Can Neurobiology Teach Us Anything about Consciousness?

Patricia Smith Churchland

Proceedings and Addresses of the American Philosophical Association, Vol. 67, No. 4 (Jan., 1994), 23-40.

Stable URL:
http://links.jstor.org/sici?sici=0065-972X%28199401%2967%3A4%3C23%3ACNTUAA%3E2.0.CO%3B2-M

Proceedings and Addresses of the American Philosophical Association is currently published by American Philosophical Association.

CAN NEUROBIOLOGY TEACH US ANYTHING ABOUT CONSCIOUSNESS?

Patricia Smith Churchland
University of California, San Diego

Presidential Address delivered before the Sixty-Seventh Annual Pacific Division Meeting of The American Philosophical Association in San Francisco, California, March 26, 1993.

I Introduction:

Human nervous systems display an impressive roster of complex capacities, including the following: perceiving, learning and remembering, planning, deciding, performing actions, as well as the capacities to be awake, fall asleep, dream, pay attention, and be aware. Although neuroscience has advanced spectacularly in this century, we still do not understand in satisfying detail how any capacity in the list emerges from networks of neurons.[1] We do not completely understand how humans can be conscious, but neither do we understand how they can walk, run, climb trees or pole vault. Nor, when one stands back from it all, is awareness intrinsically more mysterious than motor control. Balanced against the disappointment that full understanding eludes us still, is cautious optimism, based chiefly on the nature of the progress behind us. For cognitive neuroscience has already passed well beyond what skeptical philosophers once considered possible, and continuing progress seems likely.

In assuming that neuroscience can reveal the physical mechanisms subserving psychological functions, I am assuming that it is indeed the brain that performs those functions—that capacities of the humans mind are in fact capacities of the human brain. This assumption and its concomitant rejection of Cartesian souls or spirits or "spooky stuff" existing separately from the brain is no whimsy. On the contrary, it is a highly probable hypothesis, based on evidence currently available from physics, chemistry, neuroscience and evolutionary biology. In saying that physicalism is an hypothesis, I mean to emphasize its status as an empirical matter. I do not assume that it is a question of conceptual analysis, a priori insight, or religious faith, though I appreciate that not all philosophers are at one with me on this point.[2]

Additionally, I am convinced that the right strategy for understanding psychological capacities is essentially reductionist, by which I mean, broadly, that understanding the neurobiological mechanisms is not a frill but a necessity. Whether science will finally succeed in reducing psychological phenomena to neurobiological phenomena is, needless to say, yet another empirical question. Adopting the reductionist strategy means trying to explain the macro levels (psychological properties) in terms of micro levels (neural network properties).

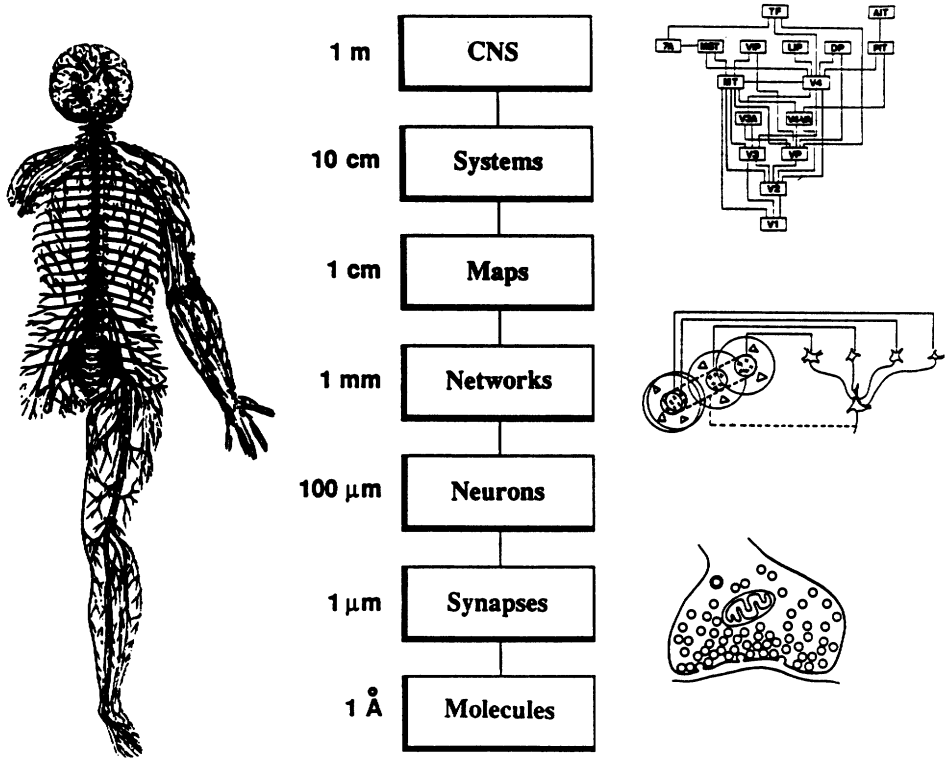

Figure 1

Schematic illustration of levels of organization in the nervous system. The spatial scales at which anatomical organization can be identified varies over many orders of magnitude. Icon to the left depicts the "neuron man," showing the brain, spinal cord, and peripheral nerves. Icons to the right represent structures at distinct levels: (top) a subset of visual areas in visual cortex; (middle) a network model proposing how ganglion cells could be connected to "simple" cells in visual cortex; and (bottom) a chemical synapse. (From Churchland and Sejnowski 1992)

The fundamental rationale behind this research strategy is straightforward: if you want to understand how a thing works, you need to understand not only its behavioral profile, but also its basic components and how they are organized to constitute a system. If you do not have the engineering designs available for reference, you resort to reverse engineering—the tactic of taking a part a device to see how it works.[3] Insofar as I am trying to discover macro-to-micro explanations, I am a reductionist. Because many philosophers who agree with me on the brain-based nature of the soul nonetheless rail against reductionism as ridiculous if not downright pitiful, it may behoove me to begin by explaining briefly what I do and, most emphatically, do *not* mean by a reductionist research strategy.[4]

Clearing away the "negatives" first, may I say that I do *not* mean that a reductionist research strategy implies that a *purely bottom-up strategy* should be adopted. So far as I can tell, no one in neuroscience thinks that the way to understand the nervous system is first to understanding everything about the basic molecules, then everything about every neuron and every synapse, and to continue ponderously thus to ascend the various levels of organization until, at long last, one arrives at the uppermost level—psychological processes. (Figure 1) Nor is there anything in the history of science that says a research strategy is reductionist only if it is purely bottom-up. That characterization is straw through and through. The research behind the classical reductionist successes—explanation of thermodynamics in terms of statistical mechanics; of optics in terms of electromagnetic radiation; of hereditary transmission in terms of DNA—certainly did not conform to any purely bottom-up research directive.

So far as neuroscience and psychology are concerned, my view is simply that it would be wisest to conduct research on many levels simultaneously, from the molecular, through to networks, systems, brain areas, and of course behavior. Here, as elsewhere in science, hypotheses at various levels can *co-evolve* as they correct and inform one another.[5] Neuroscientists would be silly to make a point of ignoring psychological data, for example, just as psychologists would be silly to make a point of ignoring all neurobiological data.

Second, by "reductionist research strategy" I do not mean that there is something disreputable, unscientific or otherwise unsavory about high level descriptions or capacities *per se*. It seems fairly obvious, to take a simple example, that certain rhythmic properties in nervous systems are network properties resulting from the individual membrane traits of various neuron types in the network, together with the way the set of neurons interact. Recognition that something is the face of Arafat, for another example, almost certainly emerges from the responsivity profiles of the neurons in the network plus the ways in which those neurons interact. "Emergence" in this context is entirely non-spooky and respectable, meaning, to a first approximation, "property of the network." Determining precisely what the network property is, for some particular feat, will naturally take quite a lot of experimental effort. Moreover, given that neuronal behavior is highly nonlinear, the network properties are *never* a simple "sum of the parts." They are some function —some *complicated* function—of the properties of the parts. High level capacities clearly exist, and high level descriptions are therefore needed to specify them.

Wherefore eliminative materialism, then? Because the existing characterization of the human brain's high level capacities, embodied in what, for want of a better term, is referred to as "folk psychology," may well be reconfigured as time and cognitive neuroscience proceed. This too is an empirical hypothesis, and one for which empirical support already exists. Reconfiguration is already underway for such categories as "memory," "attention" and "reasoning."[6]

The possibility of nontrivial revision and even replacement of existing high level descriptions by 'neurobiologically harmonious' high level categories is the crux of what makes eliminative materialism *eliminative*.[7] By 'neurobiologically harmonious' categories, I mean those that permit coherent, integrated explanations from the whole brain on down through neural systems, big networks, micronets, and neurons. Only the strawman is so foolish as to claim that there are no high level capacities, that there are no high level phenomena.[8] In its general aspect, my point here merely reflects this fact: in a profoundly important sense we do not understand exactly what, at its higher levels, the brain really does. Accordingly, it is practical to earmark even our fondest intuitions about mind/brain function as revisable hypotheses rather than as transcendental absolutes or introspectively given certainties. Acknowledgment of such revisability makes an enormous difference in how we conduct psychological and neurobiological experiments, and in how we interpret the results.

II NAYSAYING THE NEUROBIOLOGICAL GOAL

Over the last several decades, a number of philosophers have expressed reservations concerning the reductionist research goal of discovering the neurobiological mechanisms for psychological capacities, including the capacity to be conscious. Consequently, it may be useful to consider the basis for some of these reservations in order to determine whether they justify abandoning the goal, or whether they should dampen our hopes about what might be discovered about the mind/brain. I shall here consider three main classes of objection. As a concession to brevity, my responses shall be ruthlessly succinct, details being sacrificed for the sake of the main gist.

A. The Goal is Absurd (Incoherent)

One set of reasons for dooming the reductionist research strategy is summed up thus: "I simply cannot imagine that seeing blue or the feeling of pain, for example, could consist in some pattern of activity of neurons in the brain," or, more bluntly, "I cannot imagine how you can get awareness out of meat." There is sometimes considerable filler between the "it's unimaginable" premise and the "it's impossible" conclusion, but so far as I can tell, the filler is typically dust which cloaks the fallacious core of the argument.[9]

Given how little in detail we currently understand about how the human brain "en-neurons" any of its diverse capacities, it is altogether predictable that we should have difficulty imagining the neural mechanisms. When the human scientific

community was comparably ignorant of such matters as valence, electron shells, and so forth, natural philosophers could not imagine how you could explain the malleability of metals, the magnetizability of iron, and the rust resistance of gold, in terms of underlying components and their organization. Until the advent of molecular biology, many people thought it was unimaginable, and hence impossible, that to be a living thing could consist in a particular organization of "dead" molecules. "I cannot imagine," said the vitalists, "how you could get *life* out of *dead* stuff."

From the vantage point of considerable ignorance, failure to imagine some possibility is only that: a failure of imagination—one psychological capacity amongst others. It does not betoken any metaphysical limitations on what we can come to understand, and it cannot predict anything significant about the future of scientific research. After reflecting on the awesome complexity of the problem of thermoregulation in homeotherms such as ourselves, I find I cannot imagine how brains control body temperature under diverse conditions. I suspect, however, that this is a relatively uninteresting psychological fact about me, reflecting merely my current state of ignorance. It is not an interesting metaphysical fact about the universe nor even an epistemological fact about the limits of scientific knowledge.

A variation of the "cannot imagine" proposal is expressed as "we can never, never know. . . . ," or "it is impossible to ever understand. . . ." or "it is forever beyond science to show that. . . .". The idea here is that something's being impossible to conceive says something decisive about its empirical or logical impossibility. I am not insisting that such proposals are never relevant. Sometimes they may be. But they are surprisingly high-handed when science is in the very early stages of studying a phenomenon.

The sobering point here is that assorted "a priori certainties" have, in the course of history, turned out to be empirical duds, however obvious and heartfelt in their heyday. The impossibility that space is non-Euclidean, the impossibility that in real space parallel lines should converge, the impossibility of having good evidence that some events are undetermined, or that someone is now dreaming, or that the universe had a beginning—each slipped its logical noose as we came to a deeper understanding of how things are. If we have learned anything from the many counterintuitive discoveries in science it is that our intuitions can be wrong. Our intuitions about ourselves and how we work may also be quite wrong. There is no basis in evolutionary theory, mathematics, or anything else, for assuming that prescientific conceptions are essentially scientifically adequate conceptions.

A third variation on this "nay, nay, never" theme draws conclusions about how the *world must actually be*, based on *linguistic properties* of certain central categories in current use to describe the world. Permit me to give a boiled down instance: "the category 'mental' is remote in meaning—means something completely different —from the category 'physical'. It is absurd therefore to talk of the brain seeing or feeling, just as it is absurd to talk of the mind having neurotransmitters or conducting current." Allegedly, this categorial absurdity undercuts the very possibility that science could discover that feeling pain is activity in neurons in the brain. The epithet "category error" is sometimes considered sufficient to reveal the

naked nonsense of reductionism.

Much has already been said on this matter elsewhere,[10] and I shall bypass a lengthy discussion of philosophy of language with three brief points. (1) It is rather far-fetched to suppose that intuitions in the philosophy of language can be a reliable guide to what science can and cannot discover about the nature of the universe. (2) Meanings *change* as science makes discoveries about what some macro phenomenon *is* in terms of its composition and the dynamics of the underlying structure. (3) Scientists are unlikely to halt their research when informed that their hypotheses and theories "sound funny" relative to current usage. More likely, they will say this: "the theories might sound funny to you, but let me teach the background science that makes us think the theory is true. Then it will sound less funny." It may be noted that it sounded funny to Copernicus' contemporaries to say the Earth is a planet and moves; it sounded funny to say that heat is molecular motion or that physical space is non-Euclidean or that there is no absolute "downness." And so forth.

That a scientifically plausible theory sounds funny is a criterion only of its not having become common coin, not of its being wrong. Scientific discoveries that a certain macro phenomenon is a complex result of the micro structure and its dynamics are typically surprising and typically sound funny—at first. Obviously none of this is positive evidence that we can achieve a reduction of psychological phenomena to neurobiological phenomena. It says only that sounding funny does not signify anything, one way or the other.

B. The Goal is Inconsistent with "Multiple Realizability"

The core of this objection is that if a macro phenomenon can be the outcome of more than one mechanism (organization and dynamics of components), then it cannot be identified with any one mechanism, and hence the reduction of the macrophenonomenon to *the* (singular) underlying micro phenomenon is impossible. This objection seems to me totally uninteresting to science. Again, permit me to ignore important details and merely to summarize the main thrust of the replies. (1) Explanations, and therefore reductions, are domain relative. In biology, it may be fruitful first to limn the general principles explaining some phenomenon seen in diverse species, and then figure out how to account for the interspecies differences, and then, if desirable, how to account for differences across individuals within a given species. Thus the general principles of how hearts or stomachs work are figured out, perhaps based on studies of a single species, and particularities can be resolved thereafter. Frog hearts, macaque hearts and human hearts work in essentially the same general way, but there are also significant differences, apart from size, that call for comparative analyses. Consider other examples: (a) from the general solution to the copying problem that emerged from the discovery of the fundamental structure of DNA, it was possible to undertake explorations of how differences in DNA could explain certain differences in the phenotype; (b) from the general solution to the problem of how neurons send and receive signals, it was possible to launch detailed exploration into the differences in responsivity profiles of distinct classes of neuron.[11]

(2) Once the mechanism for some biological process has been discovered, it may be possible to invent devices to mimic those processes. Nevertheless, invention of the technology for artificial hearts or artificial kidneys does not obliterate the explanatory progress on actual hearts and actual kidneys; it does not gainsay the reductive accomplishment. Again, the possibility that hereditary material of a kind different from DNA might be found in things elsewhere in the universe does not affect the basic scaffolding of a reduction on this planet. Science would have been much the poorer if Crick and Watson had abandoned their project because of the abstract possibility of Martian hereditary material or artificial hereditary material. In fact, we do know the crux of the copying mechanism *on Earth*—namely, DNA, and we do know quite a lot about how it does its job. Similarly, the engineering of artificial neurons and artificial neural nets (ANNs) facilitates and is facilitated by neurobiological approaches to how real neurons work; the engineering undertakings do not mean the search for the basic principles of nervous system function is misguided.

(3) There are always questions remaining to be answered in science, and hence coming to grasp the general go of a mechanism, such as the discovery of base-pairing in DNA, ought not be mistaken for the utopian ideal of a complete reduction—a complete explanation. Discoveries about the general go of something typically raise hosts of questions about the *detailed* go of it, and then about the details of the *details*. To signal the incompleteness of explanations, perhaps we should eschew the expression "reduction" in favor of "reductive contact." Hence we should say the aim of neuroscience is to make rich reductive contact with psychology as the two broad disciplines co-evolve. I have experimented with this recommendation myself, and although some philosophers warm to it, scientists find it quaintly pedantic. In any case, "reductive contact" between molecular biology and macrobiology has become steadily richer since 1953, though many questions remain. Reductive contact between psychology and neuroscience has also become richer, especially in the last decade, though it is fair to say that by and large the basic principles of how the brain works are poorly understood.

(4) What, precisely, are supposed to be the progammatic sequelae to the multiple realizability argument? Is it that neuroscience is *irrelevant* to understanding the nature of the human mind? Obviously not. That neuroscience is *not necessary* to understand the human mind? One cannot, certainly, deny that it is remarkably useful. Consider the discoveries concerning sleep, wakeness, and dreaming; the discoveries concerning split brains, humans with focal brain lesions, the neurophysiology and neuroanatomy of the visual system, and so on. Is it perhaps that we should not get our hopes up too high? What, precisely, is "too high" here? Is it the hope that we shall discover the general principles of how the brain works? Why is that too high a hope?

C. The Brain *Causes* Consciousness

Naysaying the reductionist goal while keeping dualism at arm's length is a manoeuvre requiring great delicacy. John Searle's strategy (Searle 1992) is to say that although the brain *causes* conscious states, any identification of conscious states with brain activities is unsound. Traditionally, it has been opined that the best the reductionist can hope for are *correlations* between subjective states and brain states, and although correlations can be evidence for causality they are not evidence for identity. Searle has tried to bolster that objection by saying that whereas *a/b* identifications elsewhere in science reveal the reality behind the appearance, in the case of awareness, the reality and the appearance are inseparable—there is no reality to awareness except what is present in awareness. There is, therefore, no reduction to be had.

Synoptically, here is why Searle's manoeuvre is unconvincing: he fails to appreciate why scientists opt for identifications when they do. Depending on the data, cross-level identifications to the effect that *a* is *b* may be less troublesome and more comprehensible scientifically than supposing thing *a* causes separate thing *b*. This is best seen by example.[12]

Science as we know it says electrical current in a wire is not caused by moving electrons; it *is* moving electrons. Genes are not caused by chunks of base pairs in DNA; they *are* chunks of base pairs (albeit sometimes distributed chunks). Temperature is not caused by mean molecular kinetic energy; it *is* mean molecular kinetic energy. Reflect for a moment on the inventiveness required to generate explanations that maintain the *nonidentity* and causal dependency of (a) electric current and moving electrons, (b) genes and chunks of DNA, and (c) heat and molecular motion. Unacquainted with the relevant convergent data and explanatory successes, one may suppose this is not so difficult. Enter Betty Crocker.

In her microwave oven cookbook, Betty Crocker offers to explain how a microwave oven works. She says that when you turn the oven on, the microwaves excite the water molecules in the food, causing them to move faster and faster. Does she, as any high school science teacher knows she should, end the explanation here, perhaps noting, "increased temperature just *is* increased kinetic energy of the constituent molecules?" She does not. She goes on to explain that because the molecules move faster, they bump into each other more often, which increases the friction between molecules, and, as we all know, friction causes heat. *Betty Crocker still thinks heat is something other than molecular KE; something caused by but actually independent of molecular motion.*[13] Why do scientists not think so too?

Roughly, because explanations for heat phenomena—production by combustion, by the sun, and in chemical reactions; of conductivity, including conductivity in a vacuum, the variance in conductivity in distinct materials, etc.—are *vastly* simpler and more coherent on the assumption that heat *is* molecular energy of the constituent molecules. By contrast, trying to make the data fit with assumption that heat is some other thing *caused by* speeding up molecular motion is like trying to nail jelly to the wall.

If one is bound and determined to cleave to a caloric thermodynamics, one

might, with heroic effort, pull it off for oneself, though converts are improbable. The cost, however, in coherence with the rest of scientific theory, not to mention with other observations, is extremely high. What would motivate paying that cost? Perhaps an iron-willed, written-in-blood, resolve to maintain unsullied the intuition that heat "*is what it is and not another thing.*" In retrospect, and knowing what we now know, the idea that anyone would go to exorbitant lengths to defend the "heat intuition" seems rather a waste of time.

In the case at hand, I am predicting that explanatory power, coherence and economy will favor the hypothesis that awareness just *is* some pattern of activity in neurons. I may turn out to be wrong. If I am wrong, it will not be because an introspectively-based intuition is immutable, but because the science leads us in a different direction. If I am right, and certain patterns of brain activity *are* the reality behind the experience, this fact does not in and of itself change my experience and suddenly allow me (my brain) to view my brain as an MR scanner or a neurosurgeon might view it. I shall continue to have experiences in the regular old way, though in order to understand the neuronal reality of them, my brain needs to *have* lots of experiences and undergo lots of learning.

Finally, barring a jump to the dualist' s horse, the idea that there has to be a bedrock of subjective "appearance" on which reality/appearance discoveries must ultimately rest is faintly strange. It seems a bit like insisting that "down" cannot be relative to where one is in space; down is down. Or like insisting that time cannot be relative, that either two events happen at the same time or they don't, and that's that. Humans are products of evolution; nervous systems have evolved in the context of competition for survival—in the struggle to succeed in the four F's: feeding fleeing, fighting, and reproduction. The brain's model of the external world enjoys improvement through appreciating various reality/appearance distinctions—in short, through common critical reason and through science. In the nature of things, it is quite likely that the brain's model of its internal world also allows for appearance/reality discoveries. The brain did not evolve to know the nature of the sun as it is known by a physicist, nor to know itself as it is known by a neurophysiologist. But, in the right circumstances, it can come to know them anyhow.[14]

D. The Problem is Beyond our Feeble Intelligence

Initially, this claim appears to be a modest acknowledgment of our limitations. In fact, it is a powerful prediction based not on solid evidence, but on profound ignorance (Colin McGinn 1990). For all we can be sure now, the prediction might be correct, but equally, it might very well be false. How feeble is our intelligence? How difficult is the problem? How could you possibly know that solving the problem beyond our reach, no matter how science and technology develop? Inasmuch as it is not known that the brain is more complicated than it is smart, giving up on the attempt to find out how it works would be disappointing. On the contrary, as long as experiments continue to produce results that contribute to our understanding, why not keep going?[15]

III Tracking Down The Neural Mechanisms of Consciousness

A. Finding a Route In

In neuroscience there are many data at higher levels relevant to consciousness. Blindsight, hemineglect, split brains, anosagnosia (unawareness of deficit), for starters, are powerful constraints to guide theoretical reflection. Careful studies using scanning devices such as magnetic resonance imagining (MRI) and positron emission tomography (PET) have allowed us to link specific kinds of functional losses with particular brain regions.[16] This helps narrow the range of structures we consider selecting for preliminary micro exploration.

For example, the hippocampus might have seemed a likely candidate for a central role in consciousness because it is a region of tremendous convergence of fibres from diverse areas in the brain. We now know, however, that bilateral loss of the hippocampus, though it impairs the capacity to learn new things, does not entail loss of consciousness. At this stage, ruling something out is itself a valuable advance. We also know that certain brain stem structures such as the locus coereleus (LC) are indirectly necessary, but are not part of the mechanism for consciousness. LC does play a nonspecific role in arousal, but not a specific role in awareness of particular contents, such as awareness at a moment of the color of the morning sky rather than the sound of the lawn sprinklers. The data may be fascinating in its own right, but the question remains: how can we get from an array of intriguing data to genuine explanations of the basic mechanism? How can we get *started*?

In thinking about this problem, I have been greatly influenced by Francis Crick. His basic approach is straightforward: if we are going to solve the problem, we should treat it as a scientific problem to be tackled in much the way we tackle other difficult scientific problems. As with any scientific mystery, what we want is a revealing experimental entry. We want to find a thread which, when pulled, will unloose a whole lot else. To achieve that, we need to devise testable hypotheses that can connect macro effects with micro dynamics.

Boiled down, what we face is a constraint satisfaction problem: find psychological phenomena that (a) have been reasonably well studied by experimental psychology, (b) are circumscribed by lesion data from human patients as well as data from precise animal microlesions, (c) are known to be related to brain regions where good neuroanatomy and neurophysiology has been done and (d) where we know quite a lot about connectivity to other brain regions. The working assumption is that if a person is aware of a stimulus, his brain will be different in some discoverable respect from the condition where he is awake and attentive but unaware of the stimulus. An auspicious strategy is to hunt down those differences, guided by data from lesion studies, PET scans, magnetoencephalograph (MEG) studies, and so forth. Discovery of those differences, in the context of neurobiological data generally, should aid discovering a theory of the mechanism.

The central idea is to generate a theory constrained by data at many levels of brain organization—sufficiently constrained so that it can be put to meaningful tests. Ultimately a theory of consciousness will need to encompass a range of

processes involved in awareness, including attention and short term memory. Initially, however, it may target a subset, such as integration across space and across time. Whether the theory falls to falsifying evidence or whether it survives tough tests, we shall learn something. That is, either we shall have ruled out specific possibilities—a fine prize in the early stages of understanding—or we can go on deepen and develop the theory further—an even finer prize. In any case, the trick is to generate testable, meaty hypotheses rather than loose, frothy hypotheses susceptible only to experiments of fancy. The trick is to make some real progress.

B. Visual Awareness

What plausible candidates surface from applying the constraint satisfaction procedure? Interestingly, the choices are quite limited. Although metacognition, introspection, and awareness of emotions, for example, are indeed aspects of consciousness, either we do not have good lesion data to narrow the search space of relevant brain regions, or the supporting psychophysics is limited or both. Consequently, these processes are best put on the back burner for later study.

Visual awareness, by contrast, is a more promising candidate. In the case of vision, as Crick points out, there is a huge literature in visual psychophysics to draw upon, there is a rich literature of human and animal lesion studies, and relative to the rest of the brain, a lot is known about the neuroanatomy and neurophysiology of the visual system, at least in the monkey and the cat. Visual phenomena such as filling in, binocular rivalry, seeing motion, seeing stereoptic depth, and so forth might reward the search for the neurobiological differences between being aware and not being aware in the awake, attentive animal. This may get us started, and I do emphasize *started*.

1. The Crick Hypothesis

Immersed in the rich context of multi-level detail, Crick has sketched an hypothesis concerning the neuronal structures he conjectures make the salient differences, depending on whether the animal is or is not visually aware of the stimulus.[17] Integration of representations across spatially distributed neural networks—the unity in apperception, so to speak—is thought be accomplished by temporal 'binding', namely synchrony in the output responses of the relevant neurons. Very crudely, Crick's suggestion is that (1) for sensory awareness, such as visual awareness, the early corteces are pivotal (e.g. visual areas V1, V2; somatosensory areas S1, S2 etc.). This makes sense of lesion data, as well as recent PET data (Kosslyn et al. 1993) and single cell data (Logothetis and Schall 1989). (2) Within the early sensory cortical areas, pyramidal cells in layer 5 and possibly layer 6, play the key role.

What good is this idea? Part of its appeal is its foothold in basic structure. In biology, the solution to difficult problems about mechanism can be greatly facilitated by identification of critical structures. Crudely, if you know "what," it helps enormously in figuring out "how." On its own, the Crick hypothesis can be

only a small piece of the puzzle. If we are lucky, however, it, or something like it, may be a *key* piece of the puzzle. This is not the time for a fuller discussion of this hypothesis. Suffice it to say that true or false, the Crick hypothesis provides a bold illustration of how to approach a problem so tricky it is often scrapped as unapproachable.

2. The Llinas Hypothesis

Another promising entry route is suggested by the differences—phenomenological and neurobiological—between sleep/dreaming/wakeness (SDW) states.[18] This entry point is attractive first because there is the familiar and dramatic loss of awareness in deep sleep, which is recovered as we awake, and is probably present also during dreaming. The phenomenon is highly available in lots of different subjects and across many species. Second, MEG and EEG techniques reveal global brain features characteristic of different states. Human and animal lesion data are important, especially as they concern deficits in awareness during wakeness. Here again I note the significance of research on blindsight, hemineglect (tendency to be unaware of stimuli in various modalities on the left side of the body), simultanagnosia (inability to see several things simultaneously), anosagnosia (unawareness of deficits such as paralysis, blinds, garbled speech and so forth).

Third, we have learned a great deal from abnormalities in and manipulation of the SDW cycle and the link to specific brain properties. Fourth, some of the global changes in state in the SDW cycle seen by macro techniques have been linked by micro techniques to interactions between specific circuits in the cortex and subcortical circuits, especially circuits in several key structures in the thalamus. Fifth, and more specifically, MEG data reveal an robust 40 Herz wave form during wakeness and dreaming.[19] The definition and amplitude is much attenuated during sleep, and the amplitude is modulated during wakeness and dreaming. Analysis of the wave from by MEG reveals it to be a traveling wave, moving in the anterior to posterior direction in the brain, covering the distance in about 12 to 13 *milliseconds*. Cellular data suggest that these dynamical properties emerge from particular neural circuits and their dynamical properties.

What does all this add up to? Based on these data, and mindful of the various high-level data, Rodolfo Llinas and colleagues (1991; 1993) have hypothesized that the fundamental organization subserving consciousness and the shifts seen in the SDW pattern are pairs of coupled oscillators, each of which connect thalamus and cortex, but each connects distinct cell populations via its own distinctive connectivity style. (Figure 2) One oscillator 'family' connects neurons in a thalamic structure known as the intralaminar nucleus, a bagel-shaped structure whose neurons reach to the upper layers of cortex to provide a highly regular fan-like coverage of the entire cortical mantle. The other oscillator 'family' connects neurons in thalamic nuclei for modality-specific information (MS nuclei) originating for example, in the retina or the cochlea, with modality specialized cortical areas (e.g. V2, S2). During deep sleep, the intralaminar neurons projecting to cortex cease their 40 Hz behavior.

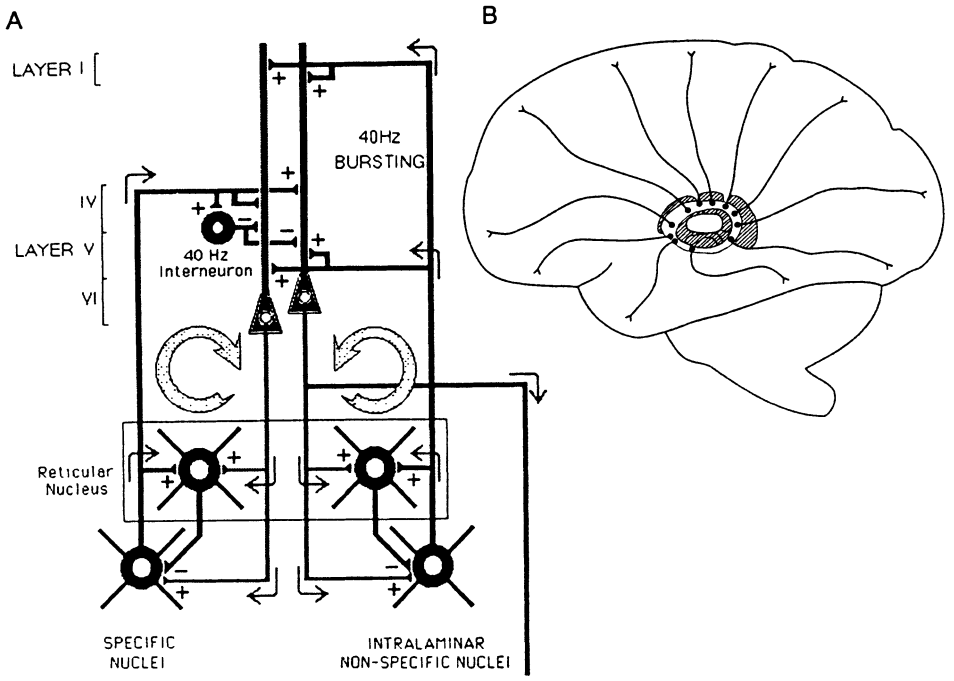

Figure 2

Schematic diagram of the circuits between the thalamus and the cerebral cortex proposed to serve temporal binding. (A) Diagram of two different types of circuit connecting thalamus and cortex. On the left, specific sensory nuclei or motor nuclei of the thalamus project to Layer IV of cortex, producing cortical oscillation by direct activation and feed-forward inhibition via 40Hz inhibitory interneurons. Collaterals of these projections produce thalamic feedback via the reticular nucleus (a kind of rind covering the thalamus). The return pathway (circular arrow with stipple) reenters this loop to specific and reticular nuclei via Layer VI cells. On the right, the second loop shows nonspecific intralaminar nuclei projecting to Layer I of cortex, and giving collaterals to the reticular nucleus. Layer V cells return oscillation to the reticular and the intralaminar nuclei, establishing a second resonant loop. The conjunction of the specific and nonspecific loops is proposed to generate temporal binding. Connectivity between the loops is seen chiefly in Layer V. (B) Schematic diagram showing the intralaminar nucleus as a circular neuronal mass (stipple shading). Other parts of the thalamus are shown in hatched shading. The intralaminar nucleus projects widely across the cortex, to Layer I. (From Llinas and Ribary 1993)

During deep sleep and dreaming, external signals to the cortex are gated by the reticular nucleus of the thalamus.

Ever so crudely, the idea is that the second oscillator 'family' provides the content (visual, somatosensory etc.) while the first provides the integrating context. In deep sleep the oscillators are decoupled; in dreaming they are coupled but the MS oscillating circuit is largely nonresponsive to external signals from the periphery; in wakeness, the oscillators are coupled, and the MS circuit is responsive to external signals.

What are the effects of lesions to the intralaminar thalamic structure (bagel)? The main profile of small unilateral lesions is neglect (unawareness) of all stimuli originating the opposite body side. Bilateral lesions result in "inanition," meaning roughly that the patient initiates no behavior and responds very poorly, if at all, to sensory stimuli or questions. Animal studies show much the same profile.

Lesions to modality-specific regions of the thalamus, by contrast, lead to modality specific losses in awareness—visual awareness, for example will be lost, but awareness of sounds, touches etc. can be normal. Intriguingly, the MEGs of Alzheimer's patients who have degenerated to a state of inanition show a dilapidated 40 Hz wave form when it exists at all. Obviously these data are not decisive, but at least they are consistent with the hypothesis.

Do the Llinas hypothesis and the Crick hypothesis fit together? Minimally, they are consistent. Additionally, they are mutually supporting at the neuron and network levels. One encouraging point is this: the two families of oscillators (MS and intralaminar) richly connect to each other mainly in *cortical layer 5* (Figure 2). From what we can tell now, those connections seem to be the chief means whereby the oscillators are coupled. The possibility entertained here is that the temporal synchrony Crick hypothesizes in neurons carrying signals about external stimuli may be orchestrated by the intralaminar-corticalcircuit. Connections between brain stem structures and the intralaminar nucleus could have a role in modulating arousal and alertness.

Many questions now suggest themselves. For example, how do the pivotal structures for awareness interface with behavior? (Or as Dennett would ask, "what happens next?")[20] More specifically, what are the connections between the intralaminar nucleus and motor structures, and between layer 5 of sensory corteces and motor structures; do the projections from the intralaminar nucleus to the cingulate cortex have a role in attention? These are questions motivated by independent data. Convergence of hypotheses is of course encouraging, but it is well to remember that it can also encourage us down the proverbial garden path. Wisdom counsels guarded optimism.

IV CONCLUDING REMARKS

Viewing matters from the mystery side of a phenomenon, solutions can seem impossible, and perhaps even unwanted. On the understanding side, however, solutions seem almost obvious and hard to miss. Why, one might wonder, did it take so long to figure out what the elements are? How could someone as brilliant as

Aristotle miss the plausibility in Aristarchus' idea that the Earth was a sphere moving around the sun? The deeper truths are all too easy to miss of course, just as it is all too easy for us to miss whatever it is that explains why animals sleep and dream, and what autism is. The problems for neuroscience and experimental psychology are hard, but as we inch our way along and as new techniques increase noninvasive access to global brain processes in humans, intuitions change. What seems obvious to us was hot and surprising news only a generation earlier; what seems confounding to our imagination is routinely embraceable by the new cohort of graduate students. Who can tell with certainty whether or not all our questions about consciousness can eventually be answered? In the meantime, it is rewarding see progress—to see some questions shift status from Mysteries We Can Only Contemplate in Awe, to Tough Problems We Are Beginning to Crack.

References

Bickle, J. (1992). "Revisionary physicalism." *Biology and Philosophy*. 7: 411-430.

Churchland, P. M. (1988). *Matter and Consciousness*, 2nd Edition. Cambridge, Mass.: MIT Press.

Churchland, P. M.. (1993a) "Betty Crocker's theory of the mind: A review of *The Rediscovery of the Mind*, by John Searle." *London Review of Books*, (In press).

Churchland, P. M. (1993b) "Evaluating our self conception." Mind and Language. (In press).

Churchland, P. M. and P. S. Churchland (1990). "Intertheoretic reduction: A neuroscientist's field guide." *Seminars in the Neurosciences*. 4: 249-256.

Churchland, P. S. (1986). *Neurophilosophy*. Cambridge, Mass.: MIT Press.

Churchland, P. S. (1987). "Replies to Comments. Symposium on Patricia Smith Churchland's *Neurophilosophy*." *Inquiry*. 29: 241-72.

Churchland, P. S. (1988). "Reduction and the neurobiological basis of conscious ness." In: *Consciousness in Contemporary Science*. Ed. A. J. Marcel and E. Bisiach. 273-304.

Churchland, P. S. and T. J. Sejnowski (1989). "Brain and cognition." In: *Foundations of Cognitive Science*. Ed. M. Posner. Cambridge, Mass.: MIT Press. 245-300.

Churchland, P. S. and T. J. Sejnowski (1992). *The Computational Brain. Cambridge*, Mass.: MIT Press.

Crick, F. H. C. (1994). *The Astonishing Hypothesis*. New York: Scribner's and Sons.

Crick, F. H. C. and C. Koch (1990). "Towards a neurobiological theory of consciousness." *Seminars in the Neurosciences*. 4: 263-276.

Damasio, A. R. (forthcoming). *Descartes' Error*. New York: Simon and Schuster.

Damasio, H. "Neuroanatomy of frontal lobe in vivo: A comment on methodology." In: *Frontal Lobe Function and Dysfunction*. Ed. H. Levin, H. Eisenberg, and A. Benton. New York: Oxford University Press. 92-121.

Damasio, H. and A. R. Damasio (1990). The neural basis of memory, language and behavioral guidance: advances with the lesion method in humans. *Seminars in the Neurosciences*. 4: 277-286.

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

Farah, M. J. (1993). "Neuropsychological inference with an interactive brain: A critique of the 'locality assumption'." *Behavioral and Brain Sciences*. (In press).

Feyerabend, P. K. (1981). *Philosophical Papers*. Vols. 1 and 2. Cambridge: Cambridge University Press.

Flanagan, O. (1992) *Consciousness Reconsidered*. Cambridge, Mass.: MIT Press.

Flanagan, O. (forthcoming) "Prospects for a unified theory of consciousness, or, what dreams are made of." In: *Scientific Approaches to the Question of Consciousness: 25th Carnegie Symposium on Cognition*. Ed. J. Cohen and J. Schooler. Hillsdale, N. J.: L. Erlbaum.

Kosslyn, S. M., N. M. Alpert, W. L. Thompson, V. Maljkovic, S. B. Weise, C. F. Chabris, S. E. Hamilton, S. L. Rauch, and F. S. Buoanno. "Visual mental imagery activated topographically organized visual cortex: PET investigations." *Journal of Cognitive Neuroscience*. 5: 263-287.

Llinas R. R. and D. Pare (1991). "Of dreaming and wakefulness." *Neuroscience* 44: 521-535.

Llinas, R. R. and U. Ribary (1993). Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences*. 90: 2078-2081.

Logothetis, N. and J. D. Schall. (1989). Neural correlates of subjective visual perception. *Science*. 245: 753-761.

Lycan W. G. (1987). *Consciousness*. Cambridge, Mass.: MIT Press.

McGinn, C. (1990). *The Problem of Consciousness*. Oxford: Blackwells.

Schaffner, K. F. (1993). "Theory structure, reduction, and disciplinary integration in biology. *Biology and Philosophy*. 8: 319-348.

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

Notes

1. See our discussion in *The Computational Brain*, Churchland and Sejnowski (1992).

2. For concordant opinions, see also Francis Crick (1994); Paul Churchland (1989); Daniel Dennett (1991); Owen Flanagan (1992); William G. Lycan (1987); John Searle (1993).

3. As P. S. Churchland and T. J. Sejnowski argued in (1989).

4. For an outstanding discussion of reductionism that includes many of the complexities I am not worrying too much about here, see Schaffner (1993).

5. P. S. Churchland (1986), *Neurophilosophy*.

6. See Churchland and Sejnowski (1992); Paul M. Churchland 1993b.

7. Or, as we have preferred but decided not to say "what makes revisionary materialism *revisionary*" (P. S. Churchland 1987). See also P. M. Churchland (1993). For a related but somewhat different picture, see Bickle (1992).

8. Ibid. See also P. M. Churchland and P. S. Churchland (1990).

9. For example, Colin McGinn (1990).

10. See for example, Feyerabend (1981).

11. See also Owen Flanagan (forthcoming).

12. In the following discussion, the ideas are mostly owed to Paul Churchland (1993a). For his discussion, see "Betty Crocker's Theory of the Mind: A Review of John Searle's *The Rediscovery of the Mind*." The London Review of Books. (In press).

13. Paul Churchland made this discovery in our kitchen about eight years ago. It seemed to us a bang-up case of someone's not really understanding the scientific explanation. Instead of thinking the thermodynamic theory through, Betty Crocker just clumsily grafts it onto on old conception as though the old conception needed no modification. Someone who thought electricity was *caused* by moving electrons would tell a comparable Betty Crocker story: "voltage forces the electrons to move through the wire, and as they do so, they cause static electricity to build up, and a sparks then jump from electron to electron, on down the wire." When I regale audiences of scientists with Betty's "microwave" explanation, the mirth is audible.

14. See P. M. Churchland (1993b).

15. See Daniel Dennett's convincing and more detailed discussion of McGinn's naysaying (Dennett 1991).

16. See especially H. Damasio and A. R. Damasio (1990); H. Damasio (1991); A. R. Damasio (forthcoming); Farah (1993).

17. This point is made in Crick and Koch (1990) and in Crick (1994).

18. See also my discussion in P. S. Churchland (1988).

19. See Llinas and Pare (1991).

20. Dennett (1992).