ORIGINAL PRINTING, **CONSCIOUSNESS - PAPERS FOR TUCSON II**
ED: S. HAMEROFF, J. LAUKES ET AL., 1997

Brainshy:
Non-Neural Theories of Conscious Experience[1]

Patricia Smith Churchland
Philosophy, University of California San Diego, Salk Institute

Abstract

Understanding how the mind/brain works is difficult for many reasons. Some are conceptual or theoretical, some involve unavailability of suitable techniques, and some are rooted in practical difficulties involved in performing needed experiments even given available techniques (e.g. for ethical or financial reasons, or because of the sheer labor or man-hours required). Quite apart from these obstacles, the idea that conscious awareness is not at bottom a neural matter motivates some theorists to call for a non-neural approach. One variation on this theme suggests that consciousness is a fundamental property of the universe based in information (David Chalmers); another sees consciousness as unexplainable in terms of brain properties because it is an "intrinsic" property (John Searle); a third (Roger Penrose) adopts the hypothesis that consciousness is a property arising out of quantum-level phenomena, below the level of neurons. Otherwise discordant, the three share the conviction that understanding the brain will not yield an understanding of consciousness. In exploring each of these possibilities, I conclude that none is sufficiently appealing -- empirically, theoretically or logically -- to merit serious research investment. At this stage, I believe the evidence strongly favors neuroscience and psychology as having the best shot at solving the problem.

---

[1] The first sections of this paper are drawn from my earlier article, **The Hornswoggle Problem** (1996), *Journal of Consciousness Studies*, and is republished here with permission. In preparing both that paper, as well as this more inclusive one, I am greatly indebted to Paul Churchland, Francis Crick, Joe Bogen, David Rosenthal, Rodolfo Llinas, Michael Stack, Clark Glymour, Dan Dennett, Ilya Farber and Joe Ramsey for advice and ideas.

I. Chalmer's Approach

A. Introduction

Conceptualizing a problem so we can ask the *right* questions and design *revealing experiments* is crucial to discovering a satisfactory solution to the problem. Asking where animal spirits are concocted, for example, turns out not to be the right question to ask about the heart. When Harvey asked instead, "how much blood does the heart pump in an hour?", he conceptualized the problem of heart function very differently. The reconceptualization was pivotal in coming to understand that the heart is really a pump for circulating blood; there are no animal spirits to concoct. My strategy here, therefore, is to take the label, "The Hard Problem" in a constructive spirit -- as an attempt to provide a useful conceptualization concerning the very nature of consciousness that could help steer us in the direction of a solution. My remarks will focus mainly on whether in fact anything positive is to be gained from the "Hard Problem" characterization, or whether that conceptualization is counterproductive.

I cannot hope to do full justice to the task in short compass, especially as the contemporary characterization of the problem of consciousness as the intractable problem has a rather large literature surrounding it. The watershed articulation of consciousness as 'the most difficult problem' is Thomas Nagel's classic paper 'What is it like to be a bat?' (1974) In his opening remarks, Nagel comes straight to the point: 'Consciousness is what makes the mind-body problem really intractable.' Delineating a contrast between the problem of consciousness and all other mind-body problems, Nagel asserts: "While an account of the physical basis of mind must explain many things, this [conscious experience] appears to be the most difficult." Following Nagel's lead, many other philosophers, including Frank Jackson, Saul Kripke, Colin McGinn, John Searle, and most recently, David Chalmers, have extended and developed Nagel's basic idea that consciousness is not tractable neuroscientifically.

Although I agree that consciousness is, certainly, *a* difficult problem, difficulty *per se* does not distinguish it from oodles of other neuroscientific problems. Such as how the brains of homeotherms keep a constant internal temperature despite varying external conditions. Such as the brain basis for schizophrenia and autism. Such as why we dream and sleep. Supposedly, something sets consciousness apart from *all* other macro-function brain riddles such that it stands alone as The Hard Problem. As I have tried to probe precisely what that is, I find my reservations multiplying.

B. Carving Up the Problem Space

The "Hard-Problem" label invites us to adopt a principled empirical division between consciousness (The Hard Problem) and problems on the "Easy" (or perhaps hard but not Hard?) side of the ledger. "Easy" presumably encom-

passes problems such as the nature of short-term memory, long-term memory, autobiographical memory, the nature of representation, the nature of sensorimotor integration, top-down effects in perception -- not to mention such capacities as attention, depth perception, intelligent eye movement, skill acquisition, planning, decision-making, and so forth. On the other side of the ledger, all on its own, stands consciousness -- a uniquely Hard Problem.

My lead-off reservation arises from this question: what is the rationale for drawing the division exactly *there*? Might not, say, attention be as hard a problem as awareness? Dividing off consciousness from all of the so-called "easy problems" listed above implies that we could understand all those phenomena and still not know... know what? How the "qualia-light" goes on? Now *that* is about as insightful a conceptualization as supposing babies are brought by storks.

What exactly is the evidence for the conviction that we could explain all the "Easy" phenomena and still not understand the neural mechanisms for consciousness? The "evidence" derives from a thought-experiment, which roughly goes as follows: we can conceive of a person, like us in all the aforementioned Easy-to-explain capacities (attention, short term memory, etc.), but lacking qualia. This person would be *exactly* like us, save that he would be a Zombie -- an anaqualiac, one might say. Since the scenario is conceivable, supposedly that makes it possible; if it is possible, then whatever consciousness is, it is explanatorily independent of those activities.[2] (Something akin to this was argued by Saul Kripke in the 1970's.)

I take this argument to be a demonstration of the feebleness of thought experiments. *Saying* something is possible does not thereby guarantee it *is* a real possibility. So how do we know the anaqualiac idea is really possible? To insist that it *must* be is simply to beg the question at issue. That is, it is to insist that the neurobiology for attention, short term memory, decision making, integration etc. could all be understood without our ever understanding consciousness. But that claim is what we want an argument *for*; it cannot be used to prove itself. As Francis Crick has observed, it might be like saying that one can imagine a possible world where gases do not get hot, even though their constituent molecules are moving at high velocity. As an argument against the empirical identification of temperature with mean molecular KE, the thermo-dynamic thought-experiment is feebleness itself.[3]

---

[2] As I lacked time in my talk at Tucson to address the "Mary" problem, a problem first formulated by Frank Jackson in 1982, let me make several brief remarks about it here. In sum, Jackson's idea was that there could exist someone, call her Mary, who knew everything there was to know about how the brain works but still did not know what it was to see the color green (suppose she lacked 'green cones', to put it crudely.) This possibility Jackson took to show that qualia are therefore not explainable by science. The main problem with the argument is that to experience green qualia, certain wiring has to be in place in Mary's brain, and certain patterns of activity have to obtain and since, by Jackson's own hypothesis, she does not have that wiring, then presumably the relevant activity patterns in visual cortex are not caused and she does not experience green. Who would expect her visual cortex -- V4, say -- would be set ahumming just by virtue of her *propositional* (linguistic) knowledge abont activity patterns in V4? Not me, anyhow. She can have propositional knowledge via other channels, of course, including the knowledge of what her own brain lacks vis a vis green qualia. Nothing whatever follows about whether science can or cannot explain qualia.

[3] In *Neurophilosophy* I suggested that a typical weakness with philosophers' thought experiments is too much thought and not enough experiment.

Is consciousness -- the problem on the "Hard" side of the ledger -- sufficiently well-defined to sustain the Hard/Easy division as a fundamental *empirical* principle? Although it is simple enough to agree about the presence of qualia in certain prototypical cases, such as the pain felt after a brick has fallen on a bare foot, or the blueness of the sky on a sunny summer afternoon, things are less clear-cut once we move beyond the favored prototypes. Some of our perceptual capacities are rather subtle, as, for example, positional sense is often claimed to be. Some philosophers, e.g. Elizabeth Anscombe, have actually opined that we can know the position of our limbs without any "limb-position" qualia. As for me, I am inclined to say I do have qualitative experiences of where my limbs are -- it feels different to have my fingers clenched than unclenched, even when they are not visible. The disagreement itself, however, betokens the lack of consensus once cases are at some remove from the central prototypes.

Vestibular-system qualia are yet another non-prototypical case. Is there something "vestibular-y" it feels like to have my head moving? To know which way is up? Whatever the answer here, at least the answer is not glaringly obvious. Do eye movements have eye-movement qualia? Some maybe do, and some maybe do not. Are there "introspective qualia", or is introspection just paying attention to perceptual qualia and talking to yourself? Ditto for self-awareness. Thoughts are also a bit problematic in the qualia department. Some of my thoughts seem to me to be a bit like talking to myself and hence like auditory imagery. On the other hand, some just come out of my mouth as I am talking to someone or affect decisions without ever surfacing as inner dialogue. So do these cases belong on the "Hard" or the 'Easy' side of the ledger? None of this denies the pizzazz of qualia in prototypical cases. The point is just that prototypical cases give us only a *starting point* for further investigation and nothing like a full characterization of the class as a whole to which they belong.

My suspicion with respect to The Hard Problem strategy is that it seems to take the class of conscious experiences to be much better defined than it is. The point is, if you are careful to restrict your focus to the prototypical cases, you can easily be hornswoggled into assuming the class is well-defined. As soon as you broaden your horizons, troublesome questions - about fuzzy boundaries, about the connections between attention, short term memory and awareness -- are present in full, what-do-we-do-with-*that* glory.[4]

Are the Easy Problems *known to be easier* than The Hard Problem? To begin with, it is important to acknowledge that for none of the so-called "easy" problems, do we have an understanding of their solution. (See the partial list on p. 2) It is just false that we have anything approximating a comprehensive theory of sensorimotor control or attention or short-term memory or long-term memory. Consider one example. A signature is recognizably the same whether signed with the dominant or nondominant

[4] As I understand Dennett, these considerations loom large in his approach, e.g. *Consciousness Explained* (1994).

hand, with the foot, with the mouth or with the pen strapped to the shoulder, or written in half-inch script or in two-foot graffiti. How is "my signature" represented in the nervous system? How can completely different muscle sets be invoked to do the task, even when the skill was not acquired using those muscles? We still do not understand the general nature of sensorimotor representation.

Notice that it is not merely that we are lacking details, albeit important details. The fact is, we are lacking important conceptual/ theoretical ideas about how the nervous system performs fundamental functions -- such as time management, such as motor control, such as learning, such as information retrieval. We do not understand how back projections work, or the degree to which processing is organized hierarchically. These are genuine and very deep puzzles, and it is unwise to "molehill" them in order to "mountain" up the problem of consciousness. Although quite a lot is known at the cellular level, the fact remains that how real neural networks work and how their output properties depend on cellular properties still abounds with nontrivial mysteries. Naturally I do not wish to minimize the progress that has been made in neuroscience, but it is prudent to have a probative assessment of what we really do not yet understand.

Carving the explanatory space of mind-brain phenomena along a "Hard" and "Easy" divide, as Chalmers proposes, poses the danger of inventing an explanatory chasm where there really exists just a broad field of ignorance. It reminds me of the division, deep to medieval physicists, between sublunary physics (motion of things below the level of the moon) and superlunary physics (motion of things above the level of the moon). The conviction was that sublunary physics was tractable, and is essentially based on Aristotelian physics. Heavy things fall because they have gravity, and fall to their Natural Place, namely the earth, which is the center of the universe. Things like smoke have levity, and consequently they rise, *up* being their Natural Place. Everything in the sublunary realm has a Natural Place, and that is the key to explaining the behavior of sublunary objects. Superlunary events, by contrast, we can neither explain nor understand -- not, at least, in sublunary terms.

This old division was not without merit, and it did entail that events such as planetary motion and meteors were considered unexplainable in terrestrial terms, as they were Divinely governed. Although I do not know that Chalmers' Easy/Hard distinction will prove ultimately as misdirected as the Sublunary/Superlunary distinction, neither do I know it is any more sound. What I do suspect, however, is that it is much too early in the science of nervous systems for this distinction to command much credence.

One danger inherent in embracing the distinction as a principled empirical distinction is that it provokes the intuition that only a real humdinger of a solution will suit The Hard Problem. Thus the idea seems to go as follows: the answer, if it comes at all, is going to have to come from somewhere Really Deep -- like quantum mechanics, or -- Wow -- perhaps it requires a whole new physics. As the lone enigma, consciousness surely cannot be just a matter of a complex dynamical system doing its thing. Yes,

there are emergent properties from nervous systems such as co-ordinated movement as when an owl catches a mouse, but consciousness must an emergent property like unto no other. After all, it is *The Hard Problem*! Consequently, it will require a very deep, very radical solution. That much is evident sheerly from the hardness of The Hard Problem.

I confess I cannot actually see that. I do not know anywhere nearly enough to see how to solve either the problem of sensorimotor control or the problem of consciousness. I certainly cannot see enough to know what one problem will, and the other will not, require a Humdinger solution.

C. Using Ignorance as a Premise

In general, what substantive conclusions can be drawn when science has not advanced very far on a problem? Not much. One of the basic skills we teach our philosophy students is how to recognize and diagnose the range of non-formal fallacies that masquerade as kosher arguments: what it is to beg the question, what a *non sequitur* is, and so on. A prominent item in the fallacy roster is *argumentum ad ignorantiam* -- argument from ignorance. The canonical version of this fallacy uses ignorance as the key premise from which a substantive conclusion is drawn. The canonical version looks like this:

> We really do not understand much about a phenomenon P.
> (Science is largely ignorant about the nature of P.)
>
> Therefore: we *do* know that:
> (1) P can never be explained *or*
> (2) Nothing science could ever discover would deepen our understanding of P. *or*
> (3) P can never be explained in terms of scientifically familiar properties of kind S.

In its canonical version, the argument is obviously a fallacy: none of the tendered conclusions follow, not even a *little* bit. Surrounded with rhetorical flourish, brow-furrowing and hand-wringing, however, versions of this argument can hornswoggle the unwary.

From the fact that we do not know something, nothing very interesting follows -- we just don't know. Nevertheless, the temptation to suspect that our ignorance is telling us something positive, something deep, something metaphysical or even radical, is ever-present. Perhaps we like to put our ignorance in a positive light, supposing that but for the Profundity of the phenomenon, we *would* have knowledge. But there are many reasons for not knowing, and the specialness of the phenomenon is, quite regularly, not the real reason. I am currently ignorant of what caused an unusual rapping noise in the woods last night. Can I conclude it must be something special, something unimaginable, something... alien ... other-worldly? Evidently not. For all I can tell now, it might merely have been a raccoon gnawing on the compost bin. Lack of evidence for something is just that: lack of evidence.

It is not positive evidence for something else, let alone something of a humdingerish sort. That conclusion is not very glamorous perhaps, but when ignorance is a premise, that is about all you can grind out of it. Now if neuroscience had progressed as far on the problems of brain function as molecular biology has progressed on transmission of hereditary traits, then of course we would be in a different position. But it has not. The only thing you can conclude from the fact that attention is mysterious, or sensorimotor integration is mysterious, or that consciousness is mysterious, is that we do not yet understand the mechanisms.

Moreover, the mysteriousness of a problem is not a fact about the problem, it is not a metaphysical feature of the universe -- it is an epistemological fact about *us*. It is about where we are in current science, it is about what we can and cannot understand, it is about what, given the rest of our understanding, we can and cannot imagine. It is not a property of the problem itself. It is sometimes assumed that there can be a valid transition from "we cannot now explain" to "we can never explain" , so long as we have the help of a subsidiary premise, namely, "I cannot *imagine* how we could ever explain ... ". But it does not help, and this transition remains a straight-up application of argument from ignorance. Adding "I cannot imagine explaining P" merely adds a psychological fact about the speaker, from which again, nothing significant follows about the nature of the phenomenon in question. Whether we can or cannot imagine a phenomenon being explained in a certain way is a psychological fact about us, not an objective fact about the nature of the phenomenon itself. To repeat, it is an epistemological fact -- about what, given our current knowledge, we can and cannot understand. It is not a metaphysical fact about the nature of the reality of the universe.

Typical of vitalists generally, my high school biology teacher argued for vitalism thus: I cannot *imagine* how you could get living things out of dead molecules. Out of bits of proteins, fats, sugars --how could life itself emerge? He thought it was obvious from the sheer mysteriousness of the matter that it could have no solution in biology or chemistry. He assumed he could tell that it would require a Humdinger solution. Typical of lone survivors, a passenger of a crashed plane will say: I cannot imagine how I alone could have survived the crash, when all other passengers died instantly. Therefore God must have plucked me from the jaws of death.

Given that neuroscience is still very much in its early stages, it is actually not a very interesting fact that someone or other cannot imagine a certain kind of explanation of some brain phenomenon. Aristotle could not imagine how a complex organism could come from a fertilized egg. That of course, was a fact about Aristotle, not a fact about embryogenesis. Given the early days of science (circa 350 BC), it is no surprise that he could not imagine what it took many scientists hundreds of years to discover. I cannot imagine how ravens can solve a multi-step problem in one trial, or how temporal integration is achieved, or how thermoregulation is managed. But this is a (not very interesting) psychological fact about me. One could, of course, use

various rhetorical devices to make it seem like an interesting fact about me, perhaps by emphasizing that it is a really *really* hard problem. If, however, we are going to be sensible about this, it is dear that my inability to imagine how thermoregulation works is *au fond*, pretty boring.

The "I-cannot-imagine" gambit suffers in another way. Being able to imagine an explanation for Pis a highly open-ended and under-specified business. Given the poverty of delimiting conditions of the operation, you can pretty much rig the conclusion to go whichever way your heart desires. Logically, however, that flexibility is the kiss of death.

Suppose someone claims that she can imagine the mechanisms for sensorimotor integration in the human brain but cannot imagine the mechanisms for consciousness. What exactly does this difference amount to? Can she imagine the former in *detail*? No, because the details are not known. What is it, precisely, that she can imagine? Suppose she answers that in a very general way she imagines that sensory neurons interact with interneurons that interact with motor neurons, and via these interactions, sensorimotor integration is achieved. Now if that is all "being able to imagine" takes, one might as well say one can imagine the mechanisms underlying consciousness. Thus: "The intemeurons do it." The point is this: if you want to contrast being *able* to imagine brain mechanisms for attention, short term memory, planning etc., with being *unable* to imagine mechanisms for consciousness, you have to do more that say you can imagine neurons doing one but cannot imagine neurons doing the other. Otherwise one simply begs the question.

To fill out the point, consider several telling examples from the history of science. Before the tum of the twentieth century, people thought that the problem of the precession of the perihelion of Mercury was essentially trivial. It was annoying, but ultimately, it would sort itself out as more data came in. With the advantage of hindsight, we can see that assessing this as an easy problem was quite wrong -- it took the Einsteinian revolution in physics to solve the problem of the precession of the perihelion of Mercury. By contrast, a really hard problem was thought to be the composition of the stars. How could a sample ever be obtained? With the advent of spectral analysis, that turned out to be a readily solvable problem. When heated, the elements turn out to have a kind of fingerprint, easily seen when light emitted from a source is passed through a prism.

Consider now a biological example. Before 1953, many people believed, on rather good grounds actually, that in order to address the copying problem (transmission of traits from parents to offspring), you would first have to solve the problem of how proteins fold. The former was deemed a much harder problem than the latter, and many scientists believed it was foolhardy to attack the copying problem directly. As we all know now, the basic answer to the copying problem lay in the base-pairing of DNA, and it was solved first. Humbling it is to realize that the problem of protein folding (secondary and tertiary) is *still* not solved. *That*, given the lot we now know, does seem to be a hard problem.

What is the point of these stories? They reinforce the message of the argument from ignorance: from the vantage point of ignorance, it is often very difficult to tell which problem is harder, which will fall first, what problem will turn out to be more tractable than some other. Consequently our judgments about relative difficulty or ultimate tractability should be appropriately qualified and tentative. Guesswork has a useful place, of course, but let's distinguish between blind guesswork and educated guesswork, and between guesswork and confirmed fact. The philosophical lesson I learned from my biology teacher is this: when not much is known about a topic, don't take terribly seriously someone else's heartfelt conviction about what problems are scientifically tractable. Learn the science, do the science, and see what happens.

II. Searle: Consciousness is an Intrinsic, Irreducible Property

Rather different neural-naysayers are John Searle and Roger Penrose. Each has a distinct but articulated mistrust of the prospects for success of the neurobiological project, though neither rejects the value of neuroscience in contributing to our understanding of consciousness. Rather, both believe that a fundamental change in science is needed to do justice to the phenomenon of conscious experience. I shall focus first on Searle's views, and in the next section, on those of Penrose and his collaborator, Stuart Hameroff.

Searle's view is that the brain *causes* conscious states, but that conscious states are not *explainable in terms of* states of the brain. Why not? According to Searle, being aware of the smell of cinnamon, for example, is *ontologically distinct* from any physical state of the brain. The basic point is that consciousness is an intrinsic property. What does this mean? To a first approximation, that it has no hidden structure, that it has no aspects not immediately available to introspection; crudely, it is what it is and not another thing. Taking Searle at his word that he is no dualist, what does "ontological distinctness" mean? Since Searle says that ontology has to do with "what real features exist in the world", what does the brain-inexplicability of mental states mean as a matter of *ontology*?

Here Chalmers might chime in bravely: "it means that conscious experience is a fundamental feature of the universe, along with mass, charge, and spin." Searle does not say precisely that, though he does seem to rub shoulders with it when he suggests that science as it currently exists is not equipped to cope with the ontological distinctness of conscious awareness. Moreover, he seems to suggest that he is indifferent between going the fundamental-feature-of-reality route and property dualism, according to which mental states are nonphysical states of the brain. He says "Whether we treat the irreducibility from the materialist or the [property] dualist point of view, we are left with a universe that contains an irreducibly subjective physical component as a component of reality."[5] (p. 117, *The Rediscovery of*

the Mind.) In this respect then, Chalmers and Searle are like-minded.[6]

What is it about mental states that makes them inexplicable in terms of brain states, even though they *are* brain states? Is it that the subjectivity of mental states is owed to the fact that we *know* something about them from the inside by having them? Apparently not, for he emphasizes repeatedly that his point is *ontological*, not *epistemological*; it [Searle's point] is "....not, except derivatively, about how we know about those features.." (p. 117, *The Rediscovery of the Mind*.)

Searle gets to his radical irreducible-to-brain-states conclusion on the back of a premise he takes to be obviously true: whereas science might find the reality behind the appearance for objective phenomena -- fire, light, life, etc. -- in the case of consciousness, the appearance *is* the reality. And if the appearance -- seeing blue, feeling pain -- is the reality, then nothing neuroscience can discover will ever show me anything about the pain that is more real than feeling it. Feeling the pain is all the reality there is to pain.

Searle's premise has an obviously true bit and probably false bit, and the second slips in under the skirts of the first. What is obviously true is that sensations are real. My pains are as quite as real as Searle's. What is troublesome is the idea that all the reality there is to a sensation is available through sheerly having it. How could you possibly know that? I suggest instead a rather simple alternative: a sensation of pain is real, but not *everything* about the nature of pain is revealed in introspection. Its neural substrate, for example. Commonly science discovers ways of taking us beyond the manifest or superficial properties of some phenomenon. Light is refracted in water -- that is observable. Light turns out to be electromagnetic radiation, a property not straightforwardly observable. Does the observable property - refraction -- cease to be real or reputable or in good standing when we discover one of light's unobservable properties? No. If a property seems to common sense to be "structureless", is that a guarantee that it *is* structureless? No.

Perhaps Searle has been led to his "irreducible-property-of-the-brain" idea as a result of his assumption that "reductionism" implies "go-away-ism". Applying this extreme characterization of reductionism to the case of conscious experience, he is misled thus: (1) if we get an explanation of conscious states in neurobiological terms, that means we have a reduction. (2) If we have a reduction, then conscious states are not to be considered real - they are eliminated. (3) But conscious states *are* real -- any idiot knows that. Conclusion: we cannot explain conscious states neurobiologically.

The undoing of this argument is the falsity of its second premise. Reductions are *explanations* -- of macrophenomena in terms of microphenomena. When physics explains, for example, that temperature *is* mean molecular KE, or electricity is movement of electrons, or light *is* electromagnetic radiation,

---

[5] In Matter and Consciousness (1988), Paul Churchland raises this very possibility, and evaluates it as being at best a highly remote possibility. Searle appears indifferent to this evaulation and its rationale, though so far as I can tell, they remain in good standing.

[6] For an earlier and quite detailed discussion of this approach, see Bennett, Hoffman and Prakash (1989).

science does not *thereby* say there is no such thing as temperature, electricity or light. Just as, when we discover that Kim Philby is in fact the Russian mole, KS, this does not imply that Kim Philby does not exist. It is just that where we thought there were two distinct men, there is in fact just one, conceived in different ways under distinct circumstances.

Eliminative discoveries, such as the explanation of burning in terms of oxygen rather than phlogiston, *do* imply the non-existence of phlogiston. That is because oxygen and phlogiston provide mutually incompatible explanations. Only *some* scientific developments are eliminativist, however. Whether a theory about some macro phenomenon is smoothly reduced to a micro level theory, or whether there is a kind of conceptual revolution, depends on the facts of the case. It does not depend on what sounds funny to time-honoured intuitions.

So far as I can tell, the main difference between Searle's (1992) view, and my view, as articulated in *Neurophilosophy* (1986) as well as in many articles, concerns whether or not it is reasonable to try to explain consciousness in neurobiological terms (e.g. P. S. Churchland 1983, 1988, 1990, 1994, 1995). I think it is worth trying. Searle does not. If, however, you took Searle's rendition of my view, you would conclude I was a blithering idiot. He repeatedly claims that I hold a "ludicrous and insane view" -- to wit, that consciousness is to be eliminated from science, that consciousness does not really exist, that there is no such thing as awareness.[7]

Do I hold this view? Not at all. In print as well as in conversation with Searle, I have consistently argued that of course the *phenomenon* is real -- I have no inclination whatever to deny the existence of awareness. I do think science may discover some surprising things about it, that scientific discoveries may prompt us to redraw some categorial boundaries, perhaps to introduce some new words and to discover a more insightful vocabulary. Whether or not that actually happens depends on the nature of the empirical discoveries.

The not-very-subtle distinction Searle ignores is between reconfiguration

---

[7] In *The Rediscovery of the Mind* (1992), Searle cites my reply to McGinn as showing where I hold insane views about the elimination of consciousness. I enclose my entire reply to McGinn as Appendix 1. I am at a loss to understand what Searle has in mind. In my first paper on consciousness, "Consciousness: The Transmutation of a Concept" (1983), as the title conveys it is the transmutation of a *concept* that I am talking about. Moreover, in that paper I suggest how we might make progress in studying the phenomenon: "In addition to the research already discussed, studies to determine the neurophysiological differences between conscious and nonconscious states, to find out what goes on in the brain during REM sleep and during non-REM sleep, to determine the order and nature of events during the various facets of attention, and how this fits into a broader scheme of intelligent brain activity, would surely contribute to a deeper understanding of what sort of business consciousness is." (p. 93) Does this sound like I want to eliminate consciousness? Where is Searle getting his claim? Is the insane idea perhaps in *Neurophilosophy*? There I say, "This research [on sleep and dreaming] raises questions about whether there are different kinds of conscious states, and different ways or different levels of being aware that we have yet to understand." p. 208. Rather mild, I would have thought. I also say, "The brain undoubtedly has a number of mechanisms for monitoring brain processes, and the folk psychological categories of 'awareness' and 'consciousness' indifferently lump together an assortment of mechanisms. As neurobiology and neuropsychology probe the mechanisms and functions of the brain, a reconfiguring of categories can be predicted." (p. 322). This obviously concerns reconfiguration of the categories, not denying the existence of the phenomenon. (See also "Reduction and the neurobiological basis of consciousness." (1988), in *Consciousness in Contemporary Science,* ed. by Marcel and Bisiach. *Where is the evidence that I want to eliminate the phenomenon of consciousness?*

of a concept and denial of a phenomenon. I envisage the possibility of lots of conceptual reconfiguration as the brain and behavioral sciences develop. Never have I come even close to denying the reality of my experiences of pain, cold etc. In a reply to critics in 1986 (sic), I wrote a section called "Eliminative Materialism: What Gets Eliminated?" There I say: "Now, to put not too fine a point on it, the world is as it is; theories come and go, and the world keeps on doing whatever it is doing. So theory modification does not entail that the nature of the world is *ipso facto* modified, though our understanding of the nature of the world is modified. So if anyone thought eliminativism means that some part of the *ding an sich* -- say, whatever it is that we now think of as 'qualia'-- is eliminated by a mere tinkering with theory, then with Campbell, I agree he is certainly confused." (p. 247). This was uttered in 1988, four years before the publication of Searle's book.

There may be another badger to be flushed from the woodpile. A few philosophers (I do not know whether this includes Searle) expect that the difference between *my* feeling a pain and *your* feeling a pain would somehow be unreal or nonexistent on the hypothesis that pain actually is a neurobiological state. As supportive illustration, the naysayer imagines a scenario in which the neural reality behind the felt reality of pain was known to be a pattern of activity in neurons. Further in the scenario, it transpires that if I saw that very pattern right now in your brain, then I too would have that pain. Conclusion: the scenario is really silly; ergo, the hypothesis is really silly.

Not so fast. The scenario is silly, but *does* it follow from the hypothesis in question? Not at all. Suppose the hypothesis is true -- pains are physical states of the brain. Then I would expect that *you* will feel a pain only if the appropriate pattern of neuronal activity exists in *your* brain; and *I* will feel it only if the activity exists in *my* brain, and so on. Whyever assume that my seeing your neuronal activity would produce that very pattern of activity in my brain? After all, I can see the pattern of neuronal activity that produces a knee jerk in you, without my own knee jerking.

Subjectivity is a matter of whose brain is in the relevant physical state such that the person feels something or sees something or hears something. As I see it, pain is a real -- physical -- state of the brain, felt as painful by the person whose brain is appropriately configured, though detectable as a pattern of activity by a neuroscienist, had we a sufficiently powerful noninvasive imaging technique to make it visible.

Adopting Searle's hypothesis that my feeling a pain is either an irreducible state of the universe (unexplainability) or a nonphysical state of the universe (property dualism) is to take a grand leap in the dark. The more sensible course would be to leap only when it is tolerably clear that the data demand it. And the data, so far anyhow, do not demand it -- they do not even make it moderately tempting. This is not to say Searle's hypothesis is definitely wrong, but only that adopting it is *very* costly in pragmatic terms, since no one has the slightest testable idea how to make it fit with what we know about physics, chemistry, evolutionary biology, molecular biology, and the rest of neuroscience.

III. Penrose and Hameroff: Quantum Gravity and Microtubules

Roger Penrose and Stuart Hameroff also harbor reservations about explaining awareness neurobiologically, but are moved by rather different reasons (Penrose and Hameroff 1995). They believe the dynamical properties at the level of neurons and networks to be incapable of generating consciousness, regardless of the complexity. For Penrose and Hameroff, the key to consciousness lies in quantum events in tiny protein structures -- microtubules -- within neurons. Why *there*? And why *quantum level* phenomena? Because the nature of mathematical understanding, Penrose believes, transcends the kind of computation that could conceivably by done by neurons and networks. As a demonstration of neuronal inadequacy, Penrose cites the Godel Incompleteness Result, which concerns limitations of provability in axiom systems for arithmetic. What is needed to transcend these limitations, according to Penrose, are unique operations at the quantum level. Quantum gravity, were it to exist, could do the trick. Granting that no adequate theory of quantum gravity exists, Penrose and Hameroff argue that microtubules are about the right size to support the envisioned quantum events, and they have the right sort of sensitivity to anesthetics to suggest they do sustain consciousness.

The details of the Penrose-Hameroff theory are highly technical, drawing on mathematics, physics, biochemistry and neuroscience. Before investing time in mastering the details, most people want a measure of the theory's "figures of merit", as an engineer might put it.[8] Specifically: is there any hard evidence in support of the theory, is the theory testable, and if true, would the theory give a clear and cogent explanation of what it is supposed to explain? After all, there is no dearth of crackpot theories on every topic from consciousness to sun spots. Making theories divulge their figures of merit is a minimal condition for further investment.

First, a brief interlude to glimpse the positive views Penrose has concerning the question of how humans understand mathematics. In 1989, he suggested as unblushing a Platonic solution as Plato himself proposed *circa* 400BC:

> "Mathematical ideas have an existence of their own, and inhabit an ideal Platonic world, which is accessible via the intellect only. When one "sees" a mathematical truth, one's consciousness breaks through into this world of ideas, and makes direct contact with it ... mathematicians commmunicate ... by each one having a *direct route to truth*. [Penrose's italics] (p. 428)

As a solution to questions in the epistemology of mathematics, Platonism is

[8] For the details behind my reservations, see Grush and Churchland (1995) and the reply by Penrose and Hameroff (1995). See also Putnam (1994, 1995), Feferman (1995), Benacerraf and Putnam (1983), Kitcher (1984). Pat Hayes and Ken Ford fmmd Penrose's mathematical argmnent to be so outlandish that they awarded him the Simon Newcombe Award in 1995 (Hayes and Ford 1995). They explain that Simon Newcombe (1835-1909) was a celebrated astronomer who insisted in various articles that manned flight was physically impossible.

not remotely satisfactory (for recent discussions, see, for example, Benacerraf and Putnam 1983; Kitcher 1984). Knowing what we now know in biology, psychology, physics and chemistry, the Platonic story of mathematical understanding is as much a fairy tale as claim that Eve was created from Adam's rib. Far better to admit we have no satisfactory solution than to adopt a "And-God-said-Lo" solution.

Let us return now to evaluating the quantum-gravity-microtubule theory of conscious experience. The figures of merit are not encouraging. First, mathematical logicians generally disagree with Penrose on what the Godel result implies for brain function (Feferman 1996; Putnam 1994, 1995). Additionally, the link between conscious experiences such as smelling cinnamon and the Godel result is obscure at best.

Now, is there any significant evidential link between microtubules and awareness? Hameroff believes microtubules are affected by hydrophobic anesthetics in such a way to cause loss of consciousness. But there is no evidence that loss of consciousness under anesthesia depends upon the envisaged changes in microtubules, and only indirect evidence that anesthetics do in fact -- as opposed to "could conceivably" -- have *any* effect on microtubules. On the other hand, plenty of evidence points to proteins in the neuron membrane as the principal locus of action of hydrophobic anesthetics (Franks and Lieb 1994; Bowdle, Horita and Kharasch 1994; Franks, this volume).

Is there any hard evidence that quantum coherence happens in microtubules? *Only that it might.* Surely the presence of cytoplasmic ions in the microtubule pore would disrupt these effects? *They might not.* Surely the effects of quantum coherence would be swamped by the millivolt signaling activity in the neuronal membrane? *They might not be.* Can the existence of quantum coherence in microtubules be tested experimentally? *For technical reasons, experiments on microtubules are performed in a dish, rather than in the animal.* If tests under these conditions failed to show quantum coherence, would that be significant? *No, because microtubules might behave differently in the animal, where we cannot test for these effects.* Does any of this, supposing it to be true, help us explain such things as recall of past events, filling in of the blindspot, hallucinations and attentional effects on sensory awareness? *Somehow, it might.*

The want of directly relevant data is frustrating enough, but the explanatory vacuum is catastrophic. Pixie dust in the synapses is about as explanatorily powerful as quantum coherence in the microtubules. Without at least a blueprint or an outline or a prospectus or *something* showing how, if true, the theory could explain the various phenomena of conscious experience, Penrose and Hameroff are offering a make-believe pig in a fantasy poke. None of this shows that Penrose and Hameroff are wrong, of course, only that the theory needs work.

Conceptual innovation is needed, and needed for a host of problems quite apart from sensory awareness. To be sure, most new ideas are bound to go the way of the three-legged trout. The idea-climate should not, of course,

be so harsh as to snuff any contender that looks outlandish. For this reason alone, I applaud the boldness of Penrose and Hameroff. Having looked closely at the details of their proposal, however, I find I am inclined to look elsewhere for ideas whose figures of merit are strong enough to invite serious investment.

VI. Concluding Remarks

1. Consciousness is a difficult problem, but for all we can tell now, it may turn out to be more tractable than other problems about neural function, such as how the brain manages to get timing right. We shall have to do the science and see.

2. Thought experiments are typically too underdescribed to give real credence to the conclusions they are asked to bear. All too often they are merely a heartfelt intuition dressed up to look like a scientifically grounded argument.

3. Consciousness might turn out to be a fundamental property of the universe, but so far there is no moderately convincing reason to believe it is. Insofar as most information processing in brains and machines is non-conscious, it is not plausible to assume that an information-based physics *per se* is the key to consciousness.

4. Consciousness might turn out to be produced by quantum coherence in microtubules, but so far there is no moderately convincing reason to believe that it is.

5. Let's keep plugging away at experimental psychology and neuroscience, trying to invent revealing experiments that will help us make progress on the problem. We need to continue developing both direct strategies (that have the neural substrate for awareness as their immediate target) and indirect strategies (that focus on perception, motor control, attention, learning and memory, emotions, etc.) with the hope that along the way, much will be revealed about awareness in those functions. We need to continue to address theoretical as well as experimental question, and to foster new ideas targetting how the brain solves problems such as sensory-motor integration, time-management, and using back projections in sensory systems to bias the perception, to fill-in, and to "see-as".

References

Benacerraf, P., and Putnam, H. (ed.) (1983). *Philosophy of mathematics; selected readings.* Englewood Cliffs, N.J.: Prentice-Hall.

Bennett, B. M., Hoffman, D. D., Prakash, C. (1989). *Observer mechanics.* San Diego: Academic Press.

Bowdle. T. A., A. Horita, and E. D. Kharasch (1994). *The Pharmacological Basis of Anesthesiology.* New York: Churchill Livingstone.

Chalmers, D. (1996). *The Conscious Mind.* Oxford: Oxford University Press.

Chalmers, D. (1995). "The puzzle of conscious experience." *Scientific American.* 273: 80-86.

Churchland, Patricia S. (1983). "Consciousness: The transmuation of a concept." *Pacific Philosophical Quarterly.* 64: 80-95.

Churchland, Patricia S. (1986) *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* Camrbidge, MA: MIT Press.

Churchland, Patricia S. (1986). "Replies to commentaries on Neurophilosophy." *Inquiry* 29: 241-272.

Churchland, Patricia S. (1988). "Reductionism and the neurobiological basis of consciousness." In: *Consciousness in Contemporary Science.* Ed. by A. M. Marcel and E. Bisiach. Oxford: Oxford University Press. 273-304.

Churchland, Paul M. (1988) *Matter and Consciousness.* Cambridge MA: MIT Press.

Churchland, Paul M. (1995). *The Engine of Reason; the Seat of the Soul.* Cambridge, MA: MIT Press.

Churchland, Paul M. (1996). "The rediscovery of light." *The Journal of Philosophy.* 93: 211-228.

Churchland, Paul M. and Churchland Patricia S. (1991). "Intertheoretic reduction: A neuroscientist's field guide." *Seminars in the Neurosciences.* 2: 249-256.

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

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

Feferman, S. (1995). *Penrose's Godelian Argument.* Psyche 2:23.

Franks, N. P. and W. R. Lieb (1994). Molecular and cellular mechanisms of general anaesthesia. *Nature.* 367: 607-614.

Franks, N. (this volume).

Grush, R. and P. S. Churchland (1995). "Gaps in Penrose's toilings." *Journal of consciousness studies.* 2:10-29.

Hayes, P and K. Ford (1995). "The Simon Newcombe Awards." *AI Magazine.* 16: 11-13.

Jackson, F. (1982). "Epiphenomenal qualia." *Philosophical Quarterly.* Vol. 32.

Kitcher, P. (1984). *The Nature of Mathematical Knowledge.* Oxford: Oxford University Press.

Nagel, T. (1974). "What is it like to be a bat?" *Philosophical Review.* Vol. 83.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford: Oxford University Press.

Penrose, R. (1994a). Interview with Jane Clark. *Journal of Consciousness Studies*, 1, 1:17-24.

Penrose, R. (1994b). *Shadows of the mind.* Oxford: Oxford University Press.

Penrose, R. and S. Hameroff (1995). "What gaps?" *Journal of consciousness studies.* 2:99-112.

Putnam, H. (1994). "The best of all possible brains?" Review of *Shadows of the Mind. New York Times Book Review.* November.

Putnam, H. (1995). Review of *Shadows of the Mind. Bulletin of the American Mathematical Society*

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

Searle, J. (1995). "The mystery of consciousness: Part II." *The New York Review of Books.* 42:

Appendix

This letter by me in response to McGinn's review of *Neurophilosophy* was published March 13, 1987, *Times Literary Supplement*:

Sir, --Galileo's telescope, and experimental evidence generally, were scorned by the cardinals as irrelevant and unnecessary. They knew that the celestial order was revealed as a matter of faith, pure reason, and sacred text. The very idea of moons revolving around Jupiter was an affront to all three. Observation of denormalizing data in the form of sun spots and Jovian satellites was therefore impossible. In his review (February 6, 1987) of my book, Neurophilosophy, Colin McGinn displays a general reaction to neuroscience that bears a chilling resemblance to that of the cardinals. For McGinn, the very idea that our intuitive convictions concerning the nature of the mind might be reassessed is "virtually inconceivable". Intuitive (folk) psychology, insists McGinn, is an "autonomous mode of person understanding" (sic), and the autonomy he claims for it keeps it sacred, shielding it from whatever might be discovered by empirical investigation. But to call an explanatory framework "autonomous" is just to cobble up a polite label for digging in and refusing to allow the relevance of empirical data. This is no more acceptable for psychology than it was for folk physics, folk astronomy, creationism vitalism or alchemy.

The main theme of the book is that if we want to understand the mind, research in neuroscience will have an essential role, as will research in psychology, ethology, computer modeling *and* philosophy. Very briefly, the reason is this: the brain is what sees, thinks, feels, and so forth, and if we want to know how it performs these jobs, we will have to look at its components and organization. Psychology is essential, because it provides constraints at a higher level, and helps the neurobiologist specify the functions to be explained by neural networks. Modeling is essential because there are properties at the level of circuits that cannot be determined at the level of single cell analysis. Co-evolution of theories at all levels, where each level informs, correct and inspires the others, is therefore the research ideology that looks most productive. At the same time, there is the empirical possibility that the result of a substantial period of co-evolution will yield a psychology and a neurobiology that look quite different from what we now work with. Some evidence in this direction is already available, as I show in several chapters of my book. Beyond the normal apprehension of things news, this prospect should not alarm McGinn, for it represents a deepening of our understanding of human nature.

What then is the role of philosophy? My view here is that philosophy is also essential to the wider project of understanding how the mind-brain works. It is, as always, the synoptic discipline: it attempts to synthesize the

existing sciences into a unified and coherent account of reality. And it is, as always, a seminal discipline: in addressing the limits of common-sense understanding, it attempts to found new sciences where none existed before. I think this role is very much in keeping with the long tradition in philosophy, as exemplified by Aristotle, Hume, Kant, James and Pierce. But I also say, "this sort of philosophy is not an a priori discipline pontificating grandly to the rest of science; it is in the swim with the rest of science, and hence stands to be corrected as empirical discovery proceeds." (p. 482)

McGinn, however, finds this conception of philosophy "absurd". He apparently wants to keep philosophy free from the taint of empirical science, pure to undertake that most subtle of tasks, the analysis of concepts. Whose concepts? The concepts of the scientifically uninformed. The trouble is, *know-nothing philosophy is dead-end philosophy*, and the divination of *a priori* truths is a delusion. Without bridges to the relevant disciplines, philosophy becomes insular, in-grown, and wanting in vigour. Such observations motivated Kenneth Craik's call in 1943 (!) for an "experimental philosophy" of mind.

The real absurdity is to make a virtue out of ignorance and to scoff at informed research as "scientism". The doctrine of keeping philosophy pure makes the discipline look silly, and it is philosophy pursued under the banner of purity that quite naturally provokes the impatience and incredulity of the wider intellectual community. Moreover, the very best research by contemporary philosophers is richly cross-disciplinary, as can be seen in the work of Ned Block, Dan Dennett, John Earman, Arthur Fine, Jerry Fodor, Clark Glymour, Adolf Grunbaum, John Haugeland, Philip Kitcher, Michael Redhead, Elliott Sober and Stephen Stich, to name a few. A willingness to cooperate across boundaries and an acute sense of the value of such exchanges is increasingly visible in this decade. This is surely a healthy development as we collectively get on with the question of how to make sense of our universe --and ourselves.