareness of being in conscious states, we would need some other reason if we are to rule out misidentification.

One reason might be that it is conspicuously difficult to come up with a case in which misidentification seems to occur. It is not easy to describe a credible case in which, though I am right that somebody consciously feels pain, my sense that I am the one that is in pain is based solely on my having misidentified that person as being me. Assume that when I have a conscious pain, I identify the individual I am aware of as being in pain with the individual that is thus aware. But when it is somebody else that actually has a pain, what would lead me to misidentify that individual as myself? What circumstances can we describe in which such misidentification might occur? One way to explain why it seems hard to specify such circumstances is to hold that misidentification in that kind of case simply is not possible.

But there is another explanation for why such cases are hard to come by, which rests on the compelling subjective sense we have that when we are in a conscious state, nothing mediates between the state one is aware of and one’s awareness of that state.

Consider a conscious pain. There is a compelling subjective sense when a pain is conscious that nothing mediates between that pain and one’s awareness of it; they seem subjectively to be of a piece. And since one is disposed to identify the individual that is in pain with the individual that is aware of pain, that awareness is tied to oneself. So it seems subjectively as though nothing comes between the self that is aware of the pain and the pain itself; the subjective sense of immediacy appears to forge an unbreakable tie between self and conscious state.

It is this appearance of an unbreakable tie between pain and oneself that makes it inviting to hold that misidentification of that self cannot occur. If one ascribes a pain to oneself, it cannot be because one is right to ascribe pain to somebody, but mistaken about whether it is oneself.
But we have no reason to think that the lack of mediation in question is more than subjective appearance. For one thing, actual immediacy is not needed to explain our subjective sense of immediacy. When we are aware of being in a conscious state, we are unaware of any mediation; we are unaware of any observation that led to our being aware of the states, and unaware of any inference that led to that awareness. And that yields our subjective sense of immediacy; since we are not aware of any mediation, it is tempting to conclude that none occurs.

But there is much in our mental functioning that we are unaware of. That is especially so for mental processes, which lead from one mental state to another; we are seldom if ever aware of such intervening processes. So our not being aware of mediation is little reason to suppose that there isn’t any. Only if we assumed that the mind is transparent to itself in respect of such mental processes could we explain apparent immediacy by appeal to actual immediacy.

It is tempting to hold that our awareness of our conscious states is always accurate about those states. And perhaps that would provide some reason to conclude that nothing mediates between those states and our awareness of them. But the subjective sense that our awareness of our conscious states is always accurate is not reliable. We know that our awareness of what mental states we are in is sometimes mistaken. The appearance that we are never mistaken stems from our never having any first-person resources by means of which we might check and correct our first-person access to our mental states. The subjective sense that consciousness is the last word on our mental life does not support the view that nothing mediates between our conscious states and our awareness of them.

Nonetheless, the subjective appearance that no mediation occurs does make it difficult to construct a credible case in which one would ascribe a conscious state to oneself solely because one misidentified somebody else as oneself. Any case we describe as involving such misidentification is hard to see as also involving a subjective appearance that nothing mediates between the conscious state and oneself.

But since the subjective appearance that nothing mediates is unfounded, we should be wary of trusting the conclusion that no such misidentification
can occur. Still, if such misidentification is possible, we should be able to describe a situation in which it is credible that it does occur. And it turns out that there is such a case. In multiple personality disorder, now better known as dissociative identity disorder (DID), there are several apparent personalities, so-called alters, that inhabit a single body, or host. Each alter is dominant at different times from the others. These alters have different memories and different conscious access to mental states, though these states typically overlap somewhat and one alter often has access to the mental functioning of others, but not conversely.

These alters function in various ways as distinct selves. As with ordinary selves, the experiences, memories, thoughts, and desires of each alter are well integrated and connect in characteristic ways with the social ties and behavior patterns distinctive of that alter. And the awareness each alter has of those states will also be well integrated into its own mental life and behavior. But one alter will sometimes have subjectively unmediated awareness of the mental states of another; one alter will disavow being in particular mental states, ascribing them instead to another alter. In such cases, one alter has HOTs, reasonably well integrated into its own mental life, that describe the other alter as being in various mental states. (For more, see Ross 1997 and McAllister 2000.)

Because some alters have access not only to their own memories and experiences, but to those of other alters as well, there is room for confusion about which state belongs to which alter. So when one alter but not another actually has a pain, it may happen that the other alter is sometimes wrong in being aware of itself as being in pain because, and only because, the alter that is aware of itself as being in pain misidentifies the other alter as itself.

What is it for conscious states that occur in a host body to belong to one alter rather than another? In ordinary cases, we rely on identifying the organism in question, but that is not available here. Nonetheless, we do have a robust commonsense test. A state belongs to one alter rather than another if it is suitably integrated with a range of other states characteristic of a particular alter, and not other alters. The range of mental states that anchor an alter’s identity would include awareness of various mental states, where both the states and the awareness of them fit together in the way characteristic of a self. So for an alter to be aware of some particular state
as belonging to itself, and not another alter, would require awareness of the state as belonging to an individual identified by appeal to the relevant range of mental states. Some means of identification must underlie each alter’s awareness of a mental state as belonging to itself.

Despite the strangeness of DID, we have no reason to doubt that the awareness of their own conscious states that alters enjoy is psychologically on a par with the awareness that occurs in ourselves. Whatever explanation we adopt for ourselves will have, with suitable adjustment, to work for them as well. And because their first-person awareness of conscious states sometimes requires identification of a relevant self, it is open that this holds also for normal cases. And where identification can occur, there is a possibility of misidentification.

DID is a controversial subject, on which theorists and clinicians do not agree. But it is natural to see the phenomenon clinicians describe as involving multiple functioning selves. And that, together with the theoretical considerations that support the occurrence of self-identification in connection with our awareness of our conscious states, gives ample reason to doubt that we are immune to error through misidentification.

IV. SELF-IDENTIFICATION AND SELF-AWARENESS

If no disposition to identify oneself occurred in connection with our awareness of ourselves as being in conscious states, such self-ascription would be immune to error through misidentification. But a disposition to identify does arguably underlie one’s awareness of one’s own conscious states and reference to oneself as being in those states. One’s awareness of one’s conscious states consists in having a HOT that one is in that state, and the reference to oneself that such HOTs make is secured by a disposition to identify the relevant self as the individual that is thinking that very HOT.

Actual identifying is not required to undermine immunity to error through misidentification. Substantive misidentification does not require an actual identification; a disposition to identify is sufficient for misidentification to be possible. However one is disposed to identify the self that one is aware of as being in the conscious state, one could instead be disposed to identify that self in some other way instead.
But even though reliance on a disposition to identify does not raise a
difficulty, perhaps the particular disposition appealed to in the foregoing
proposal does. For the relevant misidentification to be possible, a
disposition is needed to identify the self one is aware of as being in a
particular conscious state. But is a disposition to identify that self merely as
the self that is thus aware sufficient? A champion of immunity might well
argue that such a thin type of identification cannot sustain the possibility of
substantive error through misidentification. If the disposition to identify
amounts to nothing more than that, perhaps substantive immunity is
sustained after all.

One reason to doubt that such thin identification can undermine immunity
is that it seems by itself not to anchor the self in any satisfactory way. When
one is in a conscious pain, one is disposed to identify the individual that one
is aware of as being in pain with the individual that is aware of the pain. But
who is that individual? Such thin identification does not by itself seem to go
beyond the uninformative Tractarian picture of the self as merely “a limit of
the world.”

Such thin identification faces another difficulty as well. There is a
compelling subjective sense that all our conscious states belong to a single
unified self, rather than there merely being one self for each conscious state.
This is reflected in Hume’s bundle metaphor itself, on which the self is
“nothing but a bundle or collection of different perceptions”; Hume casts
his very denial of awareness of a self in terms of a metaphor of unity. And it
is what Kant insisted on in positing a “transcendental unity of self-
consciousness” (B132). But the thin identifying of the self each conscious
state belongs to with the self that is aware of that particular state cannot
account for such unity and such bundling.

But that thin identification is not all we have to go on. We identify
ourselves in ways that rely on a large and heterogeneous collection of
factors, ranging from highly individual considerations to others that are
fleeting and mundane. We appeal to location in time and place, current
situation, bodily features, the current and past contents of our conscious
mental lives, and various psychological characteristics and propensities,
indeed, to all the properties we take ourselves to have. The factors that
figure in our identifying ourselves are mostly theoretically uninteresting and
have little systematic connection. There is no magic bullet by which we identify ourselves, whether in connection with conscious states or in any other respect. There is only a vast and heterogeneous 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.

Each of these factors reflects some belief one has about oneself, ranging from what one’s name is, where one lives, and what one’s physical dimensions and location are to what the current contents are of one’s consciousness. We can be in error about any of these beliefs about ourselves; indeed, we could be in error about most of them. One could be wrong about all one’s personal history, background, and current circumstances. One might even be mistaken about one’s location relative to other objects if, for example, one lacked relevant sensory input or the input one had was suitably distorted.

One can be wrong even about what conscious states one is currently in. One may take oneself, in a distinctively first-person way, to have beliefs and preferences that one does not actually have and to lack those one does. And one may be wrong about the sensations or emotions one is aware of having.

But how does this multitude of ways in which we identify ourselves help? How do these many ways of identifying oneself give substance to the disposition to thinly identify the self one is aware of as being in conscious states with the self that is thus aware?

Identifying oneself always 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 first-person remarks that express those thoughts. One picks out the individual those first-person thoughts are about by reference to a diverse collection of contingent properties, such as those mentioned above. The reference any new first-person thought makes to oneself is secured by appeal to the descriptive content of other, prior first-person thoughts, which gradually enlarges the stock of self-identifying thoughts available to secure such reference.
This applies to HOTs, which are themselves cast in the first person, no less than to other, more mundane first-person thoughts. Each HOT refers to a self as being in a particular mental state, and we are in every case disposed to identify that self with the individual that thinks the HOT. But we are also disposed to identify that self with the individual that all our other first-person thoughts refer to. So the disposition to thinly identify the self one is aware of as being in a particular state with the individual that is thus aware is supplemented by the disposition to identify that individual with the referent of a multitude of other first-person thoughts.

First-person thoughts are cast in terms of the mental analogue of the first-person pronoun, ‘I.’ That includes HOTs, whose content is that I am in a particular state. And ‘I’ and its mental analogue function somewhat as do proper names. We routinely take distinct tokens of a proper name all to refer to the same individual unless something overrides that default assumption. Similarly, we take all tokens of the mental analogue of ‘I’ in our first-person thoughts also to refer to the same individual. It is by no means easy, moreover, to override this default assumption for the mental analogue of ‘I,’ though as suggested in §III, this seems to happen in cases of DID. ‘I’ and its mental analogue refer to whatever individual says or thinks something in first-person terms, and we also take them to refer to one and the same individual across various first-person thoughts or speech acts.

G. E. M. Anscombe (1975) has urged that ‘I’ does not function as proper names do, maintaining that the first-person thought that I am standing, does not predicate the concept standing of a subject, but instead simply exhibits a wholly unmediated conception of standing. But if that were so, we would be unable to explain the incompatibility of my thought that I am standing with your thought that I am not. Similarly for awareness of one’s conscious states; my HOT that I am in pain is incompatible with your thought that I am not.

The reference HOTs make to a self rests on a disposition to identify the self each HOT refers to as being in a particular mental state as being the self that thinks that HOT. But your thought that I am not in pain presumably makes no appeal, nor are you in any way disposed to appeal to the self that is aware of being in pain. So if my disposition to identify the self my HOT refers to as being in pain did not connect with more substantive ways of
identifying myself, there would be nothing that could ground the incompatibility of my HOT that I am in pain with your third-person thought that I am not.

But because HOTs are first-person thoughts, my disposition to identify the self that is in pain as the thinker of a HOT relies for additional identifying characteristics on the battery of other first-person thoughts I have. We are disposed to take our first-person thoughts and speech acts as referring to the same individual from one first-person thought to another. Since my HOTs are cast in the first person, my disposition to take the mental analogue of ‘I’ to refer to a single individual provides a basis for the default assumption that my conscious states all belong to a single self.

Although the disposition to identify the self that each HOT refers to cannot sustain our subjective sense of a single unified self, the pervasive practice of taking first-person thoughts to refer to a single self does. Each awareness of a mental state as belonging to a self that is referred to in the first person is tied to every other such awareness. And each such awareness is backed by a disposition to identify the self one is aware of as being in the state in question as the self that is thus aware. Our default practice with first-person thoughts explains the compelling subjective sense we have of the unity of consciousness.

This 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 and their mental analogues, but we seldom have any actual thought to the effect that such co-reference obtains. Such a thought would arise only if something seemed to challenge or override the default of taking tokens of the same type to refer to the same thing. The same holds for talking or thinking about oneself using different tokens of ‘I’ or its mental analogue. We are disposed to take all our first-person thoughts to refer to a single individual, though it is rare that one has a thought that has that explicit content.

Explaining our conscious sense of unity in dispositional terms fits with our subjective experience. We seldom have any occurrent awareness of ourselves as unified, though we are disposed to regard our conscious states
as belonging to a single unified self. That our subjective sense of unity is dispositional fits also with Kant’s having described the unity of self-consciousness as transcendental; it is something we must assume for our actual mental lives even to be possible. And he maintains only that the representation ‘I think’ must be able to accompany all other representations, not that it actually does (B131–2; cf. B406).

We are disposed to take our first-person thoughts to be about the same individual from one thought to another, and the HOTs in virtue of which we are aware of our conscious states are first-person thoughts. This enables identifying of the self one is aware of as being in various conscious states in ways that go well beyond the thin identification as the self that is aware of those conscious states. We are disposed to identify the self our HOTs refer to as the self that figures in a multitude of heterogeneous, contingent first-person thoughts that have nothing to do with what conscious states we are in.

Indeed, it is hard to imagine a case in which somebody is aware of being in pain but is wholly unable to identify in any other way the individual that person is aware of as being in pain. What sort of awareness could it be that is not sustained by some independent way of identifying the individual one is aware of as being in pain?

These informative ways of self-identification, which go beyond what conscious states one is aware of, underwrite the conceptual ties that hold between my HOTs about my own conscious states and your thoughts about what mental states I am in. My HOT that I am in pain contradicts yours that I am not in pain because we are both disposed to identify the self in question by suitably overlapping batteries of heterogeneous, contingent thoughts about myself. My disposition to identify the self I am aware of as being in various conscious states with the self that is thus aware is buttressed by a multitude of other, more mundane first-person thoughts. The result is a substantive identification of the self that one is in each case aware of as being in a conscious state. And that substantive identification undermines immunity to error through misidentification.

There is nothing special or mysterious about the way we are aware of our mental states or the individual they belong to that issues in our subjective sense that our conscious states all belong to a single self. That subjective
sense results simply from our commonsense assumption that the heterogeneous collection of ways in which we identify ourselves combine to pick out one individual, that the mental analogue of ‘I’ in all our first-person thoughts refers to a single self.

Nonetheless, some theorists have urged that since the identification of everything apart from ourselves is ultimately relative to our awareness of ourselves, that awareness must be unique and special (e.g. Shoemaker 1968; Lewis 1979; and Chisholm 1981, ch. 3, esp. 29–32). And if that is so, we would be unable to identify ourselves by appeal to a heterogeneous battery of properties relying on the multitude of first-person thoughts we have about ourselves. Self-awareness would then provide a fixed point, independent of any need to identify and by appeal to which we identify all other things.

But we seldom do appeal to ourselves when we identify other things. Identifying objects perceptually does depend on the perceptual relationship one has to the objects we perceive, but no self-awareness underlies such perceptual identifying. Indeed, one can identify perceived objects even when the perceiving is not at all conscious. One’s bodily location does provide the origin for the coordinate system within which one locates and identifies other objects. But that coordinate system reflects the sensory field of the relevant perceptual modality, and does not derive from some awareness of oneself. We may occasionally identify things perceptually relative to ourselves, but that is rather unusual; we typically identify the things we perceive relative to a larger scheme of things that contains the target object. And when appeal to that larger framework fails to identify for whatever reason, nothing about the way we are aware of ourselves independently of that larger framework will come to our rescue.

Each HOT represents the mental state it is about as belonging to some individual. One secures informative reference to that individual by way of a large collection of other, heterogeneous first-person thoughts. And we identify in that way the self to which each HOT ascribes its target as being the same from one of an individual’s HOTs to another. 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” (Appendix, 400).
Each of us is aware of our conscious states as all belonging to a single self. But by itself that does not show that the self that we are aware of our conscious states as belonging to actually exists. Perhaps one’s HOTs do not, after all, refer to the same self for all of an individual’s HOTs. The mental analogue of ‘I’ in each HOT refers to whatever individual thinks that thought. But perhaps that individual is different for a particular person’s HOTs from one HOT to another or between groups of HOTs; perhaps there is nothing underlying whatever functional integration obtains that we can reasonably call a self. But whatever the case about those questions, the natural practice of identifying a self by appeal to the referent of an individual’s first-person thoughts is sufficient to explain the compelling sense each of us has of a unified center of consciousness.

I am grateful to Myrto Mylopoulos for helpful comments on an earlier version.

1 As “the form of a natural body having in it the capacity of life” (Aristotle, 1907, 412a20, 49).

2 And, indeed, everything about mathematics and logic. See Descartes (1984, 17).

3 Existence claims are truth evaluable. So perhaps impressions and ideas represent perceptible properties as existing, appealing to Hume’s contention that the idea of existence “makes no addition” to “the idea of what we conceive to be existent” ([1739] 2000, i, ii, 6, 48). But representing something as existing requires being able to represent it as not existing, something that impressions and ideas cannot do.


5 Sensing involves only mental qualities, and no intentional content, whereas perceiving involves both. So a higher-order perception of a mental state could make one aware of a self that a higher-order perception took the relevant states to belong to. But higher-order perceptions, like higher-order sensations, would have to exhibit higher-order mental qualities, and there is reason to believe that there are no higher-order mental qualities. Nor, pace Lycan (2004), is our higher-
order awareness of conscious states in other ways more like perceiving
than like the having of thoughts. See Rosenthal (2004b, §§2–3).

6 In John Perry’s (1979) useful term.

7 Shoemaker (1968). Page references below are to the reprinted version.

8 Perry (2000). There is an extensive literature on this phenomenon; see,
e.g., Geach (1957); Prior (1967); Castañeda (1968); Anscombe (1975);
Lewis (1979); Boër and Lycan (1980); and Chisholm (1981, chs. 3–4).

9 And it is indirect discourse that matters in specifying the intentional
content of HOTs.

HOTs need not be conscious thoughts to make one aware of the mental
states they are about. For one thing, the target states do not inherit the
property of being conscious from the HOTs that are about them; rather,
HOTs make one aware of being in the target states, and that is what it is
for a state to be conscious. A state is conscious if one is aware of being
in that state in a suitably unmediated way. There is, moreover, no
difficulty in a non-conscious state’s making one aware of something;
indeed, this is just what happens in subliminal perception. Subliminal
perception has distinctive psychological effects, e.g., in one’s
subsequent expectations, decisions, and psychological processing.
These effects would not be possible unless the subliminal perceiving
made one aware of the thing perceived, though not of course
consciously aware of it. On subliminal perception, see, e.g., Marcel
(1983a, 1983b) and Breitmeyer and Ögren (2006).

11 This disposition to identify the individual that each HOT refers to
explains how HOTs can refer to oneself in an essentially indexical way
and yet seldom be conscious. But why think that we have any such
disposition? The question about who it is that has a HOT is a question
as to which individual is aware of being in a particular mental state.
And it will not address the force of that question to specify the
individual by way of such individuating properties as the having of a
particular name or location. If this question does arise for an individual
that has a HOT, the initial response will doubtless be simply that it is I
that has the HOT. But if pressed by oneself or another for more, it will
then be natural to reply that the HOT refers to the individual that has
that HOT.
For such a claim, see, e.g., Shoemaker (1968); Chisholm (1976, ch. 1, §5 and 1981, ch. 3, esp. 29–32); and Lewis (1979).

“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” (Kaplan 1989, 533). See also Lewis (1979, 543).

More precisely, perhaps the mental states of non-linguistic creatures and human infants are conscious only in respect of their qualitative character and not, in addition, in respect of their intentional properties. The qualification about features that are mental is to allow the possibility that mental states are each identical with some physical state, in which case each mental state would have, in addition to its mental properties, some physical properties that are not mental.


The classic study is Nisbett and Wilson (1977). See also Johansson et al. (2005); Hall et al. (2010); Frith (2007); and Schnider (2008).

E.g., Kriegel (2009) and Gennaro (2006). It is unlikely that any higher-order awareness can be intrinsic to the state it is about. For one thing, that awareness typically occurs slightly after the state itself; see, e.g., Libet (2004, ch. 2) and Libet et al. (1979). Timing to one side, the higher-order state must be assertoric to result in one’s being aware of the first-order state, and the first-order state will often have a non-assertoric mental attitude or, in the case of sensations, none at all. And a difference in mental attitude points to distinct states.

As Kriegel (2009, ch. 4, §II) and Gennaro (2004, §2) urge. For convincing counter-arguments see Weisberg (2008).

Shoemaker (1968, 8). Shoemaker urges that such immunity applies even when I take myself to be performing some action. See also Evans (1982a) and Bermúdez (1998).

Shoemaker appeals to Hume’s statement that “I never catch myself at any time without a perception, and never can observe anything but the perception” ([1739] 2000, i, iv, 6, 165; emphasis Hume’s).

Cf. what P. F. Strawson labels the no-ownership view of the self (1959, 95–98), and Strawson’s reasons for seeing Wittgenstein as having held that view (1959, fn. on p. 95).
In (2005, ch. 13, §6) and (2004a, §4) I argued against immunity to error through misidentification based in part on an alleged counter-example. I am grateful to Myrto Mylopoulos for noting that the case I described in those places is not a counter-example and for pressing other points, which together led to my developing the present treatment. What I say here is similar in spirit to my earlier discussions, though with substantial differences in the argument.

Though even immediacy cannot guarantee reliability without other substantive assumptions.

The newer designation supplanted the older one in DSM-IV (1994).

If a pain or other mental state is integrated into the mental life and behavior of one alter but not another, it counts as belonging to the first and not the second.

Some theorists have speculated that even in normal cases there is something like multiple selves that must be suitably integrated within the overall psychological functioning relevant to a single individual person. On the possibility of something like multiple selves, see, e.g., Gazzaniga (1985) and Ornstein (1986).

As Anscombe imagines (Anscombe 1975, 58).

The representation ‘I think’ is a non-empirical (B132) or transcendental (B401, A343) representation.

It is convenient to posit a single self when an individual’s mental life and behavior is suitably well integrated. Similarly, we countenance multiple alters in DID when such integration breaks down and there are instead several independent streams of mental states in a single individual, each of which is relatively self-contained and well integrated.

For challenging doubts and useful references about whether there is normally a single, unified self for each individual, see Klein (2010).
References


Berger, Jacob (manuscript). “Intentionalism and representational qualitative character.”


Byrne, Alex 2005. “Introspection.” *Philosophical Topics* 33, 1: 79–104.


Fiala, Brian (manuscript) “The phenomenology of explanation and the explanation of phenomenology.”


Gordon, Robert M. 1995. “Simulation without introspection or inference from me to you.” In Martin Davies, and Tony Stone (eds.) Mental


activation during the observation and experience of touch.” *Neuron* **42**: 335–46.


Loar, Brian 2003. “Phenomenal intentionality as the basis for mental content.” In Martin Hahn and Bjorn Ramberg (eds.) Reflections and


Rust, Joshua and Schwitzgebel, Eric (in preparation). “Ethicists’ and non-ethicists’ responsiveness to undergraduate emails.”


Schwitzgebel, Eric 2010. “Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief.” *Pacific Philosophical Quarterly* 91: 531–53.


Tunik, Eugene, Lo, On-Yee, and Adamovich, Sergei V. 2008. “Transcranial magnetic stimulation to the frontal operculum and supramarginal gyrus


CONSCIOUSNESS AND THE SELF: NEW ESSAYS

JEELOO LIU and JOHN PERRY