Reduction and the Neurobiological Basis of Consciousness

Patricia Smith Churchland

Introduction

The idea that mental phenomena might be reducible to neurobiological phenomena has seemed to some people to be outrageous, or at least deeply implausible. The reasons for the opposition are varied, the more traditional view being that mental phenomena are properties of a non-physical substance. Though not without support in some circles, substance dualism has come to seem a most remote empirical possibility, given what else we know about the brain, about evolution, and about biology generally.

Nevertheless, philosophers have only rarely been persuaded by the reductionist strategy, not because they are substance dualists, but because, in one way or another, mental categories have seemed necessarily non-biological or anyhow wholly unsuitable to explanation in neurobiological terms. Currently in vogue are two new kinds of dualism: property dualism and theory dualism. Although neither of the new forms takes the mind to be a non-physical substance, both envision an unbridgeable division between mental events and neurobiological states.

Property dualism essentially says that mental states have special properties that are not explainable in terms of brain states. It is especially appealing to those whose rejection of reductionism is based on a fascination with consciousness and the peculiarly experiential aspects of mental states (for example, see Sperry 1980; Nagel 1974; Jackson 1982).

Theory dualism is the favourite of those who focus on the semantic dimension of mental representation, and who see belief-desire (intentional) explanations at the heart of theories in cognitive science (for example, see Fodor 1975, 1981; Pylyshyn 1984). The new breed of dualists are typically materialists, in the sense that they agree that any particular mental state or process is a state or process in the nervous system, but they also believe that psychology is an autonomous level of explanation. On this view, psychological theory will not reduce to neurobiological theory, and hence psychological properties are not explainable in terms of the properties of nervous systems.
The new dualism, both the property and the theory version, draws on several crucial assumptions concerning the nature of reduction, and the conditions to be satisfied in order for a reductive research ideology to be reasonable. In my view, much of the appeal of the new dualism derives from misconceptions concerning what sort of business intertheoretic reduction is. In what follows I shall offer an account of intertheoretic reduction drawn from the work of historians and philosophers of science, followed by a discussion of the possibility of a reductive explanation of consciousness (see also Schaffner 1976; Churchland 1979; Hooker 1981).1

Intertheoretic reduction2

‘Reduction’ has come to acquire a whole range of connotations. To some, it reflects the goal that any sound science ought, in the long run, to aim for; to others, it represents the misguided idea that knowledge of fine-grained detail will automatically lead to knowledge of macro-properties, or the equally misguided idea that the study of macro-properties is a waste of time. Before addressing the question of reductionism with respect to neuroscience, it will be useful to have a more precise account of what reduction is. The basis of the analysis I shall offer consists of examples drawn from the history of science, where we can see in examples at arm’s length what got reduced to what, and the conditions to be satisfied for reduction to be achieved.

Reduction is first and foremost a relation between theories. Simply put, one theory is said to reduce to another theory when the first is explained in terms of the second. Statements about some phenomenon reducing to another phenomenon (or set of phenomena) are derivative upon the more basic claim that the theory which characterizes the first reduces to the theory which characterizes the second.

For example, when it is claimed that light has been reduced to electromagnetic energy, what this means is that (a) the theory of optics has been reduced to the theory of electromagnetic radiation, and (b) the theory of optics is reduced in such a way that it is appropriate to identify light with electromagnetic radiation. Similarly, when we entertain the question ‘Is light reducible to electromagnetic radiation?’; the fundamental question is whether the theory of optics is reducible to the theory of electromagnetic radiation.

In specifying more exactly how theories must be related in order for one to reduce the other, the logical empiricists offered this: the reduced theory must be logically derived from the reducing theory plus some extra stuff. The extra stuff included boundary conditions, limiting assumptions, approximations, and, most crucially, bridge principles. It was the function of bridge principles to connect properties comprehended by the reduced approximations, and, most crucially, bridge principles. It was the function of bridge principles to connect properties comprehended by the reduced theory to properties comprehended

---

1 For a more complete discussion of these issues, see Churchland (1986).
2 This discussion is based on a more extended treatment in Churchland (1986).
by the reducing theory. In the most straightforward case, the bridge principles would identify the properties in the reduced theory with properties in the more basic theory. This specification of reduction, essentially involving the derivation of the old theory and an identification of properties, new and old, was both elegant and appealing, and it came to be regarded as definitive (Nagel 1961).

However, as historians of science examined the dynamics of science, they found it necessary to disensnare themselves from a background myth abetting logical empiricism; namely the myth that science is mainly a smoothly cumulative, orderly accretion of knowledge. Science, it appeared, was a rather more turbulent affair. Sometimes one theory was substantially corrected before a reduction was possible; sometimes one theory was displaced outright rather than smoothly reduced. The problem for the logical empiricist account was that theories invariably had to be corrected and modified to get something derivable from the basic theory. With modifications accomplished, it was of course not the old theory itself which was derived, but a corrected version of the old theory (Feyerabend 1963).

Theories range themselves on a spectrum of how much correcting and re-vamping they require to be rendered suitable for reduction. Some, for example the theory of optics, required relatively little correction in order to reduce to electromagnetic theory; in other cases (thermodynamics to statistical mechanics), the corrections were greater; but in some there was so much correction that almost nothing, save a few low-level, homey generalizations, could be retained. The phlogiston theory of combustion is one such example. Here the correction required was so massive that it seemed appropriate to think of the old ontology as displaced entirely by the new theoretical ontology; that is, we now say that there is no such thing as phlogiston, not that phlogiston reduced to some compound containing oxygen. The spectrum accordingly contains smoothly reduced theories at one end, and theories which were largely eliminated at the other end, and there are intermediate cases in between.

This re-evaluation of the business of reduction turns out to be no mere nicety appended for the sake of historical accuracy. Rather, it has important implications for what we look for in the reductive futures of interrelated sciences. The reason is this: whether identifications of properties specified by macro- and micro-theories are forthcoming depends on the extent of the correction and revision required. In the event of considerable revision or a fragmentation of the macro properties, identifications go by the board. For example, ‘impetus’ in Aristotelian mechanics cannot be identified with anything in Newtonian mechanics, and, in turn, the ‘mass’ of Newtonian mechanics cannot be identified with the ‘relativistic mass’ of relativity theory, though the two are analogues. As it now seems, the genes of early transmission genetics are not identifiable with segments of DNA as specified in modern molecular genetics. As Kitcher (1982) has remarked concerning this latter case, it has turned out that there is no science of the gene, only the science of genetic material.

The general point is this: if capacities, entities, and properties as specified by a given macro-theory fragment as science proceeds, then the identifications which figure so prominently in the traditional account of reduction turn out to have no place in the reductive portrait of that theory. Instead we see recon-
figurations, adjustments, and some times quite revolutionary revisions. Identifications are typical only of the favoured few reductions at one end of the spectrum. They are the badge of the theories lucky enough to reduce smoothly; but, in such cases, the reduced theory must be sufficiently correct in all its essentials that revision and correction is minimal.

This brings us to the second major point, a point which was suppressed for simplicity in the foregoing discussion. Reduction of one theory to another is usually the final stage in a complicated courtship. Earlier phases involve the co-evolution of the theories, where each provides inspiration and experimental provocation for the cohort theory and where the results of each suggest modifications, revisions, or constraints for the other. As the theories co-evolve, they gradually knit themselves into one another, and points of reductive contact are established and elaborated.

Typically, it is during this gradual co-evolutionary development of theories that the corrections and extensions to both theories are made, and from such theoretical interanimation may ultimately emerge a unified theoretical framework. At this late stage, the reduced theory will have an explanation—of its properties, laws, entities, and so forth—in terms of the reducing theory. However, as Francis Crick has observed, by the time we get to the point of being able to sit down and effect the derivation of one theory to another, most of the really exciting science is over. By then, the inspired modelling, the wild and woolly imaginative forays, the wall-tumbling experiments, the fitting and revising and revolutionizing, are pretty much behind us.

Admittedly, this is something of an exaggeration, inasmuch as scientific theories are ever incomplete and there is always fun to be had. None the less, what is right about this view is that it emphasizes the importance of the co-evolutionary process in achieving a reductively integrated theory. The logical empiricists, in focusing selectively on the final products of a long history of theoretical co-evolution, overlooked the dynamics of theory revision. This is a serious oversight, since it is frequently in a theory’s evolution that the major reductive links are forged and the major revisions—categorial and ontological—are wrought. Since 1953, transmission genetics and molecular genetics have embarked upon such a co-evolutionary development, each correcting and constraining the other, and we can see it continue in the developments of thermodynamics and statistical mechanics, even after the successful reduction of the gas laws. (Hooker 1981).

Reductive achievements sometimes fall short of the complete reduction of one theory to another because the available mathematics are insufficient to the task. Thus quantum mechanics has succeeded in explaining the macro-properties of only the simplest of atoms, and whether more will be forthcoming depends on developments in mathematics. Some people tend to want to make a lot of this (e.g. Popper in Part I of Popper and Eccles 1977), but it has not seemed to me to be a lesson of any great significance anywhere, and not, so far as I can tell, for the reductive future of psychology and neuroscience. In the case of quantum mechanics, the mathematical limitations do not entail that the macro-properties of the more complex atoms are emergent in some spooky sense, but only that mathematical limitations mean we cannot now explain them. Whether the appropriate mathematics will ever be developed is very
much an open question, but the important thing in any case is that the general outlines of the reductive story are in place. Moreover, we have no reason at this stage for assuming that a reductive programme in neuroscience will be stopped dead in its tracks because the enabling mathematics peters out.

Consciousness, co-evolution, and reduction

In the domain at issue, namely the mind-brain, what are the theories whose co-evolution might eventually lead them towards reductive integration? In a very loose sense, they are psychological theories concerning macro-phenomena and neurobiological theories concerning micro-phenomena. The sense is very loose for several reasons: (1) both sciences are still in their infancy, and theories of the required kind still have a long co-evolutionary haul ahead of them, and (2) both sciences are finding it increasingly necessary to address intermediate levels of mind-brain organization. For psychology, one major difficulty is that it is still far from clear what the macro-capacities and macro-properties are which need, ultimately, to be explained neurobiologically; for neurobiology, one major difficulty is that it lacks theories of higher levels of organization— theories which specify the representational and computational properties of cell assemblies and, in turn, of collections of cell assemblies. That is, neurobiology needs a theory of what is going on above the level of the single cell.

Much of the discussion concerning the reduction of mental phenomena to neurobiological phenomena is conducted on the assumption that a reduction requires the identification of mental phenomena, as understood within the framework of folk psychology, with neurobiological phenomena, as understood within contemporary neuroscience. Given the foregoing discussion concerning reduction in general, we may assume that a reduction of this kind is improbable unless the categories and their embedding framework are fundamentally correct.

Should the psychological categories undergo major redrawing and reconfiguration, should our understanding of the psychological capacities be revamped quite radically, then certainly it will not be the original, old-time folk categories which will figure in any reduction to neurobiology. Rather, it will be the newly drawn categories in a newly configured psychological theory. And if I am right in envisaging a co-evolutionary development of the psychological and neurobiological theories, then reductive ‘feelers’, at the very least, are on the cards. Instead of talking in general terms about the reduction of psychology to neuroscience, I want now to focus more narrowly on one domain within the wider psychological framework, namely consciousness.

The question before us is this: can we get a reductive explanation for consciousness? That is, can we understand in neurobiological terms what it is for an organism to be conscious? A straightforward tactic for a materialist is to search for the neurobiological mechanism which results in an organism being conscious, secure in the knowledge that substance dualism is, on empirical grounds, flatly down and out, and hopeful that, in the long run, neuroscientific techniques will reveal the inner secret of what seems terribly mysterious about
consciousness. Nor is it difficult to conjure the air of deep mystery about consciousness. Look at a slice of nervous tissue, watch the oscilloscope during recording from a single cell, trace the circuit in the sea hare Aplysia that permits the animal to habituate to harmless stimuli; and notice the phenomenological qualities of feeling delight or sympathy, or of seeing blue or hearing a sigh: these seem infinitely beyond the explanatory reach of neurobiology. How on earth can feeling a pain result from ions passing across a membrane?

Nevertheless, our current bafflement does not of itself show that no neurobiological understanding is forthcoming. The nature of light, fire, the heavens, of reproduction and life have also seemed intractably mysterious, yet we now understand quite a lot about them. Pre-scientific intuitions are often the products of a wider framework which is itself a skewed model of reality, and this framework sometimes contributes to bewilderment rather than to clarification. Part of the task of science is to press on even at the risk of shocking our intuitions and revising our pre-theoretic frameworks.

Is it just a blind materialist faith that consciousness is amenable to neurobiological explanation? No, because everything we know about biology, evolution, neuroscience, physics, and chemistry suggests that substance dualism cannot be right, and that mental states are states of the brain. How we see, hear, walk, catch a ball, plan, and problem-solve are operations of the physical brain. For none of these achievements do we have a complete explanation, and they remain, one and all, mysterious. Catching an outfield fly is every bit as mysterious, it seems to me, as consciousness. But in all these instances psychology and neuroscience have made progress, and we can begin to see what neurobiological explanations of behaviour will look like. It is, I suppose, possible that we will not after all find satisfactory neurobiological explanations for these things, perhaps because the job is too hard, or even perhaps because they actually are the effects of operations in ‘spooky stuff’. Logic does not absolutely rule that out, but science does make it highly improbable, and I prefer to go with the probabilities.

The materialist strategy, accordingly, is to seek the neurobiological substrate for consciousness. One way to conceive of the task is on the model of the search for and discovery of the structure of DNA. As is well known, this discovery was profoundly important for genetics and for molecular biology in general, minimally because it allowed us to begin to really explain how traits could be transmitted from parents to offspring. But the implications went much further than that. Since all living things have DNA (except viruses, assuming they are living, which have only RNA), it also yielded a clue to the bond between all living things. In addition, understanding the genetic code provided the key to understanding the microbiological substrate for evolutionary processes. Inspired by this analogy, we may envision the possibility of discovering the neurobiological mechanism of consciousness, and such a discovery might similarly serve to extend explanations, provoke research, and perhaps yield a clue to the bond between human consciousness and that of other animals. To underestimate the case, such a discovery would be of tremendous importance.

A research vision can have a powerful shaping influence on how people think about a problem and organize their experiments. Thinking about the neurobiological basis for consciousness on the model of looking for the structure
of DNA has much to recommend it, principally because it is very direct: consciousness must be a property of the brain, so let us look for the neuronal configurations that produce it. Hard the problem may be, but let us at least find out how hard. Even granting these singular virtues, it may nevertheless be profitable to examine this research vision with greater circumspection.

An important question to ask is this: what precisely was the phenomenon or capacity explained by discovering the structure of DNA? What exactly was the problem solved? When we focus this closely, the answer is that first and foremost it solved the copying problem. It explained how genes could be copied, and how, even when a mutant allele was produced by irradiation, it could nevertheless be copied in the gametes. Most basically, it explained how DNA could replicate genetic information. Discovering how genetic information could be replicated was, of course, an important piece in the puzzle concerning the relation between genotype and phenotype, and how phenotypic traits are transmitted. But many additional questions would first have to be answered before the larger puzzle could be solved. For starters, much would have to be learned about the role of enzymes, of transfer RNA and messenger RNA, and so forth. It was only a very small, if crucial, piece of the puzzle of reproduction of organisms. On that topic, much still remains to be learned about cell differentiation and embryological development. Notice, however, that the copying problem itself was quite well-defined; it was a fairly specific and constrained problem. Moreover, research in classical genetics provided a substantial and experimentally integrated framework within which to understand the general phenomenon of genetic transmission and gene expression, and biochemistry also provided a rich theoretical basis within which to conceive of hypotheses and to test them experimentally.

Thus the discovery of the genetic code was made within the context of a substantial macro-level theory (classical genetics), a substantial micro-level theory (biochemistry), and a specific, reasonably well-defined phenomenon which needed to be explained (replication), and in the vicinity of other highly constrained problems concerning the relation between chromosomal material and phenotypic traits.

Given the research vision based on the discovery of the genetic code, the question to be asked is this: Is there comparable infrastructure to support the neurobiological investigation of consciousness, or at least enough to motivate experimental research on its neurobiological basis? Probably not. One problem concerns the background micro-structural theory. Although we know quite a lot at the neurobiological level about how individual neurons work, we know very little about the dynamics of circuits—of how neural networks achieve their effects. But the more serious problem is this: at the psychological level we do not have anything remotely as well-worked out for consciousness as the principles of classical genetics were worked out for genes and genetic transmission. Consequently, the explanandum—consciousness—is not at all well-defined or circumscribed. In the case of gene replication (the explanandum for DNA structure), scientists had a pretty good handle on what the phenomenon was that needed to be explained; by contrast, in the case of consciousness, even the range and nature of the phenomena to be explained is notoriously unclear and vaguely indicated.
To make matters worse, what we now lump together as ‘consciousness’ may be not so much a unitary phenomenon admitting of a unitary explanation, but a rag-bag of sundry effects requiring a set of quite different explanations. This want of specificity is very serious, for if we do not know what the phenomena are for which we are looking for an explanation, launching our investigation is rather Pickwickian. We may find ourselves echoing the inimitable words of Raymond Smullyan (1983, p. 19): ‘now that we have lost sight of our goal, we must redouble our efforts’. These ideas need considerable discussion to justify, however, and so far I have made claims without justification. That is the focus of the next section.

**Consciousness: what is the explanandum?**

What are the phenomena to be explained? It seems easy enough to start answering this question, and the favoured place to start is awareness of sensations. Experiencing colours, pains, sounds, smells, tastes, and so forth are typically treated as paradigmatic cases of what it is to be conscious (Dennett 1988, this volume). Consciousness, accordingly, may sometimes be thought of as that general state in which we are aware of sensations: visual, auditory, tactile, olfactory, etc. There are other, more subtle, receptors which yield sensations we may be conscious of. Proprioceptors give us awareness of the position of limbs, the vestibular apparatus gives us awareness of acceleration and balance.

There can also be awareness of visceral circumstance. Sacks’ (1985, chapter 3) description of Christina, whose neuropathy had left her without any proprioceptive information, teaches us how important is one’s body image and of knowing without looking where the limbs are. Of course when one is conscious, one is aware only of a subset of signals relayed from peripheral receptors. So already there is an interesting question: What are the differences in the state of the conscious brain when it is aware of an input signal and when it is not so aware?

This question provokes us to see that there are other aspects to consciousness, less ‘juicy’ (immediately salient) than experiencing sensations, but important as we consider the function of consciousness. Planning, deliberating, and deciding appear to be central. For certain tasks, the organism needs to plan, perhaps to imagine the task before actually performing it, to pay attention to certain sensory signals and not others, to prepare, lie in wait, anticipate, figure things out, and to search for sensory signals. In general, sensori-motor control is an absolutely fundamental function of the nervous systems of animals. Animals are essentially movers, and their motor behaviour must fit the circumstances or they will fail to reproduce.

In some role, yet to be determined, short-term memory appears to figure in the story of what it is for an organism to be conscious, at least because reacting to recently occurring events is an essential part of sensori-motor control. We need to take into account events that happened a few moments earlier, or minutes earlier, in order to respond. Even more mysterious, we are
also aware of time—of temporal duration, sequences, ordering, of what is long ago and what is not so long ago. And we are aware of space, of where things are in relation to one another, where oneself is—at home, away from home, etc.

We are additionally aware of ourselves as a self—as a thing distinct from other objects, as having a certain identity or coherence through time. A person is aware of specific internal states, desires, hopes, motives, intentions, images, imaginings, and day-dreams. Sometimes a person is aware of even more complex internal states, such as that someone looks familiar, or that he has forgotten someone’s name, or that he is confused about the effect of reserpine on aminergic cells. Remembering sometimes involves awareness—recollecting the spelling of ‘pneumonia,’ the location of the barn relative to the creek, whether Washington had a beard. A person can be aware of various emotions, such as embarrassment, anxiety, satisfaction, disgust, etc. When we are aware of these things, we can (normally) report them, and this seems to be a central feature of human introspection. In sleep, we are conscious, in the sense that we are arousable and not in coma, but during some stages of sleep, awareness seems entirely absent, while during REM (rapid eye movement) sleep we are aware of dreams but not of the real external world.

This canvas and catalogue of the various aspects of consciousness is done against the background question: What is the explanandum? One cannot help noticing two prominent things: (1) the features considered in the realm of the phenomenon are remarkably diverse, and (2) words such as ‘experience,’ ‘awareness,’ ‘introspection’ are used to specify the phenomenon, but they are as equally ill-defined as ‘consciousness.’ Before I consider how not to lose heart in the face of such confusing complexity, I want first to make three philosophical observations.

**Defining the words**

There is a fatal temptation to try to deal with the problem of the vagueness of ‘consciousness’ and the related set (awareness, experience, introspection, contemplation, reflection, etc.) by giving stipulative definitions, guided perhaps by certain deep and sometimes quirky hunches about the nature and function of consciousness in nervous systems. The difficulty is that, if we are not clear about the phenomena that are meant to be captured under ‘consciousness,’ stipulative definitions will not help significantly. What is needed is a genuine *theory*, one which has real predictive and explanatory power, fits in with other parts of psychology and neurobiology, and is experimentally based. Without such surrounding structure, new definitions tend to look arbitrary. Moreover, as a matter of sociological fact, they are inevitably counter-examples to death by simple appeals to the core use of the relevant words, to show, in effect, that the new definitions take us beyond the core use with no significant gain in empirical insight or utility.

The scenario, familiar to anyone who has dipped into the literature or attended conferences on consciousness, goes like this:

**Proposal:** Let us distinguish between two kinds of consciousness, one which involves only (say) discrimination of a physical difference (call this C₁), and C₂, which additionally carries the ability to report.
Reply: But that means thermostats (bit of litmus paper, etc.) are conscious!

Proposal: Well, yes, that is a consequence of my definition, but of course thermostats are only C1 — conscious in this narrow sense.

Reply: Fine, but we want to know about the actual phenomenon, about consciousness itself, not about some other phenomenon you have decided to attach the word to.

The general problem is that, unless the new definitions are empirically motivated, no one can see much point in buying into the extensions or re-categorizations. The idea that if only we could get the words correctly defined then we would understand the phenomenon is seductive but misguided. The words will come to have a more precise meaning as they are more deeply embedded within the framework of an empirical theory. This is not to say that no clarification of meaning is useful, but only that, in the absence of additional empirical hand-holds, what we can accomplish through new definition and re-definition alone is actually quite limited. Recognizing that significant meaning change accompanies rather than precedes empirical discovery, some philosophers have called the ‘define-the-words-first’ strategy the heartbreak of premature definition. The point is that if we understood more about the phenomenon we would know what to say ‘consciousness’ means.

Natural kinds
The list of elements that go into the pot when we are talking about consciousness and awareness is very diverse, and I want now to dwell on that diversity in order to raise the possibility that there may not in fact be a unitary phenomenon at all. That is, there may not be a single type of neuronal configuration which is the substrate of all those cases, unifying the apparent diversity. Intuitively, perhaps, it seems that consciousness is a natural kind, in the sense that it is a unitary sort of phenomenon, yielding to a single, integrated, unified explanation. The paradigm case of a non-natural kind is ‘gems’ or ‘weeds’. Gems turn out to be whatever rare and not very useful stones a given culture chooses to value as precious, and hence will vary as a function of culture, scarcity, etc. Weeds are plants that grow vigourously in a given locale and which a particular gardener person happens not to want in his garden, and what is a weed to one gardener is a prized botanical specimen to another.

By contrast, ‘muscle cell’, ‘gamete’, ‘protein’, ‘acid’, ‘electron’ are natural kinds—or, at least, at this stage of science we believe them to be. There are generalizations and counter-factuals that hold true of acids in virtue of acids having the chemical properties and micro-structure they actually have, where these properties are not relativized to the preferences, whims, interests, accidental collections of persons who happen to interact with the acids.

The possibility I want to entertain is that within the very broad class of states we call being aware or being conscious there are subclasses which are amenable to different neurobiological explanations (Churchland 1985a).

---

1 For a more complete discussion of these issues, see Churchland (1986).
That we do in fact use a common expression to cover the diversity is not in itself a decisive consideration, since it is a common feature of folk theories that categorization may be tied to ‘appearances’, and the transition to scientific theories involves re-categorization, as Nature is carved at her proverbial joints. Thus ‘fire’ was used to classify not only burning wood, but also the activity on the sun and various stars (actually fusion), lightning (actually electrically induced incandescence), the Northern lights (actually spectral emission), and fire-flies (actually phosphorescence). As we now understand matters, only some of these things involve oxidation, and some processes which do involve oxidation, namely rusting, tarnishing, and metabolism, are not on the ‘Fire’ list. The classification of these latter processes with the burning of carbon-rich materials was really very surprising, because pre-scientifically the intuition might have been that being very hot was criterial for being the same type of process as wood-burning. It is only when we understand the deeper nature of the phenomenon that we begin to see how the old classification was skewed. Thus the case of fire illustrates both how the intuitive classification can be re-drawn, and how the new classification can pull together superficially diverse phenomena once the underlying theory is available.

Could it be that ‘consciousness’ in a similar manner is a classification which could be re-drawn as we discover more about the nature of the mind-brain? Perhaps. I have already touched on the diversity in kinds of awareness, but further points should be made. There are some rather obvious differences between, for example, (a) being aware of seeing blue and (b) being aware of ambivalent feelings about hang-gliding, and (c) being aware that one has a tendency to jump to conclusions. The first point is that introspective awareness is internally generated, whereas the visual experience of blue is standardly dependent on receptor stimulation. There are no receptors for introspection, though external stimuli may of course play some role in initiating the process. But my thought that ‘there is no axiomatization of arithmetic that is both consistent and complete’ is not caused by an external event isomorphic to the thought.

Planning, choosing, and so forth probably involve long-term memory and short-term memory in a manner quite different from experiencing the colour blue. A bird’s plan to build a nest may involve use of past history about sites, predators, availability of materials, and so forth. Awareness of movement in one’s peripheral vision probably is less connected with memory, but it is not much like a ‘hot’ sensation such as seeing blue either. Awareness of time presumably has a special relation to memory not shared by tactile experiences or by body-image awareness.

We can note that having a blue visual experience is not (usually) under voluntary control, except as we do a visual search for blue things. Choosing, reflecting, and deliberating do all seem to involve a different degree of voluntariness. It is not that when I want to plan my trip this summer I do an introspective search for a ‘trip experience’, comparable to searching the external environment for a ‘blue experience’ or a ‘smells-like-a-violet’ experience; it is rather, to put it crudely, that I create the plan as I go. Sense of self, on the other hand, does not seem to have a large voluntary component, though of course we do undertake self-improvement and we self-exhort, and so forth.
But the self is just there. Awareness of being ill, of being well again, of having changed in character, of feeling energetic, tired, or fed up seem different yet again. Some of them at least appear to require the relevant background concepts (e.g. awareness of being ‘fed up’), others seem more minimal in that regard, and some involve language essentially, such as my thought about the Godel result. Dreaming experiences are in certain respects very unlike waking experiences. As Hobson et al. (1987) put it, waking and dreaming seem on the opposite ends of three continua:

(i) stability-instability (of orientation—where one is in space and time);
(ii) congruity-incongruity (of context—what is going on, who belongs, what belongs together);
(iii) confidence-uncertainty (of concept—who the characters are, what the objects are).

Awareness in the dreaming state normally also lacks a self-reflective dimension. Characteristically, one is not surprised or amazed even by very stark incongruity, or by orientational instability. Were such dream events to occur in waking life, we should certainly be aware of and puzzled by their bizarreness. A visit from a long-dead grandfather arriving on a flying fish during a baseball game would, in waking life, be cause for considerable consternation; in a dream, we pretty much just take it as unexceptional (Hobson 1984). Some ‘lucidity’ can be achieved with practice, so that it is possible for a dreamer to recognize that he is dreaming and that the dream events are not really happening. However, this sort of control and integration of dream data and waking data is not easy to maintain, requires considerable effort, and tends to slip rather quickly. It can occur in the untrained dreamer, but, although the brief lucid period may be memorable, it is in fact a very rare occurrence.

At this stage it is not clear exactly what is implied by this really quite impressive diversity. We simply do not know enough to decide whether the diversity is essentially an articulation of the same basic organizing principles, nor whether we have a genuinely mixed bag, nor, in general, how to conceptualize what is going on. It is, however, important to be alerted both to the amorphous nature of the explanandum and the possibility that consciousness is not a single, unified kind, for it provokes us to focus more closely on the question of research strategies. At bottom, the interesting question for research is this: Where can we get our empirical hooks in and begin to get answers constrained by both psychological and neurophysiological data? I suspect the diversity helps to emphasize the importance of not adopting an exclusively top-down research strategy to investigate the nature of consciousness. Important things have been and remain to be discovered at the psychological level, but research conducted exclusively at that level will not enable us to get at the principles of operation, or at the significance and function of the variety of phenomena collected under the rubric of conscious awareness.

A useful result, therefore, of wallowing in the diversity and dwelling on the ill-defined nature of ‘consciousness’ is that it provokes us to find a more narrowly circumscribed domain—a domain where a phenomenon can be reasonably well-defined and where there are neuroscientific techniques for
investigating the neurobiological basis of the phenomenon. (Certainly I do not want to wallow in the mysteries of consciousness merely for the sake of wallowing.) Left open until much later is the question of how the favoured phenomenon does or does not connect with other aspects we pre-theoretically include as part of consciousness. That is, the strategy is to find a domain, relevant to the consciousness, where we have something roughly comparable in specificity and surrounding theoretical structure to the copying problem in genetics. The domain which perhaps comes closest is the cycle of waking-synchronized sleep-REM sleep, and this will be discussed in the next section.

Denormalizing data
If, following Kuhn’s (1962) helpful simplifications, we think of ‘normal science’ as those periods in a science’s development where there is a governing paradigm, and hence a relatively stable categorial framework, then ‘denormalizing data’ are those data which threaten not merely to falsify a local hypothesis, but also to undermine our confidence that the governing framework is basically correct. In the clinical neurology literature, there are studies which collectively have denormalizing implications for the categories of ‘consciousness’, ‘awareness’, and related members of the group. These data are denormalizing because they are so profoundly at variance with the conventional wisdom regarding the nature of consciousness that they make us wonder whether the conventional wisdom is itself just shared misconception and thus whether the categories in question fail to do justice to the real nature of the phenomena.

For example, it is generally assumed to be dead obvious that if someone can report on some visual aspect in the environment then he must be consciously aware of it. Yet the blindsight studies (Weiskrantz et al. 1974; Weiskrantz 1986, 1988, this volume) show that subjects who are blind in certain parts of their visual field as a result of lesions to their visual cortex may none the less be able to make such reports. That is, they can point with great accuracy to where in the blind area of their visual field a light is shining, indicate whether a light is diagonal or horizontal, and so forth. Since there are pathways from the retina to brain areas other than the visual cortex, for example to the superior colliculus, this is not a miracle, but it does present a challenge to our naive assumptions about the connection between reportability and consciousness awareness.

Commissurotomized subjects provide denormalizing data for that ‘obvious’ assumption according to which the conscious self is a single, unanalysable unity. On this assumption, if the self reports a conscious experience, there is no other part of the self which could be unaware of that experience. It is now well known that data can be presented to one hemisphere of a split-brain patient so that the other hemisphere will be unaware of the data (Gazzaniga and LeDoux 1978; also see Churchland 1986). Multiple personalities, striking and extraordinary as they are, also suggest that there is much to be learned about the so-called unity and coherence of the self.

Denial syndromes are especially puzzling, and perhaps the most remarkable is blindness denial (Anton’s syndrome). Sometimes a subject who suffers a sudden onset of blindness as a result of a lesion to the visual cortex will behave as though he is unaware that he cannot see. These patients insist that their
vision is fine, though they may hint that the lights are a bit dim, and they confabulate coherently, if erroneously, in answer to questions about what is in their environment. Unlike patients who know they are blind, these patients do not adjust, and persist in bumping into furniture. This syndrome seems almost impossible, because, intuitively, we take it as part of the very concept of consciousness that if one is not having visual experience then one is aware, unequivocally and without inference or reasoning, that one is not having visual experiences.

Hemineglect also has a denormalizing effect on what we conventionally understand about the nature of awareness. It seems obvious that so long as one is capable of having visual experiences in both visual hemifields, one will be aware of both hemifields and of experiences in both hemifields. We naively assume that if you are visually aware, then you are visually aware, and awareness does not just stop at some point in space. Yet that is exactly what happens for patients with hemineglect, whose brain lesion is frequently in the right parietal cortex. They tend not to be aware of left hemispace, and thus do not groom the left side of the body, do not eat food on the left side of the plate, do not notice events in left hemispace, do not complete the left side of drawings and so on (Heilman 1985). Moreover, as Bisiach and Luzzatti (1978) have shown, hemineglect patients also show a deficit in the capacity to form images of left hemispace.

Somnambulism challenges the assumption that consciousness and control must go hand in hand. It is part of the conventional wisdom that the conscious self is in control, and what we are in control of we are also conscious of. Somnambulism does not occur during the dreaming phase of sleep, but during synchronized sleep, when we apparently are not consciously aware at all—of the environment or of dreams. There are as well common, everyday occurrences of the decoupling of consciousness and control, such as when a skill is automatized, allowing control either without attention or with intermittent attention.

This list of denormalizing data is by no means complete, but I want to conclude it here with the observation that not all the denormalizing data are pathological. The regular appearance of confabulation in ordinary, everyday explanations of one’s behaviour suggests that we do not have anything like unmediated access to our desires, beliefs, decisions, or intentions. As the work of social psychologists (see Nisbett and Ross 1980) demonstrates, in controlled experiments subjects often give explanations for a choice, a preference, a decision when we know that their explanation is in fact incorrect. For example, women shoppers in a shopping mall were asked to choose a pair of panty hose from a table. Unknown to the women, all the hose were identical. When their choices were analysed, a pronounced position effect was evident, in that the women tended to choose hose in the rightmost position. However their self-generated explanations of their choices referred to (non-existent) differences, such as having greater sheerness, a preferred colour, and so forth. None cited the dominant causal factor, namely that they tended to choose objects in the rightmost position. This confabulation seems to be normal, inveterate, and habitual, and does not involve anything like Freudian repression, nor is it done with deliberate or conscious awareness. There are many other studies showing
similar results, and they strongly suggest that theorizing about one’s own motives, intentions, and other internal states is not anomalous, but commonplace. As Paul Churchland (1985b) puts it, these assorted paradoxical cases in neurology and psychology stand to folk psychology as near-luminous velocities stand to Newtonian mechanics, as very large masses stand to Newtonian gravitational theory, and as very low or very high pressures stand to the classical gas laws. That is, when applied to phenomena outside the comfortable bubble of ‘familiar’ cases, the categorical and explanatory resources of the old theory prove entirely inadequate. Folk psychology seems to work as well as it does so long as its range of application is narrowly confined—confined to cases where it works.

Waking, sleeping, and dreaming

If there is a domain relevant to consciousness which has sufficient supporting infrastructure and surrounding theory to enable experimental discovery, it is the sleep–dream–awake cycle. There are a number of reasons why this looks like the kind of domain where research will be fruitful. First, there are very striking differences at the introspective and phenomenological level. In deep sleep, one is seldom aware at all; in dreaming, one is aware of internally generated images; while being awake is the paradigmatic conscious state, where we are aware of external events and the brain displays complex sensori-motor control. These factors suggest that the cycle is relevant to consciousness, pre-scientifically understood.

Second, there are robust behavioural criteria for identifying the different states, which means that objective measures may be made and animal models can be used. It also means that many invasive neurophysiological and anatomical techniques can be used to determine correlations between physiological and psychological conditions.

Third, neuroscientific techniques are beginning to reveal the microstructural properties distinguishing the three states and the neuronal generator for producing shifts in state. Experimental data show that waking, sleeping, and dreaming as described at the psychological level correlate with large-scale functional states involving wide areas of the brain; i.e. with what researchers in

---

4 Gazzaniga and LeDoux (1978) discuss the confabulation of split-brain subjects whose left hemispheres try to explain behaviour initiated by the right when the right hemisphere was given an instruction by the experimenter. Even when the subjects know the experimental situation, and know that as a split-brain subject one hemisphere does not have access to the information of the other, confabulation—natural, facile, coherent confabulation—is typically observed. What does not happen is that the subject’s left hemisphere will say: ‘Well, I don’t know why I walked to the door, but I guess my right hemisphere must have got information unavailable to me and went ahead with its plan.’ Instead, a likely story will be presented: ‘I wanted to get a drink, so I went to walk out.’

5 Stated thus, the explanatory power of folk psychology seems to be rather crudely trivialized. Admittedly, any theory quite properly suppresses a good deal of empirical evidence as irrelevant and inevitable ‘noise’, if the theory is deeply wrong, however, the only way to identify relevant and refuting ‘signals’ in that ‘noise’ is by actively exploring alternative theoretical approaches. Moreover, it is deeply suspicious if a theory’s bacon is regularly saved by rejecting troublesome data as noise; at some point one has to take seriously the possibility that the theory does not work very well even in its favoured domain.
the field call ‘behavioural states’ of the central nervous system (Hobson 1978; Hobson and Steriade 1986; Hobson et al. 1986). The general concept of ‘state’ as it is being developed in sleep and dreaming research will probably be a pivotal conceptual advance as we address other problems such as focal attention, moods, mental set, control, and the voluntary/involuntary distinction.

Rapid eye movement sleep is typical of mammals, with the probable exception of the voluntary breathers such as dolphins and whales (Crick and Mitchison 1986), and, rather curiously, the echidna (Crick and Mitchison 1983). Electroencephalogram (EEG) studies have shown that the brain of the sleeping organism exhibits a rhythmic pattern in shifting from light sleep, to deep sleep, to REM sleep, and around again. As the sleeping period progresses, the REM periods lengthen and the non-REM (NREM) sleep periods shorten (Fig. 13.1). The time constant for states varies as a function of development, and infants spend most of their sleep time (about 80 per cent or eight hours) in REM sleep, which decreases until about adolescence (25 per cent or one-and-a-half hours) when it levels out. There is a small upward swing around age 40 years, declining steadily thereafter. REM-deprived subjects often show REM rebound (longer REM periods when permitted), but some subjects treated for depression with REM-suppressing drugs do not show these effects.

Fig. 13.1.
Ultradian sleep cycle of NREM and REM sleep shown in detailed sleep-stage graphs of 3 human subjects (A), and REM sleep periodograms of 15 human subjects (B). In polysomnograms of A, note the typical preponderance of the deepest stages (III and IV) of NREM sleep in the first two or three cycles of the night; REM sleep is correspondingly brief (subjects 1 and 2) or even aborted (subject 3). During the last two cycles of the night, NREM sleep is restricted to the light stage (II), and REM periods occupy proportionally more of the time, with individual episodes often exceeding 60 minutes (all 3 subjects). Same tendency to increase REM sleep duration is seen in B. In these records, all of which begin at sleep onset, not clock time, note the variable latency to onset of first (usually short) REM sleep epoch. Thereafter inter-REM period length is relatively constant. For both A and B time is in hours. [Reproduced from Hobson and Steriade (1986). Neuronal basis of behavioural state control. In Handbook of Physiology, (ed. V. Mountcastle) pp. 701-823. American Physiological Society.]
There are five, principal, identifying indicators of REM sleep: muscular atonia, desynchronized EEG, which resembles the waking EEG, irregular muscular twitching, PGO (pontine-geniculo-occipital) waves, and rapid eye movements (Fig. 13.2). The evidence, derived principally by arousing subjects and asking for reports, indicates that subjects dream during REM sleep (summarized in Hobson et al. 1986.)

Salient psychological features of REM sleep are familiar: the dreamer is unaware of external events and is aware of internally generated events. It is probably very significant for understanding the neurobiological basis that dream events are virtually never remembered unless the dreamer is awakened. While we may recall rather trivial events in waking life, even highly charged emotional events in dreaming are not recalled unless the dreamer is awakened (Hobson 1984). Also significant is the bizarreness of dreams and the lack of

![Fig. 13.2.](image)

Characterization of behavioural states in humans. State of waking, NREM sleep, and REM sleep have behavioural, polygraphic, and psychological manifestations. Sample tracings of 3 variables used to distinguish state are also shown: electromyogram (EMG), which is highest in waking, intermediate in NREM sleep, and lowest in REM sleep; and electroencephalogram (EEG) and electro-oculogram (EOG), which are both activated in waking and REM sleep and inactivated in NREM sleep. Each sample record is approximately 20 seconds. The three lower channels describe other subjective and objective state variables. [Reproduced from Hobson and Steriade (1986). Neuronal basis of behavioural state control. In Handbook of Physiology, (ed. V. Mountcastle) pp. 701-823. American Physiological Society.]
insight of the dreamer into his actual state (see earlier discussion, p.12; Hobson et al., 1987). These are features at the psychological level that a neurobiological account ought to explain. As we shall see, impressive progress has been made in providing just such explanations.

Sensory input is gated during NREM sleep at the thalamo-cortical level via hyperpolarization, but in REM sleep it is gated at the periphery. Motor output is blocked during NREM sleep by disfacilitation of brainstem neurons, but, by contrast, during REM sleep there is active inhibition of motor neurons by neurons in the pons. This produces a kind of paralysis, which can be abolished by making a small lesion near the locus coeruleus, whereupon the animal will move about during REM sleep, apparently in accordance with dream events. There are ten recorded patients (all male) who have REM without atonia. They behave in accordance with the motor demands of their dream narrative, and consequently crash into walls and furniture. This behaviour contrasts quite markedly with somnambulists who typically manœuvre quite well, open doors, and so on.

Sensory input is evidently not completely blocked or completely unprocessed, however, since adult sleepers show considerable navigational and orientation facility (they do not fall out of bed, they tend not to crash into each other, etc.) and sleepers can be aroused from synchronized sleep (also called deep sleep or S-sleep) or REM sleep by specific auditory stimuli, such as a whimper of a baby, even though they may be deaf to louder sounds the brain knows it can safely ignore. Arousal from S-sleep is normally easier than arousal from REM sleep.

The neurophysiological level has been investigated using mainly three types of technique: (1) recordings for correlation studies, with the aim of determining the behaviour of populations of neurons under different behavioural states, (2) lesions, to determine the effect of removing a section of tissue, and (3) pharmacological intervention, most recently, by micro-injections of specific chemicals to mimic, enhance, or disrupt the function of relevant endogenous neurochemicals. Neuroanatomical techniques have been used to determine projection paths of various neuronal groups, and neuropharmacological techniques are used to determine the kinds of neurochemicals involved in the crucial structures and the role of those chemicals in the sleep-dream cycle.

So far the neurobiological story is incomplete, and I shall present only the barest outline of the theory⁶, but even that much yields a picture that underscores the fruitfulness of experimental investigation of a seemingly unapproachable mystery concerning consciousness (Hobson and Steriade 1986). The main anatomical structure involved is the brain stem and, even more specifically, the reticular formation in the pons (Fig. 13.3). It now appears that there are three principal elements underpinning behavioural states and their cycles. The first is the aminergic neurons in the locus coeruleus (in the pons). They are very active during the waking state, then they decrease their activity with the onset of S-sleep, and they have a very low level of activity during REM sleep. There are also aminergic neurons in the dorsal raphe nucleus which may

---

⁶ The best review articles are Hobson (1984), Hobson and Steriade (1986), and Hobson et al. (1986).
have a modulatory role in the states. Second, there is a special group of neurons with huge cell bodies in the tegmental field (hereafter referred to as ‘FTG neurons’) which are very active during REM sleep. These cells are believed to release acetylcholine as a neurotransmitter and have receptor sites for acetylcholine. Injecting the centre of this area with carbachol, an acetylcholine agonist, produces REM sleep in the animal (Steriade and Hobson 1976). The third element is that in general the activity of cholinergic neurons in the brain stem is high both in waking and in REM sleep, but low in S-sleep (Figs. 13.4 and 13.5).

The inactivity of the aminergic neurons during REM sleep begins to make sense when we see that they project diffusely all over the brain (Fig. 13.6), including the thalamus, the hypothalamus, and all over the cortex. A decrease in their activity suggests that their dynamics is an important element in the account of the REM state and how it differs from the awake state. In the waking state, the locus coeruleus is crucially involved in alerting and orienting the brain to particular sensory inputs.

Does the reduction of the aminergic output from the pons mean that in sleep and dreaming the brain just loaf and is generally inactive? Not at all, and this is what is especially intriguing about the diminution of aminergic activity. Neuronal activity in REM sleep and in waking is comparable, and activity during NREM sleep is down by only about 20 per cent. This includes such ostensibly unlikely places as the cerebellum, which busily hums away during
REM sleep. It is the ratio of aminergic activity to activity involving other neurochemicals in those regions receiving aminergic projections that distinguishes REM sleep from the waking state. The acetylcholine/monoamine\(^7\) ratio is substantially higher during REM sleep, and this is a rather subtle effect which does not show up on gross measures of neuronal activity.

The functional significance of this change in ratio is not yet understood, but because aminergic neurochemicals have an important role in learning in the systems so far studied, a leading hypothesis is that the major effect of decreased aminergic activity is a reduction in plasticity (Flicker et al. 1981). As Hobson (1984, p. 250) suggests, ‘it is the postsynaptic instruction about the messages—rather than the messages themselves—that differentiates the waking from the dreaming sleep state.’ Under conditions of decreased aminergic activity, cortical neurons may process as usual whatever information they receive, but the network does not learn and remember. Since the alerting and

---

\(^7\) The monoamines are norepinephrine (NE) from neurons in the locus coeruleus, and serotonin (5-HT) from neurons in the dorsal raphe nucleus.
orienting mechanism of the locus coeruleus is damped, attentional and reflective operations are also largely eliminated.

During REM sleep, increased activity by cholinergic cells in the pons activate the thalamus and the sensory cortex, and this activity shows up in polygraph recordings as the PGO (pontine-geniculate-occipital) waves. They also activate oculo-motor and vestibular nuclei in the brain stem. The cortex does not know whether the inputs it gets are generated internally or not, it just processes the data it gets (Hobson 1984), trying, as one might say, to make sense of whatever it gets. In the absence of external data to constrain the information that is received by the cortex, bizarre and haphazard images are bound to occur. Copies of the signals sent to the oculo-motor muscles from brain-stem structures (what is called ‘effference copy’) probably reach cortical areas, and may be a significant component in the information the cortical networks try to incorporate into a coherent picture (Fig. 13.7).

The brain-stem regions are very complex, and their architecture does not readily suggest functional hypotheses about what the networks are doing. Certainly many questions about the sleep-wake cycle remain. For example, is there a general utility resulting from the increased activity of cholinergic FTG cells during REM sleep? Why are there not just two states—S-sleep and being awake? What is the role of the other neurochemicals in the reticular formation, such as the peptides? Most pressingly, perhaps, precisely how does decreased

---

**Fig. 13.5.**
Schematic diagram showing the relation between sleep-wake stages as measured by the EEG, and activity of selective brain stem cells. REM sleep is indicated by a solid black line and appears in the third panel. The locus coeruleus cells begin to decrease in activity during deep sleep and show least activity during REM sleep. The giant pontine cells show the reverse pattern of behaviour. [Reproduced from Hobson et al. (1977). *An Experimental Portrait of the Sleeping Brain*, for the Carpenter Center for the Visual Arts, Harvard University. ©J. Allan Hobson and Hoffman-La Roche Inc.]
activity in aminergic cells of the locus coeruleus and the raphe affect the processing and network function of cells to which they project? Although we do not yet have answers to these questions, several points are worth emphasizing: (1) the questions are far more specific and well-defined than they could have been even 30 years ago, and (2) the result is that testable hypotheses can be generated.

Fig. 13.6.
Schematic diagram showing the very diffuse projection paths of neurons in the locus coeruleus of a rat brain. Upper figure is a view from the side of a section made through the middle of the brain (sagittal section); lower figure is a top view of the brain (horizontal section). DCB = dorsal catecholamine bundle. CTT = central tegmental tract. MFB = medial forebrain bundle. [Reproduced from Angevine and Cotman (1981) Principles of Neuroanatomy, by kind permission of Oxford University Press.]

Accordingly, I think there is much that is satisfying and inspiring in the research on the sleep-wake cycle. At the very least—and this is no mean thing—we have got our empirical hooks into a phenomenon that appears to be a global state (behavioural state) of the nervous system. If we conceive of the behavioural state in terms of a state space, the dimensions of which will include activity of aminergic cells, the neuro-modulatory role of peptides such as histamine and substance P, the activity of cholinergic cells, the state of motor neurons, sensory gating activity, and so forth, then we can begin to get a feel for what it is to be awake or in REM sleep in terms of the shape of a volume of the relevant state space. Naturally it is disappointing that we do not yet have physiological techniques for tapping into events at the network or circuit level, for we need to know not only the cellular but also the network effects of decreased aminergic activity. Nevertheless, computer modelling may be of considerable help in this regard, and new techniques may become available.

The general approach to global states also raises the question of how we might modify connectionist models (Sejnowski and Hinton 1985; Kienker et al. 1986; Lehky and Sejnowski, in press) to accommodate such features as global state control and the modulation of behavioural states. Although existing network models generally do not include such features, the network devised by Kienker et al. (1986) to solve figure-ground problems does have the kind of attentional mechanism which serves to bias the network. This enables the network to switch figure and ground, as in the familiar face-vaseline diagram.

If we are expecting to find in the neurobiology of sleep and dreaming a mechanism that turns the light off and turns it on again, then the odds are that we shall be frustrated. The data available so far indicate that the neurobiological
explanation is not going to look remotely like that, and consequently that being awake and being asleep should not be understood on that model. Instead, the theory will be a network-to-network account, involving the role of sensory input in constraining the sequence of internally generated images, of the functioning of short-term memory when sensory input is admitted, of emotions, of appropriate motor assembling, short-term memory, long-term memory, and goodness knows what else. If it is even roughly right, this theory indicates that we will have to re-think our pre-scientific intuitions concerning what it is to be conscious.

Concluding remarks

Where then does this leave the question of the neurobiological reduction of consciousness? It is possible, for all I have said, that consciousness will reduce smoothly, in the classical sense, to neurobiological phenomena; that is, in the way that light reduced to electromagnetic radiation (theory of optics to theory of electromagnetic radiation), or temperature in a gas to mean molecular kinetic energy. But smooth reductions are atypical in the history of science, and more commonly old theories are corrected, revised, and modified in various degree as points of reductive contact are made. For a theory to reduce smoothly, its categorial framework must be pretty much correct.

While the ways for a categorial framework to get it wrong are legion, getting it right is a much smaller target. As points of reductive contact are made between psychology and neurobiology, and as the two fields co-evolve, we can expect modification and corrections to the received wisdom at all levels of description. In such fashion, the various sciences of the mind-brain will likely converge upon unified explanations. Perhaps not of ‘consciousness’, for in the evolved framework that may have gone the way of ‘caloric fluid’ or ‘vital spirit’ (see pp. 274-277 above).

The prospect that consciousness might not yield to a classical reduction should not, of course, give any comfort to a dualist. For the classical reduction may be missed, not because there is spooky stuff or spooky properties, but because the folk psychological categories lack sufficient integrity to stick. In that case, the reduction is either a revisionary reduction—of revised psychological categories to neurobiological categories—or, more radically, an outright replacement of the old folk notion of consciousness with new and better large-scale concepts.
Acknowledgements

I am particularly indebted to Allan Hobson for teaching me about the importance of behavioural states, and to Francis Crick for provoking me to think about the appropriate domain for neurobiological investigation. Others whose conversation helped me in various ways are Edoardo Bisiach, Paul Churchland, Daniel Dennett, Warren Dow, Patricia Kitcher, Philip Kitcher, Tony Marcel, Terry Sejnowski, Steve Stich, and David Zipser.

References


