In *The Philosophy of Marjorie Grene*, in *The Library of Living Philosophers*, ed. Randall E. Auxier and Lewis E. Hahn © *David M. Rosenthal* 

### PERSONS, MINDS, AND CONSCIOUSNESS

There is a recurring tension in philosophical thinking having to with the relations between wholes and their parts. Should we seek to understand things in terms of combinations of constituent parts from which they are, or might be, constructed? Or must we, instead, explicate those parts in terms of the wholes to which they belong? Both strategies have long histories and strong advocates.

This tension is especially vivid in cases in which we describe some phenomenon by reference to some distinctive unity or unifying function. An example of this is the problem of saying what it is for something to be a person. One difficulty in thinking about what it is to be a person stems from Descartes's famous argument in Meditation VI that, because the self is an unextended, thinking substance, it is distinct from anything bodily.1 Persons are beings with mental capabilities, and that mental functioning is essential to their nature. But, as Marjorie Grene has forcefully and eloquently insisted, we cannot understand what it is to be a person as long as we see its mental functioning as that of a distinct Cartesian substance. We can understand the nature of persons only if we see that mental functioning as firmly rooted in the rest of reality.2 It is not just that we must see persons as having both mental and bodily natures, as P. F. Strawson and others have argued.3 As Grene rightly argues, we must represent these mental and bodily natures as functionally and inextricably unified.

Grene goes farther, however, in also arguing that we cannot capture the distinctive functional unity of persons in terms of the independent categories of mind and body. We must, she urges, rethink the issues "in terms of new categories, primarily in terms of the category of the 'person' rather than of either body or mind." We must try to "go between the horns of the traditional dilemma and to espouse neither matter nor mind, nor both of them, as [our] fundamental concepts."4

Here I am less certain. I believe that we need not start with the concept of a person, but can instead explain what it is for a creature to be a person by reference to various distinctive aspects of that creature's mental functioning. The issue here is not the ontological question of whether persons have distinct mental and physical parts; few today would maintain any such view. The issue, rather, is the methodological question about how to understand what it is for something to be a person. Can we understand what it is to be a person in terms of various aspects of the mental and bodily functioning of persons? Or must we posit a basic category of persons, which cannot in turn be fully understood in terms of the kinds of mental and bodily functioning that go into being a person?

In what follows, I sketch my reasons for believing that we can understand the concept of person in terms of the mental and bodily functioning characteristic of persons. In section I, I argue that we can understand mental functioning only as inextricably bound up with the interactions the relevant body has with its environment. This is not just because mental states are special cases of bodily states; it would be so in any case. But not all creatures that function mentally qualify as persons. So section II takes up the question of what it is that distinguishes persons from other creatures with mental endowments. It turns out that these distinguishing features all have to do with the special way in which persons are conscious of their thoughts, feelings, desires, and perceptions. Since we must understand consciousness to understand personhood, section III briefly sketches an account of what it is for mental states to be conscious, an account that avoids the Cartesian gulf between mind and matter. Section IV explains the special way in which persons are conscious of their mental states, which involve our sense of the unity of consciousness and the self. and section V considers several objections that explanation.

# I. Mind, Body, and Environment

One reason for the continuing appeal of Cartesian approaches to understanding the mind is the difference between what it's like for one to feel, perceive, desire, or think something and how we think about somebody else's being in such mental states. It seems obvious to many that the mental properties in virtue of which one knows of oneself that one is in pain, for example, are of a wholly different kind from the properties in virtue of which one knows that somebody else is in pain. And similarly for perceiving, believing, desiring, and all other mental states.

Not all mental states occur consciously. It is now widely

recognized that thoughts and desires occur without being conscious; people believe and desire many things without being in any way at all conscious of their having those beliefs and desires. There is also compelling reason to hold that even bodily sensations, such as pains and aches, and perceptual sensations, such as visual sensations of red, can occur without being conscious. In subliminal perception and peripheral vision, and in experimental results such as those involving masked priming,5 perceptual sensations occur without the subject's being at all conscious conscious of them. It even sometimes happens that we can tell from somebody's behavior, say, from a slight limp, that the person is in pain, even though the person may have been wholly unaware of it. And no mental state is conscious if the subject is in no way whatever aware of being in that state.

The recognition that mental states occur without being conscious causes us to rethink the apparent disparity between firstand third-person points of view about what it is for somebody to be in a mental state. The first-person point of view about mental states operates only when those states are conscious. That firstperson point of view is cast in terms of what it's like for one to be in a mental state, and when a mental state is not conscious there is nothing at all that it's like for one to be in it. When the mental state one is in is not conscious, one has access to that state in just the way others might have access to it, by inferring from one's behavior and from the occurrence of environmental stimuli.

Any particular mental state, however, can be conscious at one time and not at another. Even when a belief or desire occurs without being conscious or a perception occurs subliminally, one can come to be conscious, in the way characteristic of our conscious states, of that belief, desire, or perception. But we can have a first-person point of view only about those of our mental states which are currently conscious. Since the very same states can be at one time conscious but not at another time, one must adopt a third-person point of view toward many of one's own mental states. Whatever disparity obtains between first- and third-person points of view is equally a difference between one's own conscious and nonconscious mental states. This underscores the need to dissolve the appearance of a gulf between first- and third-person approaches. We must find some way to explain how it is the very same mental phenomena to which we have access in the two ways.

Pressure to reconcile first- and third-person points of view about mental states makes so-called functionalist accounts of such

states especially appealing. Such accounts define mental states in terms of the characteristic causal connections each type of state has with behavior, stimuli, and other mental states. On the version of this approach developed by David Lewis, the relevant connections are those we can extract from the folk-psychological platitudes that constitute our shared knowledge and assumptions about all the various types of mental state. These commonsense platitudes encapsulate general background knowledge not only about causal ties, but also about how the various mental states get grouped into types and other matters that help define them. Because the general background information defines each type of state in part by reference to the connections states of that type have with other types of state, we cannot define any state solely in terms of the behavior and stimuli that is relevant to it. Rather, we will cast the definition of each type of state in terms of ties that state has not only with behavior and stimuli, but with mental states of many other types.6

Since these functional definitions of mental states rely on the platitudes that constitute our shared folk-psychological knowledge, they will include shared knowledge about our first-person access to mental states. What it's like for one to be in conscious states of each type will include information about how the various mental states seem to resemble and differ from one another and how we taxonomize them from a first-person point of view. All of this will figure in the resulting functional descriptions of mental states. Lewis argues, moreover, that these functional descriptions define our concepts of each type of mental state. So our very concepts of the different types of mental state characterize not only those states in terms of environment and behavioral factors, but in terms of both first-person considerations as well. They represent all our mental states both from a first-person point of view and as inextricably embedded within nonmental reality.

Lewis develops his functionalist characterization of mental phenomena in the context of an argument for mind-body materialism. The discovery that certain bodily states realize the functional description our folk-psychological platitudes generate would sustain mind-body materialism. But the functional characterization of mental phenomena, by itself, is wholly independent of both dualism and materialism. It is a theory, rather, of the nature of mental states. And, since the theory relies on nothing but our shared folk-psychological knowledge, none of our commonsense intuitions about the mental can undermine it.

Not everybody writing today would accept Lewis's functional

characterization of the mental states. Some theorists, indeed, would characterize at least some types of mental state in terms free of any essential ties with nonmental reality. Such characterizations of mental states, especially common for qualitative states, also make the reconciliation of first- and third-person points of view about mental states difficult, if not impossible.

But the persistence of such quasi-Cartesian approaches should not encourage us to think that we can overcome the Cartesian gulf between matter and mind and between first- and third-person only by adopting a basic category, such as that of a person. Lewis's functionalism shows that we can put those Cartesian oppositions to rest even if we operate with relatively traditional notions of mind and body.

Indeed, even for those who reject Lewis's functionalist approach, the dominant theoretical approaches today tend to characterize mental states in ways that minimize the Cartesian gulf. Many contemporary explanations of what it is for an intentional state to have content, for example, appeal in some ineliminable way to environmental factors.7 Since content is essential to a state's being intentional, such externalist theories of mental content characterize intentional states in terms of such environmental considerations. Such characterizations are again independent of issues about dualism and materialism.

# II. Persons and the Mental

Functionalism in the style of Lewis shows that we can characterize mental functioning in ways that reconcile first- and third-person descriptions and locate mental functioning inextricably within nonmental reality. But, by itself, that does not obviate the need for adopting a basic category of a person independent of the traditional categories of mind and body. For that, we must also show that we can understand what it is to be a person by appeal only to those traditional notions, without any primitive category of a person.

As Harry G. Frankfurt noted some years ago, traditional discussions often overlook the difference between what it is to be a person and what it is simply to be a creature with mental capabilities.8 But there is more to being a person than just having a mental life. Many nonhuman animals that have relatively little in common with persons nonetheless have mental capabilities; indeed, psychologists continue to discover that the mental capacities of

such nonhuman animals are astonishingly elaborate. Nor is the concept of a person simply the concept of a human being; though it is doubtful that any other terrestrial animals are persons, we all recognize the possibility of encountering nonterrestrial creatures that do count as persons.

Frankfurt's own account of what is special about persons appeals to an individual's ability to have second-order desires, desires that one have or not have some first-order desire. Forming such higher-order desires, he urges, involves identifying oneself with one, rather than another, of one's first-order desires. And he sees the ability to identify oneself in this way with one's first-order desires both as essential to the process of deciding and also as what is distinctive of being a person.

Doubtless the ability to identify oneself as an individual with certain desires is a crucial aspect of what it is to be a person. But self-identification by way of first-order desires is hardly the only way in which we identify ourselves by reference to our mental states. We also identify ourselves as individuals who think certain things, have various memories, and have characteristic feelings. All these self-identifications, moreover, play an important role in what it is for a creature to be a person. It would at best be an oddly limited person whose self-identification by way of its mental states was limited to being an individual that had certain desires.

Frankfurt's appeal to second-order desires does, however, point toward a more satisfactory account of what it is to be a person. Second-order desires presumably influence what first-order desires an individual will have, and thereby impose a measure of order and unity on that individual's mental life. Such order and unity in one's mental life are an important aspect of what it is to be a person. In influencing what first-order desires one will have, moreover, one's second-order desires will often lead one to reflect on one's desires and on what desires one wishes to have. In part this is because one's second-order desires will influence one's first-order desires when one desires not to have some firstorder desire that one actually has or to have some first-order desire that one lacks. Such conflict between first- and secondorder desires will often prompt one to reflect on both desires and the tension between them. And the ability for such reflectiveness is also central to what it is for a creature to be a person.

Second-order desires, however, are not the only sources for the mental unity and order and the self-reflectiveness that are characteristic of persons. One may come to reflect on one's mental states because one is puzzled about a practical problem or an intellectual issue, and one may reflect on one's perceptual experiences because of one's enjoyment of a situation or an aesthetic experience. Whenever one thinks seriously about something or attends to one's past or present experiences, one is likely to enhance the sense of unity and order in one's mental life.

Many creatures that do not qualify as persons have mental states and processes that exhibit an impressively high degree of integration and coherence. Integration and coherence of mental function is necessary to get around in the world successfully, and especially to interact with one's conspecifics. So it cannot simply be such integration and coherence that distinguishes persons from other creatures. What sets persons apart is rather that they are conscious of their mental functioning as being coherent and they are conscious of it as all belonging to a single individual. It is not mental unity and coherence that distinguishes persons, but consciousness of mental unity and coherence. Both the selfreflectiveness and the mental unity characteristic of persons are matters of how persons are conscious of their thoughts, feelings, and perceptions.

More must be said about these special ways of being conscious of one's mental states. As noted earlier, no mental state counts as being conscious if the individual who is in that state is in no way conscious of the state. But the routine way we are ordinarily conscious of our conscious states falls well short of the self-reflective, unified consciousness that is distinctive of persons. Section IV takes up the question of what this special self-reflective, unified consciousness consists in. First, however, we must address a Cartesian challenge the about consciousness of mental states generally.

Consciousness is the mark Descartes appeals to as distinctive of mental functioning. Thus he defines thoughts in the Geometrical Exposition of the *Second Replies* by saying that "the word 'thought' applies to all that exists in us in such a way that we are immediately conscious of it" (*AT* VII, 160). "[N]o thought," he reiterates in the *Fourth Replies*, "can exist in us of which we are not conscious at the very moment it exists in us" (*AT* VII, 246). Indeed, consciousness very likely underlies Descartes's very distinction between mind and body; it is presumably the unity of consciousness that underlies the Meditation VI claim that mind is indivisible, in contrast with the indefinite divisibility of body. Consciousness, it seems, is the fundamental property of mental reality which sets it apart from body.

This suggests a difficulty in explicating what is distinctive of persons in terms of the way persons are conscious of their thoughts, feelings, and perceptions. If we explain mind in terms of consciousness, there is nothing left by appeal to which we could, in turn, explain consciousness; consciousness will lie outside the net of naturalist explanation. And, if consciousness sets mind apart from all physical reality, the appeal to consciousness in constructing an understanding of personhood will bring the Cartesian gulf between mind and body along with it. The only way to avoid that gulf then would be, as Grene urges, to "go between the horns of the traditional dilemma and to espouse neither matter nor mind, nor both of them, as [our] fundamental concepts."

We need not follow Descartes, however, in defining mind in terms of consciousness; we can, instead, go in the opposite direction, and explain consciousness in terms of mental states that are not themselves conscious. That will allow us to operate with a concept of consciousness that implies no Cartesian gulf, so we can appeal to it in constructing our understanding of personhood without commitment to such a gulf. If we then also adopt a functionalist account of mind, we can explain both consciousness and personhood in a way that locates mental states firmly in the context of nonmental, physical reality. The following section, therefore, develops an account of consciousness in terms of which this can all be done.

# III. Consciousness and the Mental

No mental state is conscious if the individual who is in that state is in no way whatever conscious of being in that state. This means that whenever a state is conscious, we are in some way or other conscious of that state; being in some way conscious of a mental state is a necessary condition for that state's being conscious. Indeed, it is notable that Descartes never writes of a thought's being conscious, but only of our being immediately conscious of a thought, though it is plain that he means by this exactly what we mean by saying of a mental state that it is conscious.

Although we are in some way or other conscious of every conscious state, not every way of being conscious of a mental state results in that state's being conscious. Suppose, for example, that I am conscious of thinking or feeling something solely because I infer that I do, or solely because you tell me and I believe you. Being conscious of a thought or feeling solely in those ways does not result in its being conscious. So we need to specify just how one must be conscious of a mental state if that state is to be conscious. If we succeed in specifying that, we will have conditions for a state's being conscious that are not only necessary, but sufficient as well.

It might seem that explicating what it is for a state to be conscious in terms of one's being conscious of the state is unavoidably and viciously circular, since it seeks to explain consciousness in terms of itself. But the phenomenon of consciousness being explained here is not the phenomenon that the explanation appeals to. A mental state's being conscious is, intuitively, a property of that state in virtue of which it figures as part of one's stream of conscious. Conscious states are those which one can report on; one can tell others when they occur and what it's like for one to be in them. One can, moreover, deliberately shift one's attention to them. When a mental state is not conscious, one can do none of those things.

A mental state will exhibit none of these marks of its being a conscious state unless one is in some suitable way conscious of the state. But we understand independently of all this what it is for one to be conscious *of* something. One is conscious of something when one has some mental reaction to it that enables one to respond to its presence. There are, accordingly, two broad ways in which we are conscious of things, by sensing them and by having thoughts about them as being present. Having a sensation of something or a thought about it as being present, moreover, makes one conscious of that thing even when that sensation or thought is not, itself, a conscious sensation or thought. If I subliminally see something, my visual sensation is not a conscious state, but having that sensation makes me conscious of the thing I subliminally see. Though I am not conscious of seeing it, my seeing it still makes me conscious of it.

More important, we understand what it is for somebody to have a thought or a sensation of something independently of understanding what it is for that thought or sensation to be conscious. So we understand what it is to be conscious *of* something independently of understanding what it is for mental states to be conscious. There is no circularity in explaining what it is for a mental state to be conscious by reference to our being conscious *of* that state. Corresponding to the two ways of being conscious of things, there are two models for how it is that we are conscious of those of our mental states which are conscious states: by sensing those states or by having thoughts about them. The first model, which posits an "inner sense"9 by which we are aware of our conscious states, has been by far the dominant theoretical approach to the question of how we are aware of those states. And, since the access we have to things by sensing them seems unmediated, a model on which we sense our conscious states would readily explain why our awareness of our conscious states also seems to be direct and unmediated.

Despite its traditional popularity, however, the inner-sense model faces grave difficulties. Perhaps the most damaging is that sensing always involves some qualitative character, but there is no distinctive qualitative character associated with the way we are conscious of our conscious mental states. No qualitative properties figure at all in connection with our conscious intentional states, such as beliefs, doubts, desires, expectations, and the like. And, although conscious perceptions, sensations, and emotions do have qualitative character, the qualities are always those of the states we are aware of, not the higher-order states in virtue of which we are aware of them. Since some qualitative character always occurs in sensing, it cannot be that we are conscious of our conscious states by sensing them.

The only alternative is that we are conscious of our conscious mental states by having thoughts about those states as being present. Because these thoughts are about other mental states, I refer to them as *higher-order thoughts (HOTs)*. A mental state is conscious if one is conscious of it by having a HOT about it, a HOT to the effect that one is in that state.10

The inner-sense model seemed to explain why the way we are conscious of our conscious states seems unmediated and direct, since sensing in general makes us aware of things that way. But HOTs can handle this as well. The only reason we have to think that we are directly aware of our conscious states is, after all, that it seems that way to us; it is just that nothing seems to mediate between those states and our awareness of them. And that often happens with the thoughts we have about things. When a thought relies on no observation and on no inference of which we are aware, nothing seems to us to mediate between the thought and what it is about. And we can stipulate that the HOTs in virtue of which we are aware of conscious mental states do not result from any inference of which we are aware.11 As noted earlier, the self-reflective, unified consciousness that distinguishes persons from other creatures goes well beyond the unreflective, seemingly automatic way in which we are ordinarily conscious of our everyday conscious states. It is plain, in the case of most conscious states, that we are aware of being in them; we can, if asked, say whether we are in them. Still, our awareness of our mental states in these cases is neither focused nor deliberate, nor, in these cases, do we not even notice that we are aware of those states. We are aware of them without noticing that we are. In the case of the self-reflective, unified consciousness that distinguishes persons, by contrast, the way we are conscious of our conscious states is deliberate and focused. We are not merely conscious of these states; we are actually conscious that we are. We are, in these cases, *introspectively* conscious of our conscious states.

The HOT model offers a compelling explanation of the difference between such introspective consciousness of our mental states and the way we are conscious of those states in the ordinary, unreflective cases. A mental state is conscious if it is accompanied by a HOT to the effect that one is in that state, a HOT based on no conscious inference. But that HOT is, itself, a mental state, and it will not be conscious unless it is accompanied, in turn, by a yet higher-order thought about it. In ordinary, unreflective cases, our HOTs are not conscious; they occur, but we are not aware of them. Introspective consciousness is the special case in which we are conscious not only of our first-order mental state, but conscious also of the HOT in virtue of which we are conscious of the target first-order state.12

This explanation of the difference between introspective and nonintrospective consciousness fits well with the intuitive data about both kinds of case. When a mental state is conscious in the ordinary, nonintrospective way, we are conscious of the state but not of being conscious of it. That is why we are unaware, in these ordinary cases, of having any HOTs; our HOTs are not, in these cases, conscious. When we are reflectively conscious of a mental state, and so conscious of it as being part of a unified fabric of conscious states, we are also aware of the way we are conscious of the state. We are aware not only that we are focusing on the state attentively and deliberately, but also that the way we are conscious of the state represents it as part of that unified mental fabric. We not only have HOTs about the state; the HOTs we have about it are conscious thoughts.13 This account of self-reflective consciousness we sometimes have of our mental states builds entirely on mental states that are not, in themselves, conscious states. A mental state's being conscious, on this account, is the relational property a mental state has of being accompanied by a HOT in virtue of which one is aware of being in that target state. And our being reflectively conscious of a state, and of ourselves as being in that state, consists in our also being aware of that HOT, itself. Having a HOT makes one conscious of oneself as being in the target state; so when a HOT is conscious, we are also aware of this way of being conscious of ourselves.

On this account, no mental state is intrinsically or essentially a conscious state. We do not explain mentality in terms of consciousness, but the other way around; consciousness is a special case of mentality. Since we can develop an informative explanation of consciousness, there can be no serious temptation to regard it as something apart from the rest of reality, and hence as nonphysical. We can thus appeal to reflective consciousness in explaining what distinguishes persons from other mental beings and still avoid commitment to the Cartesian gulf between mind and body.

The foregoing account of consciousness also fits comfortably with Lewis's functionalist theory about the nature of mental states generally. It is part of our folk-psychological understanding of a mental state's being conscious that one is conscious of that state. So a functionalist account that extracts defining causal relations from our folk-psychological platitudes will characterize a state's being conscious in that way. The result is a fully naturalist model for reflective consciousness that anchors it firmly in the context of the interactions between a creature's mental functioning, its bodily constitution, and its physical environment.

As noted earlier, many who reject such functionalism still characterize intentional states by reference to nonmental factors. That is because they hold that intentional content is determined in part in terms of those things in the physical environment to which intentional states refer. But this account may seem to raise a problem for the HOT model of consciousness. On such an externalist account of content, the content of a thought is a function of what environmental objects or substances the thought is connected to. Whether that content is about water or the macroscopically indistinguishable twin water of Hilary Putnam's famous twin-earth thought experiment would then depend on which of the two substances the actual environment contains.14 But the content of the HOT in virtue of which a first-order thought is conscious is only about that first-order thought. So its content would be a function not of what substance occurs in the environment, but just what mental state it is about. But the way HOTs characterize their target states determines the way we are conscious of them; we are conscious of being in the type of state a HOT describes. So how, if the content of our thoughts is in part a function of what they are about, could one be conscious of one's first-order thought *as* a thought about water, and not a thought about twin water?15

One possibility is that the HOT gets its content, insofar as that pertains to environmental objects and substances, from its target. But there is another possibility as well. An intentional state's actual content need not be the same as the content we are conscious of that state as having. The way we are conscious of our conscious states is a function of the way HOTs characterize them; so when an intentional state is conscious, we are conscious of being in a state with the content the HOT represents its target as having.

Suppose, then, that one has a conscious thought about water; its being about water is a function of some connection the thought has with actual water in the environment. But the HOT one has about that thought, in virtue of which it is conscious, may well represent the thought in a way that is neutral with respect to the distinction between water and twin water. One would not be conscious of the thought *as* a thought specifically about water, as against twin water, but as a thought that is about some substance that could, for all one knows, be either. Indeed, it is likely that that is how we are actually conscious of many of the thoughts we have about water, at least those not cast, directly or not, in terms of chemistry.16 Conscious states are seldom, if ever, conscious in respect of all their mental properties. So intentional states need not be conscious in respect of content that is as fine grained as the content they actually have.

### IV. Selves and the Unity of Consciousness

Descartes saw mind as nonphysical because the physical, being spatial in nature, is indefinitely divisible, whereas mind is indivisible. Even if there were, contrary to Descartes's own view, physically indivisible atoms, the atoms would be conceptually divisible. But mind might seem to be no less divisible, not spatially, but into mental parts. At any given moment, a person is in many mental states; why not, then, regard a person's mind as constituted by those states, and hence divisible into them? If the mind is, as Hume held, "nothing but a mere heap or collection of different perceptions," each perception can, as he observed, "be consider'd as separately existent."17

But persons do not seem to be mere heaps or collections of mental states. The conscious states of persons seem, rather, to be unified somehow into a single consciousness. But it may also seem as though the HOT model of consciousness just sketched embraces a mental atomism like that of Hume. If each conscious state owes its consciousness to a single accompanying HOT, how can any sense arise of the unity of consciousness? Why would all our conscious states seem to belong to a single, unifying self?18

One unifying force in our mental lives is the way sensory experiences occur in relation one to another. Consider visual experience; we can locate each visual sensation in relation to every other, each appearing to be to the right or to the left or above or below every other such sensation. Similarly with all the other sensory modalities. And we calibrate this apparent spatial location from one sensory modality to another, so that a sound, for example, may appear to occur at the same place as a sight. Such cross-modal calibration appears to bind the various sensory fields together into a single modality-neutral field.

But the problem about the unity of consciousness is not to explain how the qualitative contents of experience are unified, but rather why we are *conscious of* those qualitative contents as being unified in a single consciousness. If a mental state's being conscious were an intrinsic property of that state, perhaps the unity of qualitative contents would carry along with it the unity of consciousness. But on the HOT model, a state's being conscious is not an intrinsic property of that state, but the relational property of being accompanied by a HOT. And we want to know why mental states that are conscious solely because each is accompanied by an individual HOT should seem to be unified in a single consciousness.

Part of the answer is that a HOT need not be just about a single target state, but can be about a number of states at once. Indeed, introspection itself suggests that we are conscious of our conscious mental states in this way. When something attracts our attention visually, we may be conscious of a small area of our visual field in considerable detail, but usually we are conscious of many different visual inputs, none in much detail, as a kind of

overview. So it is reasonable to suppose many target qualitative states will be conscious in virtue of a single HOT that is about all of them, and about all of them as occurring in some suitable spatial array. Something similar happens with hearing, and even to some degree with tactile awareness. Doubtless a single HOT often has as its targets qualitative states from distinct modalities, as when we at once consciously see and hear a particular event. And our often being conscious of many sensory inputs as a group helps engender in us a sense that our qualitative states are unified, in that they belong to a single consciousness. The HOT model can readily explain this by supposing that, with many HOTs, each makes us conscious of large clusters of qualitative states.

HOTs operate in another way to produce in us a sense of the unity of consciousness. A HOT is a thought to the effect that one is in a particular mental state or cluster of such states. Such reference to oneself is required because a thought makes one conscious of things only when it represents them as being present, and the only way for a mental state to be present is for it to be a state of oneself. So the content of each HOT must make reference to a self that is in the state or states in question.

But this requirement cannot, by itself, give rise to a sense of the unity of consciousness, since the self to which each HOT refers might, for all we have shown so far, be different from one HOT to the next. And no sense of unity will result from each HOT's referring to a self unless it also seems that the self each HOT refers to is the same for all HOTs. And it may well seem that individual HOTs cannot achieve this sense of sameness.

We identify ourselves as individuals in a broad variety of ways that have little systematic connection. Some of these ways appeal to memories of personal history, while others rely on contingent facts about one's current location and situation. There is no magic bullet in virtue of which we identify ourselves, just a vast but loose collection of considerations, each of which, taken individually, is relatively unimpressive. But the combination is enough to identify oneself in any case in which the question may arise; any time one has a first-person, that is, a thought about oneself, one can, if pressed, specify the individual that thought is about.

Such self-identification operates in the first instance with the mundane first-person thoughts each of us has about ourselves, about what one's name is and where one lives, what one's personal history is, what one likes and dislikes, and so forth. Whenever one has a new first-person thought, one secures the reference to oneself that occurs in that thought by appeal to the referent of these other, self-identifying first-person thoughts. And this sometimes adds to the stock of self-identifying thoughts one uses to secure reference to oneself.

HOTs are also first-person thoughts, and the same process extends to them. We appeal to this broad, heterogeneous collection of contingent considerations to specify the individual each HOT represents its target state as belonging to. We take this heterogeneous collection to pick out the same individual from one case to another. And because that applies to our HOTs, it forms the basis for the sense we have that our conscious mental states are unified as belonging to a single individual. Our sense of the unity of consciousness does not result from something special about the the way we are conscious of our conscious mental states. Rather, it is an extension of the everyday assumption we operate with that, for each of us, the heterogeneous collection of ways in which we identify ourselves go together to pick out a single individual.

The everyday assumption that our first-person thoughts all refer to one individual may suffice for some sense that our conscious states occur in a single consciousness. But that sense of unity will not, itself, be conscious unless some of one's HOTs are themselves conscious thoughts. And this is just what happens in the reflective consciousness distinctive of persons. When one is introspectively conscious of one's mental states, one is conscious not only of those states, but also of being conscious of them.

On the HOT model, that means becoming conscious of one HOTs, each of which represents its target state or states as belonging to some individual. Since one secures the reference to that individual by way of one's heterogeneous collection of self-identifying thoughts, one identifies the individual to which each HOT assigns its target as being the same for all HOTs. So, when one is conscious of one's HOTs, one becomes conscious of them as assigning their targets to some single individual. One thereby becomes conscious of oneself as a center of consciousness.19 We need not posit an indivisible Cartesian soul or any special kind of consciousness to explain traditional intuitions about self-consciousness and our sense of the unity of consciousness.

There are two main characteristics that distinguish persons, the ability persons have to be reflectively conscious of their mental states and the sense they have that their mental states are unified by belonging to a single center of consciousness. The foregoing considerations explain why these two characteristics go together. Being reflectively conscious of one's mental life consists in being not just conscious of one's conscious states, but conscious, in addition, that one is conscious of them. On the HOT model, one not only has HOTs, in virtue of which one's mental states are conscious; one's HOTs are, themselves, sometimes conscious. And being conscious of one's HOTs enables one to become aware of one's conscious states as belonging to a single self.20

## V. Problems about the Self

It might be objected that the appeal to a heterogeneous collection of contingent properties cannot do justice to the way one's firstperson thoughts, including HOTs, refer to oneself. Mistakes are always possible in identifying oneself; one might, for example, think that one is Napoleon. As I have argued elsewhere, such error is even possible in the case of HOTs; one can be conscious of oneself as being in mental states that one is not actually in.20

But despite the possibility of these kinds of error, it has been argued that there is another way in which none of one's firstperson thoughts can be mistaken. Though I can be mistaken in thinking I am Napoleon, it seems that I cannot in such a case be mistaken about who it is that I think is Napoleon. And if I think that I am in pain or that I believe a certain thing, I cannot be mistaken about who it is that I think is in pain or has that belief. One's first-person thoughts are, in Sydney Shoemaker's phrase, "immune to error through misidentification," misidentification, that is, with respect to reference to oneself.22

Such immunity conflicts, however, with the foregoing account of self-identification. Identifying the individual a first-person thought refers to by appeal to some heterogeneous collection of contingent properties plainly leaves open the possibility of error through misidentification. If I identify actually myself by reference to a heterogeneous collection of properties, I might do so erroneously. This holds for the HOTs in virtue of which we are conscious of our conscious states no less than for our other firstperson thoughts.

When I have a conscious pain, I cannot erroneously think that the individual that has that pain is somebody distinct from me, though I can of course be wrong about just who I am. How can we capture this elusive distinction? When I have a conscious pain, I am both in pain and conscious of being in pain. The error I cannot make is to think that the individual who is conscious of the pain is distinct from the individual that has the pain of which some individual is conscious.

The HOT model allows for a natural explanation of this immunity from error. HOTs represent their targets as being states of the same individual that thinks the HOT. The HOT I have about a conscious pain represents that pain as belonging to the very same individual as the individual that thinks the HOT, itself. Thinking of the conscious pain as belonging to an individual distinct from me would mean thinking that the individual that has that HOT is distinct from the individual that HOT represents as being in pain. And that would conflict with the way the HOT represents things.23

The respect in which I cannot represent my conscious pain as belonging to somebody distinct from me consists in my being unable to represent the pain as belonging to somebody other than the individual who is conscious of the pain. The idea that some special immunity to error is involved here results from an illicit assimilation of mental states to one's consciousness of them.

Although I cannot think that the individual that has that pain is somebody distinct from me, I can be mistaken about just who it is that I am. Being mistaken about that is simply going wrong in how I identify the individual who is conscious of the pain. And that is possible because such identification consists in my identifying the individual who has the HOT in virtue of which my pain is conscious as the individual picked out by some heterogeneous collection of contingent properties. And I can be wrong about whether the individual who has the HOT about the pain is the same as the individual, if any, that has all those properties.

These considerations help also with another quandary about the self. It has often been noted that I can think that I, myself, have a particular property without thereby thinking that some individual has that property, even when I am the individual the first thought is about. Consider John Perry's vivid example, in which I see a trail of spilled sugar from somebody's grocery cart. My thought that somebody's grocery cart is spilling sugar does not imply the thought that my grocery cart is spilling sugar, even if it happens to be my cart.24 This disparity between the two types of thought is sometimes taken to show that contingent properties cannot underwrite the reference to oneself that occurs in one's first-person thoughts. Whatever contingent properties one appeals to, a thought that identifies an individual as having those properties will still not be a thought about oneself, as such.25 But it is unclear what it actually is for a thought to be about one, as such. One natural possibility is that a thought's being about me is just its being about the individual identified by reference to the huge collection of contingent properties in terms of which I think about myself. If there is something more to a thought's being about me, as such, it needs independent explication, not just insistence on the difference.

There is, of course, a strong sense we have that a thought's being about me, as such, is something more than its being about an individual identified by reference to some collection of contingent properties. But we can explain that sense without crediting the intuition.26 Mental states are conscious, when they are, in virtue of their being accompanied by HOTs, and each HOT represents its target as belonging to the individual who also thinks the HOT in question. Something similar happens with my thoughts that are about me, as such. Suppose I think that I have some property, P. That thought's being about me, as such, consists in its representing the individual that has property P as being the very individual who thinks that thought. And no collection of properties figures in securing that co-reference. If I think that an individual that has some collection of properties also has property P, my thought will not automatically represent the property as belonging to the very individual who thinks that thought.

The present approach to personhood and the self also suggests a natural account for the puzzling phenomenon of Multiple Personality Disorder.27 Consider a highly simplified, artificial case. An individual seems to have two selves, each with different thoughts, desires, and experiences. The first self, moreover, seems to know nothing of the second, though the second seems to know about all the thoughts, desires, and experiences of the first.

In part, such cases must involve an individual's having two partially disjoint series of thoughts, desires, feelings, and experiences. These collections will partially overlap when it comes to beliefs about a lot of shared background information, as well as matters pertaining to location and the environment; and they will overlap in other ways, as well. But they will diverge about many other things. The states in each collection fit with one another in a reasonably coherent way, though those in one group often do not fit comfortably, if at all, with those in the other.

But that is not, by itself, enough to explain the sense we have of two selves' being present in the individual. For that, we

must posit two disjoint sets of HOTs the individual has about the various first-order states. In the first condition, the individual has HOTs about thoughts, experiences, and desires that occur in only one of the two partially disjoint collections. In the second condition, the individual's HOTs are about the states in the other partially disjoint collection.

We identify all our first-person thoughts, HOTs included, by reference to some heterogeneous collection of contingent properties. In the imagined case of an individual that appears to have two selves, that individual will use two, partially disjoint collections of contingent properties to identify the individual that its first-person thoughts are about. That applies equally to how the individual identifies who it is that is conscious of its current conscious states. The individual appeals, in the two conditions, to different self-descriptions in identifying who it is that is conscious of its conscious states.

This explains how the second apparent self can have access to the first self's states without those states being conscious, for the second self. In the second condition, the individual has many conscious thoughts about the desires, feelings, and experiences that occur in the first condition. It may even be that these conscious thoughts rely on no inference of which the individual is then aware. Still, the individual assigns all the states those thoughts are about to somebody identified as distinct from the individual that has those conscious thoughts.

There is a feature of persons so far not discussed which is often thought to be central to what it is to be a person. Persons do not simply do things; they have a sense of themselves as being in some important way free in performing many, at least, of their actions. Their actions seem, from a first-person point of view, not to be causally determined in the way other events are. Whether or not such actions actually have causal antecedents that determine their occurrence, we experience many of the actions we perform as being, in some significant way, free.

Even when we experience our actions as free, however, we also experience them as resulting from conscious desires and intentions we have. But, although we are conscious of desires and intentions that seem to cause our actions, we typically are not also conscious of anything as causing those desires and intentions.28 Because we are typically conscious of our desires and intentions as being spontaneous and uncaused, we experience the resulting actions as also being free and uncaused. Sometimes we are conscious of a desire or intention as resulting from other, earlier mental states; when we consciously deliberate, for example, we are aware of our desire as being due to that process of deliberation. But the conscious chain does not continue indefinitely. There is always an antecedent intentional state we are conscious of but for which we are not conscious of any cause, and we will accordingly be conscious of it as being spontaneous and uncaused.

None of this shows that any of our desires and intentions are actually uncaused. We are conscious of only relatively few of our mental states; so there will always be some mental antecedent of which we fail to be conscious. Still, the result is that we are always conscious of our desires and intentions and, indeed, our intentional states generally as being up to us. If we experience them as being caused at all, we experience them as resulting from a causal sequence of intentional states whose initial member we are conscious of as being uncaused. The sense we have of free agency results from the way we are conscious of our conscious desires and intentions. Here, again, what is distinctive of persons is a matter of the way they are conscious of their mental lives.

# NOTES

1 *Oeuvres de Descartes*, ed. Charles Adam and Paul Tannery, Paris: J. Vrin, 1964-75, vol. VII, 71-90, esp. pp. 78-86; see *The Philosophical Writings of Descartes*, by René Descartes, translated by John Cottingham, Robert Stoothoff, Dugald Murdoch, and Anthony Kenny, 3 volumes, Cambridge: Cambridge University Press, 1984-91, vol. 2, which includes page numbers for Adam and Tannery. Quotations are from this translation.

2 E.g., Marjorie Grene, "Knowledge, Belief, and Perception," in Grene, *A Philosophical Testament*, Chicago: Open Court, 1995, 9-27, p. 14, and "People and Other Animals," in Grene, *The Understanding of Nature: Essays in the Philosophy Of Biology*, Dordrecht: D. Reidel Publishing Company, 1974, 346-360.

3 P. F. Strawson, "Persons," in *Minnesota Studies in the Philosophy of Science* II, ed. Herbert Feigl, Michael Scriven, and Grover Maxwell, Minneapolis: University of Minnesota Press, 1958, pp. 330-353; reprinted with slight changes as ch. 3 of P. F. Strawson, *Individuals: An Essay in Descriptive Metaphysics*, London: Methuen & Co., Ltd., 1959, 87-116; and "Self, Mind and Body," *Common Factor* IV (Autumn 1966): 5-13, reprinted in David M. Rosenthal, *The Nature of Mind*, New York: Oxford University Press, 1991, 58-62.

4 "People and Other Animals," 347.

5 For early work on masked priming, see Anthony J. Marcel, "Conscious and Unconscious Perception: Experiments on Visual Masking and Word Recognition," *Cognitive Psychology* 15 (1983): 197-237, and "Conscious and Unconscious Perception: An Approach to the Relations between Phenomenal Experience and Perceptual Processes," *Cognitive Psychology* 15 (1983): 238-300.

6 David Lewis, "An Argument for the Identity Thesis," *The Journal of Philosophy* LXIII, 1 (January 6, 1966): 17-25; reprinted with additions in Lewis, *Philosophical Papers*, vol. I, New York: Oxford University Press, 1983, pp. 99-107; "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy* L, 3 (December 1972): 249-258; "Mad Pain and Martian Pain," in *Readings in Philosophy of Psychology*, ed. Ned Block, vol. I, Cambridge, Massachusetts: Harvard University Press, 1980, pp. 216-222; reprinted in Lewis, *Philosophical Papers*, vol. I, pp. 122-130, with postscript, "Knowing What It's Like," pp. 130-132.

Lewis's functionalist approach has been usefully developed in a number of writings by Sydney Shoemaker; see especially "Functionalism and Qualia," *Philosophical Studies* XXVII, 5 (May 1975): 292-315; reprinted, slightly revised, in *Readings in the Philosophy of Psychology*, vol. I, ed. Ned Block, Cambridge, Massachusetts: Harvard University Press, 1980, pp. 251-267; and other articles in Shoemaker, *Identity, Cause, and Mind: Philosophical Essays*, Cambridge: Cambridge University Press, 1984, and in Shoemaker, *The First-Person Perspective and Other Essays*, Cambridge: Cambridge University Press, 1996.

7 For a sample, see Tyler Burge, "Individualism and the Men-

tal," *Midwest Studies in Philosophy* IV (1979): 73-121; "Individualism and Psychology," *The Philosophical Review* XCV, 1 (January 1986): 3-45; Fred Dretske, *Explaining Behavior: Reasons in a World of Causes*, Cambridge, Massachusetts: MIT Press/Bradford Books, 1988; and Jerry A. Fodor, "A Theory of Content," in Fodor, *A Theory of Content and Other Essays*, Cambridge, Massachusetts: MIT Press/Bradford Books, 1990.

8 Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy* LXVIII, 1 (January 14, 1971): 5-20.

9 In Kant's useful phrase: *Critique of Pure Reason*, transl. and ed. Paul Guyer and Allen W. Wood, Cambridge: Cambridge University Press, 1998, p. 174, A22/B37. Locke uses the phrase "internal Sense" (*An Essay Concerning Human Understanding*, edited from the fourth [1700] edition by Peter H. Nidditch, Oxford: Oxford University Press, 1975, II, i, 4, p. 105. See also D. M. Armstrong, "What is Consciousness?", in Armstrong, *The Nature of Mind*, St. Lucia, Queensland: University of Queensland Press, 1980, 55-67, p. 61; and William G. Lycan, *Consciousness and Experience*, Cambridge, Massachusetts: MIT Press/Bradford Books, 1996, ch. 2, pp. 13-43.

10 I have developed this HOT model in "Two Concepts of Consciousness," *Philosophical Studies* 49, 3 (May 1986): 329-359; "The Independence of Consciousness and Sensory Quality," in *Consciousness: Philosophical Issues, 1, Consciousness,* 1991, ed. Enrique Villanueva, Atascadero, California: Ridgeview Publishing Company, 1991, pp. 15-36; "Thinking that One Thinks," in Martin Davies and Glyn W. Humphreys, eds., *Consciousness: Psychological and Philosophical Essays,* Oxford: Basil Blackwell, 1993 197-223; "A Theory of Consciousness," in *The Nature of Consciousness: Philosophical Debates,* ed. Ned Block, Owen Flanagan, and Güven Güzeldere, Cambridge, Massachusetts: MIT Press, 1997, pp. 729-753; and "Consciousness and its Expression," *Midwest Studies in Philosophy* XXII (1998): 294-309.

11 We need not include the provision about observation, since an observation would figure in one's being aware of a mental state one is in only if one inferred one's being in that state from the observation. 12 For more on introspective consciousness, see my "Introspection and Self-Interpretation," *Philosophical Topics* (Winter 2001), forthcoming.

13 I argued earlier that sensing always involves mental qualities and no higher-order qualities occur when we are conscious of our conscious states. So we are not aware of our conscious states by inner sense. An advocate of inner sense could reply that, since we are not in ordinary cases conscious of how we are aware of our conscious states, higher-order qualities could occur without our knowing it. But in introspecting we do become aware of the way we are conscious of our conscious states, and even then we are aware of no higher-order mental qualities.

14 Barring refinement about individuals' switching worlds, and the like. See Hilary Putnam, "The Meaning of 'Meaning'," *Minnesota Studies in the Philosophy of Science* VII, ed. Keith Gunderson, Minneapolis: University of Minnesota Press, 1975, pp. 131-193, and "Meaning and Reference," *The Journal of Philosophy* LXX, 19 (November 8, 1973): 699-711.

15 This type of worry has been raised independently of the HOT model, as a problem for reliability of first-person access to our intentional states on a externalist theory of content. See Donald Davidson, "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* LX, 3 (January 1987): 441-458, and Tyler Burge, "Individualism and Self-Knowledge," *The Journal of Philosophy* LXXXV, 11 (November 1988): 649-663; both reprinted in *Self-knowledge*, ed. Quassim Cassam, Oxford: Oxford University Press, 1994, pp. 43-64 and 65-79, respectively.

16 Davidson and Robert Stalnaker, "On What's in the Head," *Philosophical Perspectives: Philosophy of Mind* III (1989): 287-316, suggest in effect that intentional content in general is often neutral in this way. The present suggestion is that, whatever is so about the intentional content of thoughts, in general, such neutrality typical holds for the intentional content of HOTs.

17 *A Treatise of Human Nature* [1939], ed. L. A. Selby-Bigge, Oxford: Clarendon Press, 1888, I, IV, ii, p. 207. Cf. Appendix, p. 634. For the famous "bundle" statement, see I, IV, vi, p. 252.

18 I am grateful to Sydney Shoemaker for pressing this question, in a paper entitled "Consciousness and Co-consciousness," delivered at the Fourth Annual Meeting of the Association for the Scientific Study of Consciousness, Brussels, July 2000.

19 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, p. 636).

20 For more on the apparent unity of consciousness, see "Introspection and Self-Interpretation," sec. VII.

21 See, e.g., my "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, sec. 5; "Consciousness, Content, and Metacognitive Judgments," *Consciousness and Cognition*, 9, 2, Part 1 (June 2000): 203-214, sec. 5; and "Metacognition and Higher-Order Thoughts," *Consciousness and Cognition*, 9, 2, Part 1 (June 2000): 231-242, sec. 4.

22 Shoemaker, Sydney, "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, p. 8. See also Gareth Evans, "Demonstrative Identification," in Evans, *Varieties of Reference*, ed. John McDowell, Oxford: Clarendon Press, 1982, pp. 142-204, esp. 179-191.

23 In "Two Concepts of Consciousness," I wrongly suggested that we could construe the very content of HOTs along these lines, as being thoughts that whoever thinks this very thought is also in the target mental state. As Thomas Natsoulas pointed out, such content seems to imply that HOTs would always be conscious, since every HOT would be, in part, about itself ("What is Wrong with the Appendage Theory of Consciousness," *Philosophical Psychology* VI, 2 [1993]: 137-154, p. 23, and "An Examination of Four Objections to Self-Intimating States of Consciousness," *The Journal of Mind and Behavior* X, 1 [Winter 1989]: 63-116, pp. 70-72). But HOTs can still represent their targets as belonging to the same individual that thinks the HOT, since the mental word 'I', in a HOT, would refer to whatever individual thinks that very thought.

24 John Perry, "The Problem of the Essential Indexical," *No–s* XIII, 1 [March 1979]: 3-21. See also Roderick M. Chisholm, *The First Person*, Minneapolis: University of Minnesota Press, 1981, chs. 3 and 4; and David Lewis, "Attitudes *De Dicto* and *De Se*," *The Philosophical Review* LXXXVIII, 4 [October 1979]: 513-543.

25 For an argument that this 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, sec. iii.

26 A satisfactory theory must explain the intuitions we have that pertain to the self, but may well not credit them all as being accurate. For a probing discussion of many aspects of the sense we have of the self, see Galen Strawson, "The Self," *Journal of Consciousness Studies* 4, 5/6 (1997): 405-28.

27 More recently also called Dissociative Identity Disorder. The theoretical issues behind the terminology are irrelevant for present purposes.

28 As always, it is crucial to distinguish the mental state one is conscious of from the event of being conscious of it, in this case, to distinguish the event of deciding from our consciousness of that event. Robust experimental findings confirm that our subjective awareness of decisions to perform basic actions occurs measurably later than the events of deciding of which we are con-

scious. See Benjamin Libet, Curtis A. Gleason, Elwood W. Wright, and Dennis K. Pearl (1983), "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness Potential)," Brain 106. Part III (September): 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 Musseler, Amsterdam: Elsevier, 1999), pp. 215-231; and Patrick Haggard and Martin Eimer, "On the Relation between Brain Potentials and Awareness of Voluntary Movements," Experimental Brain Research, 126, 1 (1999): 128-133.