2. The Penrose-Hameroff Approach.

Perhaps the most ambitious attempt to create a quantum theory of consciousness is the one of Roger Penrose and Stuart Hameroff. Their proposal has three parts: The Gödel Part, The Gravity Part, and the Microtubule Part.

The Gödel Part, which is due to Penrose, is an effort to use the famous Gödel Incompleteness Theorem to prove that human beings have intellectual powers that they could not have if they functioned in accordance with the principles of classical physical theory. Establishing this result would reaffirm a conclusion of the von Neumann formulation of quantum theory, namely that a conscious human being can behave in ways that a classical mechanical model cannot. Penrose's argument, if valid, would yield this same conclusion, but within a framework that relies not on quantum concepts, which are generally unknown to cognitive scientists, but rather on Gödel-type arguments, which are familiar to some of them.

The general idea of Penrose's argument is to note that, due to the mathematically deterministic character of the laws of classical physics, the output at any specified finite time of any computer that behaves in accordance with the classical laws should in principle be deducible, to arbitrarily good accuracy, from a finite-step procedure based on a finite set of mutually consistent rules that encompass the laws of arithmetic. But then a human being that can be adequately modeled as a classical computer should be able to know, at any finite time, the truth *only* of those statements that can be deduced from a finite-step computation based on the finite set of rules that govern that computer. Yet Gödel-theorem-type arguments allow real mathematicians to know, given *any* finite set of consistent logical rules that encompass the laws of arithmetic, the truth of mathematical statements that cannot be deduced by any finite-step proof based on those rules. This seems to imply that a real mathematician can know things that no classical physics model of himself could ever know, namely the truth of statements that his classical computer simulation could not establish in a finite time.

Filling in the details of this argument is not an easy task. Penrose spends the better part of five chapters in "The Emperor's New Mind," (Penrose, 1989) and some 200 pages in “Shadows of the Mind” (Penrose, 1994) explaining and defending this thesis. However, the Harvard philosopher Hillary Putnam challenged Penrose’s conclusion in a debate appearing in the New York Times Review of Books, (Putnam, 1994) and numerous logicians have since weighed in, all, to my knowledge, claiming the non-validity of Penrose’s argument. Thus the Gödel Part of the Penrose-Hameroff approach cannot now be regarded as having been successfully established.

The Gravity Part of the Penrose-Hameroff approach addresses a key question pertaining to the quantum dynamics: exactly *when* do the sudden “quantum jumps” occur? In von Neumann’s theory, if construed ontologically, these jumps would occur when the neural correlates of conscious thoughts become sufficiently well formed. But von Neumann gives no precise rule for when this happens.

The lack of specificity on this point is a serious liability of the von Neumann theory, insofar as it is construed as a description of the ontological mind-matter reality itself. That difficulty is the basic reason why both the original Copenhagen formulation and von Neumann’s extension of it eschew traditional ontological commitments, and hew rather to the pragmatic position that it is the development of practical relationships between empirical findings and theoretical concepts that constitutes the essence of science, rather than shakey speculations about the ultimate nature of reality. The pragmatic position is that theoretical ideas that optimally provide reliable practical relationships between human experiences constitute, themselves, our best *scientific* understanding of “reality.” Added ontological superstructures are, according to this viewpoint, superfluous metaphysical musings, not true science, for, to the extent that these additions go beyond optimal theoretical prescriptions of valid connections between human experiences, they cannot be tested empirically.

Penrose wants to provide an ontology that has “real quantum jumps.” Hence he must face the issue: when do these jumps occur. He seeks to solve this problem by linking it to a problem that arises when one attempts to combine quantum theory with Einstein’s theory of Gravity.

Einstein’s theory of gravity, namely General Relativity, is based of the idea that space-time is not a rigid flat structure, as had previously been thought, but is a *deformable medium*, and that the way it is deformed is connected to the way that matter is distributed within it. This idea was developed within the framework of classical physical theory, and most applications of it are made within a classical-physics idealization. But serious problems arise when the quantum character of “matter” is brought in. For, according to orthodox quantum theory, a particle, such as an electron or an ion, has no well defined location: its location is completely described by a smeared out “probability cloud.” But if the locations of the material particles are not well defined then, according to General Relativity, neither is the form of the space-time structure in which the particle structures are imbedded.

Penrose conjectures that nature abhors uncertainty in the structure of space-time, and that when too much ambiguity arises in the space-time structure a quantum jump to some less ambiguous structure will occur. This “principle” allows him to tie quantum jumps to the amount of uncertainty in the structure of space-time.

There is no compelling reason why nature should be any more perturbed by an uncertainty in the structure of space-time than by an uncertainty in the distribution of matter. However, by adopting the principle that nature finds intolerable *excessive ambiguity in* the *structure of space-time* Penrose is able to propose a specific rule about when the quantum jumps occur.

Penrose’s rule depends on the fact that Planck’s constant gives a relationship between energy and time: this constant divided by any quantity of energy gives a corresponding interval of time. Thus if an energy associated with a possible quantum jump can be defined then a time interval associated with that potential jump becomes specified.

To identify the pertinent energy consider a simple case in which, say, a small object is represented quantum mechanically by a small cloud that divides into two similar parts, one moving off to the right, the other moving off to the left. Both parts of the cloud are simultaneously present, and each part produces *a different distortion* of the underlying spacetime structure, because matter is distributed differently in the two cases. One can compute the amount of energy that it would take to pull apart, against their gravitational attraction, two copies of the object, if each copy is located at the position specified by one of the two clouds. If one divides Planck's constant by this "gravitational energy" then a time interval associated with this distortion of space-time into these two disparate structures becomes defined. Penrose proposes that this time interval is the duration of time for which nature will *endure* this bifurcation of its space-time structure into the two incompatible parts, before jumping to one or the other of these two forms.

This conjectured rule is based on two very general features of nature: Planck's universal constant of nature, and the Newton-Einstein universal law of gravitation. This universality makes the rule attractive. But no deeper reason is given why nature must comply to this rule.

Does this rule have any empirical support?

An affirmative answer can be provided by linking it to Hameroff's belief that consciousness is closely linked to the *microtubular sub-structure of the neurons.*

It was thought at one time that the interiors of neurons were basically structureless fluids. That conclusion arose from direct microscopic examinations. But it turns out that in those early studies the internal substructure was wiped out by the fixing agent. It is now known that neurons are filled with an intricate structure of *microtubules*.

Each microtubule is a cylindrical structure that can extend over many millimeters. The surface of the cylinder is formed by a spiral chain of tubulin molecules, with each circuit formed by thirteen of these molecules. The tubulin molecule has molecular weight of about 110,000 and it exists in two slightly different geometric (i.e., configurational) forms. Each tubulin molecule has a single special electron that can be in one or the other of two relatively stable locations. The molecule will be in one or the other of the two configurational states according to which of these two relatively stable locations this special electron is occupying.

Hameroff is an anesthesiologist, and he noted that there is close correspondence between, on the one hand, the measured effects of various anesthetics upon consciousness, and, on the other hand, the capacity of these anaesthetics to diminish the ability of the special electron to move from one stable location to the other. This suggests a possible close connection between consciousness and the configurational activity of microtubules.

This putative linkage allows an empirical test of Penrose's rule to be made.

Suppose, in keeping with the case considered by Penrose, you are in a situation where one of two possible experiences will probably occur. For example, you might be staring at a Necker Cube, or walking in a dark woods when a shadowy form jumps out and you must choose "fight" or "flight," or perhaps you are checking your ability to freely choose to raise or not raise your arm. Thus one of two alternative possible experiences is likely to occur. Various experiments suggest that it takes about half a second for an experience to arise. Given this time interval, Penrose's formula specifies a certain corresponding energy. Then Hameroff can compute, on the basis of available information concerning the two configurational states of the tubulin molecule, how many tubulin-molecule configurational shifts are needed to give this energy.

The answer is about 1% of the estimated number of tubulin molecules in the human brain. This result seems reasonable. Its reasonableness is deemed significant, since the computed fraction could have come out to be perhaps billions of times smaller than, or billions of times greater than, 100%. The fact that the computed value is in "the ballpark" supports the idea that consciousness may indeed be closely connected to tubulin configurational activity.

Given this rather radical idea – it was previously thought that the microtubules were merely a construction scaffolding for the building and maintenance of the physical structure of the neurons – many other exotic possibilities arise. The two configurational forms of the tubulin molecule mean that it can hold a "bit" of information, so maybe the microtubular structure forms the substrate of a complex *computer* located within each neuron, thus greatly expanding the computational power of the brain. And maybe each such computer is in fact a "quantum computer." And maybe these quantum computers are all linked together to form one giant brain-wide quantum computer. And maybe these hollow micro-tubes form wave guides for quantum waves.

These exotic possibilities are exciting and heady ideas, and they go far beyond what conservative physicists are ready to accept, and far beyond what the 1% number derived from Penrose's rule actually supports, which is merely a connection between consciousness and microtubular activity, *without the presence* of the further stringent "coherence" conditions required for the functioning of a quantum computer. Quantum computation requires an effective isolation of the quantum informational waves from the surrounding hot, wet, noisy environment. But interaction of the computational waves with the environmental degrees of freedom tends to destroy very quickly the delicate "interference effects" that underlie the quantum computation.

The simplest system that exhibits a behavior that depends strongly on quantum interference effects, and for which the maintenance of *coherence* is essential, is the famous "double-slit experiment." When photons of a single wave length are allowed to pass, one at a time, through a pair of closely spaced narrow slits, and are later detected by some suitable detection device, one finds that *if the photonic system is not allowed to perceptibly influence any environmental degree of freedom* on its way to the detection device then the pattern of detected events depends on an *interference* between the parts of the beam passing through the two different slits. This pattern is very different from what it would be if the photon were allowed to perceptibly disturb, the surrounding environment. A disturbance of the environment produces a "decoherence" effect: a weakening or disappearance of the interference effects.

This condition "perceptibly disturb" is a weak condition: if even *one* particle in the environment is disturbed by a discernible amount then the coherence is lost, and the quantum interference effect will disappear.

Since the medium in which the putative quantum information waves are moving involves different conformational states of huge tubulin molecules of molecular weight ~110,000, it would seemingly be exceedingly hard to ensure that the passage of these waves will not disturb even one particle of the environment by a discernible amount.

Max Tegmark wrote an influential paper in Physical Review E that mathematically supports the intuition of most physicists that the macroscopic coherence demanded by the Penrose-Hameroff requirement that the microtubular conformal states form the substrate of a quantum computer that extends over a large part of the brain cannot be realized in a living human brain. Tegmark concluded that the coherence required for macroscopic quantum computation would be lost in a ten trillionth of a second, and hence should play no role in consciousness. This paper was widely heralded, but Hagan, Hameroff, and Tuszynski wrote a rejoinder in a later issue of the same journal that pointed out several flaws in Tegmark's paper. Corrections of these errors lengthened the coherence time by 8 or 9 orders of magnitude, thus bringing the situation into a regime where the non-equilibrium conditions in a living brain might become important: energetic biological processes might conceivably intervene in a way that would make up the still-needed factor of ten thousand. However, the details of how this might happen were not supplied. Hence the issue is, I believe, still up in the air, with no detailed explanation available of how the macroscopic quantum coherence could be maintained in a functioning human brain.

It must be stressed, however, that these exotic "quantum computer" effects are not necessary for the emergence of strong quantum effects within the general framework supplied by the combination of Penrose's rule pertaining to gravity and Hameroff's claim of the importance of microtubules. According to von Neumann's general formulation, the state of the brain - or of the microtubular part of the brain - is adequately represented by what physicists call the "reduced density matrix" of that subsystem. This representation depends only on the variables of that subsystem itself – i.e., the brain, or microtubular array - but nevertheless takes adequate account of the interactions of that system with the environment. It keeps track of the quantum coherence or lack thereof. Penrose's rule can be stated directly in terms of the "reduced density matrix," which displays, ever more clearly as the interaction with the environment grows, the two alternative states of the brain – or of the microtubular array – that nature must choose between. This reduced density matrix representation shows that the powerful decoherence effect produced by strong interactions with the environment actually *aids* the implementation of Penrose's rule, which is designed to specify *when* the quantum jump occurs (and perhaps to which states the jump occurs). The capacity of the brain to be or not to be a quantum computer is a very different question, involving enormously more stringent conditions. It thus is important, for logical clarity, to separate these two issues of the requirements for *quantum computation* and for *quantum jumps*, even though they happen to be interlocked in the particular scenario described by Penrose and Hameroff,

----

Hagen, S., Hameroff, S, & Tuszynski, J. (2002). Physical Review E65, 061901-1 – 061901-11.

Hameroff, S. & Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules: a model for consciousness. J. Consciousness Studies 3, 36-53.

Penrose, R. (1986). *The emperor's new mind*. New York: Oxford.

Penrose, R. (1994). *Shadows of the mind*. New York: Oxford.

Putnam, H. (1994). Review of Roger Penrose, Shadows of the Mind, *New York Times Book Review*, November 20, p.7. Reprinted in AMS bulletin:
www.ams.org/journals/bull/pre-1996data/199507/199507015.tex.html

Tegmark, M. (2000). Importance of quantum decoherence in brain process. *Physical Review* E61, 4194-4206.