Chapter 1. The Self-Representational Theory of Consciousness

1. A Theory of Consciousness at First Pass

This book presents a theory of consciousness. This very first section offers a first pass at the theory, the rest of this opening chapter a second pass, and the remainder of the book the third pass.

When I have a conscious experience of the blue sky, there is something it is like for me to have the experience. In particular, there is a bluish way it is like for me to have it. This "bluish way it is like for me" constitutes the phenomenal character of my experience. Phenomenal character is the property that makes a phenomenally conscious state (i) the phenomenally conscious state it is and (ii) a phenomenally conscious state at all.

The bluish way it is like for me has two distinguishable components: (i) the *bluish* component and (ii) the *for-me* component. I call the former *qualitative character* and the latter *subjective character*. To a first approximation, phenomenal character is just the compresence of qualitative character and subjective character. To a second approximation, there is a more specific division of conceptual labor between qualitative and subjective character: a phenomenally conscious state's qualitative character is what makes it the phenomenally conscious state it is, while its subjective character is what makes it a phenomenally conscious state at all. Thus, my conscious experience of the blue sky is the conscious experience it is in virtue of its bluishness, but it is a conscious experience at all in virtue of its for-me-ness.

A full theory of phenomenal consciousness would include accounts of both qualitative and subjective character. But the account of subjective character is of greater philosophical significance. For subjective character is what makes something a phenomenally conscious state *at all*, and the mystery of consciousness is in the first instance the problem of understanding why there is phenomenal consciousness *at all*.

The theory of consciousness to be presented here revolves around the idea that phenomenally conscious states have qualitative character in virtue of representing

environmental features and subjective character in virtue of representing *themselves*. More specifically, the qualitative character of a conscious state consists in its representing certain response-dependent properties of external objects; its subjective character consists in its representing itself in a suitable way. Thus, my experience of the blue sky represents both the sky and itself. It has its qualitative character (bluishness) in virtue of representing the sky's property of being disposed to elicit a certain appropriate response in appropriate respondents; it has its subjective character (for-me-ness) in virtue of representing itself in an appropriate way.

It follows that what makes something a phenomenally conscious state (at all) is suitable self-representation. The reasoning is this: a phenomenally conscious state's subjective character is what makes it a phenomenally conscious state (at all); such a state has subjective character in virtue of suitably representing itself; therefore, it is a phenomenally conscious state (at all) in virtue of suitably representing itself. What accounts for there being phenomenal consciousness at all, then, is the fact that some mental states self-represent in the right way. I call this the *self-representational theory of consciousness*, or *self-representationalism* for short.

2. The Concept of Phenomenal Consciousness

It is often felt that different theories of consciousness have tended to target different phenomena under the heading of "phenomenal consciousness." To some, this suggests that many disputes among theories of consciousness are purely verbal, resting on different senses of the terms in play. On this "pluralist" view of things, genuine progress would require, in the first instance, careful distinguishing and systematic labeling of the plurality of phenomena that go by the name "consciousness."[1]

I want to agree that different theories of consciousness have targeted different phenomena, yet insist that the pluralist attitude is somewhat misleading. Thus, it is natural to hold that so-called representationalists about consciousness have effectively offered a theory of something like what I have called qualitative character, while higher-order theorists have targeted something like what I have called subjective character.[2] At the same time, although at one level philosophers may have chased different phenomena, at another there is only one phenomenon that preoccupies them, namely, whatever property generates in them the sense of mystery as they

contemplate the nature of consciousness and its place in the world. The term "phenomenal consciousness" has come into wide usage as a tag for that property.

With this in mind, I propose the following terminological regime. The term "consciousness" is of course an ordinary, everyday term to be understood intuitively. We can then use this intuitive understanding in fixing the reference of the technical term "*phenomenal* consciousness" with the following rigidified definite description: "the property F, such that, in the actual world, F is responsible for the mystery of consciousness." This description can be precisified in a number of ways. For example, suppose we read "responsible" causally, and construed "mystery" in terms of the sense that the facts of consciousness are not deducible from the physical facts. Then we could offer the following precisification: "phenomenal consciousness is the property F, such that, in the actual world, F causally produces (in the suitably reflective subject, say) the sense that the facts of consciousness cannot be deduced from physical facts." Other precisifications will attend different understandings of some of the elements in the rigidified definite description we have started with. But all will construe *phenomenal* consciousness as whatever property is the source of the mystery surrounding consciousness.

My contention is that this is the most accurate and most profitable way to fix our ideas with respect to the notion of *phenomenal* consciousness. It is the most accurate because it captures the reason the notion has come into wide usage in philosophical circles, and it is the most profitable because it allows us to extract the substantive core in many debates on consciousness that seem to be an unclear mix of the substantive and the verbal.

For example, when representationalists about consciousness debate higher-order theorists, there is a feeling that at one level they are concerned with different properties but also that at another level there is a genuine disagreement between them. This feeling is vindicated and explicated by the characterization of phenomenal consciousness proposed here. Although representationalists and higher-order theorists may offer accounts of different properties, in that the former target qualitative character and the latter subjective character, there is also a substantive disagreement between them, namely, over which of these properties is responsible for the mystery of consciousness. Representationalists seem wedded to the idea that qualitative character is the property which, in the actual world, is the source of mystery. Higher-order theorists hold that subjective character is that property.

The notion of phenomenal consciousness is sometimes introduced by reference to the "what it is like" locution. While there is something about that locution that resonates, it is probably unlikely to be of much help to the uninitiated, and does not illuminate the function served by the term "*phenomenal* consciousness" in the relevant discourse. Most importantly, it is of no help in extracting the substantive core of partially confused disputes, since typically disputants disagree on what kinds of state have something it is like to be in. Thus, higher-order theorists typically maintain that there is nothing it is like to be in a mental state with a qualitative character but no subjective character, whereas representationalists hold that there is. Intuitions regarding the appropriate applicability of what-it-is-like talk seem to lie downstream of philosophical theorizing at this level of nuance. My contention is that the notion of phenomenal consciousness is best introduced in terms of the aforementioned rigidified definite description, or more generally in terms of whatever property is responsible for the sense of mystery that surrounds consciousness.

Note well: the term "mystery" in the definite description should be taken to denote *prima facie* mystery rather than *ultima facie* mystery. In speaking of the mystery of consciousness, I do not mean to imply that consciousness is an *ultimate* mystery, i.e., one that will not succumb to eventual demystification.[3] I *do* mean to suggest, however, that it is a *genuine* mystery, that is, that our sense of mystery upon contemplating the nature of consciousness is rational and appropriate and is not based on sheer confusion.

The existence of this genuine *prima facie* mystery can be appreciated by dualists and physicalists alike. David Chalmers, a dualist, uses what he calls the "hard problem" of consciousness, in contrast with the "easy problems" (Chalmers 1995a, 1995b, and 1996), to bring it out. Joseph Levine, a physicalist, uses the label "explanatory gap" to denote an associated problem (Levine 1983). I will discuss the nature of the mystery, and the relationship between these labels for it, in Chapter 8. For now, I only wish to stress that, regardless of wider motivation, many philosophers have a sense that consciousness involves a certain mystery not found in other aspects of nature, and use the term "phenomenal consciousness" to refer to whatever produces this mystery.

The kernel of truth in the pluralist attitude is that, in the theory of consciousness more than elsewhere, different explanatory edifices are often matched with different conceptions of the phenomenon in need of explanation. But it does not

follow, and is a mistake to suppose, that all conceptions of the phenomenon in need of explanation are equally good. On the contrary, some misidentify the source of the mystery of consciousness, while others identify it correctly. Thus the theory of consciousness is essentially a package deal comprising an account of the explanans alongside an account of the explanandum.[4] Both accounts are supposed to *get things right*. The account of the explanandum is supposed to get things right by identifying correctly the source of the *prima facie* mystery surrounding consciousness; the account of the explanans is supposed to get right what ultimately demystifies that source. On this view of things, unlike on the pluralist approach, homing in on the explanandum is not just a matter of judicious stipulation, but involves substantive claims. Correlatively, debates over the nature of consciousness are rarely if ever purely verbal. When we are tempted to claim they are, it is because we forget that non-verbal disputes in the theory of consciousness can concern not only the nature of the explanans but also the nature of the explanandum, that is, can concern which property is responsible for the mystery of consciousness.

If we adopt the characterization of phenomenal consciousness in terms of the aforementioned rigidified definite description, it will probably fall out that there is only one property denoted by "phenomenal consciousness." This is to be distinguished from (a) the view that there is *more* than one such property and (b) the view that there is *less* than one (i.e., no) such property.

For there to be *less* than one property denoted by "phenomenal consciousness," there would have to be no single property systematically responsible for our sense of mystery. The sense of mystery would have to be elicited on different occasions by unrelated properties or be altogether grounded in a web of confusions. While this is a coherent position, it is not antecedently plausible: a very compelling argument would have to be adduced if we are to take it seriously. And in one sense, even that scenario would not show that there is less than one property denoted by "phenomenal consciousness," only that the one property is *disjunctive*.

For there to be *more* than one property denoted by "phenomenal consciousness," there would have to be multiple properties each of which is independently responsible for the sense of mystery that descends upon us when we contemplate the way consciousness fit into the natural world. This, again, is antecedently implausible. It is perhaps plausible that multiple properties *conspire* to generate this sense of mystery. But this would not mean that "phenomenal

consciousness" has several denotata, only that it has a *conjunctive* denotatum. If, for instance, qualitative and subjective character have to conspire to mystify us, then the denotatum of "phenomenal consciousness" is the conjunctive property of being both qualitative and subjective. What is required for "phenomenal consciousness" to have several denotata (e.g., to be ambiguous as between qualitative character and subjective character) is that several distinct properties suffice individually to generate the mystery in an overdetermining fashion (e.g., that qualitative and subjective character mystify us severally and independently).[5] This is not altogether impossible, but nor is it *prima facie* plausible. And it is only under those conditions that the pluralist attitude portrayed at the opening would be well-founded.

The main point is that statements about what needs to be explained in the theory of consciousness are better thought of as substantive claims about what generates the mystery of consciousness than as bare stipulations. Taking them to be stipulations only fudges over the source of the philosophic anxiety surrounding consciousness. There could be endless accounts of different stipulated notions of consciousness, but the only ones that would speak to that philosophic anxiety would be those that target whatever it is that generates the mystery of consciousness in the first place.

3. The Structure of Phenomenal Character

Turning, then, to the nature of the explanandum, consider my conscious experience as I look out my window at the blue sky. My experience has many properties. The great majority are completely uninteresting: it is happening on a Tuesday, it is occurring 648 years after the death of the tallest man of the fourteenth century, etc. My experience *has* those properties, but we are uninterested in them, partly because we already understand them perfectly well.

Some of my experience's properties are more interesting. Thus, the experience has a strong impact on working memory; it implicates neural activation in V4; it carries information about the sky.[6] These broadly psychological properties are more interesting, partly because our understanding of them is presently imperfect. At the same time, we have a good sense of how a comprehensive explanation of these properties would – will – go. We do not stand mystified by them.

Getting closer to something that may genuinely mystify us, there is the fact that there is something it is like for me to have my experience.[7] In particular, there is a bluish way it is like for me to have it. As I said at the opening, we can distinguish two *aspects*, or *components*, of this "bluish way it is like for me." There is, on the one hand, the *bluish* component, which I call the experience's *qualitative character*, and on the other hand, the *for-me* component, which I call the experience's *subjective character*.[8]

When a colorblind person looks at the same sky under the same conditions, there is a psychologically real difference between her experience and mine. The psychologically real dimension along which our experiences differ is what I call qualitative character. We need not commit at this stage to any particular view of qualitative character. Without pretending that this provides an analysis, let us only say the following: qualitative properties "correspond" to sensible features in the environment.[9] There is a potential qualitative property of a given creature's conscious experience for every property in the creature's environment that is sensible by that creature. An experience's qualitative character at a time may be thought of as the set, or sum, of all qualitative properties the experience instantiates at that time.

As for subjective character, to say that my experience has subjective character is to point to a certain *awareness* I have of my experience. Conscious experiences are not states which we may *host*, as it were, unawares. Freudian suppressed states, sub-personal states, and a variety of other unconscious states may occur within us completely unbeknownst to us, but the intuition is that conscious experiences are different. A mental state of which one is completely unaware is not a conscious experience. In this sense, my conscious experience is not only *in me*, it is also *for me*.[10]

One reason it is useful to draw the distinction between qualitative and subjective character is that, as already noted, philosophical theories of consciousness have tended to fall into two groups, one targeting qualitative character, the other targeting subjective character.[11]

There are several possible views about the interrelations among phenomenal, qualitative, and subjective character. This matter will be discussed in greater detail in Chapter 2. For now, we can just say that four central positions stand out. One view is that phenomenal character is identical with qualitative character; this seems to be the view, e.g., of typical representationalists. Another view is that it is identical with

subjective character; this seems to be the view, e.g., of higher-order theorists. Yet another view is that phenomenal character is identical with something like the compresence, or conjunction, of qualitative and subjective character. Finally, there is the view that phenomenal consciousness bears no constitutive relations to either qualitative or subjective character.

There are two views I am attracted to. The first is less plausible but very clear – that phenomenal character is identical to subjective character (i.e., that subjective character is the source of the mystery of consciousness). The second is more plausible but also more complicated – that phenomenal character is some kind of complex property involving both subjective and qualitative character. It is hard to argue about such matters, though I will discuss them in more detail in Chapter 2 and the Appendix. For now, and merely by way of gesturing at the sensibility behind the views I am attracted to, consider the following.

Regarding the first view, arguably, the (genuine) puzzlement over the fact that a bunch of neurons vibrating inside the skull are associated with a yellowish qualitative character is no different from that surrounding the fact that a bunch of atoms lurching in the void are associated with a yellow color. It is no more surprising that neurons can underlie yellowishness than that atoms can underlie yellowness.[12] And yet the latter is not deeply mystifying. By contrast, the mystery surrounding the fact that a bunch of neurons vibrating inside the skull are associated with a for-me-ness has no parallel in the non-mental realm. It is surprising that neurons can underlie for-me-ness, and there is no analogous surprise regarding what atoms can underlie. This phenomenon is thus more suitable to capture the distinctive mystery presented by consciousness.[13]

As for the second view, I am attracted to it by considering that when we say that a mental state is phenomenally conscious, there must be (a) something that makes it the phenomenally conscious state *it is* and (b) something that makes it a phenomenally conscious state *at all*. For the reason adduced in the previous paragraph and other reasons, I hold that subjective character is what makes a mental state phenomenally conscious *at all*. But this leaves somewhat unclear what makes it the phenomenally conscious state *it is*, once it is phenomenally conscious at all. One natural thought is that it is the state's qualitative character that makes it the phenomenally conscious state it is. The upshot is the view that qualitative character provides the identity conditions of phenomenality while subjective character provides

its existence conditions: what makes a mental state phenomenally conscious at all (rather than a non-phenomenal state) is its subjective character, what makes it the phenomenally conscious it is (rather than another) is its qualitative character.

According to this Division of Conceptual Labor thesis, my experience's qualitative character (its bluishness) determines *what* phenomenally conscious state it is, whereas its subjective character (its for-me-ness) determines *that* it is a phenomenally conscious state. Thus for a mental state to *become* phenomenally conscious, it must acquire a subjective character; once it *has*, so to speak, become phenomenally conscious, it is its qualitative character that makes it the specific type of phenomenally conscious state it is.[14]

One odd feature of this position, as stated, is the complete divorce of identity and existence conditions. Typically, what makes something the F it is and what makes it an F at all are not completely unrelated, but are rather are rather related as determinate to determinable. What makes something a car at all is that it is an auto-mobile machine (plus bells and whistles), whereas what makes it the car it is is its being an auto-mobile machine *of a certain make*. Being an auto-mobile machine of a certain make is a determinate of being an auto-mobile machine. Likewise, if the bluish qualitative character of my experience is what makes it the phenomenally conscious state it is, then we should expect that what makes it a phenomenally conscious state at all is that it is has *some* qualitative character. Here subjective character falls out of the picture altogether – implausibly.

The solution is to think of the key feature of my conscious experience as *bluish-for-me-ness* (rather than as a feature factorizable into two components, bluishness and for-me-ness). In a way, this is to hold that subjective and qualitative character are not separate properties, but in some (admittedly problematic) sense are aspects of a single property.[15] Attractively, bluish-for-me-ness is a determinate of for-me-ness (or "something-for-me-ness," if you will). This generates the desired result that the existence and identity conditions of phenomenality are related as determinable to determinate.

This is the position I favor on the structure of phenomenal consciousness, and I will develop it more in Chapter 2. It casts phenomenal character as a matter of a complex compresence of qualitative and subjective character, which are not entirely separable. Accordingly, a theory of phenomenal consciousness must address each. Chapter 3 develops an account of qualitative character and Chapter 4 one of

subjective character. The following chapters focus on aspects of subjective character. The reason is that in the picture we ended up with, there is still a more central place for subjective character, since it constitutes the existence condition of phenomenality. This is important because although the deep mystery of consciousness does have to do with why and how conscious episodes differ from each other, it is much more concerned with why and how there are conscious episodes *to begin with*, that is, why there exists such a thing as consciousness *at all*.[16]

4. A Narrow Representational Account of Qualitative Character

The purpose of a reductive theory of phenomenal consciousness is to account for phenomenal character in non-phenomenal terms. If the non-phenomenal terms employed in the account are purely physical, the result is a *physicalist* theory of consciousness. But it is not definitional of a reductive theory that it is physicalist. What is definitional is that it accounts for the phenomenal in non-phenomenal terms.

The primary goal of this book is to provide a reductive and hopefully physicalist account of phenomenal character. A reductive account of phenomenal character would comprise two parts: a reductive account of what makes a phenomenally conscious state the phenomenally conscious state it is, and an account of what makes a phenomenally conscious state a phenomenally conscious state at all. Given the Division of Conceptual Labor thesis from the previous section, the first part would be constituted by a reductive account of qualitative character, the second by a reductive account of subjective character.

In Chapter 3, I defend a representationalist of qualitative character: a conscious experience's qualitative character is constituted by (an aspect of) its representational content. On such a view, whatever else my experience of the blue sky represents, it also represents features of the sky. Naturally, the sky has many different features, and accordingly my experience instantiates many different representational properties in relation to it. On the account I will defend, only a very specific subset of these representational properties are constitutive of its qualitative character. In particular, only the representation of certain response-dependent properties of the sky can be constitutive of my experience's qualitative character.

How are we to understand the relevant response-dependent properties? Answering this question is one way of fleshing out one's specific version of *response-*

dependent representationalism (as I will call the view). Much of Chapter 3 will be dedicated to homing in on the relevant response-dependent properties. In any event, one result of this kind of response-dependent representationalism is that it casts the representational content constitutive of qualitative character as *narrow content*.[17] Conscious experiences may well have wide representational properties, but those are not constitutive of their qualitative character.

My main concern in the book, however, will be the reductive account of subjective character. There are two kinds of reason for this focus. The insubstantial reasons are (a) that the literature strikes me as much more advanced when it comes qualitative character than to subjective character, (b) that a narrow representationalist view very similar to the one I will defend has already been defended in the literature (e.g., in Shoemaker 1994a), and (c) that the newer ideas I have to offer pertain mostly to subjective character. The substantial reasons have to do with the fact that subjective character is, on the view presented here, what makes a phenomenally conscious state such a state *at all*, and that this is the deeper nexus of the philosophical mystery surrounding consciousness. That is, the existence (as opposed to identity) condition of phenomenality is the more pressing aspect of consciousness to account for, and subjective character is what needs to be understood in order to do so.

5. A Self-Representational Account of Subjective Character

The central thesis of the book is that what makes something a phenomenally conscious state at all, what constitutes its subjective character, is a certain kind of *self-representation*: a mental state has phenomenal character at all when, and only when, it represents itself in the right way. All and only phenomenally conscious states are suitably self-representing. Thus, whatever else a conscious state represents, it always also represents itself, and it is in virtue of representing itself that it *is* a conscious state. This is self-representationalism.

Although out of the limelight in modern discussions of consciousness, self-representationalism has quite a venerable history behind it. One of its early clear and explicit endorsements is by Franz Brentano, who, in his *Psychology from an Empirical Standpoint*, wrote this (1874:153-4):[18]

> [Every conscious act] includes within it a consciousness of itself. Therefore, every [conscious] act, no matter how simple, has a double object, a primary and a secondary object. The simplest act, for example the act of hearing, has as its primary object the sound, and for its secondary object, itself, the mental phenomenon in which the sound is heard.

Thus, the auditory conscious experience of a bagpipe sound is intentionally directed both at the bagpipe and at itself. Against the background of a representational conception of intentionality, this would commit one to the thesis that all conscious states are self-representing.[19] Brentano may have adopted this view from Aristotle. Although Aristotle is not nearly as explicit as Brentano, he does write in the *Metaphysics* that conscious "knowing, perceiving, believing, and thinking are always of something else, but of themselves on the side."[20]

But what does it mean for a mental state to represent itself? Answering this question is one way of fleshing out self-representationalism. Different answers will result in different versions of the view. I will offer my own take in Chapters 4 and 6. In any event, pending a compelling argument to the contrary, there is no reason to suspect that there is something deeply unintelligible about the notion of a self-representing state.

Familiar forms of self-representation occur in linguistic expressions.[21] The token sentence "This very sentence is written in Times New Roman" is self-representing. As it happens, it is a true sentence, and it itself is a constituent of its truthmaker.[22] On the plausible suppositions that in order to be true, a sentence must make semantic contact with its truthmaker, and that semantic contact is achieved through, perhaps even constituted by, representation of the truthmaker's constituents, this true sentence must represent its truthmaker's constituents, which include it itself.[23,24] I am not citing this form of sentential self-reference as an accurate model of conscious-making self-representation. I am adducing it merely by way of addressing a potential worry to the effect that the very notion of self-representation is somehow deeply unintelligible

It is useful to see self-representationalism as the upshot of two claims. The first is that all phenomenally conscious states are conscious in virtue of being *represented* (in the right way). The second is that no phenomenally conscious state is conscious in virtue of being represented by a *numerically distinct* state, that is, a state other than itself. It follows from these two claims that all phenomenally conscious

states are phenomenally conscious in virtue of being represented (in the right way) by a mental state that is not numerically distinct from themselves, that is, in virtue of self-representing (in the right way).

More formally, we might state the master argument as follows. For any phenomenally conscious state C,

1) C is conscious in virtue of being suitably represented;
2) It is not the case that C is conscious in virtue of being represented by a numerically distinct state; therefore,
3) C is conscious in virtue of being suitably represented by itself; that is,
4) C is conscious in virtue of suitably representing itself.[25]

The argument is clearly valid, so the two premises do entail self-representationalism. The question is whether they are true. In the next two sections, I sketch the main motivation for each.

6. Phenomenal Consciousness and Inner Awareness

The first premise in the master argument for self-representationalism is the claim that a conscious state is conscious in virtue of being suitably represented. Impressionistically put, the general motivation for this premise is the thought that it is somehow essential to a conscious state that its subject be aware of it. Conscious states are not states that just happen to take place *in* us, whether or not we are aware of their taking place; they are also *for* us, precisely in the sense that there is something it is like *for* us to have those states. Mental states that merely occur *in* us, but of which we are completely unaware, are not conscious experiences.[26]

Let us call awareness of external features and objects in one's environment or body *outer awareness*, and awareness of internal events and states in one's own mental life *inner awareness*. The thought under consideration is that inner awareness is somehow essential to phenomenal consciousness. For starters, a mental state is phenomenally conscious only if its subject has inner awareness of it. But more strongly, the *right kind* of inner awareness is not only a necessary condition but also a sufficient condition for phenomenality. It is this inner awareness that ultimately makes the mental state phenomenally conscious at all. For it is when the subject has

this inner awareness of it that the state acquires subjective character, and subjective character is what makes a state phenomenally conscious at all.

It may be thought doubtful that one is always aware of one's concurrent conscious experiences. As I have my conscious experience of the blue sky, I am not attending to myself and my experience in the least. The focus of my awareness is on the blue sky, not my experience of it. However, it does not follow from the fact that my experience is not in the focus of my awareness that it is not in my awareness at all. While the focus of our awareness is most manifest to us, there are a host of objects and features that routinely lie in the *periphery* of awareness. Thus, as I stare at the laptop before me, the laptop occupies the focus of my visual awareness. But at the periphery of my visual awareness are a number of objects lying on the far edge of my desk: a pen, a coffee mug, a copy of the *Tractatus*.[27] As I will argue in Chapter 5, the focus/periphery distinction applies not only to visual awareness, but to all awareness, including inner awareness. My claim is that in the case of my experience of the sky, although I am not focally aware of the experience itself, but rather of the sky, I am nonetheless peripherally aware of the experience itself. That is, the experience combines focal outer awareness of the sky with peripheral inner awareness of itself.

Plausibly, being aware of something is just representing it in the right way. For a subject to be aware of a rainbow *just is* for her to harbor the right sort of mental representation of the rainbow. This representational treatment of awareness extends to inner awareness as well. Thus to be aware of, say, a growing anxiety *just is* to harbor a mental representation of one's anxiety. From the representational treatment of awareness and the thesis that inner awareness is essential to phenomenal consciousness, it follows that conscious states are essentially states of which their subject has a mental representation.

In summary, the motivation behind the claim that conscious states are conscious in virtue of being represented in the right way derives from the thought that conscious states are states we are aware of, albeit peripherally, and that such awareness is a form of representation. We may thus present the following sub-argument for the first premise of the master argument. For any conscious state C of a subject S,

1) C is conscious in virtue of S's being suitably aware of C;

2) For S to be suitably aware of C is for C to be suitably represented by S; therefore,
3) C is conscious in virtue of being suitably represented by S; therefore,
4) C is conscious in virtue of being suitably represented.

The next phase of the master argument is the move from being represented to being self-represented.

7. Inner Awareness and Self-Representation

When a conscious state is represented, either it is represented by itself, or it is represented by a state other than itself, a numerically distinct state. To the extent that conscious states are conscious precisely in virtue of being represented in the right way, there are three possibilities: (a) all conscious states are conscious in virtue of being represented by themselves; (b) all conscious states are conscious in virtue of being represented by numerically distinct states; (c) some conscious states are conscious in virtue of being represented by themselves and some in virtue of being represented by numerically distinct states. Self-representationalism holds that (a) is true. The argument will be pursued in Chapter 4. Here let me offer only a preliminary sketch of a central part of that chapter's argumentative strategy.

One way to argue for (a) above is by elimination, that is, by arguing against (b) and (c). It is rather implausible, on the face of it, that some conscious states are conscious in virtue of being represented one way while others are conscious in virtue of being represented another way. One would expect a certain underlying unity, at a reasonable degree of abstraction, among conscious states. Indeed, if conscious states form a natural kind, as they probably do, they should boast a relatively concrete "underlying nature" common to all of them. These considerations militate against (c).

The argument against (b) proceeds by destructive dilemma. When a mental state is represented by a numerically distinct state, the numerically distinct state must be either *conscious* or *unconscious*. Let us consider the following dilemma: are all conscious states conscious in virtue of being represented by numerically distinct conscious state, or not? If they are not, then (b1) some or all conscious states are conscious in virtue of being represented by numerically distinct *unconscious* states. (This is effectively the view of higher-order theories.[28]) If they are, then (b2) all

conscious states are conscious in virtue of being represented by numerically distinct *conscious* states. The argument against (b) consists in showing that neither (b1) nor (b2) is plausible.

The case against (b2) is straightforward: (b2) leads to infinite regress. If every conscious state was necessarily represented by a numerically distinct conscious state, then the occurrence of a single conscious state would implicate an infinity of mental states. But this is doubly implausible: it fails to offer an explanatory account of what makes a conscious state conscious, and it is empirically implausible (perhaps impossible).[29]

As for (b1), the main argument against it is somewhat more delicate, and will be only outlined here.[30] In the previous section, I said that conscious states are states we are aware of. But how do we know that we are aware of our conscious states? That is, on what basis do we come to subscribe to the thesis that conscious states are states we are aware of? It seems incorrect to say that we subscribe to the thesis purely on the strength of third-person evidence. It is not as though we have consulted experimental data which pointed to the existence of such awareness. Rather, we subscribe on the strength of some kind of first-hand, first-person knowledge of our awareness of our conscious states. However, we would not have such knowledge of awareness if the awareness were grounded in unconscious representations. For our knowledge of unconscious states is always third-person knowledge. So if our awareness of our conscious states was grounded in unconscious states, our knowledge of it would be the sort of third-person knowledge it does not in fact seem to be.

To repeat, this is only a sketch of a line of argument against (b1) that will be fleshed out and defended in much greater detail in Chapter 4. Note, however, that since (b1) is in fact the view of higher-order theories, many of the arguments in the literature against higher-order theories apply. Those too will be rehearsed and/or developed in Chapter 4. Perhaps the best known of those arguments is the argument from "targetless" higher-order representations. According to (b1), what makes a conscious state conscious is that it is targeted by a numerically distinct, hence higher-order, representation. But what happens when a subject has a higher-order representation that misrepresents not only the *properties*, but also the very *existence*, of a lower-order target? Higher-order theory seems committed to saying that the subject is having no conscious experience, even though what she is going through is subjectively indistinguishable from the having of a conscious experience. This means,

quite absurdly, that there is no conscious state the subject is in, but there is something it is like for her to be in the conscious state she is not in.[31]

If the above considerations are correct, then neither (b2) nor (b1) can be made to work, which, given the non-viability of (c), entails that (a) is the only viable position: all conscious states are conscious in virtue of being represented by themselves. To recapitulate, the presented argument for this turns on two basic ideas. First, it is implausible that a conscious state is conscious in virtue of being represented by an *unconscious* state. But if it is conscious in virtue of being represented by a conscious state, the representing conscious state cannot be numerically distinct from the represented conscious state, on pain of vicious regress or disunity. It follows that the representing and represented conscious states are one and the same, that is, that conscious states are self-representing. We may thus present the following sub-argument for the second premise of the master argument. For any conscious state C,

1) It is not the case that C is conscious in virtue of being unconsciously represented; and,
2) It is not the case that C is conscious in virtue of being consciously represented by a numerically distinct state; therefore,
3) It is not the case that C is conscious in virtue of being represented by a numerically distinct state.

If we connect the argument of this section with the argument of the previous section, we obtain the following partial fleshing out of the master argument. For any conscious state C,

1) C is conscious in virtue of S's being suitably aware of C;
2) For S to be suitably aware of C is for C to be suitably represented by S; therefore,
3) C is conscious in virtue of being suitably represented by S; therefore,
4) C is conscious in virtue of being suitably represented;
5) It is not the case that C is conscious in virtue of being unconsciously represented; and,
6) It is not the case that C is conscious in virtue of being consciously represented by a numerically distinct state; therefore,

7) It is not the case that C is conscious in virtue of being represented by a numerically distinct state; therefore,
8) C is conscious in virtue of being suitably represented by itself; that is,
9) C is conscious in virtue of suitably representing itself.

The master argument can be further fleshed out, of course. The present formulation is intended mainly to give a sense of the basic pull of the self-representational approach to phenomenal consciousness.

8. Plan of the Book

The combination of narrow representationalism about qualitative character and self-representationalism about subjective character delivers a fully representational account of phenomenal character. It is an unusual representational account in many ways, but nonetheless it is true, on this account, that what makes a conscious experience the conscious experience it is, and a conscious experience at all, is the nature of its representational content. The experience's non-representational properties play no constitutive role (though they may play a causal role) in making it a conscious experience.

The book divides broadly into two parts. The first consists of Chapters 2-4, which develop and defend the account of phenomenal character sketched here. Chapter 2 develops and defends my conception of the explanandum of the theory of consciousness, Chapter 3 my specific version of a representational account of qualitative character, and Chapter 4 the self-representational account of subjective character. The second part of the book consists in Chapters 5-8, which zoom in on central aspects of subjective character – and develop the self-representational approach to them. Chapter 5 discusses the phenomenological nature of consciousness, Chapter 6 offers an ontological assay of consciousness, Chapter 7 hypothesizes on the scientific nature of consciousness, and Chapter 8 considers the prospects for a physicalist reduction of consciousness.

The express goal of the book as a whole is to make the case for the self-representational theory of consciousness. But in the end, what I can reasonably hope is to convince the reader that the theory should be taken seriously, at least on a par with extant philosophical theories of consciousness. The idea that phenomenal

consciousness is injected into the world when subjects' internal states acquire the capacity to represent themselves has more initial appeal to it, it seems to me, than what we have become accustomed to in cotemporary discussions of consciousness. The burden of this book is to show that this initial appeal does not dissipate, but on the contrary deepens, upon closer examination. In other words, the goal is to show that the self-representational theory of consciousness is not only antecedently attractive, but can be fleshed out in a theoretically compelling way – no less compelling, at any rate, than the known alternatives.[32]

---

[1] The term pluralism is used by Block (1995) to refer to his own distinction between four different phenomena which are sometimes targeted by the theory of consciousness, and which he calls phenomenal consciousness, access consciousness, monitoring consciousness, and self-consciousness.

[2] Representational theories hold that a conscious state's phenomenal character is constituted by (an aspect of) its representational content. For representational theories, see mainly Dretske 1995 and Tye 1992, 1995, 2000, and 2002. A more complex version is developed in Shoemaker 1994a, 1994c, and 2002. All modern versions are inspired to some extent by the so-called transparency of experience: the only introspectively accessible properties of conscious experiences are their representational properties (see Harman 1990). Higher-order theories hold that a conscious state's phenomenal character is constituted by its figuring in the representational content of a numerically distinct higher-order state. The most worked out version of higher-order theory is no doubt David Rosenthal's (1986, 1990, 1993, 2002a, 2005). For other versions, see mainly Armstrong 1968, Lycan 1996 and 2004, Carruthers 1998 and 2000, and Van Gulick 2001, 2004, and 2006. Most or all of these views are inspired by the so-called transitivity principle: conscious states are states we are conscious (or aware) of (see Rosenthal 1986). Interestingly, the work of some fierce critics of representational and higher-order theories betrays a commendation (implicit or explicit) of these theories for at least targeting the right phenomenon. This can be seen, for instance, in Block's (1990, 1996) criticism of representational theories and Levine's (2001) criticism of higher-order theories. In the present terminology, Block seems to hold that the phenomenon that needs to be targeted is qualitative character, whereas Levine believes it is subjective character.

[3] The view that it *is* an ultimate mystery *has* been defended, most notably by Colin McGinn (1989a, 1995, 1999, 2004).

[4] To be sure, the explanandum of all theories of consciousness is, by definition, phenomenal character, the property that makes a conscious experience (i) the conscious experience it is and (ii) a conscious experience at all. But already at the level of description, accounts differ on the nature of this property.

[5] Even under these conditions, it could be argued that phenomenal consciousness ought to be identified with a single property, namely, the *disjunctive* property such that each of its disjuncts is one of those overdetermining sources of mystery. Arguably, this view is mandated by the use of the definite article in the rigidified description.

[6] Of course, some philosophers may deny any of these claims, but it is premature at this point to address particular philosophical doctrines of the sort that would have to be involved. I list these as reasonable properties we can expect in a pre-philosophical mood conscious states to have.

[7] The locution, in its present usage, is due to Farrell (1950), but was brought into wide usage through Nagel (1974). As I said in the previous section, the expression may not be of much use in introducing the notion of phenomenal consciousness to the uninitiated. But I am here addressing myself to the initiated, and anyway am not (yet) claiming that what-it-is-like-for-me is the same as phenomenal consciousness.

[8] There is a question as to whether the qualitative and the subjective are really separable. Even if they are not separable in reality, however, they are certainly separable "in thought." That is, there is a *conceptual* distinction to draw here even if no *property* distinction corresponds. I will say more on this in Chapter 2.

[9] We may construe the environment here to include the subject's body.

[10] There is much more to be said about the nature of subjective character, or for that matter qualitative character. And more will be said in Chapter 2. On the other hand, I find that anything one says in this area is contestable, and to that extent amounts to substantive claims. For this reason, I do not wish to expand on the nature of subjective and qualitative character in the context of introducing them. I say just enough to get the reader a sense of what I am talking about.

[11] Thus, representational theories seem to target qualitative character, while higher-order theories seem to target subjective character. As I argued in the previous section, this chasm should not be thought of as a mere verbal dispute, but as a substantive dispute about the source of the mystery of consciousness.

[12] This observation is due to Shoemaker (1994a). It should be noted, however, that some philosophers have made this observation in the service of the converse claim: that the explanatory gap between consciousness and neural matter is nothing but the explanatory gap between the manifest image of external objects and their scientific image (Byrne 2006).

[13] This consideration is quick and dirty and I do not intend to lean heavily on it at this stage. It is merely a characterization of my tendencies in this area. A more systematic approach to the issues that arise here will come to the fore at the end of Chapter 2.

[14] I am using here the language of process purely metaphorically. There is no real process in which a mental state first becomes a conscious experience at all and later – after a time lag – becomes a particular type of conscious experience. Obviously, the two occur simultaneously. In fact, under most reasonable understandings of "occur", they are the one and the same occurrence.

[15] After all, even though my experience of the sky is *for me*, not every aspect of it is for me in the relevant sense. Plausibly, the only aspect of it that is for me in the relevant sense is its bluishness.

[16] The mystifying fact is primarily that there are some events in this world that seem categorically different from the brute, blind, inanimate proceedings of the unconscious realm. Thus the mystery concerns in the first instance what makes something a conscious state *in the first place* and only derivatively what makes it the conscious state it is *given that it is one*. This suggests that subjective character ought to be the central target of the theory of consciousness, in that the demystification of consciousness would require most centrally the demystification of subjective character.

[17] I am appealing here to the distinction, originally drawn by Putnam (1975), between *narrow* and *wide* content. Narrow content is content that is fully determined by the non-relational properties of the subject of the state whose content it is. Wide content is content that is *not* so determined. At least this is one way to draw the distinction. Narrow representationalism about qualitative character, then, is the view that the qualitative character of a conscious experience is constituted by aspects of the experience's content that are fully determined by the non-relational properties of the subject. I first defended a version of this view (under the name "internalist representationalism") in Kriegel 2002a.

[18] I am quoting from the 1973 English translation. The view is first introduced by Brentano in Section 7 of chapter II ("Inner Consciousness") in Book 2, which is entitled "A Presentation and the Presentation of that Presentation are Given in One and the Same Act." In this section, Brentano canvasses his conception of conscious experiences as self-representational.

[19] One could consider a representational conception of intentionality either as a tautological thesis or as a substantive but true thesis, depending on one's understanding of these terms. I have seen both in the literature. In recent philosophy of mind it is more common to treat such a conception as a tautology, but there are exceptions (e.g., Cummins 1989).

[20] 1074b35-6. I am using here Caston's (2002) translation. At least one scholar has suggested to me that "on the side" is not the best translation, and "as a byproduct" would be more appropriate. It is not clear that this makes a substantial difference for our present purposes. There are also passages in *De Anima* that can be read as endorsing self-representationalism. For a sustained self-representationalist interpretation of Aristotle, see Caston 2002. Caston notes that an interpretation of Aristotle along these lines is also to be found in a dissertation on the unity of mental life in Aristotelian philosophy, written under Brentano's supervision by one J. Herman Schell.

[21] There are also more rigorous mathematical models of self-representation, especially in nonwellfounded set theory (see Williford 2006 and Landini ms).

[22] If one prefers to think of truthmakers as states of affairs, one could say that this sentence's truthmaker is the state of affairs consisting in its instantiation of the property of being written in Times New Roman.

[23] Somewhat unmelodiously, I am using the phrase "it itself" as an indirect reflexive. I will say much more about the semantic peculiarities of indirect reflexives in Chapters 4 and 9. For seminal work on this matter, see Castañeda 1966 and 1969.

[24] Let me emphasize that.

[25] Here and in what follows, I am leaving the quantifier outside the indented presentation of the argument for stylistic reasons, but it should be clear that the presentation in the text is equivalent to the following:

1) For any conscious state C, C is conscious in virtue of being suitably represented;
2) For any conscious state C, it is not the case that C is conscious in virtue of being represented by a numerically distinct state; therefore,
3) For any conscious state C, C is conscious in virtue of being suitably represented by itself; that is,
4) For any conscious state C, C is conscious in virtue of suitably representing itself.

[26] There are a number of issues and potential objections that arise quite immediately for this line of thought. But I am offering it here not as an official argument but as a gesture toward the kind of thought that motivates the idea that conscious states are necessarily represented. The "official" argumentation will be offered in Chapter 4.

[27] These objects are presented rather vaguely and inaccurately in my experience, but they are presented nonetheless. Thus, I would never be able to make out, on the basis of my experience alone, that the book is the *Tractatus*. But it would be obviously erroneous to infer that I am therefore altogether unaware of the book. Rather, we should say that I am aware of it, not *focally* however, but *peripherally*. These issues will be discussed more fully in Chapter 5.

[28] More specifically, contemporary higher-order theories hold that *most* conscious states are conscious in virtue of being suitably represented by unconscious states, but *some* are conscious in virtue of being suitably represented by conscious states. The latter are conscious states that are being explicitly introspected. The former are all other conscious states, and obviously constitute the great majority of conscious states.

[29] As Leibniz says, if this were the case, we would never get passed the first thought. He writes: "It is impossible that we should always reflect explicitly on all our thoughts; [otherwise] the mind would reflect on each reflection *ad infinitum*… It must be that…eventually some thought is allowed to occur

without being thought about; otherwise I would dwell forever on the same thing." (Quoted in Gennaro 1999: 355-6.)

[30] I will develop it more fully in Chapter 4 below. I first floated this argument in Kriegel 2003a.

[31] Again, I will elaborate on this, and other problems for higher-order theories, in Chapter 4.

[32] For comments on a previous draft of this chapter, I would like to thank David Bourget, Jordi Fernández, Terry Horgan, Bill Lycan, Ken Williford, and an anonymous referee for OUP. I have also benefited from relevant conversations with David Chalmers, Brendan Jackson, Derk Pereboom, and Daniel Stoljar.