B. Quantum Theory and Human Consciousness

Leaders: Stuart Hameroff, University of Arizona[1]
Roger Penrose, University of Oxford
October 22, 1998

On October 22, 1998, the third in a series of study group meetings was held at RAND. The series, sponsored by the Defense Advanced Research Projects Agency (DARPA), focused on social and political governance questions arising from the impacts of the information and biological revolutions. This paper presents a summary of ideas and thoughts presented at the study group meeting.

Questions about consciousness and intelligence arise in the context of both the information and biological revolutions. To understand these phenomena, the session continued the study group's inquiry into the human brain. The goal was to understand the physical limitations of the brain, how it evolved to its current capabilities, and the origins of conscious thought and intelligence. This information is needed for putting any technological enhancements to the human brain in context that might be researched and attempted. In addition, this information is key to understanding whether computers or networks of computers can ever achieve a kind of conscious intelligence.

Presentation by Stuart Hameroff and Roger Penrose

The study group leaders invited Stuart Hameroff and Roger Penrose to present their research on the brain and human consciousness, and discuss the role of the computer in enhancing concepts of human intelligence. Professor Hameroff is a professor in the Departments of Anesthesiology and Psychology at the University of Arizona, and a physician on staff at the University Hospital. Professor Penrose is Rouse Ball Professor of Mathematics at the University of Oxford. He is the recipient of a number of awards, including the 1988 Wolf Prize (which he shared with Stephen Hawking for their research into the understanding of the universe), the Dannie Heinemann Prize, the Royal Society Royal Medal, and the Albert Einstein Prize. His 1989 book, *The Emperor's New*

[1]Additional information, as well as more detailed citations, is available on Dr. Hameroff's Web site: http://www.u.arizona.edu/~hameroff/ .

Mind, was a best-seller and won the 1990 Rhone-Poulenc Science Book Prize. His latest works include *Shadows of the Mind* (1994), *The Nature of Space and Time* (1996) (with Stephen Hawking), and *The Large, the Small and the Human Mind* (1997).

Together, Professors Hameroff and Penrose have developed a theory of consciousness. They propose that quantum theory and a newly proposed physical phenomenon, quantum wave function, are essential for consciousness and occur in cytoskeletal microtubules and other structures within the brain's neurons. Several papers on this theory can be found at Dr. Hameroff's Web site: http://www.u.arizona.edu/~hameroff/.

The Problem of Consciousness

Conventional explanations portray consciousness as an emergent property of classical computerlike activities in the brain's neural networks. While there is some disagreement as to the particular point of origin, the prevailing view among scientists in this camp is that (1) patterns of neural activity correlate with mental states; (2) synchronous network oscillations of neuronal circuits in the thalamus and cerebral cortex temporarily binds information; and (3) consciousness emerges as a novel property of computational complexity among neurons.

However, these approaches appear to fall short in fully explaining certain enigmatic features of consciousness, such as

- the nature of subjective experience, or "qualia"—our "inner life" (Chalmers' "hard problem," 1996)
- the binding of spatially distributed brain activities into unitary objects in vision, and a coherent sense of self, or "oneness"
- the transition from preconscious processes to consciousness itself
- noncomputability, or the notion that consciousness involves a factor that is neither random nor algorithmic, and that consciousness cannot be simulated (Penrose, 1989, 1994, 1997)
- free will
- subjective time flow.

Brain imaging technologies have demonstrate the anatomical locations of activities that appear to correlate with consciousness but may not be directly responsible for consciousness.

How do neural firings lead to thought and feelings? The conventionalist (also called functionalist, reductionist, materialist, physicalist, and computationalist) approach argues that neurons and their chemical synapses are the fundamental units of information in the brain and that conscious experience emerges when a critical level of complexity is reached in the brain's neural networks. The basic idea is that the mind is a computer functioning in the brain (brain = mind = computer).

However, in fitting the brain to a computational view, such explanations omit incompatible neurophysiological details:

- widespread apparent randomness at all levels of neural processes (is it noise or underlying levels of complexity?)
- glial cells (which accounts for some 80 percent of the brain)
- dendritic-dendritic processing
- electrotonic gap junctions
- cytoplasmic / cytoskeletal activities
- living state (the brain is alive!).

A further difficulty is the absence of testable hypotheses in emergence theory. No threshold or rationale is specified; rather, consciousness "just happens."

Finally, the complexity of individual neurons and synapses is not accounted for in such arguments. Since many forms of motile single-celled organisms lacking neurons or synapses are able to swim, find food, learn, and multiply through the use of their internal cytoskeleton, can they be considered more advanced than neurons? Are neurons merely simple switches, or are they something more?

Microtubules

Activities within cells ranging from single-celled organisms to the brain's neurons are organized by a dynamic scaffolding called the cytoskeleton. A major component of the cytoskeleton is the *microtubule*, a hollow, crystalline cylinder 25 nm in diameter. Microtubules are, in turn, composed of hexagonal lattices of proteins, known as *tubulin*.

Microtubules are essential to cell shape, function, movement, and division (Figure B.1). In neurons, microtubules self-assemble to extend axons and dendrites and to form synaptic connections, then help to maintain and regulate synaptic activity responsible for learning and cognitive functions (Figure B.2).

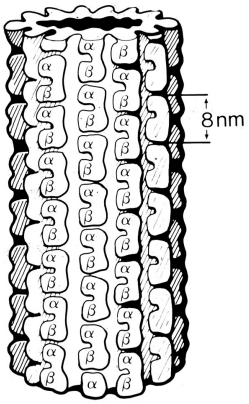

Figure B.1—Crystallographic Structure of Microtubules

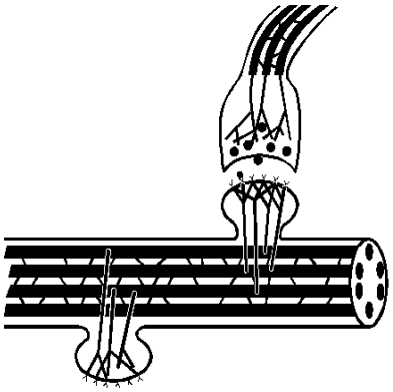

Figure B.2—Schematic View of Two Neurons Connected by Chemical Synapse. Axon terminal (above) releases neurotransmitter vesicles, which bind receptors on postsynaptic dendritic spine. Cytoskeletal structures microtubules ("MTs"—thicker tubes) are visible within the neurons, as well as actin, synapsin, and others that connect MTs to membranes.

While microtubules have traditionally been considered to be purely structural elements, recent evidence has revealed that mechanical, chemical, and electrical signaling and communication functions also exist, the result of microtubule interaction with membrane structures by linking proteins, ions and "second-messenger" signals, and voltage fields, respectively.

Current models propose that tubulins within microtubules undergo coherent excitation, switching between two or more conformational states in nanoseconds. Dipole couplings among neighboring tubulins in the microtubule lattice form

dynamic patterns, or "automata," which evolve, interact, and lead to the emergence of new patterns. Research indicates that microtubule automata computation could support classical information processing, transmission, and learning within neurons.

Microtubule automaton switching offers a potentially vast increase in the computational capacity of the brain. While conventional approaches focus on synaptic switching at the neural level, which optimally yields about 10^{18} operations per second in human brains ($\sim 10^{11}$ neurons per brain, with $\sim 10^{4}$ synapses per neuron, switching at $\sim 10^{3}$ sec^{-1}), microtubule automata switching can explain some 10^{27} operations per second ($\sim 10^{11}$ neurons with $\sim 10^{7}$ tubulins per neuron, switching at $\sim 10^{9}$ sec^{-1}). Indeed, the fact that all biological cells typically contain approximately 10^{7} tubulins could account for the adaptive behaviors of single-celled organisms, which have no nervous system or synapses. Rather than simple switches, then, it seems that neurons are actually complex computers.

Theories of Consciousness: Panexperiential Philosophy Meets Modern Physics

Still, greater computational complexity and ultrareductionism to the level of microtubule automata cannot address the enigmatic features of consciousness—in particular, the nature of the conscious experience. Something more is required. If functional approaches and emergence are incomplete, perhaps the raw components of mental processes (or "qualia") are fundamental properties of nature (like mass, spin, or charge). This view has long been held by panpsychists throughout the ages.

For example, Buddhists and Eastern philosophers claimed a "universal mind." Following the ancient Greeks, Spinoza argued in the 17th century that some form of consciousness existed in everything physical. The 19th century mathematician Leibniz proposed that the universe was composed of an infinite number of fundamental units, or "monads," with each possessing a form of primitive psychological being. In the 20th century, Russell claimed that there was a common entity underlying both mental and physical processes, while Wheeler and Chalmers have maintained that there is an experiential aspect to fundamental information.

Of particular interest is the work of the 20th century philosopher Alfred North Whitehead, whose panexperiential view remains most consistent with modern physics. Whitehead argued that consciousness is a process of events occurring in a wide, basic field of protoconscious experience. These events, or "occasions of

experience," may be comparable to quantum state reductions, or actual events in physical reality (Shimony, 1993). This suggests that consciousness may involve quantum state reductions (a form of quantum computation). But in what medium do such "occasions" occur?

Whether protoconscious experience, or qualia, could exist in the empty space of the universe depends upon how space is defined. Historically, space has been described as either an absolute void or a pattern of fundamental geometry. Democritus and the Michaelson-Morley results argued for "nothingness," while Aristotle ("plenum") and Maxwell ("ether") rejected the notion of emptiness in favor of "something"—a background pattern. Einstein weighed in on both sides of this debate, initially supporting the concept of a void with his theory of special relativity but then reversing himself in his theory of general relativity and its curved space and geometric distortions—the space-time metric. Could protoconscious qualia be properties of this fundamental metric?

Quantum Computing and Consciousness

At extremely small scales, space-time is not smooth, but quantized. Quantum electrodynamics and quantum field theory predict virtual particle-waves (or photons) that pop into and out of existence, creating quantum "foam" in their wake. Lamoreaux verified presence of virtual photons in space-time in 1997. In 1971, Roger Penrose modeled this granularity as a dynamic web of quantum spins. These "spin networks" create an array of geometric volumes and configurations at the Planck scale (10^{-33} cm, 10^{-43} secs), which dynamically evolve and define space-time geometry. If spin networks are the fundamental level of space-time geometry, they could provide the basis for protoconscious experience. Thus, particular configurations of quantum spin geometry would convey particular types of qualia, meaning and aesthetic values. A process at the Planck scale (e.g., quantum scale reductions) could then access and select configurations of experience.

If protoconscious information is embedded at the near-infinitesimal Planck scale, how could it be linked to biology? Penrose's answer is to extend Einstein's theory of general relativity (in which mass equates to curvature in space-time) down to the Planck scale. Specific arrangements of mass are, in reality, then specific configurations of space-time geometry. Events at the very small scale, however, are subject to the seemingly bizarre goings-on of quantum theory. A century of experimental observation of quantum systems has shown that, at least at small scales, particles (mass) can exist in two or more states or locations simultaneously. Penrose views this phenomenon of quantum superposition as

simultaneous space-time curvature in opposite directions—a separation or bubble in underlying reality.

Superposition and subsequent reduction, or collapse, to single, classical states may have profoundly important applications in technology, as well as toward the understanding of consciousness. In the 1980s, Benioff, Feynman, Deutsch, and other physicists proposed that states in a quantum system could interact (via entanglement) and enact computation while in quantum superposition of all possible states (i.e., "quantum computing"). While classical computing processes bits (or conformational states) as 1 or 0, quantum computations involve the processing of superpositioned "qubits" of both 1 and 0 (and other states) simultaneously.

Quantum theory also predicts that two or more particles, if once together, will remain somehow connected, even when separated by great distances. This "entanglement" enables quantum computing to achieve a nearly infinite parallel computational ability. Thus, quantum computers, if they can be constructed, will be able to solve important problems (e.g., factoring large numbers) with efficiency unattainable with classical designs (Shor, 1994).

Results, or solutions in quantum computing, are obtained when, after a period of quantum superposition, the qubits "collapse," or reduce to classical bit states. As quantum superposition may only occur in isolation from the environment, reduction may be induced by breaching isolation. But what about quantum superpositions that remain isolated—for example, Schrödinger's mythical cat, which is both dead and alive? This is the famous problem of wave function collapse, or quantum state reduction.

Roger Penrose's Objective Reduction (OR)

So how or why do quantum superimposed states that avoid environmental interactions become classical and definite in the macroworld? Many physicists now believe that some objective factor disturbs the superposition and causes it to collapse. Penrose proposes that this factor is an intrinsic feature of space-time itself: quantum gravity. According to Penrose's interpretation of general relativity, quantum superposition—a separation in mass from itself—is equivalent to separation in underlying space-time geometry, or simultaneous space-time curvatures in opposite directions.

Penrose argues that these separations in fundamental reality, or "bubbles," are unstable—even when isolated from the environment—and will reduce spontaneously and noncomputably to a specific state at a critical threshold of

separation or "decoherence," thereby avoiding the need for "multiple worlds." This objective threshold is defined by the indeterminacy principle:

$$E = h / T$$

where E is the gravitational self-energy of the superposed mass separated from itself; h is Planck's constant divided by 2p, and T is the coherence time until collapse occurs. Thus, the size and energy of a system in superposition, or the degree of space-time separation, are inversely related to the time T until reduction. (E can be calculated from the superposed mass m and the separation distance a. See Hameroff and Penrose, 1996a.)

Assuming isolation, the following masses in superposition would collapse at the designated times, according to Penrose's objective reduction:

Mass (m)	Time (T)
Nucleon	10^{7} years
Beryllium ion	10^{6} years
Water speck	
• 10^{-5} cm radius	Hours
• 10^{-4} cm radius	1/20 second
• 10^{-3} cm radius	10^{-3} seconds
Schrödinger's cat (m = 1 kg, a = 10 cm)	10^{-37} seconds

If quantum computation with objective reduction occurs in the brain, enigmatic features of consciousness could be explained:

- By occurring as a self-organizing process in what is suggested to be a panexperiential medium of fundamental space-time geometry, objective reductions could account for the nature of subjective experience by accessing and selecting protoconscious qualia.
- By virtue of involvement of unitary, entangled quantum states during preconscious quantum computation and the unity of quantum information selected in each objective reduction, the issue of binding may be resolved.
- Regarding the transitions from preconscious processes to consciousness itself, the preconscious processes may equate to the quantum superposition-computation phase, and consciousness to the actual, instantaneous objective reduction events. Consciousness may then be seen as a sequence of discrete events (e.g., at 40 Hz).

- As Penrose objective reductions are proposed to be noncomputable (reflecting influences from space-time geometry that are neither random nor algorithmic), conscious choices and understanding may be similarly noncomputable.
- Free will may be seen as a combination of deterministic preconscious processes acted on by a noncomputable influence.
- Subjective time flow derives from a sequence of irreversible quantum state reductions.

In what types of brain structures might quantum computation with objective reduction occur? If these events occur in the brain, they would be expected to coincide with known neurophysiological processes with appropriate time scales. For consciousness, then, T should be in range of tens to hundreds of milliseconds.

Event	T (ms)
Buddhist "moment of awareness"	13
"Coherent 40 Hz" oscillations	25
Electroencephalogram alpha rhythm (8 to 12 Hz)	100
Libet's sensory threshold (1979)	500

Objective reduction events in this time frame would require a mass on a nanogram scale. So, what is m?

Are Proteins Qubits?

Biological life is organized by proteins. By changing their conformational shape, proteins are able to perform a wide variety of functions, including muscle movement, molecular binding, enzyme catalysis, metabolism, and movement. Dynamical protein structure results from a "delicate balance among powerful countervailing forces" (Voet and Voet, 1995). The types of forces acting on proteins include charged interactions (such as covalent, ionic, electrostatic, and hydrogen bonds), hydrophobic interactions, and dipole interactions. The latter group, also known as van der Waals forces, encompasses three types of interactions:

- permanent dipole–permanent dipole
- permanent dipole–induced dipole
- induced dipole–induced dipole.

As charged interactions cancel out, hydrophobic and dipole–dipole forces are left to regulate protein structure. While induced dipole–induced dipole interactions, or London dispersion forces, are the weakest of the forces outlined above, they are also the most numerous and influential. Indeed, they may be critical to protein function. For example, anesthetics are able to bind in hydrophobic "pockets" of neural proteins and ablate consciousness by virtue of these London forces. London force attraction between any two atoms is usually less than a few kilojoules; however, since thousands occur in each protein, they add up to thousands of kilojoules per mole, and cause changes in conformational structure.

If proteins are qubits, assemblies of proteins in some type of organelle or biomolecular structure could act as a quantum computer. So which biological structures are best suited for objective reduction? Ideal structures would

- be abundant
- be capable of information processing and computation
- be functionally important (e.g., regulating synapses)
- be self-organizing
- be tunable by input information (e.g., microtubule-associated protein orchestration)
- be periodic and crystal-like in structure (e.g., dipole lattice)
- be isolated (transiently) from environmental decoherence
- be conformationally coupled to quantum events (e.g., London forces)
- be cylindrical waveguide structures
- have a plasmalike charge-layer coating.

While various structures or organelles have been suggested (e.g., membrane proteins, clathrins, myelin, presynaptic grids, and calcium ions), the most logical candidates are microtubule automata.

The Penrose-Hameroff Orchestrated Objective Reduction Model

The Penrose-Hameroff Orchestrated Objective Reduction (Orch OR) model proposes that quantum superposition–computation occurs in microtubule automata within brain neurons and glia. Tubulin subunits within microtubules act as qubits, switching between states on a nanosecond (10^{-9} sec) scale, governed by London forces in hydrophobic pockets. These oscillations are "tuned" and "orchestrated" by microtubule-associated proteins (MAPs),

providing a feedback loop between the biological system and the quantum state. These qubits interact computationally by nonlocal quantum entanglement, according to the Schrödinger equation, with preconscious processing continuing until the threshold for objective reduction (OR) is reached (E = h / T). At that instant, collapse occurs, triggering a "moment of awareness," or a conscious event—an event that determines particular configurations of Planck-scale experiential geometry and corresponding classical states of microtubule automata that regulate synaptic and other neural functions. A sequence of such events could provide a forward flow of subjective time and "stream" of consciousness. Quantum states in microtubules may link to those in microtubules in other neurons and glia by tunneling through gap junctions, permitting extension of the quantum state throughout significant volumes of the brain.

From E = h / T, the size and extension of Orch OR events that correlate with subjective or neurophysiological descriptions of conscious events can be calculated:

Event	T (ms)	E
Buddhist "moment of awareness"	13	4×10^{15} nucleons (4×10^{10} tubulins / cell ~ 40,000 neurons)
"Coherent 40 Hz" oscillations	25	2×10^{15} nucleons (2×10^{10} tubulins / cell ~20,000 neurons)
EEG alpha rhythm (8 to 12 Hz)	100	5×10^{14} nucleons (5×10^{9} tubulins / cell ~5,000 neurons)
Libet's sensory threshold (1979)	500	10^{14} nucleons (10^{9} tubulins / cell ~1,000 neurons)

But how could delicate quantum superposition–computation be isolated from environmental decoherence in the brain (generally considered to be a noisy thermal bath), while also communicating with the environment? One possibility is that quantum superposition–computation occurs in an isolation phase that alternates with a communicative phase (Figures B.3 through B.5). One of the most primitive biological functions is the transition of cytoplasm between a liquid, solution ("sol"), phase and a solid, gelatinous ("gel"), phase due to assembly and disassembly of the cytoskeletal protein actin. Actin sol-gel

transitions can occur at 40 Hz or faster and are known to be involved in neuronal synaptic release mechanisms.

Mechanisms for enabling microtubule quantum computation and avoiding decoherence long enough to reach the OR threshold may include

- sol-gel transitions
- plasma phase sleeves (Sackett)
- quantum excitations, ordering of surrounding water (Jibu / Yasue / Hagan)
- hydrophobic pockets
- hollow microtubule cores
- laserlike pumping, including environment (Frohlich / Conrad)
- quantum error correcting codes.

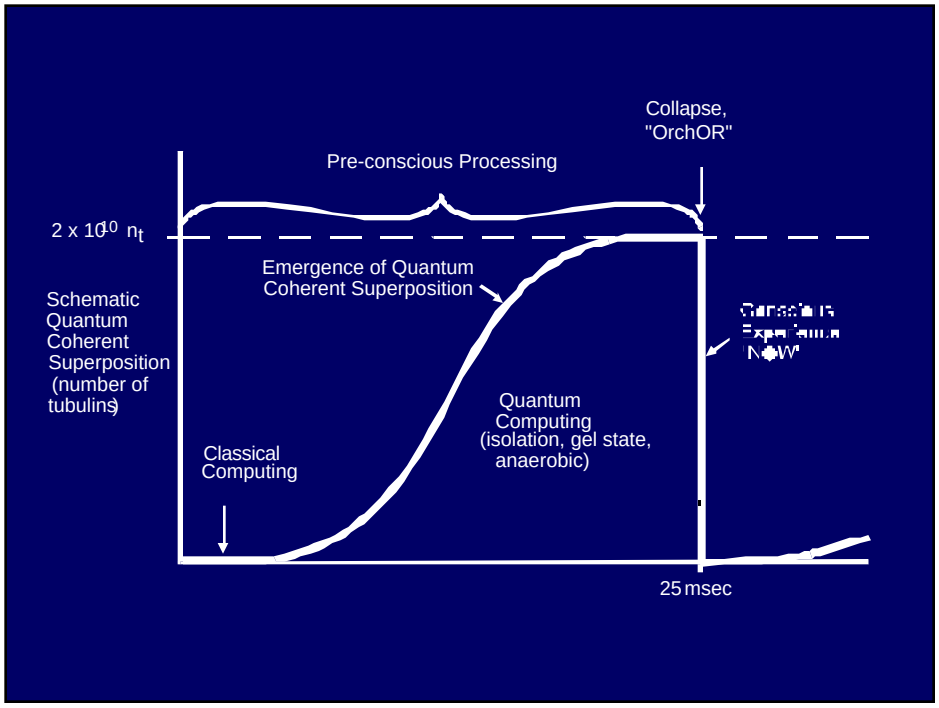

Figure B.3—Schematic Graph of Proposed Preconscious Quantum Superposition (number of tubulins) Emerging Versus Time in Microtubules. Area under curve connects superposed mass energy E with collapse time T in accordance with E=h/T. E may be expressed as nt, the number of tubulins whose mass separation for time T will self-collapse. For T = 25 msec (e.g. 40 Hz oscillations), nt = 2 x 1010 tubulins.

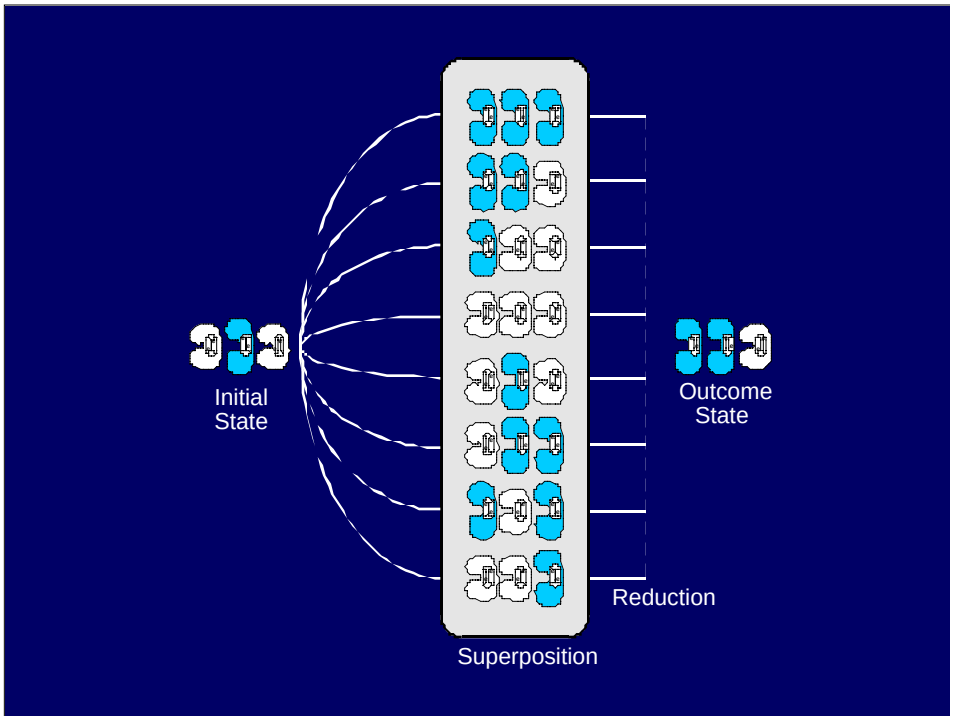

Figure B.4—Schematic of Quantum Computation. Three tubulins begin (left) in initial classical states, then enter isolated quantum superposition in which all possible states coexist. After reduction, one particular classical outcome state is chosen (right).

Orchestrated Objective Reduction, Cognition, and Free Will

Quantum computation with objective reduction may be associated with cognitive activities. While classical neural-level computation can provide a partial explanation, the Orch OR model allows for far greater information capacity and addresses issues of conscious experience, binding, and noncomputability consistent with free will. Such functions as face recognition and volitional choice may require a series of conscious events arriving at intermediate solutions. Preconscious processing of information occurs in the form of qubits, or superposed states of microtubule automata. As the threshold for objective reduction is reached, these qubits collapse to definite states and become bits, resulting in a conscious experience of recognition or choice.

The problem in understanding free will is that human actions seem neither totally deterministic nor random. In Orch OR, reduction outcomes involve a factor that is “noncomputable.” The microtubule quantum superposition evolves

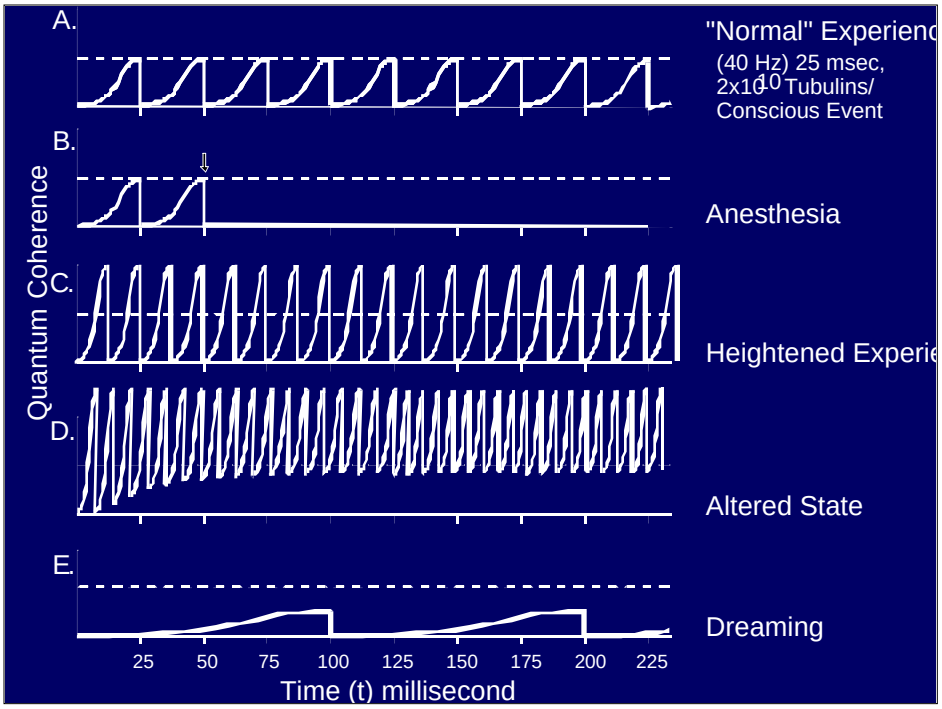

Figure B.5—Quantum Superposition Entanglemenin Microtubules for Five States Related to Consciousness. A. Normal 30 Hz experience. B. Anesthesia: anesthetics bind in hydrophobic pockets and prevent quantum delocalizability and coherent superposition. C. Heightened Experience: increased sensory experience input increases rate of emergence of quantum superposition. Orch OR threshold is reached faster and Orch OR frequency increases. D. Altered State: even greater rate of emergence of quantum superposition due to sensory input and other factors promoting quantum state (e.g. meditation, psychedelic drug). Predisposition to quantum state results in baseline shift and collapse so that conscious experience merges with normally sub-conscious quantum computing mode. E. Dreaming: prolonged sub-threshold quantum superposition time.

linearly (analogous to a quantum computer) but is influenced at the instant of collapse by hidden nonlocal variables (quantum-mathematical logic inherent in fundamental space-time geometry). The possible outcomes are limited (or probabilities are set) by neurobiological feedback (MAPs). The precise outcome (our "chosen" action) is determined by effects of the hidden logic on the quantum system poised at the edge of objective reduction. This could explain why people generally do things in an orderly, deterministic fashion, but occasionally their actions or thoughts are surprising, even to themselves.

Consciousness and Evolution

When in the course of evolution did consciousness first appear? Are all living organisms conscious, or did consciousness emerge more recently (e.g., with language or toolmaking)? The Orch OR model (unlike other models of consciousness) is able to make a prediction as to the onset of consciousness. Using E = h / T, the feasibility of consciousness for different organisms can be explored.

A single-celled organism (e.g., a paramecium, with E = 10^7 tubulins and T = 50,000 ms) would be unlikely to achieve consciousness, whereas a nematode worm (e.g., C. elegans with E = 3 x 10^9 tubulins per cell and T = 133 ms) might possess the biological complexity to understand what it is like to be a worm. Is it mere coincidence that these organisms were prevalent at the Cambrian "explosion," a burst of evolution 540 million years ago. Did these creatures possess the first consciousness? Did primitive consciousness (via Orch OR) accelerate evolution?

Would consciousness be advantageous to survival—above and beyond intelligent, complex behavior? The answer appears to be "yes." Noncomputable behavior (i.e., unpredictability, intuitive actions) is likely to be beneficial in predator-prey relations. The conscious experience of taste may promote the search for food; the experience of pain may promote the avoidance of predators; and the pleasurable qualia of sex may promote reproduction. So, what is it like to be a worm? Absent a sensory apparatus, associative memory, and a complex nervous system, such a primitive consciousness would be a mere glimmer, a disjointed smudge of reality. But qualitatively, at a basic level, it would be akin to ours.

What about future evolution? Will consciousness occur in computers? The advent of quantum computers opens the possibility. However, as presently envisioned, quantum computers will have insufficient mass in superposition (e.g., electrons) to reach the threshold for objective reduction due to environmental decoherence. Still, future generations of quantum computers may be able to realize this goal.

Conclusions

Brain processes relevant to consciousness extend downward within neurons to the level of the cytoskeleton. An explanation of conscious experience requires (in addition to neuroscience and psychology) a modern form of panprotopsychism in which protoconscious qualia are embedded in the basic level of reality, as

defined by modern physics. The Penrose model of objective reduction connects brain structures to fundamental reality, leading to the Penrose-Hameroff model of quantum computation with objective reduction in microtubules. The Orch OR model is consistent with known neurophysiological processes, generates testable predictions, and is the type of fundamental, multilevel, and interdisciplinary theory that may account for the mind's enigmatic features.