Editorial Note: The two papers that follow hark back in a sense to the early days of brain-body biology, when observation and logic were the available tools for constructing a theory. Prior to the development of tools and techniques for experimentation on the nervous system, correct prediction of outcome was the best available confirmation of a theory. Not surprisingly, at the time it was the mathematician-philosophers who were doing the inquiry and building the theories. Notable among them was Christian Wolff in his *Prolegomena to Empirical Psychology* and *Rational Psychology.* The former book was originally published in 1732 and the latter in 1734. Wolff's contribution was to relate the method of studying psychology to that of physics. He sought to establish laws of sensation, memory, emotion, understanding, and behavior. Johann Friederich Herbart further developed Wolff's measuring methods, and Wilhelm Wundt, who considered Wolff "the most influential psychological systematist among moderns," used Wolff's ideas to develop his psychophysics.

More recently, it has again been physicists studying complexity who are among those who are beginning to develop noninvasive experimental methods to test theories of brain organization and function.

The paper by Sommerhoff and MacDorman that follows presents a theory about consciousness, perhaps the most elusive but basically an important function of the brain.

The subsequent article by Rotenberg presents a theory of monoamine metabolism in relation to REM sleep to explain the antidepressive effects of drugs based on what he calls a "search activity" concept, i.e., the effect of changing attitudes or behavior on brain monamines and calcium.

STEWART WOLF

An Account of Consciousness in Physical and Functional Terms: A Target for Research in the Neurosciences

GERD SOMMERHOFF AND KARL MACDORMAN

Abstract—The neurophysiology of mental events cannot be fully understood unless that of consciousness is understood. As the first step in a top-down approach to that problem, one needs to find an account of consciousness as a property of the biological organism that can be clearly defined as such. However, if it is to deliver what must be expected of it, it should address what is commonly meant by the word *consciousness.* Unless the last condition is satisfied, the theory will fail to deliver what must ultimately be expected of it.

Although current interest lies mainly in the higher functions of consciousness, such as its role in language and social relationships, the common usage of the word relates to modes of awareness that are not denied to creatures lacking language or social

Address for Correspondence: Gerd Sommerhoff, Trinity College, Cambridge, U.K., e-mail <gfs109@phx.cam.ac.uk>. Karl MacDorman, Wolfson College, Cambridge, U.K. e-mail address: <karl.macdorman@cl.cam.ac.uk>.

Integrative Physiological and Behavioral Science, April–June, 1994, Vol. 29, No. 2, 151–181.

relationships. The basic features to be covered include awareness of the surrounding world, of the self, and of one's thoughts and feelings; the subjective qualities of phenomenal experience (qualia); the conditions a brain event must satisfy to enter consciousness; and the main divisions of mental events, such as sensations, feelings, perceptions, desires, volitions, and mental images.

In the first four chapters we argue that these basic features of consciousness can all be accounted for in terms of just three categories of internal representations, each supported by the empirical evidence and each accurately definable in physical and functional terms. In the fifth, and last, chapter we take a closer look at two of the categories and what these in particular suggest as the most relevant lines of research in the contemporary spectrum of the neurosciences.

Key Words—Consciousness, mental models, mind-brain, qualia, representation, self-awareness, vision.

1. Introduction

1.1 The Elusiveness of Consciousness

Dennett (1991) has described the problem of consciousness as "just about the last surviving mystery." Other mysteries, such as those remaining in cosmology or atomic physics, may not have been finally solved, he says, but at least they have been tamed. "With consciousness, however, we are still in a terrible muddle. Consciousness stands alone today as a topic that often leaves even the most sophisticated thinkers tongue-tied and confused." In the words of McGinn (1991), it is "the hard nut of the mind-brain problem."

Consciousness is central to an understanding of the mind and of the intrinsic difference between voluntary and involuntary behavior. Yet, some modern compendia of the cognitive sciences do not even mention consciousness in the index (Garner, 1985; Posner, 1989). Although experimental psychologists are often reluctant to take up such broad questions, a comprehension of the whole is needed if psychology is to be more than a vast collection of data with *ad hoc* explanations. Thus Searle (1990) describes consciousness as "the most important feature of the mind" (our mental life being composed entirely of what is either conscious or potentially conscious), and asks how this faculty came to be so badly neglected in those very disciplines that are officially dedicated to its study. He notes the many books that have *consciousness* in the title but offer no theory of consciousness. Nor do they offer a scientific definition of consciousness. These two conditions are linked, for a scientific definition of consciousness presupposes a theory of consciousness.

The state of confusion is mirrored in the radically different views held by some scientists about what can be conscious. At one extreme there are those who believe that if consciousness includes self-consciousness, then only human beings have it, for they believe that only the ability to put things into words creates a personal sense of conscious reality (Gazzaniga & LeDoux, 1978). At the opposite extreme Blackmore (1988) argues that consciousness can be attributed to any system that interacts with its environment. Thus, she concludes that a thermostat has consciousness, but a thermometer does not.[1]

It is perhaps not surprising, therefore, that some scientists have come to take a rather skeptical view of the whole topic. Witness the concluding remarks of the entry under *consciousness* in the *Macmillan Dictionary of Psychology* (Sutherland, 1989): "Consciousness is a fascinating but elusive phenomenon: It is impossible to specify what it is, what it does, and why it evolved. Nothing worth reading has been written on it."

1.2 How Fragmentation Has Hindered the Growth of a Unified Scientific Theory of Consciousness

One obstacle in the way of a unified scientific theory of consciousness has no doubt been the vagueness of the notion itself and the latitude this has given to its interpreters. But a greater obstacle, perhaps, has been the exponential growth of specialized fields of research, for this has made it increasingly difficult to see the forest for the trees. Baars and Banks (1992) list more than 40 fields of investigation done under labels that imply such things as conscious experience, voluntary control, or self-awareness.

This, in turn, has produced a confusing multitude of different approaches and conceptual frameworks, each tending to view consciousness through the prism of its own premises—from the philosophical, psychological, and sociological to the computational and information-processing, not to mention the connectionist, neuropsychological, neurophysiological, and psychodynamic. Even chaos theory has entered the field. It has also created the temptation to identify consciousness with just one particular aspect of it, such as the selective function of focal attention (Neisser, 1976), global mapping (Edelman, 1989), the experience of sensations (Humphrey, 1992), its "constructive" (Mandler, 1984) or holistic (Sperry, 1987) nature; its role as a "supervisory system" (Shallice, 1988), as a "global workspace" of limited capacity (Baars, 1988), as a bond between the individual and the community (Barlow, 1987), or as an instrument for coping with novelty (Mangan, 1991); its "capacity to form non-algorithmic judgments" and to "defeat the restraints of computability" (Penrose, 1989); or its redescription of cognitive activities in a series of higher-level languages (Karmiloff-Smith, 1987). The differences can be so great that the reader is left wondering whether the authors are talking about the same thing.

The big divide in approaches to consciousness is between, on the one hand, the natural-science approaches (biological, physiological, computational), with their strict methodology, coherent framework, and conceptual accuracy and, on the other, the approaches that allow themselves greater latitude in the choice of their concepts, accepting both metaphors and undefined concepts whose meaning and tacit premises are often far from clear. An example is the "explanation" of consciousness in abstract metaphorical terms offered by Dennett (1991).

The main thrust of Dennett's book is against the metaphor of a central observer in the brain. The metaphor he substitutes is that of a hierarchy of humunculi that compose "drafts" of what is happening in the outside world, which they then "edit" on a running basis in the light of the current sensory data. However, the very dependence of this "explanation" on metaphors reflects the distance that still separates it from a scientific explanation covering the same ground, such as that suggested in this paper.

1.3 Overview and General Commentary

Our aim is to explain consciousness in a top-down approach as a property of the biological organism that can be clearly defined in terms of physical and functional relationships—in a language, therefore, that is essentially that of physiology. The importance of this aim lies in the fact that this is a necessary first step toward the ultimate goal of explaining consciousness in neural terms. Thus the paper seeks to clarify and articulate what the neurosciences need to target in their search for the neural correlates of consciousness. But it goes no further. Hence, the different theoretical approaches that may be followed in pursuit of that target, such as those listed above, lie beyond its scope. Although

in §5.2 we look at the computational approach, we do so only to make a particular point about the brain's internal representation of the surrounding world. The same applies to the simple neural circuits discussed in §5.8.

A top-down approach is needed because consciousness is a high-level function of the brain operating as an integral system: a causally emergent systems property, so to speak. If this top-down approach is to deliver what must ultimately be expected of it, it needs to cover all the basic features of what is commonly meant by the word *consciousness.* For reasons to be explained in Chapter 2, we take these features to be awareness of the surrounding world and of one's thoughts and feelings, the subjective character of direct experience, and the major divisions of mental events, such as sensations, feelings, perceptions, desires, volitions, and mental images.

We shall suggest that all these features can be accounted for by just three basic categories of internal representations formed by the brain, each supported by the empirical evidence and each clearly definable in physical and functional terms. More specifically, we shall suggest that consciousness may be identified with one of these categories operating on the other two.

Since our account of consciousness also covers the general characteristics of direct experience, including its subjectivity, it challenges the view that the subjectivity of conscious experience precludes a neurophysiological account of consciousness (Nagel, 1974; Searle, 1992).

The empirical evidence for those three categories of internal representations is not one of our problems. Their existence is beyond question at least in the human case. Our problem is how they may be defined in physical and functional terms and how they may be related to the phenomenon of consciousness. We cannot *prove* that consciousness relates to them in the manner asserted in this paper and briefly indicated above. But we aim to show that the case for this interpretation of consciousness is strong enough to suggest it as a hypothesis that needs to be taken seriously in the neurosciences.

Because the paper follows a natural science approach, it is subject to certain constraints to which attention needs to be drawn from time to time. For example, they restrict the range of questions that the paper can be expected to address, ruling out metascientific and metaphysical questions. Thus the only sense in which we need to worry about the theory being reductionist concerns the extent to which it succeeds in covering all the basic features of consciousness and thus avoids offering merely a Procrustean solution of the problem. Again, scientific theories consist of hypotheses designed to explain the phenomena under investigation. It is immaterial how these hypotheses are arrived at, be it by induction or deduction, or merely by way of an inspired guess or bold lateral jump of the imagination. What matters is that they should succeed in explaining what they claim to explain; that they are supported by the empirical evidence; and that they should have the conceptual precision on which the natural sciences insists before any statement is accepted as a hypothesis worthy of consideration and genuinely applicable to the data. Unequivocal theories cannot be constructed from ill-defined material. And these are also the criteria by which the hypotheses should be judged that are advanced in this paper. Hence it is essential that their key concepts are accurately defined in objective terms.

Apart from offering in effect a general account of the relation between mind and brain, the paper also addresses the question of what philosophers call the "intentionality" of mental events, the fact that most mental events (beliefs, thoughts, desires, intentions—but not sensations) have a representational content: they are *about* something. In addition it of-

fers an account of the general conditions a brain event must satisfy to enter consciousness and of the general relation between conscious and unconscious mental phenomena. Although it is mainly concerned with the human brain, it also addresses the question of consciousness in animals and why consciousness should have evolved in the first place.

One final remark. If our interpretation of consciousness is accepted, it could be used as the basis for a precise scientific or "technical" definition of *consciousness* to standardize the use of the word in science and act as an expedient substitute for the fuzzy notions of consciousness that have often been a cause of confusion. Such substitutions are not exceptional in science. The physicist's definition of *work* and *energy* are typical examples of theory-bound, precise definitions of words that are also in common use. However, there is a difference in our case: it does not matter for physics how closely its definitions of *work* and *energy* match the common usage of those words. It uses them merely as labels for particular theoretical variables. In our case, however, it would be shifting the goal posts if one were to introduce a "technical" notion of consciousness that failed to take account of what the word means in everyday life, e.g., the basic faculties that are lost when someone is knocked out by a blow to the head.

1.4 Functional Descriptions

Since we use some functional descriptions, a word must be said about their substance because this, too, can be a source of confusion. According to Searle (1990), for example, the functional description of a biological organ or process, such as the role of the heart in the circulation of the blood, is "simply one of the causal levels described in terms of our interests." Searle sees no difference in the facts asserted by *the heart pumps the blood,* and *the function of the heart is to pump the blood.* Thus he concludes that functional descriptions are not a legitimate level of scientific ontology.

We dispute this. If you are told that the function of the spark plugs in an engine is to ignite the fuel mixture, you are told not only their causal effects but also that they owe their structure and very presence in the engine to their capacity to have these causal effects. This is a statement in which the causal level of the action is fused with the causal level of this component's origin or development. It has nothing to do with the speaker's interests. The spark plugs have other causal effects as well: their action causes a noise. But no one would say that the production of that noise was a *function* of the spark plugs. The same remarks apply to the functional descriptions of biological organs or processes that are the standard fare of the physiological sciences.

2. The Meaning of Consciousness

2.1 What Phenomena Are We Talking About?

The common notion of consciousness does not presuppose a social context. Few would deny that an abandoned baby can have consciousness. Nor is consciousness commonly conceived as something that presupposes language. We do not deny it to beings lacking language, such as the infant and the deaf-mute—nor to higher animals, for that matter. The very notion of suffering in animals implies that they have consciousness. Without it, they could suffer no more than we do under an anaesthetic.

Following Edelman (1989), we shall call this nonlexical, nonpropositional level of

consciousness the *primary* consciousness. This basic level needs to be understood before one can understand the levels that may be added by language and propositional knowledge. The present paper will look at this level only.

Since we want to remain true to the common usage of the word *consciousness,* we shall take its dictionary definition as our starting point. Here we find *consciousness* fairly uniformly defined as an awareness of one's surroundings, of the self, and of one's thoughts and feelings. This does not take us very far, for in neat circularity the same dictionaries are likely to define *awareness* as *being conscious of.* But it does suggest the scope of what is involved.

It is easy to see that this *awareness* of the surrounding world is of the nature of not only a *comprehensive* but also a *coherent* and continuously *updated* internal representation of that world: a representation that covers not only objects and events individually but also their mutual relationships and connectedness in time and space. Thus, as you wake up in the morning and recover your senses, things gradually begin to fall into place: objects assume identities and you also see them in their spatial relations; you know, in fact, where you are, in what room, house, and town; you know what lies beyond the door; you know what you would find in the wardrobe or the chest of drawers—all adding up to a comprehensive, coherent, and continuously updated internal representation of the surrounding world. "Where am I?" is the revealing first question likely to be asked when consciousness returns after an accident. However, this representation is more than just a cognitive map, for you also know various properties of the perceived objects. You know how to open the door, what effort would be required to lift the chair, and that the room's walls are immobile. Body knowledge, including knowledge of posture and movement, also returns as you wake up. And it returns in a form that is fully integrated with your representation of the surrounding world. You know the stiffness is in the leg; you know how to touch the sore spot on your shoulder; and you feel the pressure of your feet on the floor. Moreover, you know that these sensations are *yours.*

2.2 Self-awareness

One of the basic features of consciousness mentioned above is awareness of one's thoughts and feelings. The key element here is an awareness of the fact that your thoughts and feelings are *your* thoughts and feelings, part of your own current condition. It is an awareness of ownership, so to speak. We shall call this the *primary self-awareness.* It will play a critical role in our analysis of the nature of consciousness (§4.2).

Since this primary self-awareness is clearly a prerequisite of self-reports, such as "I saw a bright flash," the psychologist's general acceptance of a subject's ability to report an event as evidence that the event was a conscious one implicitly acknowledges the primary self-awareness as an integral feature of consciousness.

In the human case it also amounts to an awareness of the unity of mind and body: the unity of what is commonly called the "self." This is what William James described as the *self-as-I,* not to be confused with the *self-as-me,* i.e., with self-consciousness in the sense of a self-concept that may include such features as the perception of oneself as one object among others, awareness of one's public persona, and self-esteem. Nor is it to be confused with the "self-concept" explored when animals are tested for their ability to recognize their image in a mirror (an ability found in the great apes but not in monkeys).

2.3 Two Standpoints

Consciousness can be viewed from either a third-person or first-person standpoint. These must not be confused. When consciousness is viewed from the *third-person* standpoint, the word is understood in an objective sense. It now denotes the critical powers that distinguish a normal waking brain from one in a state of sleep or coma, such as the representational powers illustrated above in our sketch of consciousness returning after sleep. And these are the powers that a scientific theory of consciousness needs to describe and explain, for the standpoint of science is a third-person standpoint. During sleep, and even in your most lucid dreams, for example, the brain remains detached from the *real* world surrounding you to an extent that involves the suspension of an updated as well as comprehensive and coherent representation of that world, partly because the sensory perceptions that are essential for maintaining such a representation are suspended.

By contrast, when viewed from a *first-person* standpoint, consciousness is understood in a subjective sense in which it denotes personal experience as directly apprehended. The subjective qualities of this experience, or "qualia," the "what it is like" to feel a pain or taste the milk, have certain general and distinctive characteristics. In agreement with Dennett (1982), we take these to be their uniqueness, intrinsicness, privacy, and direct apprehensibility in consciousness.

There is no way in which this "feel" of your experiences can be actually conveyed to other people, or described in objective terms—the "what it is like" *to you* to have a toothache. You may be able to describe a particular sensation or feeling in ways that enable others to identify a similar category of experience in their own life and sympathize with you accordingly, but this still does not convey to them the feeling itself. In fact, it is impossible for one creature to know what another creature's experience is like. This was the thrust of Nagel's seminal essay "What Is It Like to Be a Bat?" (1974). Nor, therefore, could any neurophysiological account of conscious experience convey what it is like to feel a pain.

As has been said, this has sometimes issued in the belief that the subjective character of conscious experience puts consciousness beyond the reach of science. We disagree. Although one has to agree with Searle that "no description of the third-person, objective, physiological facts would convey the subjective, first-person character of the pain" (from which he concludes that consciousness is irreducible), one also has to realize that science is not in the business of *conveying* anything. That is the work of the artist. The business of science is to explain what from a third-person standpoint appear to be the main features of the phenomena in which it is interested. And as regards the qualia, we can take these main features to be the four general characteristics listed above. Our explanation of these features is in §4.6.

Some scientists believe on philosophical grounds that consciousness lies beyond the reach of science. They may, for example, subscribe to the dualist view of the mind/body relation (Eccles in Eccles & Popper, 1977) or to an epiphenomenalist view (Velmans, 1990). We believe that people tend to arrive at these philosophies if they view consciousness from a first-person standpoint in the mistaken belief that they are viewing it from a third-person standpoint. Our rejection of the dualist philosophy accords with Ryle's (1949) compelling arguments that the notion of a "ghost in the machine" rests on a category mistake.

3. The Brain's Internal Representation of the Surrounding World

3.1 The Meaning of "Internal Representation"

Since the concept of internal representation figures in our hypotheses, it needs to be defined. First, because scientific theories need accurate concepts. Second, because the word *representation* can be used in a variety of senses.

To be successful, the responses of living organisms need to be appropriately related to the external world, and often in a global context. Hence it is to the advantage of the organism if, from the sensory inputs, the brain can form some kind of internal representation or "model" of that world that allows it to make predictions and informed behavioral decisions. Craik (1943) first formulated in modern terms that the behavior of the higher orders of life can, in fact, be explained only on the assumption that they form such a model. Research-supported arguments for the existence of such models in the higher species were notably added by Sokolov (1963).

How is "representation" to be understood in this particular context? The question is important because any event that carries information, such as an instrument reading, could be called a "representation" of what it informs about. Clearly, the model of the world that Craik had in mind is not of this kind. He defined it as "any physical or chemical system which has a similar relation-structure to that of the processes it imitates." Such representations may be described as *analog* representations or *functional isomorphisms.* They contrast with *symbolic* representations, an expression conventionally taken to denote representations whose own relation-structure does not resemble that of the entities represented (e.g., crown as symbol of royalty).

A definition given by Dennett is interesting because it attributes also a dynamic character to the representations and thus comes close to the kind of representations we shall come to describe. "By a 'representational system' we will mean an active, self-updating, collection of structures organized to 'mirror' the world as it evolves" (Dennett, 1982).

All such representations have both a *structural* and a *functional* aspect. The first relates to their composition; the second to their role in the overall performance of the brain. Both are important, but for our definition we shall focus on the functional one, for what the brain accepts as a representation of an actual object or event, may in fact be a *mis*representation. The dim figure that on a dark night you take to be a man may, in fact, be a small tree. Yet, this misinterpretation of the sensory inputs still *functions* as a representation of the object concerned, because you respond to the apparition as being that of a man. Hence, the decisive factor that makes a collection of structures in the brain a representation lies in the fact that it is *used* as such. We shall therefore adopt a functional definition. The structural aspects will be considered later.

> *Definition*: A collection X' of structures in the brain acts as an internal *representation* of an object (event, situation) X, if responses that need to be appropriately related to X are processed in the brain as if they were responses that need to be appropriately related to X'.

(In simplistic terms: in its responses to X, the brain is guided by X'. By a response that is "appropriately related" to a given object, we mean a response whose relation to that object satisfies a requirement arising out of the goal(s) the subject is currently set to pursue. For a formal definition in terms of *directive correlation,* see Sommerhoff (1974).

It needs to be stressed that this definition of *representation* is solely intended to serve the present enquiry. It is not a dictionary definition or an improvement on the sense in which other authors have used the word in different contexts.

We shall also deal with representations of absent or imaginary objects, events, or situations. We can retain our definition for this case on the understanding that X does not now denote a real object (event, situation) that X' needs to match but a set of conditions that the creative act of the imagination seeks to satisfy (as when we try to picture a unicorn, or a machine that will perform a certain task).

3.2 Coherent Internal Representations of the World and Body-in-the-World

By a *coherent* representation of the surrounding world we mean a representation, such as we normally possess in our waking life, that covers not only the different objects and their properties individually but also their mutual relationships, especially their spatial and temporal relationships. With the exception of the rare out-of-body experiences, this representation of the outside world (and the body's place in that world) is coherently coupled with a coherent representation of the body itself, of its posture, movement, and a variety of other somatic variables, such as the location of a painful stimulus. This is commonly known as *body knowledge* or the *body schema* and primarily appears to involve the parietal cortex. At the conscious level it amounts to an awareness of the body from the inside, so to speak. It correlates *inter alia* the sensory and motor processes that specify a posture internally and externally: if you move your arm with closed eyes you will know where to expect it when you open them again. At this level, too, the unity of the body appears to be represented: if the afferent nerves from an arm are severed, the arm is experienced by the subject as a foreign object.

The comprehensiveness as well as the coherence of the world-model is important because it enables the subject's responses to take account of the global context, of the total situation. As you work on your word processor, your attention may be focused on the keyboard or the screen, but only because at the time your world-model contains nothing that called more urgently for attention. You will ignore the knocking noise you hear, for example, because your world-model has absorbed the fact that it is not a knock on your door but merely a carpenter at work on the floor below. In short, although your actions may be governed by the attended fraction of the world-model, the world-model as a whole is still effective, albeit at a different level, e.g., in the control of attention.

3.3 A Note on the Empirical Evidence

The existence of particular representations about the outside world can be explored in experimental subjects without necessarily having to rely on the subjects' verbal reports. There are innumerable ways in which nonverbal human behavior shows how the subject perceives his or her world. To a lesser extent this is also true of animals, and sometimes the evidence can point no further than the existence of a cognitive map in the animal's head, as might be shown, for example, by the ability to find shortcuts or make detours (Gallistel, 1990).

For more comprehensive world-models, especially in the human case, valuable pointers spring from the fact that events covered by the model will be *expected* events, hence the limits of that model reveal themselves by the occurrence of events that are unexpected by the subject, and such occurrences tend to elicit characteristic reactions of "surprise" that

often have observable components. This fact, of course, is extensively used in the study of the development of perception in infants. For example, the more unexpected an event, the greater is its power to attract the infant's attention. This may be observed in the direction and duration of the infant's gaze. Changes in heart rate or perceptible signs of distress may also serve as signs.

Such investigations have shown that reality-mapping expectancies manifest themselves already at an early age. For example, infants as young as one month have shown surprise when they have seen a screen being slowly moved in front of a toy, and the toy is not there when the screen is subsequently removed (Bower, 1971).

3.4 The Brain's World-Model Needs Servicing

The brain's current model of the surrounding world needs to be kept up to date and abreast of current changes if it is to remain useful for behavior. We have our senses to do that. We also have error signals to help us: lack of detail or discrepancies between the model and reality reveal themselves in the occurrence of unexpected events (including unexpected outcome of our actions), and our brain is sensitive to such occurrences. If the error signals are of sufficient magnitude, we may have to switch attention to the field concerned and search for more information in order to correct the deficiencies. In short, the brain's world-model needs to be *serviced.*

It is therefore expedient to distinguish between two kinds of activity: that of *using* the brain's model of the world (e.g., in the performance of some task) and that of *servicing* it (e.g., dealing with gaps or errors that have been shown up by the occurrence of unexpected events). It is a dichotomy well known to decision theory: the dichotomy between *using* information and *gathering* it, between *exploitation* and *exploration.*

At times, of course, the claims made by these two kinds of activity may conflict, especially claims on the subject's attention. Thus the occurrence of an unexpected noise behind your back creates a conflict between turning around and attending to this event or continuing to keep your eye on the task you have in hand. In such cases priorities need to be decided by some central processing mechanism: one, for example, that can override your initial impulse to turn your head. That applies to animals as well. Every sparrow is regularly faced with the alternative of either attending to a movement perceived at the periphery of its vision or continuing to peck at the crumbs it has found. We shall return below to the need for such a central processing mechanism.

(In the present paper we understand by *attention* a selective modulating function that enhances the influence of the selected field over the current processes in the brain, elevating it as a cue for action, for example. The selective processes may thus occur at different cognitive levels and the selection may also be either under conscious or unconscious control.)

3.5 The Need to Expand the Brain's World-Model by a Coherent Representation of the Fact That It Is Part of the Current State of the Organism

The fact that the brain's model of the world is part of the current state or condition of the organism is such an important part of the global context, of the total situation, that for maximum benefit it needs to be added to and integrated with the brain's world-model. By this we mean that the resulting representation of the total situation should be not only comprehensive in the sense required but also coherent, i.e., one that covers the relation in

space and time of the brain's world-model and the world it models. The following are two particular reasons for this expansion (others will become apparent in §4.2).

First, the choices that have to be made at any time between *servicing* the world-model and *using* it must be able to override the decisions that would flow from the pursuit of either the one or the other activity. Hence they need to be made at what amounts to a *higher level of control.* Now, in §3.2 we mentioned the importance of the brain's decisions being governed by a comprehensive and coherent internal representation of the global context, of the total situation. And at this higher level of control the context that has to be taken into account now includes not only the current state of the external world and body-in-the-world but also the current state of an internal and nonsomatic entity, namely that very model of the world and the claims it makes in consequence of the occurrence of unexpected events. And if this global representation is yet to be a fully coherent one, a fully integrated one, it needs to include a representation of how this internal and nonsomatic entity, this internal model, is related in space and time to the world it models. Hence for maximum benefit, the brain needs a representation of the fact that the model is part of the organism—part of its current state. (To understand how precisely this is to be conceived, the reader should turn again to the definition of *internal representation* given in §3.2). Second, this expansion of the brain's world-model also fulfills another role: it is also a representation of the standpoint from which the world-model "views" the world.

In normal circumstances this expansion of the global world-model would not itself need servicing, for it is normally unlikely to be erroneous. Hence we need not additionally postulate the existence of a representation of the fact that this extension, too, is part of the current state of the organism. This, in effect, is the cutoff point that avoids an infinite regress of the type "I am aware that I am aware that I am aware . . . "

4. A Representational Account of Consciousness

4.1 Three Basic Categories of Internal Representations

The topics discussed up to this point have prepared the ground for an explicit interpretation of the nature of primary consciousness in terms of particular categories of brain events and their causal or functional relationships. Our key concept is that of *internal representation* as defined in §3.1.

As an initial hypothesis we postulate a power of the brain to form three basic categories of internal representations, the latter term being understood in the physiological sense defined in §3.1. For brevity, they will be called representations of categories A, B, and C, respectively. Consciousness will presently be interpreted in terms of these categories.

Category A. Representations of actual objects or events, including their properties and relations, that amount to comprehensive and coherent representations of the current structure and properties of the surrounding world and of the organism's place in that world. This was discussed at length in §3.2. In addition, we take this category to include the coherent representation of the somatic parameters that form the "body schema" mentioned there.

Category B. Representations not of actual objects, events, or situations but of absent or imaginary ones, such as those that enter our consciousness as "mental images." Part of their biological importance lies in the fact that either at the conscious or unconscious level they may operate as representations of the intended goal of an action, thus playing a part analogous to that of the reference input to a servo-mechanism.

Category C. Representations that represent internal states or stimuli as part of the current state of the organism in the brain's integral model of the world. The states or stimuli covered are assumed to include

- Representations of category A or B. In respect to category A, these representations of category C thus produce the very *expansion* of the brain's world-model that we discussed in §3.5 and whose biological value was explained at the time.
- States or stimuli that are not already covered by representations of category A (viz., in the body schema) but are yet of behavioral or motivational significance—such as the stimuli generated by hunger or thirst, or particular components of the general impact made by a sensory stimulus or a representation.

Involved though they may seem, none of these representations of category C are in fact difficult to conceive if one adheres strictly to the functional definition of representation given in §3.1.

4.2 Consciousness Interpreted in Representational Terms

According to the central hypothesis we now propose, the main features of (primary) consciousness can be adequately accounted for in physiological terms by identifying consciousness with representations of category C, and thus identifying the contents of consciousness with the totality of brain states that are embraced by representations of this category.

We advance this interpretation of consciousness for a number of reasons:

1. We assert that the identification of consciousness previously described enables us to explain all the basic features of consciousness listed in our opening statements (§1.5). in §§4.3–4.6 we shall deal with these in turn. Moreover, since the theoretical constructs we have introduced are all explicitly defined in physical and functional terms, our interpretation could be used as the basis for a standard scientific definition of consciousness.
2. In §3.5 we explained the adaptive value and biological rationale of the additions made by the representations of category C to the brain's world-model. Since this applies to solitary and nonlinguistic creatures as well as to social and linguistic ones, our interpretation establishes a biological rationale for consciousness that applies also to animals and might help to explain the evolution of consciousness as a historical phenomenon.
3. What the representations of category C add to the brain's processes amounts in effect to what we have described in §2.2 as *the primary self-awareness.*
4. The representations of category C also result in an internal representation of the current state of the organism in which the "somatic" variables of the body knowledge are fully integrated with the "nonsomatic" variables formed by the representations of category A and B. Hence they establish the unity of a "self" that comprises both body and mind, where by "mind" we here mean the brain's conscious representations (cf. §4.5).
5. By virtue of what they comprise, the representations of category C are a plausible candidate for the top-level control system discussed in §3.5. We shall follow up

this conclusion by assuming this control system to be the basis of all *voluntary* activity (§4.5).

It is important to note that the three categories of representations on which we have based our interpretation of consciousness are not just representations of different kinds of things. They relate to distinct dynamic configurations in the brain, with distinct input, output, and feedback channels. The suggestion lies close at hand that the distinctive level of activity of category C is reflected in the observed time lag of 0.5 sec. between a stimulus being applied to the brain and conscious awareness of it (Libet, 1966). A similar value has been observed for perceptual integration times at the conscious level (Blumenthal, 1977).

4.3 Foreground and Background Consciousness

When you wake up in the morning, are the things you see but do not attend to and the areas of the room beyond your field of vision part of your returning consciousness? Indeed they are. The unattended part of the visual scene is present in consciousness, locating in the world-model the attended part of the scene, telling you where to find particular things should you want them, and so on. Again, when you watch a football match your attention may be focused on the ball, but you are nevertheless conscious of the location of the ball relative to the field, of the total scene, of who is playing whom, and of the current score.

4.4 Unconscious Representations ("Subliminal Awareness")

We are never conscious of the individual neural events that enter into the mechanisms that process the retinal inputs, nor, of course, of millions of other individual events occurring in the brain. In our terms, they are not individually covered by representations of category C.

However, what is commonly called "the unconscious" denotes a more restricted category of subliminal events, namely, events that are *potentially* conscious and thus tend to be described in terms of the form they would have in their conscious mode ("unconscious wishes," "repressed desires," "unconscious fears," and the like). For that reason they are also often (and confusingly) called unconscious "mental" events, although *mind* in its original (and still preferable) sense denotes just the contents of consciousness. The present theory offers a ready account of "the unconscious" in the form of representations of category A or B that fail to be embraced by representations of category C. This clearly satisfies the condition of their potential consciousness and also that of conforming to the common divisions of mental events in their conscious mode. In saying this, we are anticipating our interpretation of the main divisions of (conscious) mental events (§§4.5 and 4.6).

There is ample evidence (reviewed in Baars, 1988) of the effective operation in the brain of subliminal representations of various kinds. It is illustrated in everyday life by the sometimes remarkable powers of our intuition and by the things we come to do semiautomatically in well-practised routines—the latter being confirmed in the laboratory by such results as Pani's (1982) demonstration that with practice a mental image used in a matching task may fade from consciousness but return again when the task is made more difficult.

Some of the relationships we have covered are shown pictorially in Figure 1.

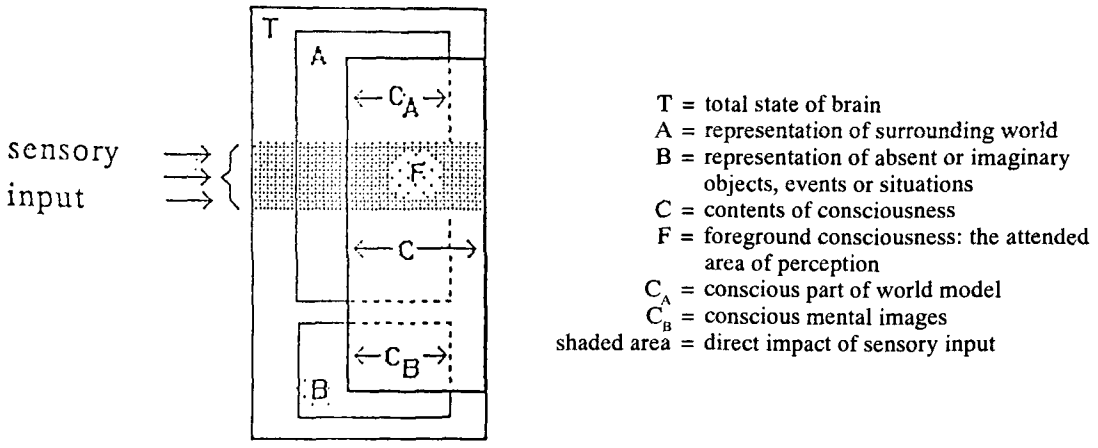

FIG. 1. Representational levels in consciousness

4.5 Interpretations of the Main Divisions of Mental Events

On the physiological interpretation of consciousness spelled out above, the main divisions of mental events may be interpreted as follows:

Conscious representations, stimuli or reactions: those that are covered by representations of category C (§4.2).

Sensations: sensory stimuli that enter consciousness by virtue of their being covered by representations of category C.

Feelings (including emotions): see §4.6 below.

Perceptions: those parts of the brain's conscious representations (hence representations embraced by category C) of the surrounding world that are currently being determined by the sensory inputs (Figure 1, §4.6).

Mental images: conscious representations (see above) of category B, not necessarily in the visual modality. For the reasons given in §2.1, thoughts taking the form of silent speech, i.e., thoughts in a symbolic mode and propositional form are not covered by the theory.

Desires: conscious representations of hypothetical situations (category B) that elicit expectancies (not necessarily articulated in consciousness) of need satisfaction or of pleasurable sensations.

Volitions: desires (as defined) that have taken control of action as conscious goal-images.

Voluntary acts: acts determined by volitions (as defined).

The will: A person's commitment to a volition.

Freedom of the will: the absence of constraint to act according to our volitions: the "freedom to act as we will" (Quine). This notion of a free will does not conflict with the assumption of causal determinism in brain-events, nor with Libet's (1985) discovery that voluntary decisions may be preceded by preparatory cortical activities.

Dreams could be interpreted as follows, subject to the reservation that there can be twilight states between sleep and wakefulness. For the purpose of internal adjustments, the organism detaches itself during sleep from the surrounding world to an extent that involves the suspension of a currently updated internal model of the real world outside. Since category C is an expansion of that model of the real world, it, too, is suspended. This leaves scope only for representations of category B, although these may be lively indeed and include pictures of the world and all the feelings of being conscious (i.e., quasi representations of category C), but both now merely imaginary. They remain unconscious (in the real sense) unless they linger in memory beyond the point at which we wake up, when they are caught by the reviving real representations of category C and thus enter consciousness as remembered images.

4.6 The Subjective Qualities of Conscious Experience

It will be recalled that by the subjective qualities of conscious experience, the qualia, we mean the "what it is like" to see a red rose, to be in pain, or to taste the milk (§2.3). *We suggest that this subjective "feel" of an experience, the qualia, may be identified with the conscious components of the total direct impact made on the subject by the sensory stimulus or perception in question.*

By the "direct impact" of a stimulus or perception we here mean the immediate subjective reactions that the event elicits; by "subjective" reactions we mean the reactions (representational or otherwise, affective or otherwise) that the stimulus elicits in consequence of both the current state and past history (including native endowment) of the individual. And by the "*conscious* components" we mean again the components that are embraced by representations of category C. We interpret *introspection* as the modulating act of attending to these conscious components.

Examples of reactions contributing to this impact would be reactions relating to the intensity of the sensory stimuli involved, affective reactions, associations or memories stirred by the event, and reactions relating to its degree of familiarity or novelty. The latter, for example, would be reflected in the orienting reactions that the event elicits.

Clearly, the conscious components of these reactions have the general characteristics listed in §2.3 for the subjective qualities of a conscious experience, for the qualia: they are ineffable, intrinsic, private, and directly accessible to the conscious subject. They are ineffable (i.e., not analyzable) because they are the cumulative product of the subject's individual history. The other characteristics speak for themselves.

On this interpretation, therefore, the qualia are not beyond the grasp of science, as is so often asserted. Although science cannot *convey* what it feels like to have a particular sensation (which cannot be the business of science), it can (as with other phenomena) explain their general characteristics and etiology.

This concludes our interpretation of the features of consciousness that were listed in Chapter 2 as those in need of explanation by a scientific theory of consciousness when

"consciousness" is understood in the common and primary sense in which it can be present also in solitary creatures and those lacking language.

4.7 Further Notes on the Empirical Evidence

Scientific hypotheses need to be supported by the empirical evidence. In our account of consciousness we have considered three categories of internal representations, briefly:

A. Representations of the surrounding world;
B. Representations of absent or imagined objects, events, or situations; and
C. Representations of the fact that the above (*inter alia*) are part of the current state of the organism.

In the human case the evidence for these three categories is so abundant that it hardly needs to be detailed. Evidence for category A was briefly considered in §3.3. To deny the existence of representations of category B, would be to deny mental images. And for representations of category C, we have compelling evidence in the human ability to make self-reports; you could not report "I see a bright light" unless you had an internal representation of the fact that this was *your* experience and not your neighbor's.

Moreover, the details of all these kinds of evidence are largely irrelevant in the context of this paper. For our main contribution to the discussion of the human case was not the assertion that those three categories of internal representations exists in the mature human brain (which is unquestionably the case) but that these categories of representations can be clearly defined in physical and functional terms. And again, the further hypothesis that consciousness can be accounted for in terms of these three categories does not hinge on empirical evidence but on our analysis of the main features of consciousness, given our understanding of what is commonly meant by the word.

On the other hand, of course, animal brains and the immature brains of young infants do raise questions of empirical evidence. Evidence for category A is here difficult to come by, and we have mentioned one of the main stratagems used in the case of human infants (§3.3). The difficulties are even greater for categories B and C. In the animal case, representations of category B can be assumed when an animal's searching behavior suggests the presence of what Tinbergen has called a "searching image," i.e., when patently error-controlled activities suggest the presence of an internal reference frame representing the intended end result of the actions (Bullock & Grossberg, 1988).

Representations of category C could in theory also be present in infrahuman species, but tests for their existence have not yet been devised. The "mirror test" assesses quite different faculties. However, the adaptive value of this category of representations (§3.5) might be considered an argument in favor of the supposition that the category may exist in at least the higher species.

5. A Closer Look at the Brain's Internal Representation of the Surrounding World and at Mental Images

5.1 Introduction

In the present chapter we take a closer look at two of the three basic categories of internal representations discussed in the last chapter, viz., the brain's internal representa-

tion or "model" of the surrounding world and mental images. Our conclusions here point to the crucial role that acquired states of expectancy or their components play in these representations. They flow from two main considerations:

1. In vision, the feature detectors that have been the main field to which the computational approach has been applied are not enough to represent the world.
2. The idea that the structure and composition of the brain's internal representation of the surrounding world may be closely related to the manner in which it comes to be acquired or updated by the brain, including the manner in which its deficiencies make themselves known to the brain.

The latter consideration is of particular interest. Shortcomings in the brain's world-model reveal themselves in the occurrence of unexpected events—for example, the unexpected outcomes of an act. Hence the scope and composition of the brain's model must in some sense be reflected in the scope and composition of the expectations that the perception of any particular situation may elicit, revise, or sustain, for example, about the outcomes of the different acts available to the organism. A state of the brain in which any expectation of the latter kind would be elicited is called a state of *conditional* expectation because these expectations are conditional on the planned or actual initiation or execution of the acts concerned.

This conclusion is of little value to a physiological approach so long as "expectation" is understood here in a psychological sense. But it is not difficult to define a state of expectancy also in physiological terms (§5.5). And an interesting question then suggests itself: Could it be that at the physiological level, the vast body of conditional or unconditional expectancies elicited, revised, or sustained by the current sensory inputs actually functions as a crucial part of the brain's internal representation of the surrounding world? Since these expectancies would largely be the product of past experience, a positive answer would be difficult to formalize in computational terms. But it would clearly be a mistake to take this into account in considering the hypothesis.

5.2 Limitations of the Bottom-up Computational Approaches

Top-down approaches, such as ours, begin with the properties and needs of the system as a whole, whereas bottom-up approaches begin with the components of the system. In the field of vision a top-down approach might thus begin with the fact that the end product of vision consists of the contribution vision makes to the brain's internal model of the surrounding world. On this view the processes of vision cannot be fully understood until the nature of this model is understood. It is here that the bottom-up computational approaches show their limitations. The particular limitations we have in mind concern the insufficiency of the feature detectors that these approaches seek to model, viz., their insufficiency as representations of the real world.

A systematic computational approach to neural functions proceeds in three stages: it begins with a formal analysis of the task (e.g., edge detection in the perception of shapes); this is followed by a search for algorithms that could produce the desired output from the given inputs; and finally the question of how these algorithms may be implemented neurally is addressed.

As an example of the bottom-up approach in computational form we may refer to the work that set the trend: David Marr's *Vision* (1982). It was the first such approach to the

mechanisms of vision that took it right up to the level of stereoscopic vision, although earlier workers, notably Julesz (1980), had already applied Fourier analysis to explain major features of visual perception in terms of neural responses to different bands of spatial frequencies.

Marr began with a top-down view but was satisfied with the conclusion that vision could be adequately explained by a hierarchy of neural structures devoted to extracting specific features of the visible world from the patterns of intensity changes occurring in the retinal stimuli. He then applied the three stages of the computational approach to these feature detectors from the bottom up, beginning with the extraction of such visual primitives as boundaries, blobs, edge segments, and orientations. For the next higher levels Marr suggested algorithms through which groups of such detected features would be discriminated, first according to size and orientation, and then according to their spatial disposition. The whole sequence of computations yielded three levels of representation, which he called the "primal sketch," the "$2^1/_2$D sketch," and the "3D sketch," the first two being based on retinal coordinates, the last on object-centered ones.

The main point here is that this computational model, and others like it, treat the brain as essentially a *symbol processor,* since the discrete outputs of these neural feature detectors are *symbolic* representations in the conventional sense that they represent something without modeling the nature or relation-structure of what they represent. And here lies the crux of the matter, for unless something is added in the brain, unless the outputs of these various feature detectors are fleshed out in this respect, they are quite inadequate guides to the underlying realities. To take a simple example, it is not enough to have a constellation of neurons that "extracts" from the visual inputs the distance from the viewer of a perceived object (for example, by producing maximal output if the object is at some specific distance); it needs to be supplemented by other patterns of neural activity that represent the *physical nature* (continuity, connectedness, causal implications) of the distance variables—especially their behavioral or ecological relevance, e.g., the physical movements needed to bridge a distance. This is a crucial part of the coherent (i.e., fully connected) representations of the outside world that the organism needs. It also applies to body variables. For example, although in various cortical regions the arrays of cells responsive to retinal inputs "map" the topography of the retina, the outputs of these cells would not amount to a representation of the actual *spatial* nature of the retina's topography unless they were supplemented by representations of its connectedness and causal or behavioral significance (e.g., representations of the eye movement needed to foveate a peripheral stimulus).

Unless the brain's internal representations are *full-bodied* in the above sense, they are as useless as a flight deck furnished merely with unlabeled and undimensioned instruments would be in the hands of a pilot lacking instruction as to their bearing on the flight of the aircraft. It is also obvious that once Marr's assumed feature detectors are supplemented in their way, the resulting representations cease to be a merely symbolic.

This problem is sometimes called the "symbol-grounding problem." It cannot be solved in computational terms if this means introducing yet further symbols. The "neural networks" of the connectionist approach, which has in recent years gained force in artificial intelligence, face a similar problem in the grounding of the inputs to the network when the approach is applied to modeling representational processes in the brain.

Nor is it plausible to argue that once the feature detectors are in place, be they innate or acquired, the subject will in due course learn the behavioral relevance of their outputs. This is putting the cart before the horse because, both at the phylogenetic and ontogenetic level,

the development of an organism's power to discriminate particular features or properties of the environment is directly or indirectly driven by the *need* to discriminate. This is true both in the postnatal development of the visual system and throughout maturity. Simply: you mainly learn what to discriminate because for one reason or another it proves to be relevant to your concerns.

However, over the past years there has been a movement away from the Marrian view of vision. It has come in the form of a growing realization that vision is not simply based on passive observations of the world, on static processes extracting information from a snapshot, such as those envisaged by Marr. Rather, vision is also based on dynamic processes directly linking it to action. This may be illustrated by the importance of movement parallax as a cue in distance perception. And it is of interest to note that whereas Marr's theory of distance perception concentrated on binocular disparity, the evidence suggests that young infants become sensitive to movement parallax first (Slater, 1989). Another familiar example is the stability of the visual scene as the eyes move in their sockets. The brain can achieve this only by taking the movement of the eyes into account. Moreover, the movement needs to be self-induced. This is readily demonstrated by the familiar observation that if your eyeball is moved passively, e.g., by pressing it with a finger, the scene dances about.

The exact extent and manner in which vision depends on movement has been a matter of dispute in the noncomputational neurosciences for more than 40 years—ever since it became clear that whether the brain interprets a moving visual input as a movement of the stimulus or a movement of the subject depends on whether self-induced movements are involved. For a brief history and cross section of the main standpoints that developed in this field, see Gyr and associates (1979).

Modern robotics, too, has adopted active vision as a necessary device, here defined as "active operation in the world in order to change the images that are being collected in a way which enhances task achievement" (Blake & Yuille, 1992).

5.3 Starting Again from the Top: (1) Act-Outcome Experiences

By a new start we here mean analysis that takes a fresh look at the tasks the brain has to perform in the field in which we are interested regardless of the ease with which these tasks may or may not be formalized. Our particular problem is how the brain comes to form the kind of *full-bodied* representations of the surrounding world that we have demanded: representations that cover not only the values of the represented variables but also their intrinsic nature, such as their continuity, connectedness, and causal implications, especially their behavioral relevance.

A logical starting point is to enquire into the ways in which such representations of the realities of the surrounding world are in practice acquired by the organism, for, from a detached standpoint, it seems plausible to assume that the nature of the brain's internal representations of the external world will be closely related to the manner in which they are acquired.

The primary sources of relevant information are, of course, the sensory inputs. But these inputs need to be interpreted in an appropriate way if a representation is to result that mirrors the physical reality of the perceived objects and their properties or mutual relationships. The notion of *interpretation* is not out of place here. Ambiguous figures like the Necker cube, Wittgenstein's "duck-rabbit," and Rubin's "vase-face" make the point. As Searle (1992) has put it: "All (normal) seeing is *seeing as,* all (normal) perceiving is *per-*

ceiving as." Sometimes this is a question of filling in missing data, as in the perception of partly occluded objects. According to Kellman and Spelke (1983), babies show this capacity already at about four months.

Although the way we see the world obviously rests on innate competences of the brain, its postnatal development depends essentially on the infant's active involvement in the world, beginning with its exploration, first of nearby space, and then of the world beyond. According to the above authors, infants have from birth a sensitivity to motion-induced information. These are act-outcome experiences; as the baby crawls about it experiences the self-initiated transformation of each scene into another. The perceived world is an assumptive world and the eye enters these processes as a prehensile, grasping, organ.

In the course of this active involvement and the resulting act-outcome experiences, the brain assimilates more and more about the relation between the sensory data and the physical realities of the surrounding world, in the first instance its the spatial properties and the intrinsic nature of spatial relationships as such. These properties, like others, reveal themselves primarily through the effect they have on the outcome of the subject's actions.

That the way we see the world in fact depends throughout life in some respects on active involvement is also shown by the importance of movement, especially self-induced movement, in the development and adaptability of the visual system. A year after Held and Hein (1962) had demonstrated that self-induced movement was essential for the development of vision in kittens, Kohler (1964) examined in humans what happens when the optical conditions are changed, for example, when inverting or left-right reversing prisms are fitted to the eyes. Systematic studies showed that in due course (which may be a matter of several days) everything will again be seen the right way up or the right way round. And movement proved to be essential for this process of readjustment. Evidently, the brain here discovered the new relationship between the visual inputs and the real world through the visual consequences of the subject's movements. Again, the movement had to be self-induced, visually cued by the subject; pushing the subject in a wheelchair did not suffice.

5.4 Act-Outcome Expectancies

The next logical step is to ask how the organism can derive lasting benefit from the act-outcome experiences discussed above. A very direct answer here is, if the act-outcome *experiences* leave their mark in the form of corresponding act-outcome *expectancies.* The acquired expectancies here at issue are *conditional* expectancies in the sense of being an acquired and enduring state of the brain that is such that *if* the respective act is performed or initiated *then* certain consequences will be anticipated by the brain.

Three considerations recommend a closer look at this answer. First, the brain is really a kind of anticipation machine. One of its most important powers resides in its ability to anticipate the consequences of the actions open to the organism, be this at a conscious or unconscious level. The anticipation of events as the results of experience is, of course, a common phenomenon throughout the higher species of the animal world. Act-outcome expectancies clearly fit this bill. Second, the empirically acquired act-outcome expectancies are as representative of the properties of the world as are the act-outcome experiences from which they are derived. Third, as already mentioned, errors or lack of relevant detail in the brain's model of the surrounding world expose themselves through failed expectancies.

How and where such expectancies are encoded in the neural networks of the brain, unfortunately, is one of the most poorly researched areas in the neurosciences. Barlow

(1991) and Földiak (1992) have suggested formal models of neural circuits that could become attuned to regularly occurring coincidences or covariations among input-related variables. These circuits, in turn, could be linked to other units in a manner that would produce an output whenever such a "familiar" covariation fails to occur. Jointly, the two systems would thus act as a novelty detector or "novelty filter" (Kohonen, 1984). But these "novelty filters" suffer from a common weakness of the bottom-up approach: they detect at a merely local level discrepancies from habituated coincidences or covariations of input-related variables, whereas in real situations the novelty of an event depends on the total context in which it occurs. Another failing (in the present context) is that the switch from the concept of failed expectancies to the concept of novelty obscures the anticipatory component of expectancies, a key element in their cortical functions.

5.5 Starting Again from the Top: (2) A Physiological Concept of Expectancies

The importance of expectancies in animal and human cognition is not a new idea; some sixty years ago Tolman (1932) introduced the notion of acquired expectancies in his explanation of maze learning in rats. And the idea has continued to surface from time to time in the discussion of the brain's internal model of the world. But the notion never firmly established itself, largely because it has generally been left undefined and thus has remained a mental concept rather than a clearly defined objective one. Hence we can justify our return to the concept of *expectancy* only on condition that we define it at the functional level.

States of expectancy, as we view them at the functional level, have two components that need to be separated. First, a state of expectancy for an event implies an anticipation of the event in the sense of a state of readiness for it—by which, in turn, we mean a state that facilitates or advances an appropriate reaction to the event in the sense defined in §3.1. Being familiar with kitchen chairs, your body anticipates the effort required to lift one, and as you start lifting, your body braces itself in anticipation. Second, if the expected event fails to occur—if despite your effort the chair will not budge—you will be surprised and wonder why. The occurrence of this unexpected event reveals an error or gap in the brain's internal model of the surrounding world and the surprise-reactions that it elicits will be the first step in the task of gathering the additional information needed to fill that gap, in this case finding the cause of the chair's resistance.

Indeed, surprise-like reactions are the brain's general response to unexpected events. They are its original responses to everything unfamiliar. Technically, they are known as *orienting reactions.*

The phase "orienting reactions" dates from the work of Pavlov, who described an animal's reactions to the occurrence of unexpected stimuli as "What-is-it? responses" or "orienting responses." These reactions cover a broad spectrum: they may range from merely a fleeting shift of attention, such as a passing glance of the eyes, at the one extreme, to arousal and startle responses at the other.[2] Physiological components of orienting reactions may include changes in pupillary size, breathing, heart rate, and electrical skin conductivity.

Definition: By a state of expectancy, conscious or unconscious, we shall mean a state of the brain that has two components: It is (a) a state of readiness for an event, i.e., a state that facilitates or advances an appropriate reaction to the event, and (b) a state inhibiting the orienting reactions that would have occurred had the event been unex-

pected. By a state of conditional expectancy we shall mean an enduring state of the brain in which the occurrence of a particular state of expectancy (as defined) is conditional on the occurrence of some particular event or set of events.

According to the above definition, therefore, a state of expectancy for an event is a state of the brain that has a characteristic set of both facilitating and inhibiting neural components, all specifically related to that event in the context in which it occurs. Note especially that, as defined above, this state of the brain need not be a conscious one.

The precise form that a state of readiness for an event may take in neural terms will obviously depend on the event in question. At the lowest level anticipatory reactions could consist of no more than an enhanced sensitivity of a particular neural unit to the occurrence of a particular neural input. For example, recordings from single neurons in the visual cortex of the monkey have shown that the occurrence of a stimulus may facilitate a response to the next (Haenny & Schiller, 1988). An example of anticipation at a higher level may be taken from the electrophysiological studies on muscle in subjects catching a ball. Lacqantini and Maioli (1989) here observed anticipatory reactions that began already when the ball was released.

At the neural level, reactions to discrepancies from anticipated events are more commonly observed than the anticipatory reactions themselves. For example, unit responses to failed expectancies have been observed in the hippocampus, and cerebellar cells have been reported that respond to discrepancies between brainstem predictions of target motion and its actual motion (Carpenter, 1988). Walter (1964) was the first to discover responses to failed expectancies also in surface potentials. In these trials the brain was habituated to a pair of tones occurring in sequence. When the second tone was then omitted, a contingent negative variation was recorded in the cortical potentials at the time when that tone was due. Since then the so-called P300 waves have become widely accepted as signifying a response to failed expectancies.

5.6 The Potential of Acquired Act-Outcome Expectancies to Represent the Properties of the Surrounding World

In this section we want to draw attention to the great extent to which a particular category of conditional expectancies, viz., acquired act-outcome expectancies, can map properties of the surrounding world and could, therefore, in theory also function as their representation.

Since the distance of a perceived object is reflected not only in the movements needed to reach it but also in the movement parallax, in the disjunctive eye movements of the required vergence, and in the changes in apparent size and occlusions as the subject moves about, there are several ways in which corresponding act-outcome expectancies map the physical character and behavioral relevance of this variable. Again, scanning movements of the eyes have outcomes that reflect spatial relationships between the fixated objects, and these relationships could in theory, therefore, be represented by the corresponding act-outcome expectancies.

The same applies to the extent to which the flat-view shape of an object is mapped by the eye movements required to follow its contours, or in jumping from salient point to salient point. To that extent, therefore, a shape could functionally be represented in the brain by the set of corresponding act-outcome expectancies, i.e., conditional expectancies relating to the eye movements that would track its contour or would take the eye from

salient point to salient point. It follows that for the mature brain an active contour following would no longer be required to perceive the distinctive features of a shape; once the physical reality of the retina's topography has become mapped in the brain in the form of conditional expectancies relating to the eye movement required to foveate any extrafoveal stimulus, the spatial relations between the elements of a contour will be reflected in the ensemble of such expectancies elicited by the ensemble of contour elements. As the theory would predict, young infants still have to rely extensively on contour following (Zaporozhets, 1965). Adults tend to use it mainly when encountering a novel shape like an inkblot. This double checking could be induced by the extra attention that the novelty of the shape attracts.

An important distinction to be made in the spatial field concerns the frame of reference defining the acts in act-outcome expectancies. This may be *egocentric,* as when it is formed by retinal, head, or trunk coordinates. But it may also be *allocentric,* e.g., object-centered. Examples of the latter would be expectancies relating to the eye movement required to shift fixation from one end of an object to the other. There is evidence to suggest that these distinctions in frames of reference may be key factors in certain phenomena of spatial neglect observed after lesions in the posterior parietal cortex (Marshall et al., 1993).

Naturally, nongeometric properties of objects can also be represented by act-outcome expectancies. The weight of an object can be represented by an expectancy of the effort required to lift it; its brittleness by the expected effect of dropping it—to cite only two of the limitless variety of object-properties that manifest themselves in the outcome of specific actions, right up to the outcomes of scientific experiments.

So far we have looked only at active learning and ignored passive learning. There are many things we learn about the world just by watching things happen. Such passive experiences of sequences of events in the surrounding world amount to what-leads-to-what experiences, which can habituate as what-leads-to-what expectancies that are representative of the perceived continuities, connectedness, or causal relations. These, too, will generally be *conditional* expectancies; for example, an expectancy that if a certain event or sequence of events occurs, then such and such will follow. (In addition, of course, we learn about the world through the spoken or written word. However, since the treatment of consciousness in this paper is confined to the primary, nonpropositional, consciousness, this contribution to the brain's model of the world need not here detain us.)

It follows that indeed the layout of the entire room in which you may be writing, including its contents and their properties, could in theory be represented in the brain by the totality of acquired what-leads-to-what expectancies that are elicited, revised, or sustained by the current sensory inputs either below or above the level of consciousness. An example would be a state of conditional expectancy relating to what you would see if you were to look behind you or were to open a drawer in your desk. According to our interpretation of mental images (§5.9), this state of expectancy could also form the basis of a conscious mental image of what you would see.

These expectancies would comprise not only the act-outcome expectancies derived from a survey of the room from your present standpoint but, suitably transformed, also those acquired in previous perceptions of the room from different standpoints, for we may assume that previously acquired expectancies, suitably updated where this applies, persist until corrected by more recent experiences. Your brain will also have learned empirically to update expectancies as you move about. Indeed, Rieser, Guth, and Hill (1986) have shown that people walking without vision from one location to another seem automatically

to update the relative direction of other positions (in our terms: update expectancies about the direction in which they would have to turn should they want to reach those positions).

The totality of acquired expectancies of how a viewpoint-dependent representation of any set of objects will transform when you move about, amounts in effects to a *viewpoint-independent* representation of the configuration of that set and could form the basis for the imaginative construction of a *cognitive map,* i.e., an internal representation of the ground plan of the configuration.

5.7 The Possible Contribution of Expectancies to the Required "Full-Bodied" Representations of the Surrounding World

In the preceding sections we have pointed out the very wide range of features or properties of the physical world whose nature and behavioral relevance can be reflected in what-leads-to-what experiences, hence also in their corresponding what-leads-to-what expectancies.

Now, in §5.2 we pointed out the need for the brain to have representations of the surrounding world that are *full-bodied* in the sense that they represent not only the current values of external variables but also their physical nature and significance, especially their behavioral relevance. The conclusion is hard to resist that *the core of these representations actually consists of the totality of acquired what-leads-to-what expectancies that the current sensory inputs elicit, revise, or sustain.* For brevity we will call this the *representational expectancy hypothesis.* States of "expectancy," of course, are here to be understood in the functional sense in which they were defined in §5.4.

The totality of acquired what-leads-to-what expectancies that enter this picture clearly form a *hierarchy* of many levels. At the lowest levels they might be just expectancies relating to the shift of a retinal stimulus anticipated by the brain for a movement of the eye. At higher levels they may relate to visual or other consequences of ego movements of various kinds. Still higher levels would be formed by acquired expectancies (conscious or unconscious) about the complex consequences of complex acts. Indeed, if the totality of a subject's past experiences is taken into account, it is hard to conceive here of an upper limit.

Those who dream of neat formal and computational models of the brain's internal representation of the surrounding world may at a first reading be put off by the difficulty of formalizing this expectational account of that representation. But they would be mistaken if they were to assume that alternative theories would make it easier, for any theory would have to explain the role of failed expectancies as indicators of errors or lack of detail in the brain's internal representation of the surrounding world, and thus also the nature and origin of the expectancies concerned, including their relation to that representation as a whole.

In the present state of our knowledge there are still far too many unknowns to decide the representational-expectancies hypothesis on the empirical evidence. This is especially true at the neural level, where far more information is needed *inter alia* about the formation of anticipatory neural reactions as the result of experience. But the hypothesis has enough logical force to suggest that future research in the neurosciences should consider it and the extent to which the brain's internal model of the world may consist of some or all levels of the hierarchy of acquired what-leads-to-what expectancies that the current sensory inputs may elicit, revise, or sustain either below or above the level of consciousness.

Different cortical areas will here be involved in different measure. At the lower end of the hierarchy, for example, cells have been found in the primary visual cortex that are

influenced by the behavioral significance of the visual stimulus concerned (Haenny & Schiller, 1988). The posterior parietal and medial superior temporal areas are also of interest here; in both areas modulations have been observed that suggest their being part of a system dealing with the sensory consequences of movement—mainly of eye movement in the parietal area but also of head movement in the temporal (Mountcastle et al., 1984). Some cells in the posterior parietal area, for example, respond specifically to particular combinations of receptive fields and eye position.

Meanwhile, it is worth noting the full explanatory powers of the representational-expectancies hypothesis:

1. Representations that have the composition suggested by the hypothesis can cover the physical nature of the represented environmental features, such as their connectedness, continuities, and causal implications, and their behavioral or ecological significance. Hence they can flesh out mere feature detectors in the manner described in §5.2.
2. Since the totality of what-leads-to-what expectancies elicited by the sensory inputs in a given situation may also contain conscious or unconscious expectancies of need satisfaction, the theory also covers the *appetitive* significance that perceived objects, events, or situations may have for the perceiver.
3. By virtue of the role it assigns to act-outcome expectancies, the hypothesis suggests a way in which sensory perceptions may be directly linked to action.
4. Because the set of features that may be possessed by an object maintain their mutual relationships when the object is moved or the perspective is changed, their being bound together in the same object can be represented in the brain by specific object-centered what-leads-to-what expectancies. Hence the theory could answer what is known as the "binding problem," well known as a hard nut for the computational approach.
5. The theory has no difficulty with a number of other aspects of vision that are hard nuts for the computational approach, such as the importance of similarities and generalizations in our recognition of objects (expectancies can be more or less specific); the interpretative nature of perception ("seeing is seeing *as*"); the powers of perceptual adaptation; and the role of self-induced movements in this adaptation.
6. The main learning capacities that the theory assumes of the brain are of a kind that has already been extensively investigated in both human and animal behavior, namely, the capacity to derive what-leads-to-what expectancies from what-leads-to-what experiences.
7. It provides for the error signals that are needed to keep the brain's model of the world up-to-date, and for the initiation of required corrections in the form of the orienting reactions that are released by the occurrence of unexpected events.
8. The theory does not exclude the existence of specific feature detectors of the kind envisaged by Marr at the lower levels of the visual process. These would now figure as nodes to which expectancies can become attached, thus supplementing them in the manner required for what we have called "full-bodied" representations (§5.2).
9. As explained in §5.6, the theory readily accommodates both ego-centered and object-centered representations of the surrounding world.
10. The theory is economical, for it enables us to apply Occam's razor to the alterna-

tive views that the brain forms some kind of model of the world, that this creates conditional expectancies, and that defects in the model are signaled by failed expectancies, for if the expectancies actually *are* the effective parts of the model, then this conventional view obviously suffers from an excess of theoretical constructs.

11. The totality of acquired what-leads-to-what expectancies elicited, sustained, or revised by the current sensory inputs satisfies for the brain's representational structures what Wittgenstein (1953) has called the "condition of adequate logical multiplicity": a capacity to accommodate the enormous variety of things that may need representing (including features of the world never before experienced in the history of the species and thus beyond the scope of hard-wired detectors).
12. Since this totality has a hierarchical structure in the sense explained earlier in this section, the theory would allow for a certain degree of modularity in the relevant brain mechanisms, depending, for example, on the degree of segregation in the brain of structures that code act-outcome expectancies for different levels of activity and for the different modalities to which the outcome expectancies may relate.
13. Because there are generally several ways in which an external variable may be represented in what-leads-to-what expectancies, the theory implies a corresponding degree of redundancy in the representations concerned. An example is the extent to which the visual deficits in a split-brain patient can be rendered innocuous by what the brain can pick up from dynamic transformations of the visual image (Sperry, 1987).
14. The theory supports the modern move away from the once-fashionable notion of the brain as a static computing machine and toward a perception of the brain as a plastic and creative system in fluid motion.
15. Finally, the theory offers an economical explanation in physical and functional terms of the relation between *seeing* and *imagining* (§5.9).

5.8 How Acquired Conditional Expectancies May Become Representations Neurally

There is still a gap to be filled in the suggestion that acquired act-outcome expectancies (or what-leads-to-what expectancies, generally) can function as the core of the brain's internal representations of the surrounding world and its features or properties. It arises from the conditional nature of the expectancies (§5.4).

Act-outcome expectancies are acquired neural states of the brain in which the performance, initiation, or planning (as the case may be) of a particular act elicits more or less specific expectancies about the outcome. But if such acquired act-outcome expectancies are to function as effective representations of some property of the world, they must remain of use also in circumstances in which those particular acts are *not* performed or initiated (or even intended). Having discovered the temperature of an object by touch, that experience must be available also in other circumstances when it is relevant, not just when you might be tempted to touch it again. That is the gap in the representational expectancies theory that still needs to be filled.

There is more than one way in which this might be achieved neurally. The simplistic theoretical model described below merely serves to illustrate the kind of neural learning changes that would make it possible. It envisages a two-tier system in which the neurons of the upper array can become attuned to particular combinations of acts and anticipated

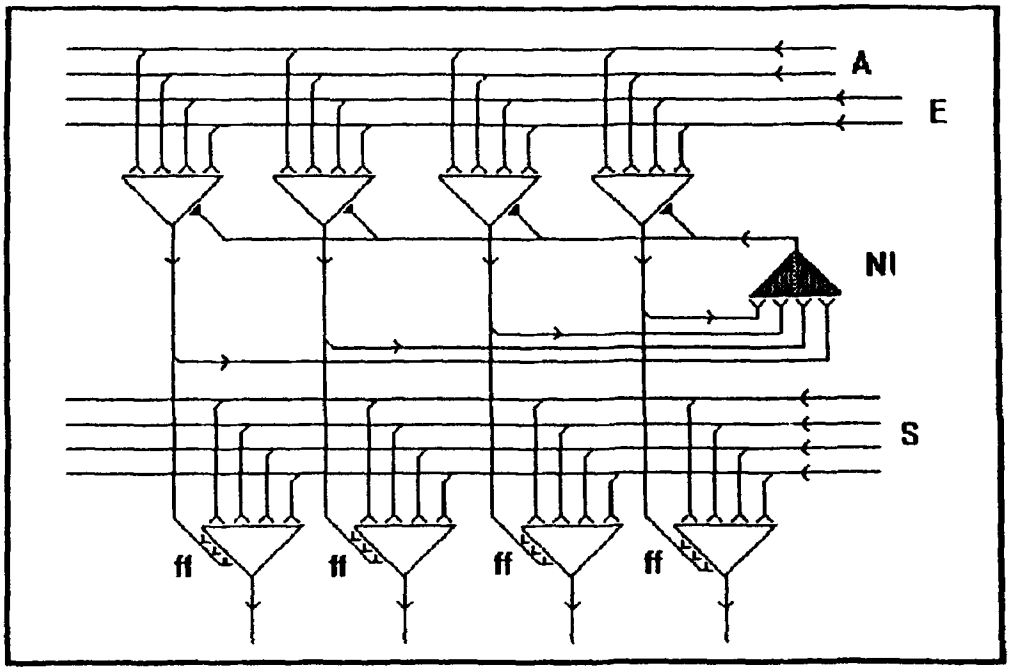

FIG. 2. Simple network to illustrate in two stages how act-outcome expectancies can become representations which are effective also when the respective acts are not performed or initiated. The acts and anticipated outcomes are represented by inputs A and E respectively, and the situations in which these expectancies are elicited by inputs S. NI is an inhibitory neuron, and ff are forcing inputs. For details see text.

outcomes, and those of the lower array enable the act-outcome expectancies to which the upper is attuned to become effective also in circumstances unrelated to the acts represented in the upper array. For clarity we have imagined these two functions to be separated, although they could be combined. Figure 2 shows a small section of the two arrays.

The input channels marked **A** in the upper array symbolize a large set of channels in which a set of alternative acts (performed or initiated) are represented by specific input patterns, and reactions related to expected outcomes (i.e., reactions that facilitate an appropriate response to the outcomes) are represented by patterns of inputs in channels **E**. Finally, a set **S** of alternative situations is envisaged in which the different acts will elicit different outcome expectancies. The situations themselves are represented by the patterns of inputs to the lower array.

The upper array is subject to recurrent inhibition via the inhibitory neuron **NI**. Given the right parameter values, the effect of this inhibition is to act as a maximum intensity filter; only the neuron with maximum stimulation will fire. In other words, the winner takes all. Furthermore, the synapses of the excitatory inputs to these neurons are assumed to be "Hebbian": the active synapses of a neuron increase in weight whenever that neuron fires. In consequence, the neurons of the upper array tend to become individually attuned to particular combinations of acts and expected outcomes. Alternative networks achieving similar effects but relying also on anti-Hebbian synapses have been suggested by Barlow (1991).

Each of the upper neurons projects to a neuron in the lower array by *forcing fibers* (**ff** in Figure 2). When activated, the commanding influence of their synapses (symbolized in the

diagram by multiple synapses) causes that neuron to fire regardless of activity in channels S.

The synapses of these S-channels on the neurons of this lower array are again assumed to be Hebbian. Eventually, they will be sufficiently powerful to take over from the forcing fibers. From that point onward, the occurrence of any particular member of the set S of situations will activate in the lower array the neurons that would have been activated by neurons in the upper array in accordance with the act-outcome expectancies pertaining to that situation if any of the relevant acts had been initiated or performed. But it would no longer depend on this being the case at the time. Hence the total pattern of activated neurons in the lower array now mirrors the total act-outcome expectancies attached to the situation concerned regardless of the state of action of the subject.

It may be noted in this regard that the neural circuits of the cerebral cortex and cerebellum are widely held to be particularly well suited for the detection of coincidences and covariations. All the main ingredients seem to be present. In the cortex the pyramidal neurons would be the key neurons. Shared inputs here are common: each afferent fiber may contact up to 5,000 pyramidal cells. The majority of the pyramidals also have recurrent collaterals, and suprathreshold stimulation of a pyramidal cell has been found to inhibit the activity of others up to a distance of 1 cm. There are less certain candidates for the forcing fibres we have envisaged. But these may not be needed if the two functions illustrated in our two separate arrays were combined in the same network.

5.9 Representations of Category B; Mental Images

Representations of category B were defined as representations of absent or imaginary objects, events, or situations. The crucial feature of these representations is that they are not tested against the current visual inputs. In consequence they can be abstract and have open termini; a leopard can be imagined without a precise location of the spots. Since they also lack the involvement of self-induced action in vision, they lack the vividness and richness of actual visual sensations.

In other respects, though, they have much in common with real visual perceptions. Their components are all derived from actual visual experiences, and they are limited by the limits of seeing: their "field of vision" spans about the same angle as in seeing, and things can be imagined only from a single standpoint. Furthermore, they seem to engage broadly the same cortical areas as direct vision. Evidence for this comes both from the patterns of cortical blood flow and from event-related cortical potentials (Farah, 1988).

Clearly, the critical difference here is the absence of checks of the representations concerned against the sensory inputs; hence also the absence of the orienting reactions that would follow a detected mismatch. If we accept the expectancy theory for the brain's internal representations of actual objects, then this critical difference suggests that the neural correlates of representations of category B may resemble those of category A, except for one crucial differenced: the second of the two components listed in our definition of expectancy (§5.5) is missing. That is to say, in imagining an object or event, the brain is in a state of *readiness* for a sensory perception, without *anticipating* that perception. Hence no surprise reactions follow if the perception fails to occur, no orienting reactions. Indeed, there would be surprise reactions if an imagined object were suddenly to materialize before your eyes.

On this supposition, therefore, representations of category B would consist of states of *readiness* but not *expectancy* for sensory perceptions in their respective modalities. This

agrees with the conclusion reached by Neisser (1976): "To imagine something that you know to be unreal, it is only necessary to detach your visual readiness from your general notion of what will really happen and embed them in a schema of a different sort."

With this in mind it would be worth considering to what extent hallucinations could be interpreted as a case in which the brain forms representations of category B, e.g., representations of a nonexistent object, but erroneously responds to these B-representations as if they were representations of category A. Because (in contrast to A-representations) B-representations consist of states of *readiness* but not *expectancy* for sensory experiences of a particular kind, the conflict between the imagined object and the actual sensory inputs fails to be registered in the brain, and so the discrepancy with reality remains uncorrected.

5.10 Concluding Remarks

We have tried to show that a top-down physiological approach can capture the central features commonly attributed to consciousness in just three theoretical constructs: three categories of internal representations, each definable in terms of physical processes and their functional relationships, and each "full-bodied" in the sense described in §5.2. The empirical evidence for the existence of these three categories is beyond question. Our contribution lies in their definition in physical and functional terms, and in our interpretation of the part they play in the constitution of consciousness.

We see the main value of this account of consciousness in the target it could set for research into the neurophysiology of consciousness. However, if accepted, it would also lend itself for a "technical" definition of consciousness to standardize the use of the term throughout the scientific community, a much-needed step. And in suggesting how the gulf between mind and brain may be bridged in physiological terms, it offers psychology a theoretical foundation in the natural sciences—one of the original aims of the behaviorists. It also explains in general terms the conditions a brain event needs to satisfy to enter consciousness. We challenged the view that the subjectivity of phenomenal experience rules out a neurophysiological account of consciousness.

The first and major part of the theory, our representational interpretation of consciousness (Chapters 1–4), can stand on its own feet. The same applies also to the second part (Chapter 5), which deals with the composition of two of the categories of representations assumed in our account of consciousness. But, here there are still too many unknowns to decide empirically the exact degree to which acquired expectancies can assume a representational role. We arrived at this suggestion on mainly logical grounds and can thus offer it merely for consideration as a particular target for research into the neurophysiology of consciousness, one part of which would be the question how states of expectancy are coded in the brain neurally.

Notes

1. Lecture given at Cambridge, UK, February 1992.
2. Some writers exclude attentional reactions from orienting reactions.

Acknowledgments

In view of the multidisciplinary significance of consciousness, we were fortunate to have found friends in a scatter of disciplines who were willing to read a draft and offer us

the benefit of their comments. Our thanks to Horace Barlow (neurophysiology), Hans-Jurgen Eysenck (psychiatry), Nick Humphrey (ethology), Bruce Mangan (psychology), Kevin Murphy (computer science), and Gysbert Stoet (AI).

References

Baars, B.J. (1988). *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.

Baars, B.J., & Banks, W.P. (1992). On returning to consciousness. *Consciousness and Cognition,* 1 (1):1–6.

Barlow, H.B. (1987). The biological role of consciousness. In C. Blakemore & S. Greenfield (Eds.), *Mindwaves.* Oxford: Basil Blackwell.

Barlow, H.B. (1991). Vision tells you more than "what is where." In A. Gorea (Ed.), *Representations of Vision: Trends and Tacit Assumptions of Vision Research.* London: Cambridge University Press.

Blackmore, S. (1988). Consciousness: Science tackles the self. *New Scientist,* 1 (4):88.

Blake, A., and Yuille, A. (Eds.). (1992). *Active Vision.* Cambridge: MIT Press.

Bower, T.G.R. (1971). The object in the world of the infant. *Scientific American,* 225:30–38.

Bower, T.G.R., Broughton, J.M., & Moore, K.M. (1970). The coordination of visual and tactual input in infants. *Perception and Psychophysics,* 8:51–53.

Blumenthal, A.L. (1977). *The Process of Cognition.* Englewood Cliffs, NJ: Prentice-Hall.

Bullock, D., & Grossberg, L. (1988). Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review,* 95 (1):49–90.

Carpenter, R.H.S. (1988). *The Movements of the Eyes.* London: Pion.

Craik, K. (1943). *The Nature of Explanation.* London: Cambridge University Press.

Dennett, D.C. (1982). In D.R. Hofstadter & D.C. Dennett (Eds.). *The Mind's I.* London: Penguin.

Dennett, D.C. (1991). *Consciousness Explained.* Boston: Little, Brown.

Eccles, J.C., & Popper, K.R. (1977). *The Self and Its Brain.* Berlin: Springer International.

Edelman, G.M. (1989). *The Remembered Present.* New York: Basic Books.

Farah, M.J. (1988). Is visual imagery really visual? *Psychological Review,* 95:307–17.

Földiak, P. (1992). *Models of Sensory Coding.* Working reports, Physiological Laboratory, Cambridge, UK.

Gallistel, C.R. (1990). Representations in animal cognition: An introduction. *Cognition,* 37:1–22.

Garner, H. (1985). *The Mind's New Science.* New York: Basic Books.

Gazzaniga, M., & LeDoux, J.E. (1978). *The Integrated Mind.* New York: Plenum.

Gyr, J., Wiley, R., & Henry, A. (1979). Motor sensory feedback and geometry of visual space: A replication. *Behavioral and Brain Sciences,* 2:59–64.

Haenny, P.E., & Schiller, P.H. (1988). State dependent activity in monkey visual cortex. *Experimental Brain Research,* 69:225–44.

Hebb, D.O. (1949). *Organization of Behavior.* New York: Wiley.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behaviour. *Journal of J. Comp. and Phys. Psych.,* 56:872–76.

Humphrey, N. (1992). *A History of the Mind.* London: Chatto & Windus.

Johnson-Laird, P.N. (1983). *Mental Models.* Cambridge: Cambridge University Press.

Julesz, B. (1980). One-, two- and three-dimensional vision. In C.S. Harris (Ed.), *Visual Coding and Adaptability.* Hillsdale, NJ: Lawrence Erlbaum.

Karmiloff-Smith, A. (1987). Beyond modularity. Transcript of an invited talk given to the Annual Meeting of the British Psychological Society, April 1987.

Kellman, P.J., & Spelke, E.R. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology,* 15:483–524.

Kohler, I. (1964). The formation and transformation of the perceptual world (Fiss, trans.). *Psychological Issues,* 3:1–173.

Kohenen, T. (1984). *Self-organization and Associative Memory.* New York: Springer.

Lacquantini, F., & Maioli, C. (1989). The role of preparation and tuning anticipatory and reflex responses during catching. In N. Wiener, & J.P. Scade, (Eds.), *Cybernetics and the Nervous System.* Amsterdam: Elsevier.

Libet, B. (1966). Brain stimulation and the threshold of conscious experience. In J. Eccles, (Ed.), *Brain and Conscious Experience.* Berlin: Springer.

Libet, B., Gleason, C.A., Wright, E.W., & Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potentials). *Brain,* 106:623–42.

Lieblich, I., & Arbib, M.A. (1982). Multiple representations in space underlying behaviour. *Behavioral and Brain Sciences,* 5:627–59.

Mandler, G. (1979). Categorical and schematic organization in memory. In C.R. Puff (Ed.), *Memory Organization and Structure.* New York: Academic Press.

Mangan, B.B. (1991). *Meaning and the Structure of Consciousness: An Essay in Psycho-Aesthetics.* Ph.D. diss. University of California. Berkeley.

Marr, D. (1982). *Vision.* New York: W.H. Freeman.

McGinn, C. (1991). *The Problem of Consciousness.* Oxford: Basil Blackwell.

Marshall, J.C., Halligan, P.W., & Robertson, I.H. (1993). Contemporary theories of unilateral neglect: A critical review. In I.H. Robertson, and J.C. Marshall (Eds.), *Unilateral Neglect: Clinical and Experimental Studies.* Hove, UK: Lawrence Erlbaum.

Mountcastle, V.B., Motter, B.C., Steinmetz, M.A., & Duffy, C.J. (1984). Looking and seeing: The visual functions of the parietal lobe. In Edelman, G.M., Gall, W.E. & Cowan, W.M. (Eds.), *Dynamic Aspects of Neocortical Function.* New York: Wiley.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review,* 38:435–50.

Natsoulas, T. (1992). Is consciousness what psychologists actually examine? *American Journal of Psychology,* 105 (3): 363–84.

Neisser, U. (1976). *Cognition and Reality.* San Francisco: W.H. Freeman.

Pani, J.R. (1982). A functional approach to mental imagery. Paper presented at the twenty-third annual meeting of the Psychonomic Society, Baltimore, MD.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* Oxford: Oxford University Press.

Posner, H. (Ed.). (1989). *Foundations of Cognitive Science.* Cambridge: MIT Press.

Rieser, J.J., Guth, D.A., & Hill, E.W. (1986). Sensitivity to perceptive structure while walking without vision. *Perception,* 15:173–88.

Ryle, G. (1949). *The Concept of Mind.* London: Hutchinson.

Searle, J.R. (1990). Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences,* 13:585–642.

Searle, J.R. (1992). *The Rediscovery of the Mind.* Cambridge: MIT Press.

Shallice, T. (1988). Information-processing models of consciousness: Possibilities and problems. In Marcel, A.J., and Bisiach, E. (Eds.), *Consciousness in Contemporary Science.* Oxford: Clarendon Press.

Slater, A. (1989). Visual Memory and perception in early infancy. In A. Slater, & G. Bremner (Eds.), *Infant Development.* London: Lawrence Erlbaum.

Sokolov, E.N. (1963). *Perception and the Conditional Reflex.* New York: Macmillan.

Sommerhoff, G. (1974). *Logic of the Living Brain.* London: Wiley.

Sperry, R.W. (1987). Split-brain and the mind. In R. Gregory, & O.L. Zangwill, (Eds.), *The Oxford Companion to the Mind.* Oxford: Oxford University Press.

Sutherland, S. (1989). *The Macmillan Dictionary of Psychology.* London: Macmillan.

Tolman, E.C. (1932). *Purposive Behavior in Animals and Men.* New York: Century.

Velmans, M. (1990). Consciousness, brain and the physical world. *Philosophical Psychology,* 3(1):77–99.

Walter, W. Grey (1964). Slow potential waves in the human brain associated with expectancy, attention and decision. *Archiv für Psychiatrie und Nervenkrankheiten* 206:309–22.

Wilkes, K.V. (1988). "—, yishi, duh, um," and consciousness. In A.J. Marcel, & E. Bisiach, (Eds.), *Consciousness in Contemporary Science.* Oxford: Clarendon Press.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford: Blackwell.

Zaporozhets, A.V. (1965). The development of perception in the pre-school child. In *European Research in Cognitive Development,* vol. 30, no. 2. Chicago: University of Chicago Press.