1 Consciousness as a Biological Phenomenon

The study of consciousness is at a stage where it is finally separating itself from pure philosophy, and rapidly becoming a new branch of empirical science. The motto in this emerging new field seems to be: "Toward a Science of Consciousness." Yet the field is still not sufficiently coherent to count as a genuine science. Therefore, greater unity is essential for the future of the field. In science, the theoretical and empirical coherence of a field is based on a well-established research program. A unified research program delivers the big picture of the order of nature in the respective domain. In this opening chapter of the book, the proposal is put forward that the unified science of consciousness should be anchored to the biological sciences. The big picture of the place of consciousness in nature would best be built on the (philosophically controversial) view that consciousness is a biological phenomenon within the brain. This view leads to a new and much clearer picture of the scientific explanation of consciousness.

The Study of Consciousness: Pretending to Be a Science, but Is It Really?

The official news from the frontline of the study of consciousness has been more or less the same for a number of years: We are steadily progressing *toward* "the science of consciousness." But what exactly is the nature of the progress made? The search for the science of consciousness was launched way back at the beginning of the 1990s, when it became increasingly clear that mainstream psychology and cognitive science had not been able (or even willing) to face the problem of consciousness. How long will it still take before we finally reach the goal, a genuine science of consciousness? A cynical observer, one who is getting impatient with the situation, might go so far as to suggest that no actual progress is being made: the whole field of consciousness studies seems to be nothing but a big philosophical and theoretical mess which is not moving in any particular direction.

I am certainly more optimistic than that about the prospects for a science of consciousness. But even so, much progress is still to be made before it will be possible to go to a bookstore and purchase something like a standard textbook of consciousness

science. That would be a clear sign that the science of consciousness is a fact, not a fiction. The transmission of scientific knowledge in established fields of science relies heavily on the advanced textbook (Giere, 1988) which as clearly as possible explains the theoretical and methodological basis of the branch of science in question and presents the central empirical approaches and core findings within one coherent framework, accepted by most of those working in the field. Conversely, the absence of a basic coherent story about the science of consciousness may be interpreted as a sign that a coherent body of scientific knowledge about the subject matter of consciousness is still hard to come by.

Why, precisely, are we still so far away from the science of consciousness? The scientific explanation of consciousness is not an entirely straightforward undertaking. It is difficult to say exactly what the task consists in, and which branches of science are supposed to take the responsibility for solving it. Resolution of the problem of consciousness has been sought in such fields as philosophy of mind, cognitive science, psychology, neuroscience, and even quantum physics. The really hard problem of consciousness—or explaining how *any* physical system could have *any* subjective, qualitative states in the first place—is still mostly thought to fall in the philosopher's domain rather than belong to empirical science. The general agreement appears to be that empirical science and experimentation can at best scratch the surface of the *easy* issues about consciousness (such as the cognitive or neural correlates of subjective experience), but the *really hard* issues (such as why anything in the physical world should have any subjective experiences at all) are best left for the philosopher to ponder. Yet, nobody expects that philosophers would actually *solve* them. The gap between the "hard" philosophical and the "easy" empirical problems of consciousness still remains wide-open (Chalmers, 1996).

Discussions in consciousness studies cover an incredibly large range of topics which leave almost no branch of science or the humanities untouched. Underlying this wide scope appear to be a multitude of dissimilar views about the basic nature of consciousness and about the empirical evidence and methodology most relevant to the theoretical explanation of consciousness.

There are some who think that one of the central issues in consciousness studies is whether consciousness is algorithmic or nonalgorithmic (e.g. Penrose, 1989), which is essentially a topic for mathematics and computational science. Some have suggested that consciousness is fundamentally related to the notion of information, which, if true, would make consciousness ubiquitous in the universe: consciousness should be present in all kinds of physical systems carrying information, even extremely simple ones (Chalmers, 1996). Some believe that the problem of consciousness is intimately tied to physical phenomena at the quantum level (e.g. Squires, 1998; Hameroff & Scott, 1998): wave-particle duality, nonlocal effects, the collapse of quantum fields, and other elusive microphysical phenomena residing at incomprehensibly small spatiotemporal scales,

way below the scales where brains, neurons, or even whole organic molecules reside. Others have argued that the key to consciousness lies at the level of specific receptors at certain types of synapses (e.g., the NMDA receptors at glutamatergic synapses [Flohr, 2000]). Moving towards still higher levels, still others have argued that there are specific "consciousness neurons" whose biophysical and neurophysiological properties differ from other neurons (Crick & Koch, 1998). Contrary to this view, some believe that consciousness cannot be defined as a property of specific single neurons but depends on the synchronized, coherent activity of large neural populations (Singer, 1994). Others have presented evidence that consciousness is associated with specific neural pathways in the brain, whereas the rest function in the absence of consciousness (Milner & Goodale, 1995). Then there are those who are reluctant to attribute consciousness to any part of the brain, but would say that consciousness is a property of the whole brain or the whole person rather than any of its parts. Consciousness is not viewed as based on particular biological phenomena in the brain, but instead on organism-environment interaction or even language and social communication (Dennett, 1991; O'Regan & Noë, 2001). The doctrines of "embodied cognition" and "enactive cognitive science" go so far as to say that mind and consciousness are not in the head but embodied in the whole organism embedded in its environment (Thompson, 2001; Thompson & Varela, 2001).

This wild variety of views, theories, and evidence argued to be relevant to the explanation of consciousness gives the impression that there are few, if any, fundamental theories or empirical findings that would form a *shared* core of the study of consciousness. If a novice enters the field and would become a consciousness researcher—a true professional in this fascinating new field—he might wonder what exactly it is supposed to include. What is the *philosophical* view of the nature of consciousness that the empirical *science* of consciousness is based on? What are the basic theories to be learned and mastered? What are the empirical results that every consciousness researcher is expected to know about? In any other field of established science, say biology, it is perfectly clear what the shared philosophical view is (and is not, say, vitalism or creationism). It is also rather obvious what a newcomer is expected to learn to become a professional in the field (simply observe which topics the standard advanced textbooks in the field cover), and what is only peripheral or totally irrelevant. One can furthermore be sure that any professional in the field will know about certain basic theories and phenomena (say, evolutionary theory, protein synthesis, or the structure of DNA), no matter what the exact specialization of that person is. But not so in the field of consciousness science. After taking a look at the massive literature published on consciousness since the early 1990s, the novice would have to wonder whether she should study quantum theory and learn about double-slit experiments, Schrödinger's cat, and the like? Or would she be better off studying the neurochemistry and neurotransmitter systems of the brain? What about single-cell response properties in the visual system of the monkey? Or the synchronization patterns of human scalp electroencephalography (EEG) during visual

perception or rapid eye movement (REM) sleep? Then there is a huge amount of data on neuropsychological patients: how important is that? What about the phenomenology of altered states of consciousness: hypnosis, dreaming, meditation, religious or aesthetic experience? Or the experimental psychology of change blindness or implicit learning? And then there are all the ancient and mysterious religious and contemplative practices of the East: perhaps they hold the key to the secrets of consciousness?

The difficulties of the would-be consciousness researcher are a reflection of the heterogeneity and incoherence that currently dominate the field of consciousness studies: there are neither any universally accepted basic theoretical principles nor any well-defined empirical core or database of consciousness science that most people working in the field would agree about and everybody would be required to master in order to be taken seriously as a professional in the field. Different people emphasize entirely different, often idiosyncratic theoretical principles and selected empirical findings; therefore, the field seems to have failed so far in establishing a shared "big picture" of consciousness. The big picture about the scientific study and explanation of consciousness would put the various theories and empirical findings into context and thereby indicate what the relative importance of different sorts of evidence is for the overall explanation of consciousness. Since such a big picture seems to be missing, consciousness studies do not (yet) have enough coherence to be called a "science" (in the sense that, say, biology or neuroscience deserves the label).

However, at the same time it appears that the study of consciousness would very much *want* to become a true, respectable science. There are large conferences being arranged that strive toward a *science* of consciousness; there is a professional organization for the *scientific* study of consciousness; there is an academic journal, *Consciousness and Cognition,* promoting a *natural-science* approach to consciousness; there have been several books published recently about the *science* of consciousness.

The Explanation of Consciousness Requires a Unified Research Program

To be worthy of the label "science" unfortunately requires more than just the frequent use of the words "science" and "scientific." It requires a *unified research program* that establishes the theoretical, philosophical, and methodological foundations of the particular science in question. Phenomena that are scientifically explained (or explainable) are explained within the context of some specific research program. Thus, to truly strive toward a science of consciousness is to seek the appropriate foundations for a coherent scientific research program on consciousness. Conversely, as long as no genuine research program on consciousness exists, no amount of throwing around the word "science" will help (in fact, too much talk might only raise the suspicion that the scientific status of the field must be somehow questionable). Instead, it might be wise to figure out what sort of thing a research program is and what would transform the mere study of consciousness into a genuine research program.

What does a coherent, systematic research program typically consist in? Philosophers of science have attempted to explicate what sort of an entity a scientific research program is supposed to be. While there have been considerable disagreements among them about this, there is also substantial agreement on the general characteristics of research programs (I will here mainly follow the ideas of Larry Laudan and Imre Lakatos).

A research program is a set of background assumptions shared by the community of researchers within that program. The background assumptions involve significant ontological and methodological commitments. These commitments can be thought of as general assumptions concerning the substantive entities and processes in the domain of study, and the appropriate methods to be used for investigating the phenomena and constructing the theories in that domain. The research program outlines the domain of application of the theories belonging to it, indicating which empirical questions are regarded as central to it and which belong to foreign domains or are only pseudoproblems. There are often several competing specific theories within a research program which present more detailed hypotheses of the ontology of the basic entities and their interaction. However, despite the variety of competing theories within a program, there is also significant integrity. The basic assumptions of the research program prevent any scientist working within the program from adopting theories that are incompatible with the metaphysics or methodology of the program. To adopt a view or to advocate a theory that is inconsistent with the core assumptions of the research program is to put oneself outside that program and to reject it.

A research program in many ways guides the research around the phenomena that fall into its domain. It constrains the interpretations that can be given to empirical findings, suggests promising lines of research, and points out suspicious lines of explanation. It allows the empirical results from a wide variety of empirical subdomains to be integrated into a coherent big picture of the subject matter.

A unified research program on consciousness would thus tell us what the relevance of the different philosophical approaches, theories, and empirical findings is to the scientific explanation of consciousness. Most importantly, it would provide us with a coherent, integrated big picture of the place of consciousness in the order of nature.

A Research Program Must Be Committed to Some Fundamental Assumptions: Not Anything Goes

Any research program needs to make *some* assumptions concerning the basic metaphysical and ontological status of the phenomena that are the objects of study. If we want to move toward a science of consciousness by aiming to establish a unified research program on consciousness, then some criteria have to be set for what sort of claims about consciousness can be made within the program, and what cannot be accepted because they are inconsistent with the core assumptions of the program.

The sad truth about the current state of consciousness studies appears to be that there is nothing even remotely resembling a unified scientific research program on consciousness. The field of consciousness studies includes a colorful variety of rival philosophical theories as well as a heterogeneous assortment of empirical studies of consciousness. While this may be a prosperous starting point on the way toward the science of consciousness, it cannot be the ultimate destination. A coherent science cannot be built on an incoherent and messy theoretical basis.

To establish a science of consciousness in the form of one widely shared and internally unified research program, at first several competing empirical research programs should emerge in the field. They will embrace dissimilar background assumptions, but each will try to develop a plausible overall account of consciousness that guides further research into, and theory about, the topic. Each program should be internally coherent and explicitly committed to some specific first principles that are both philosophically plausible and empirically fruitful. While empirical fruitfulness may be a feature that can be assessed in some objective manner in the long run, the plausibility of particular metaphysical commitments often comes down to subjective intuitions and tastes rather than objective facts, once internal conceptual coherence is established. Thus, what is plausible to one may be outrageous to another. In any case, coherent and plausible fundamental commitments—whatever their precise content may be—will pave the way for a true science of consciousness. The research program that in the long run shows the greatest overall explanatory power should be the most inviting candidate to develop into a unified science of consciousness.

Hence, if we want the diverse studies of consciousness to become a unified science in the long run, some degree of philosophical and theoretical freedom must be given up, at least by those who practice empirical research and commit themselves to a particular research program. Furthermore, any research program is associated with a specific empirical context. The phenomena under study must lie on some specific spatiotemporal scale and at some particular level of complexity in the world. The sciences that study phenomena at those scales and levels, or close to them, must be the ones most relevant to the study of consciousness. I am inclined to believe that if there is any proper place for a scientific research program on consciousness, it will be found in the empirical sciences that directly study the interface between neural and mental phenomena. Any viable research program on consciousness should build primarily on the phenomena, vocabulary, and methodology relevant and acceptable to those sciences, and try to connect the empirical study of consciousness with them.

This cognitive neuroscience approach to consciousness seems to me almost self-evident, but many others seem to have taken a different course. There have been serious attempts to show that the brain and neuroscience are irrelevant for explaining mind and consciousness (argued by classical cognitive science, and more recently by some forms of philosophical representationalism) or that quantum physics is highly relevant (argued by several researchers recently in consciousness studies).

The point here is not that we could somehow know in advance that, say, quantum mechanics, shamanism, or robotics is not going to be relevant for the explanation of consciousness. They could well be, but so could a myriad of other somewhat far-fetched fields of human knowledge. The point is, rather, that we do not know in advance that they *would* be highly relevant or, indeed, relevant at all. Any claim for their relevance must be empirically anchored to actual phenomena of consciousness and to actual data, experiments, testable hypotheses, and predictions.

Consciousness: A Theoretical Problem for the Biology of the Mind

Until recently, the neural-mental interface was relatively untouched by the psychological and biological sciences, save for the isolated attempts in neuropsychology and physiological psychology to cross the borderline. During the 1990s, a new overall approach changed it all. Cognitive neuroscience is the branch of science that explicitly operates at the interface between the biological and the mental realms, and seeks global theoretical integration between these domains. Cognitive neuroscience has been characterized as the empirical science invented to close the explanatory gap and to solve the most fundamental problem of modern science: the gap between biological processes and the processes of mind (Gazzaniga, 1995a). Remarkably, in this spirit cognitive neuroscience has been boldly declared to be "the biology of the mind" (Gazzaniga, Ivry, & Mangun, 2002).

If there already exists an empirical field of inquiry devoted to the biology of the mind, then the mind-body problem (or the brain-consciousness problem) can no longer be the sole property of an empirically isolated philosophy of mind. I suggest that consciousness should primarily be regarded as a theoretical problem of those special sciences where it *must* figure in, where it simply cannot be avoided or ignored. Therefore, the problem of consciousness is now becoming a central problem in the *philosophy of cognitive neuroscience.* Consequently, questions about whether or not science can explain consciousness should be understood as fundamental questions about the scope and limits of cognitive neuroscience.

The Hard Core of the Biological Research Program

If the philosophical mind-body problem is built into the very foundation of cognitive neuroscience, then the scientific research program on consciousness should have its roots in the same terrain. It should be a *biological* research program above all.

A research program is a metatheoretical entity that includes a set of paradigmatic assumptions, sometimes called the "hard core" of the program. What might be an appropriate pretheoretical assumption for a biological research program on consciousness? What kind of ideas should constitute the hard core of the program? I propose that the biological research program on consciousness should start with a few very simple,

crystal clear, empirically plausible basic assumptions that are consistent with the overall approach in cognitive neuroscience as a branch of the biological sciences. Those assumptions should include some general metaphysical commitments as to the nature of the phenomenon, and a clear idea of where in the physical world the phenomenon can be found. My suggestion is the following:

Subjective phenomenal consciousness is a real, natural biological phenomenon that literally resides within the confines of the brain.

Let me explain briefly what these assumptions imply. To say that consciousness is a *real* phenomenon is, first, to deny that consciousness is a mere illusion (i.e., to deny that we are zombies or that there is no important difference between us and nonconscious zombies—the notorious Dennettian view), and second, to deny that consciousness is merely some sort of causally impotent epiphenomenon. To say that consciousness is *real* implies the following claims: (1) consciousness exists, (2) it exists in time and space (i.e., conscious phenomena have temporal duration and spatial location), (3) it has some causal powers. I adopt the philosophical view that *to be real is to have causal powers* (Kim, 1992; a defense of this definition of "real" is, however, outside the scope of this book). The fundamental assumption that consciousness is a real phenomenon thus entails the claim that the phenomenal realm is causally potent, and that it has causal powers distinct from the nonphenomenal (e.g., purely neurophysiological) realm.

To say that consciousness is a *natural* phenomenon is to deny that consciousness is a supernatural or mystical phenomenon, in principle beyond a natural-science approach. It is also to deny that consciousness is only an observer-relative interpretation or social construction whose existence is dependent on conditions external to the biological organism to which the phenomenon is attributed. Consciousness exists in its own right, and the fact that there are conscious phenomena in the world is not ontologically dependent on any other facts apart from ones concerning the biological systems—brains—in which consciousness is realized. To have a fully functional human brain is a sufficient condition for having consciousness, regardless of what else there is in the world or what the causal history or surrounding context of the fully functional brain happens to be.

To say that consciousness *literally resides within the confines of the brain* is to characterize the spatial location of the phenomenon in the natural world; to determine unambiguously where we should look for the principal explanandum in the science of consciousness. Note that this statement by no means implies that consciousness is localized in a *single* area or point inside the brain (which is very unlikely); it only postulates that consciousness is somewhere within the confines of the biological reality of the brain rather than anywhere outside that reality. There are many who reject this view because their fundamental background assumptions are different from mine. To accept this internalist view is to deny that consciousness resides in, or is necessarily dependent

on, any physical or logical relations between the brain and its surroundings (a claim defended by externalists, e.g., Dretske, 1995; Tye, 1995), and it is to deny that consciousness resides only at the "personal" level of organism-environment interaction (a position defended by Dennett [1991] and O'Regan & Noë [2001]); it is to deny that consciousness is somehow "projected" or "reflected" by the brain into the ordinary physical or some ill-defined nonphysical space (cf. Velmans, 1996, 2000; McGinn, 1999; J. Smythies, 1994b, 2003) and to deny that consciousness somehow pervades the whole body or is literally embodied in the physical body outside the brain (Thompson, 2001). Like other biological phenomena, say protein synthesis or postsynaptic potentials, consciousness is, according to biological realism, located within the confines of the very biological system in which it is realized; it has no supernatural powers to escape from within the depths of the brain.

I suggest that these fundamental assumptions could establish the core of a unified biological research program on consciousness. This research program is consistent with the biology of the mind, but may have implications concerning the general metaphysical commitments of future cognitive neuroscience. Whatever the commitments might turn out to be, they should be worked out in such a manner as to be consistent with the core assumptions of the biological research program on consciousness. Cognitive neuroscience cannot turn its back on consciousness anymore; therefore, the theoretical and philosophical commitments of cognitive neuroscience have to accommodate the reality of consciousness.

The task for us now is to follow the implications of the proposed fundamental assumptions and thereby work out the theoretical foundations of the science of consciousness.

Explanation in the Biological Sciences Is Different from Explanation in Physics

If consciousness is a theoretical problem in the biology of the mind, and the proper research program is a biological one, then the question about explaining consciousness should be placed in that particular context. Thus, to understand what it takes to explain consciousness, we should find out what scientific explanation in biology requires. How are complex biological phenomena in general explained in the biological sciences? In other words, what constitutes an "explanation" in these branches of science? If that can be figured out, then it is possible to place the problem of consciousness in the proper explanatory context and ask, Could consciousness be explained by using a similar explanatory strategy? That is the appropriate question to be asked as far as the explanation of consciousness is concerned. Only by incorporating the problem of consciousness into the biology of mind will it be possible to evaluate whether consciousness constitutes some sort of inexplicable anomaly for science. Only by seeing consciousness in the proper explanatory context can we begin to understand whether consciousness is explicable, to what extent it is explicable, and, if some of its aspects must remain mysterious, why exactly that might be the case.

In current philosophy of science there are two somewhat opposing views of scientific explanation (Salmon, 1989): the *unification approach,* which is mostly based on how explanation works in physics; and the *mechanistic approach,* which applies especially to the biological sciences. The unification approach is a descendant of the classical deductive-nomological (or D-N) view in the philosophy of science. It seeks laws of nature that are as universal as possible. Universal physical laws would provide the basis of a systematized, coherent overall picture of how the world works. According to this framework of explanation, scientific knowledge consists of language-like representations in theories and laws (such as equations, mathematical formulas, or expressions and propositions in some other types of exact formal languages). Consequently, the reductive explanation of a phenomenon consists of a deduction of the phenomenon from a lower-level explanatory theory or general law, with the help of "bridge principles." These bridge principles connect the vocabulary of two theories by identifying a term in an old theory (say, "water") with a term in the new, lower-level explanatory theory ("H_2O").

The D-N view of explanation is closely related to the standard model of theory reduction which has not only been popular in the philosophy of science but has also dominated the philosophical discussions of mind-body reduction (Kim, 1998). The bold idea behind the reductive explanation of mental phenomena is that every term describing psychological phenomena could be identified with a term in (future) neuroscience, and thus psychology would reduce to neuroscience. Hence, in modern philosophy of mind it has been deemed as of utmost importance to establish whether there are "laws" or "bridge principles" that connect the mental and physical domains in a reductive manner, as required in classical D-N explanation. If reductive psychophysical laws cannot be found, then perhaps there are "fundamental" psychophysical laws that link psychological terms with biological terms without reducing the former to the latter. Or perhaps no laws between the physical and the psychological are to be found—perhaps the psychological domain is "anomalous," that is, not to be connected to the biological domain by any "lawlike statements."

This traditional view of scientific explanation and theory reduction seems to have little to do with actual scientific practice in the biological sciences, including the neurosciences. When neurobiologists explain complex biological phenomena found in the nervous system (which they are perfectly capable of doing, at least once in a while), they do not construct deductions or logico-mathematical derivations of laws, or connect predicates of two theories with bridge principles. Explanation in neurobiology is not a logical proof procedure; nor is explanatory knowledge represented in a purely propositional or linguistic form. Furthermore, universal laws of nature comparable to those found in physics are virtually unheard of in the biological sciences.

Still, philosophers of mind almost invariably accept the D-N approach and apply it to the explanation of consciousness. But if consciousness belongs in the scope of the

biology of the mind, then there is every reason to believe that the appropriate explanatory strategy for consciousness is going to resemble the successful explanations in the biological sciences rather than the physical sciences. Unless we have a deep understanding of the nature of explanation in the biological sciences, we are in no position to evaluate whether science could or could not explain consciousness. A more naturalistic conception of explanation in biology is urgently needed. Such a conception is provided by the framework of the so-called mechanistic approach. The term "mechanistic" is slightly misleading, though; thus, I shall call this explanatory framework by the more proper term, *multilevel explanation.*

Multilevel Explanation: The Proper Explanatory Strategy for the Biology of the Mind

Multilevel explanation sees explanatory knowledge not clothed in language-like representations, laws, or logical deductions, but as an understanding of the *hidden mechanisms* by which nature works. Proponents of this view (e.g., Bechtel, 1994; Machamer, Darden, & Craver, 2000) have pointed out that the traditional notion of a universal law of nature has few, if any, applications in neurobiology and molecular biology. Biological knowledge is not primarily represented in universal laws, or other propositional expressions in formal languages and mathematical symbols. Biologists typically first *identify* and *describe* an interesting system or phenomenon at one *level of organization* in nature and then try to figure out what the *components* of this system are, how they interact, and how they produce the effects that can be observed at the level of the whole system. When they go about doing this task, they try to take the system apart, or visualize it better with the help of various research instruments, in order to figure out what the components and microstructures of the system are. From these data biologists attempt to build *an idealized multilevel model* of the system, the purpose of which is to show the general structure and function of the system across several levels of organization. The model may be only partially if at all clothed in linguistic representations. Instead, all kinds of visualizable diagrams and figures, scale models, and graphical representations such as charts, maps, and computer simulations can often best depict the component structures of, and their mutual interactions within, the biological system in question (Bechtel, 1994; Bechtel & Richardson, 1992).

The biological research program on consciousness should utilize the same explanatory scheme that applies to the multilevel explanation of other complex biological phenomena. Placing consciousness, the most fascinating of all biological phenomena, into the multilevel framework would then reveal what kinds of questions must be resolved in order to explain consciousness. One big—and seemingly unanalyzable—problem would be thus decomposed to a set of smaller explanatory subproblems, each of which could then be attacked separately. Furthermore, the framework might help the researchers to see which subproblems can be studied by a particular method or experimental

approach, and how the different subproblems connect to the big picture of the subject domain. It would also help the researchers to evaluate which phenomena are central and which peripheral, and what the role of a particular empirical phenomenon is in contributing to the big picture. A general explanatory scheme on consciousness would thus serve the function of integrating the various empirical approaches and relating them to each other in a meaningful way.

The Hierarchical Structure of the World: Levels

The mechanistic view in biology depicts natural phenomena as parts of complex hierarchical causal networks. A central notion in this framework is the notion of "levels." Biological phenomena are realized at different levels of complexity in nature. The notions I will use to capture this structure, inherent in nature, are *levels of organization* and *levels of description.*

Hierarchical structure is evident in the case of biological systems. Thus, before we try to fit consciousness into this scheme, let us take a brief look at some of the levels of complexity elsewhere in biology.

We may start at the level of biochemistry which is concerned with the compounds, their chemical properties, and the multitude of chemical reactions involved in the building of living organisms. Small organic molecules are first built by autotrophs, using simple natural oxides. The molecules are then further assembled into macromolecules: amino acids into proteins, monosaccharides into polysaccharides, and so on. At the level of molecular biology, there are macromolecules and the complexes they form inside living cells. Macromolecules are capable of combining into multimacromolecular complexes, out of which the structural components found inside cells (e.g., filaments, tubules, and membranes) are built. Macromolecules, by way of their complex three-dimensional (3D) structure, are capable of causally contributing to the correct performance of otherwise highly unlikely chemical reactions within the cell. Enzyme molecules, almost all of which are proteins, catalyze thousands of chemical reactions by binding only the molecules that are the natural substrates of a specific reaction. Despite all this complexity already found at the level of the building blocks of life, we have to move to a higher level of organization in order to find an actual living system. That is the level of the complete cell, the basic unit of life. A single cell has all that is needed for life (DeDuve, 1991). At this level, we find a complex system that is separated from its environment by a boundary, the cytomembrane, and contains a multitude of subsystems such as the nucleus, the cytoskeleton, mitochondria, various organelles and so on. These subsystems carry out the various functions that constitute life: biosynthetic manufacturing of the constituents of the cell, extracting energy from the environment and converting it to work; multiplying, and so on. Pluricellular living systems such as trees, moths, and humans are composed of huge numbers of specialized cells joined together and organized into systems at levels higher than the individual cell (e.g., tissues and organs).

A Hierarchy of Levels of Organization

Thus, biological organisms can be seen as consisting of a multitude of distinct but interconnected levels. Phenomena at different levels in the hierarchy reside at different spatiotemporal scales or grains. I call the different levels of complexity in nature the *levels of organization.* These levels are taken as *ontological,* that is, as real, causally potent natural phenomena that exist out there, independent of our scientific practices or conceptual systems. A level of organization has a specific orderly structure of entities and it realizes specific activities or functions. The structure appearing at one level can be understood as composed of parts that reside at a lower level of organization. However, the interrelationships between parts and wholes are not always simple or straightforward, for a process or function thought to take place entirely at one level (e.g., a specific chemical reaction chain) may in fact be possible only in the context of structures that reside at higher levels of organization (macromolecular enzymes, membrane properties) (Bechtel & Richardson, 1992). For the sake of simplicity, we may think of the levels of organization in biological systems (such as ourselves) as composing a *hierarchical* system, with basic physical phenomena residing at the bottom levels, chemical and biochemical phenomena at the intermediate levels, and biological and physiological phenomena at the higher levels (Bechtel & Abrahamsen, 1991).

The entities and properties in this multilayered model of the world can be thought of as forming a micro-macro hierarchy (figures 1.1 and 1.2). Entities and properties reside at specific levels of organization. The same properties and entities cannot be realized at levels lower or higher than where they actually reside. Cells reside at one particular level of organization. Parts of cells are not themselves cells (but organelles, macromolecules, etc.); wholes formed by a multitude of cells are not cells themselves (but tissues, organs, and multicellular organisms).

Higher-level properties and entities can be called *macro* properties and *macro* entities, emphasizing that they reside at a higher level of organization than their *micro* constituents. But because they are in such a way constituted by lower-level phenomena, we may also say that they are *micro-based* properties and entities. An important point about *micro*-based *macro* phenomena at higher levels of organization is that *they have their own causal powers that go beyond the causal powers of their microconstituents;* they have novel causal powers not had by any of the lower-level phenomena that constitute them (Kim, 1998).

The idea of the world as a hierarchy of levels of organization is a fundamental assumption in the multilevel explanatory framework. Therefore, if this is the framework that applies to the explanation of biological phenomena in general, then also consciousness as a biological phenomenon should be seen as residing at a specific level of organization in nature. And if consciousness is a biological phenomenon within the confines of the brain (another fundamental assumption of the research program), then there must be some specific level of organization and some specific spatiotemporal grain in the brain where consciousness resides. Furthermore, if consciousness is a micro-based

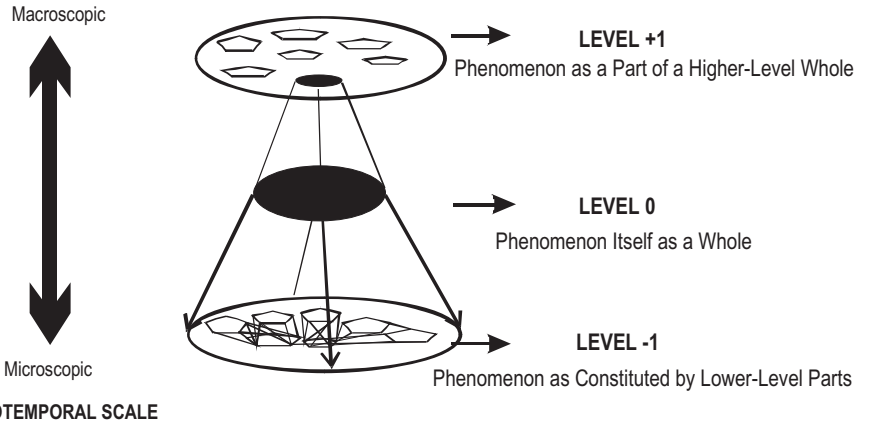

Figure 1.1
The hierarchy of levels of organization. The framework of multilevel explanation depicts biological phenomena as residing at different spatiotemporal scales in the overall micro-macro hierarchy of nature. Entities at contiguous levels bear part-whole relations to each other.

macrophenomenon along with other biological phenomena, it is bound to have unique causal powers not had by other phenomena in the brain, not even by those neural systems that are the microconstituents of consciousness itself.

A key idea that runs through this book is the following: We should adopt the working hypothesis that consciousness is a part of the hierarchical structure of the biological world. More precisely, consciousness constitutes one particular level of organization among all the other levels in nature. Thus, we may *reconceptualize* consciousness as *the phenomenal level of organization* in the brain. The characterization of consciousness as one particular biological level of organization is in harmony with the core assumptions of biological realism. Furthermore, it would clarify several problems that have surrounded consciousness. Most of them can now be reconceptualized in terms of "levels": How does the phenomenal level of organization relate to other levels of organization in the brain? How could the phenomenal level be measured or observed empirically? How could it be conceptualized or modeled theoretically? What are the causal powers of entities and properties residing at the phenomenal level of organization? I explore possible answers to these questions in the course of this book.

Note that the reconceptualization I propose above is not a *reduction* of consciousness to something else. I am not proposing that consciousness, this subjective world-*for-me,* would be anything other than what it is: a world of subjective experience. My proposal simply tries to find the proper place for the subjective world-for-me in the wider context of the physical world, by placing it among other biological levels of organization. This working hypothesis guides our further research efforts when we try to figure out how it is possible that a biological level of organization in the brain manifests itself as a sub-

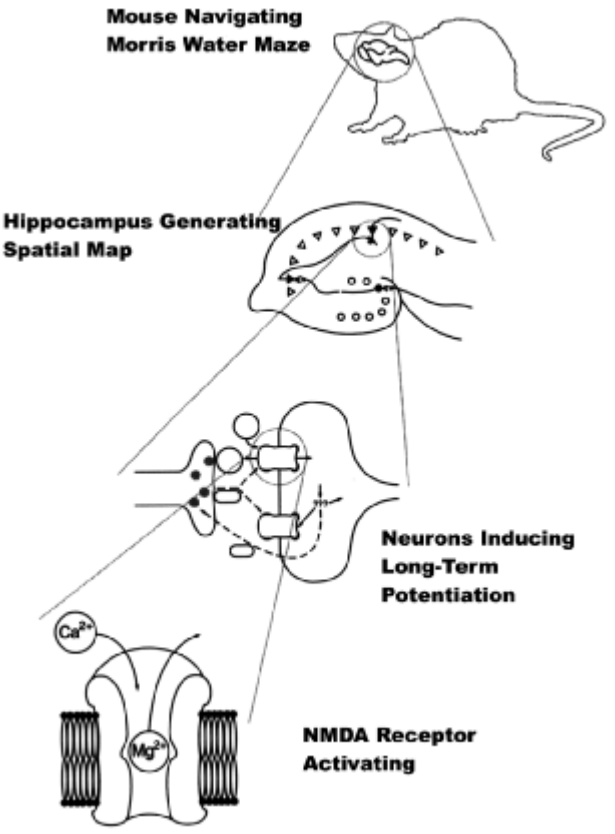

Figure 1.2
A hierarchy of levels in neuroscience. This concrete example from neuroscience shows the levels of hierarchical organization involved in the mechanism of spatial memory. (From "Discovering Mechanisms in Neurobiology: The Case of Spatial Memory" by Carl F. Craver and Lindley Darden from *Theory and Method in the Neurosciences,* Peter K. Machamer, Rick Grush, and Peter McLaughlin, Eds., © 2001 by University of Pittsburgh Press. Reprinted by permission of the University of Pittsburgh Press.)

jective world-for-me. By applying the multilevel framework I am not trying to get rid of consciousness; on the contrary, I am trying to preserve it in all its glory, only exposing its proper place in nature.

The Structure of Biological Science: Levels of Description

We should now define another notion of "levels" that is central to multilevel explanation, but that should not be confused with the levels of organization in nature. Multilevel explanatory models of natural phenomena are human constructions believed to *reflect* the hierarchical structure of the levels of organization in nature. There are different scientific disciplines and subdisciplines specialized in describing the phenomena

that reside at distinct levels of organization in nature: for example, quantum physics, particle physics, inorganic chemistry, organic chemistry, biochemistry, molecular biology, cell biology, histology, and so on. Let us call the conceptual and theoretical systems, produced by humans to describe nature, the *levels of description.* These levels, unlike the levels of organization, are dependent on and constructed by human activity and mostly clothed in linguistic representations (as in mathematical equations or laws or as verbal descriptions in textbooks) or graphical representations (as in models or visualizations or computer simulations of complex systems).

If biological realism is on the right track in assuming that consciousness is a biological level of organization in the brain, then the science of consciousness should deliver the corresponding *phenomenal level of description.* This description should first systematically capture the central features of the phenomenal domain. Next, the entire phenomenal level should be placed in the context of a multilevel explanatory model. The multilevel model should show how the phenomenal level is related to biological phenomena at other levels in the causal-mechanical networks of the brain.

The Network of Biological Explanation

To explain a biological phenomenon is to make its existence and behavior intelligible within the framework of biological levels of description. A full multilevel explanation of a biological phenomenon requires that the explanandum be located within the overall causal network that (metaphorically speaking) "surrounds" it. The network has several dimensions or pathways, involving different spatial and temporal directions in the causal network (Craver & Darden, 2001). The different dimensions of a full multilevel explanation include at least the following (figure 1.3): a downward-looking explanation, describing what constitutes the phenomenon; an upward-looking explanation, describing how the phenomenon acts as an element of higher-level mechanisms; a backward-looking explanation, specifying the causal pathway that brought about or modulated the phenomenon; and a black-box description, specifying at an abstract level the input-output functions that take place at the interface between the phenomenon and its surroundings.

Constitutive Explanation

Constitutive explanation involves moving downward in the hierarchy and looking at a smaller-scale spatiotemporal grain or a lower level of organization. It is accomplished by showing that the phenomenon appearing as an integrated whole at one level can be decomposed to its constituent parts and their causal interactions at the immediately underlying, lower levels of description. Constitutive explanations describe the lower-level mechanisms that the entity is composed of, or whose activity, when taken as a whole, simply *is* the phenomenon to be explained by the description of the mechanism.

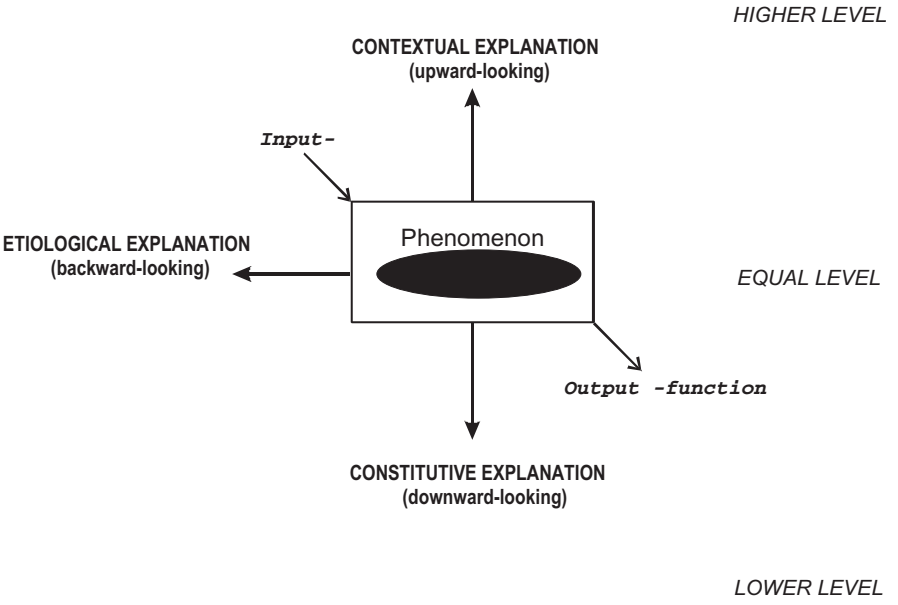

Figure 1.3
The dimensions of the causal-mechanical network. The full mechanistic explanation of a biological phenomenon involves different subtypes of explanation. The *downward-looking* and *upward-looking* explanations move across different spatiotemporal scales to lower and higher levels of organization, and the *backward-looking* explanation moves backward in time, tracing the causal chain of events that modulated or causally interacted with the phenomenon to be explained. These different, partial explanations correspond to different spatiotemporal directions in the causal-mechanical network that surround the phenomenon. The explanandum becomes intelligible when the phenomenon itself and its connections to the causal network surrounding it can be described in detail, without gaps.

An illuminating example from biology is the cell. A single cell, taken as a whole, is an independent living unit. The constitutive explanation revealing what makes the cell tick descends to the immediately lower levels of organization where the different parts of the cell and their causal interactions are to be found.

When the same strategy is applied to the explanation of consciousness there is, first of all, the phenomenon to be explained (the explanandum): consciousness. It resides at some specific level of organization in the brain (the *phenomenal* level). The lower-level, nonconscious neurophysiological mechanisms, whose activity as a whole constitutes consciousness, reside at a lower level of organization in the brain (figure 1.4).

The current search for the direct NCC appears to be the empirical approach to the constitutive explanation of consciousness. However, the notion "neural correlates of consciousness" requires considerable clarification. The relationship between an explanandum and its lower-level constituents must be stronger than mere correlation, for correlation is not an explanatory relationship. The cell membrane, the nucleus, chromosomes, cell

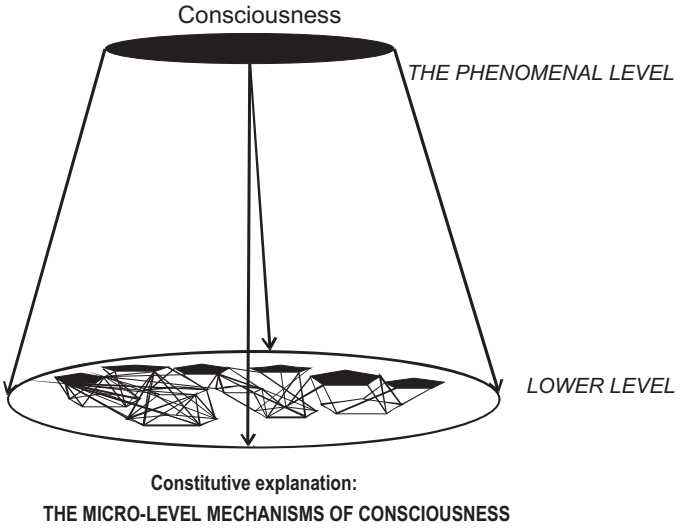

Figure 1.4
Constitutive explanation. This direction of explanation moves *downward,* to a smaller spatiotemporal scale than the scale where the explanandum lies as a whole. Constitutive explanation describes the immediately lower levels of organization (the microlevel constitutive mechanisms). Those mechanisms themselves lack the holistic features of the explanandum (in this case, consciousness). Nonetheless, when the constitutive explanation is completed, then the lower-level mechanisms can be seen to nonmagically form the higher level of organization where the explanandum itself resides.

organelles, and other microscopic parts of the cell are not merely the biological *correlates* of life, but crucial *microlevel constituents* that explain *why* the whole system is alive. In the same vein, the constitutive explanation of consciousness should describe such part-whole or mereological relationships between the lower- and higher-level phenomena that make their hierarchical connection truly explanatory.

Etiological Explanation

An *etiological explanation* is an account of the causal history of the explanandum (figure 1.5). It starts with the phenomenon to be explained, then looks backward in the temporal dimension along the causal chain to figure out what the preceding causes are that brought the phenomenon about. A distinction should be made, however, between the etiological explanation of the immediate (or proximate) causes of the phenomenon, the ontogenetic (or developmental) causes, and the evolutionary (or ultimate) ones.

In the case of visual consciousness, for example, the etiological story about the immediate causes of visual experiences—say, my perception of Orion's Belt in the night sky—would have to describe how the light (after traveling through space for years) entered my eyes, striking at the retina, causing a sequence of electrophysiological changes

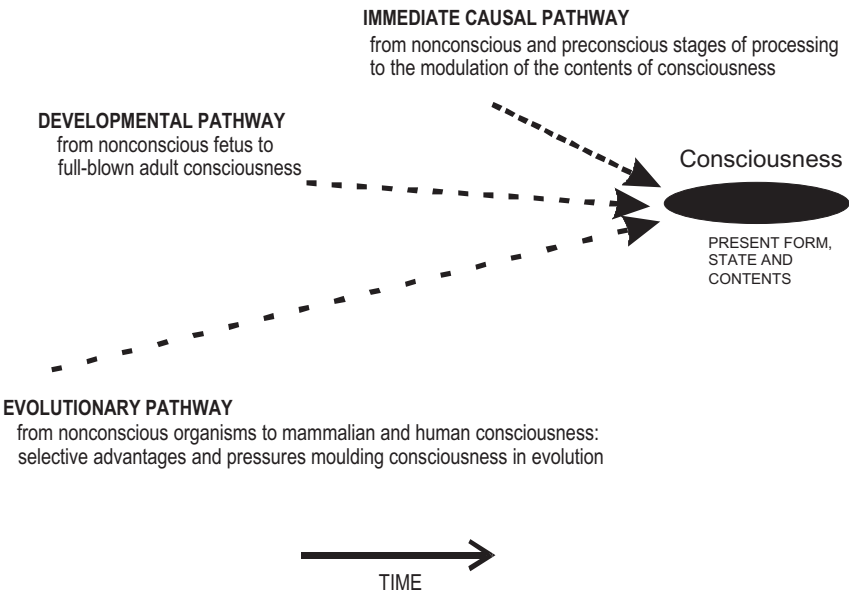

Figure 1.5
Etiological explanation. In this type of explanation, the causal chains that led to the occurrence of the explanandum are traced back. Etiological explanation looks *backward* in the temporal dimension to describe the causal pathways that *preceded* the current state of the explanandum and *causally modulated* it. Three different etiological explanations can be demarcated on the basis of how far backward in time these explanations look: (1) The short or immediate causal pathway describes how phenomenal content in consciousness was brought about by the immediately preceding external stimuli that brought about changes in sensory organs and sensory input pathways, eventually leading to changes at the phenomenal level in the brain. (2) The developmental (ontogenetic) pathway describes how the nervous system is being wired during ontogeny, at some point giving rise to the phenomenal level in the infant brain, and then developing toward full-blown adult consciousness. This story will be enormously complex, involving accounts of genetic expression in the fetal brain and of the modulatory influences of internal and external stimulation to the developing brain. (3) The evolutionary (phylogenetic) pathway looks even further backward in time and describes how, at some point during evolutionary history, the emergence of nervous systems that included the phenomenal level solved a greater number of adaptive problems than nonconscious neural systems were able to solve. Furthermore, the etiological story also describes how the organization of the phenomenal level gradually became a finely tuned species-specific adaptation (at least in mammals). The phenomenal level promoted the reproductive success of different species by constructing exactly the kind of subjective world that was useful for guiding behavior in the particular type of behavioral environment and ecological niche where each species evolved.

in the pathway from retina to V1 and from there to other parts of the visual cortex and into complex feedforward-feedback loops of neural activity, eventually causing the emergence of a well-organized visual percept in my visual consciousness.

The field known as *psychophysics* provides sorts of etiological explanations for subjective sensations by describing the lawlike relationships that obtain between physical stimulus characteristics and subjective experience: say the principles or "laws" involved when a proximate stimulus at the retina is transformed into (or rather how it causally modulates) visual phenomenology. The full story about the immediate etiological pathway should explain how the regularities and laws of stimulus-sensation relationships are physiologically implemented within the sensory input channel.

It should be made clear that the entities and activities themselves that participate in the chain of the immediately preceding causes are *not* conscious or even parts of consciousness. They are nonconscious entities and activities either entirely outside the organism, or nonconscious (preconscious) biological events inside the organism but entirely outside consciousness. *Consciousness does not and could not leak backward to its own immediate etiological pathway;* if it did, that pathway itself would be a part of consciousness rather than a chain of preceding causes leading to changes in consciousness.

The confusion between what constitutes consciousness itself and what constitutes its (nonconscious or preconscious) etiological pathway is one of the most frequent sources of error in the explanation of consciousness. Here is one way to define their difference accurately: The *constitutive* basis of consciousness includes those lower-level entities on which consciousness is *ontologically dependent;* consciousness could not exist without them existing.[1] By contrast, the *etiological* pathway includes entities on which consciousness is *causally dependent:* entities which can causally modulate, influence, and interact with consciousness. Causal dependence does not entail ontological dependence, however: I can causally interact with a stone without being ontologically dependent on it. Thus, consciousness can in principle exist even in the absence of the entities and activities that make up its etiological pathway, while it cannot, even in principle, exist without the entities and activities that make up its constitutive basis.

Again, let us consider an illuminating example from biology. First, the double helix DNA molecule could not exist without the nucleotides and the base pairs (A-T) (G-C) existing: they are microlevel constituents of the DNA double helix. Thus, DNA is *ontologically dependent* on them: it is impossible to find or synthesize (or even imagine) a DNA molecule without nucleotides and base pairs. Second, all the DNA in my cells and yours is *causally dependent* on mitosis (or meiosis): every single DNA molecule in our cells was originally constructed in one of those processes. Yet, DNA in general is not ontologically dependent on mitosis or meiosis. We can imagine that a thread of DNA could be synthesized in the laboratory through some entirely different method. There could in principle be DNA molecules without mitosis or meiosis being involved in their causal history. Therefore, DNA is not ontologically dependent on mitosis or

meiosis. Just like the DNA molecule, consciousness also depends ontologically only on its constitutive mechanisms, not on the preceding etiological pathways where the mechanisms happened to have come about or by which consciousness happened to have been causally modulated.

The etiological story that tells how the constitutive mechanisms of consciousness came about during individual development, or *ontogeny* describes (in the case of visual consciousness) how the basic neuroanatomical structure of the visual pathways and cortical areas was constructed in the fetus by an immensely complex sequence of gene expression and how the system then matured in interaction with the actual visual environment after birth. Consciousness is causally but not ontologically dependent on this etiological pathway. A brain equipped with consciousness could conceivably result from some other type of process than normal development (I will have more to say about such philosophical thought experiments later, especially in chapter 8).

The etiological account of the *ultimate evolutionary causes* of visual consciousness would have to describe the adaptational, evolutionary story that describes how sensitivity to light increased the fitness of distant ancestral species millions of years ago, leading among other things to the emergence of the primate visual system. It answers the question, Why does visual consciousness exist in the first place; why was it selected for? In biology these evolutionary explanations are sometimes called "accounts of the ultimate causation of the phenomenon."

I shall call the first type of etiological explanation *the immediate causal pathway* to consciousness; the second type, the *ontogenetic etiological pathway* to consciousness; and the third type, the *evolutionary pathway to consciousness.* A full explanation of consciousness should make it possible for us to trace backward on each of these types of etiological pathways that lead to the phenomenon to be explained.

Contextual Explanation

Contextual explanation describes the role of the entity at a higher level of description (figure 1.6). It answers such questions as: How is the phenomenon related to other ones at higher levels of organization where the phenomenon is just one element or subsystem in a complex mechanism? What role does it fulfill in the overall functional economy of the higher-level system?

DNA molecules in the chromosomes of the cell are not themselves "alive" in any manner whatsoever. "Life" does not exist at the level of single molecules. Still, the DNA plays a crucial role at the higher level of organization, in interaction with RNA and amino acids, thus contributing to the higher-level processes that render the cell as a whole a living system. The contextual explanation of DNA literally puts the molecule in the context of the causal networks in which the molecule is involved. In a similar vein, a DNA sequence becomes the gene for some phenotypic feature only in the context of the whole organism.

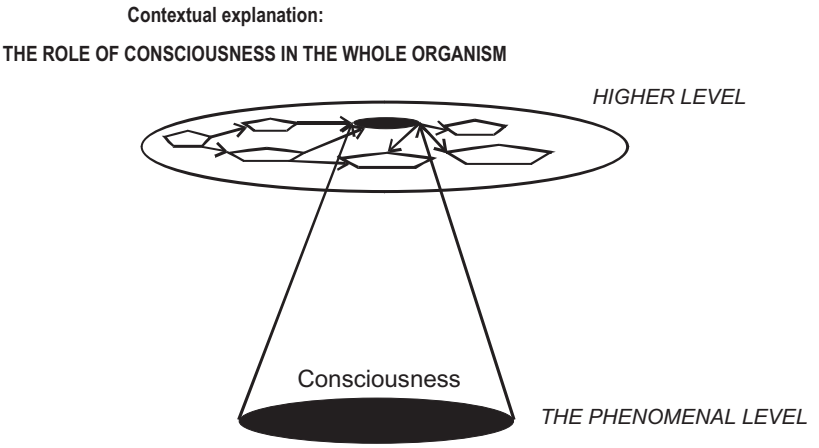

Figure 1.6
Contextual explanation. This dimension of explanation zooms *upward,* away from the level of organization where the explanandum itself lies, to higher levels of organization at coarser spatial and temporal scales. Thus it describes how the explanandum relates to and interacts with other phenomena at the higher levels. In the case of consciousness, the contextual explanation describes the special role that the phenomenal level plays at the behavioral level of organism-environment interaction, in interaction with several nonconscious sensory, cognitive, and motor systems. Although the phenomenal level alone cannot realize sensory perception or externally observable behavior but needs to interact with many nonconscious mechanisms, the contribution of the phenomenal level to the higher level of organism-environment interaction appears to be absolutely crucial.

The contextual explanation of consciousness should place the phenomenal level in the context of the causal networks in the brain. How does the phenomenal level causally interact with various nonconscious phenomena "surrounding" it in the network? Say, preconscious sensory processing that modulates the phenomenal level by providing input to that level, or "postconscious" motor programming that is influenced by conscious intentions to initiate actions. In a multilevel explanation, consciousness should be connected to various nonconscious biological systems in the brain and in the whole physical body of the biological organism, which, together with consciousness, enables coherent sequences of adaptive behavior and successful organism-environment interaction.

Input-Output Description

Sometimes it is useful just to describe the operation of the phenomenon from start conditions to finish conditions without yet specifying the micro-level mechanisms. Such descriptions are called *input-output functions* (figure 1.7). They focus on some specific activities that take place at the boundary between the phenomenon and its environment. Such activities may be represented at an abstract level, mathematically for example, and at that level the input-output function can be described explicitly and computationally.

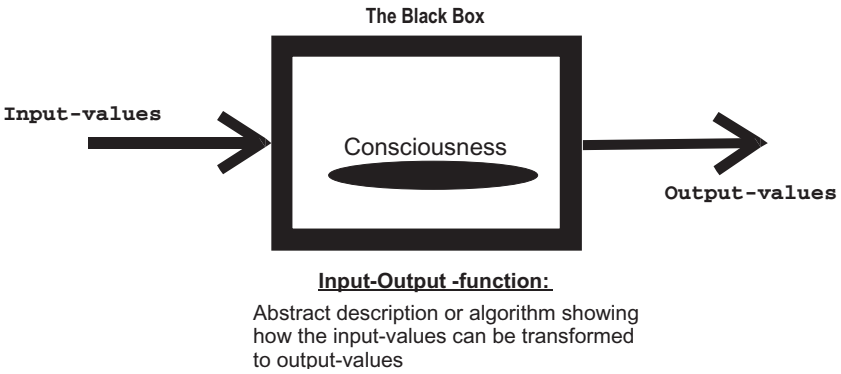

Figure 1.7
Input-output function. This description captures, at an abstract level, the formal features of the input-output transformations performed by the explanandum. These transformations may be described computationally, in an explicit computational model, by determining the algorithms that carry out the transformations.

An algorithm describing a particular input-output function may be realized in a computer program as a *simulation* of the function. It should be noted, however, that the algorithm itself is an abstract entity, not identical to the phenomenon, and in fact, the algorithm becomes a very different kind of mechanism when it is written in a programming language and instantiated in computer hardware (Machamer et al., 2000). Computational modeling and simulation in biology, neuroscience, and cognitive science depend on our ability to provide explicit input-output descriptions of biological and psychological phenomena.

The input-output function of (some aspect) of consciousness describes, at an abstract level, the activities that take place at the conscious-nonconscious interface. It specifies the algorithm or function which transforms the input values (describing sensory stimulation) to the output values (standing for some features of the conscious sensation or perception). An example that perhaps comes rather close to something like this concerns the relationship between stimulus luminance (the physical intensity of visible light striking the retina) and perceptual brightness (a feature of subjective visual experience). The input values of the function represent physical luminance, the output values represent perceived brightness, and the task is to describe an algorithm that transforms the input values to the output values in a manner closely matching empirical data from psychophysical experiments. When an accurate simulation of the brightness-from-luminance transformation is developed, some might go so far as to say that now the algorithms "used" by the brain to solve this problem have been found.

Even if we had an accurate input-output function available for some aspects of consciousness, it would be unwise to treat it as *the explanation* of the phenomenon. An input-output function treats the phenomenon as a mere black box and reveals nothing about

its internal constitution or microlevel mechanism. An algorithm that simulates the transformation from luminance to brightness does not tell us what the conscious visual experience of brightness is, or what its underlying biological mechanisms are. Certainly, an explicit function or algorithm will constrain the set of mechanisms that would be regarded as acceptable constitutive explanations for the phenomenon, but it cannot replace the microlevel explanations, or supersede the full multilayered model of the phenomenon. [2]

Multilevel Models and Consciousness

The explanation of a phenomenon is completed with the construction of a multilevel mechanistic model of the phenomenon. Such a model describes the explanandum across several different levels. The conceptual structures (theories and models) in neurobiology are really descriptions of such multilevel causal networks forming complex hierarchies of mechanisms (figure 1.8).

In a multilevel mechanistic model the phenomenon is situated at some specific level of description within the multidimensional causal network. An important part of the task is the *identification of the correct level of organization* at which to understand par-

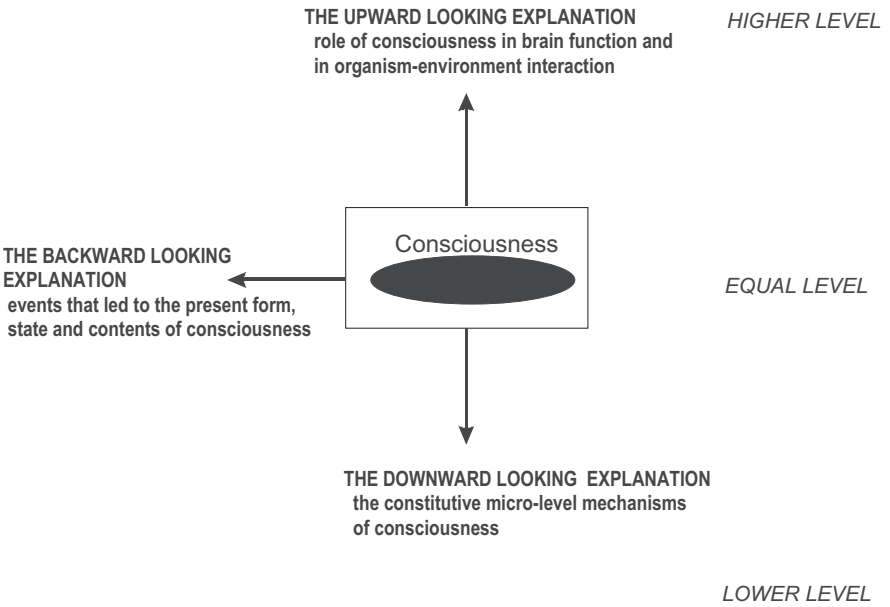

Figure 1.8
The multilevel model of consciousness. The complete explanatory model should include a description of consciousness itself (the explanandum) and then link it to the causal networks surrounding the explanandum.

ticular causal interactions: how the explanandum interacts with other phenomena, what the lower-level components of the explanandum are and how they interact at the microlevel (Bechtel, 1994). To achieve this goal, the multilevel model should include three types of explanation. These different types of explanation correspond to three different directions with respect to which the phenomenon can be situated in the mechanical reticulum: (1) The *downward-looking explanation* specifies the *constitutive mechanisms* of the phenomenon at the lower levels of description; (2) the *upward-looking explanation* specifies *the role function of the phenomenon at the higher levels of organization;* (3) the *backward-looking explanation* (the immediate or proximal causes, the ontogenetic or developmental causes, and the ultimate or evolutionary causes) specifies the etiological causal pathways or the temporally preceding events that led to or causally modulated the phenomenon.

The point of the above discussion is this: the explanation of consciousness should be seen in the context of multilevel explanation. As far as cognitive neuroscience is concerned the problem of consciousness boils down to the questions: Is it possible to locate consciousness within the multilayered causal-mechanical networks that reside in the brain? Can we find a level of organization in the brain to which consciousness naturally corresponds? Can we discover and fully describe the constitutive mechanisms, the etiological pathways, and the contextual role functions of consciousness?

If there is a place to be found for consciousness within the multilevel framework of biological phenomena, and if the causal network surrounding consciousness can be reconstructed, then consciousness can be scientifically explained by the biological sciences. If not, then we must either give up the basic assumption that consciousness is a real, natural biological phenomenon belonging within the scope of biological science, or at least the view that it can be explained by such science. The aim of this book is to find out whether this sort of multilevel explanation of consciousness is philosophically and empirically feasible.

Summary

The current state of confusion in the study of consciousness can be overcome only if theoretically and empirically unified research programs emerge in the field. One potential candidate for a unified research program, based on the philosophy called "biological realism," has been outlined in this chapter. It treats consciousness as a biological phenomenon: a particular level of organization in the brain. It ties the study of consciousness to the context of the biological sciences, especially cognitive neuroscience. Thus, consciousness should be explained in the same manner as other complex biological phenomena are explained. The explanatory framework that applies to the biological sciences appears to be significantly different from the traditional concepts of

"explanation" in the philosophy of science. In the biological context, the overall explanation of consciousness consists of several smaller subtasks that all contribute to building a multilevel model of the explanandum. This picture of the explanation of consciousness confers new unity to the field, as it sorts out and defines the different questions that we should try to answer when seeking an explanation of consciousness. Thus, the biological approach might be able to establish a unified research program in the study of consciousness.