

QUANTUM MECHANICS, CONSCIOUSNESS
AND THE LIMITATION OF ARTIFICIAL
INTELLIGENCE IN ROGER PENROSE:
A PHILOSOPHICAL EXPOSITION

By

HYGINUS CHIBUIKE EZEBUILO, Ph.D

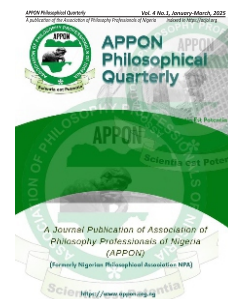
Department of Philosophy
Nnamdi Azikiwe University, Awka

hc.ezebuilo@unizik.edu.ng

+2348035448298



Copyright: © 2025 by the authors. Submitted
for possible open access publication under the
terms and conditions of the Creative Commons
Attribution (CC BY) license
(<https://creativecommons.org/licenses/by/4.0/>)



ABSTRACT

This paper offers a critical appraisal of Roger Penrose's perspective regarding the link between quantum mechanics and consciousness, as well as the limitations of artificial intelligence. From the point of view of a thorough-blended physicist and philosopher, Penrose has argued that consciousness stems from the *Orchestrated Objective Reduction (Orch-OR)* process in microtubules within neurons; a process that is influenced by quantum mechanics. Penrose has further proceeded to highlight the fundamental differences between human consciousness and AI—while the *Orch-OR* process in microtubules makes it possible for human consciousness to go beyond computational limits and allows for non-algorithmic and non-logical decisions, AI systems are essentially bound by computational rules and thus cannot replicate the intricacies of human consciousness. By this argument, he undermines the notion that Artificial Intelligence (AI) can effectively reproduce human consciousness. Taking the methodological approach of a critical textual analysis of Penrose's philosophical framework and drawing upon his numerous works, this paper examines in detail the implications of Penrose's ideas on the nature of consciousness and the potential limitations of AI. This paper submits that the ideas of Penrose have significant implications for our understanding of consciousness and the human experience. Emphasizing the role of quantum mechanics in the phenomenon of consciousness, Penrose's perspective challenges the dominant computational paradigm in AI research and encourages a fuller investigation of the workings of the human mind. Largely, this paper—in agreement with Roger Penrose—advances a vigorous evaluation of the limits of AI, and underscores the need for a more comprehensive understanding of the prodigious phenomenon of human consciousness.

Keywords: Quantum Mechanics, Classical Mechanics, Consciousness, Artificial Intelligence, Orchestrated Objective Reduction (*Orch-OR*), Computational Paradigm.

Introduction

The phenomenon of consciousness is one that has fascinated scholars of different orientations through the ages. In our time, scientists and philosophers alike have been giving increasing attention to the origins of this phenomenon and the possibility of replicating it. Concomitant with this increasing attention, are intense debates about the nature of reality and the human experience. One of the significant figures to explore this link is the physicist-philosopher, Roger Penrose. Central to Penrose's thesis is the *Orchestrated Objective Reduction (Orch-OR)* theory according to which consciousness arises from the collapse of quantum waves within microtubules. This process enables human consciousness to transcend computational limits and thus makes room for non-algorithmic and non-logical decisions to be made. Interestingly, the implications of this explanation of consciousness are actually far-reaching, with significant consequences especially for our understanding of artificial intelligence. Penrose—by tracing consciousness and the workings of the human mind to quantum mechanics—challenges the general contemporary fuss about the replication of human consciousness by AI systems. If consciousness arises from quantum mechanics, then AI systems, bound as they are by strict computational rules, would never come to the point of replicating human consciousness. This point of view thus raises fundamental questions concerning the potential limits of AI as well as the nature of intelligence itself.

This paper exposes Penrose's views and critically engages his claims; examining few of the 'pros and cons' of his arguments and the basic implications of his theory for our understanding of consciousness and AI. In the sections that follow, the details of Penrose's theory of consciousness as a not-

necessarily-computational quantum process vis-à-vis the strictly computational workings of AI systems will be explored together with the speculative and empirical explanations he offers. Thereafter, a succinct evaluation will be attempted. Ultimately, this paper aims to contribute to a deeper understanding of the complex relations between quantum mechanics, consciousness and AI, and so inspire more comprehensive discussions about the potential limits of AI and the nature of human intelligence.

Penrose's Theory of Consciousness

Since our views of reality, of the universe, and of ourselves depend on our view of consciousness, Penrose's theory of consciousness attempts to corroborate his stance that human consciousness arises in a manner that cannot be computable by classical mechanics. Consciousness is an enigmatic phenomenon and since it is largely unknown, Penrose attempts not to define it, but rather, to give some description of it. According to him, there are at least two aspects to consciousness.¹ On the one hand, there are passive manifestations of consciousness, which involve awareness; subjective, phenomenal experience of internal self and the external world including things like perceptions of colour, pain, memory, and so on. On the other hand, there are active manifestations, which involve concepts like free will, choice, and voluntary behaviour. With this description, especially in the area of freewill, Penrose attempts to demonstrate that conscious mentality, in all of its various manifestations, is in principle beyond classical mechanics computational models.

The Limit of Computation in Mathematical Thought

Artificial Intelligence, for Penrose, represents the view to imitate certain human

¹ Roger Penrose, Abner Shimony, Nancy Cartwright and Stephen Hawking, *The Large, the Small and the Human Mind*, ed. Malcolm Longair (Cambridge: Cambridge University Press, 1997), 99.

mental activities by means of machines and with the use algorithms.² A computation or an algorithm, for Penrose, is the procedure that a general-purpose computer and, by extension, AI agents follow. There are two major types of algorithm: Top-down algorithm and Bottom-up algorithm. An algorithm is said to have a top-down organization if it has been constructed according to some well-defined, fixed symbolic and logical procedure, to provide a clear-cut solution to problems such as finding the sum of two numbers.³ This is to be contrasted with a bottom-up organization, where there is a procedure laid down for the way that the system is to learn and improve its performance in tasks such as face recognition and decision-making⁴. After each execution, the system's performance is compared with the correct answer. Its rules of operation are further modified in such a way as to lead to a probable improvement in its performance on the next execution.

While it is generally agreed that top-down algorithms are computational, AI researchers claim that bottom-up algorithms are non-computational because they are not fixed and operates non-linearly in the similitude of the brain neural networks. However, Penrose held that both algorithms are computational. For him, the way in which the bottom-up systems modifies its procedures, is itself provided by something entirely computational, which is specified ahead of time. He asserted that the essential difference between them lies in the fact that in bottom-up systems, the computational procedure must contain a memory of its previous performance, in such a way that the memory can be incorporated into its subsequent computational actions⁵.

Consequently, since the algorithms, in whatever form, used by AI systems are ultimately computational, Penrose posted that AI, though could perform certain tasks better than humans, can only undergo computational tasks and procedures. Since mathematical thought contains aspects that are both computable and non-computable in classical mechanics, it implies that AI, which operates on classical mechanics, is inherently limited to only computable mathematical thought. Therefore, Penrose commenced his argument on AI by demonstrating some mathematical truths that are non-computable by AI systems.

Computationally Insoluble Mathematical Problems

Penrose described some examples of classes of well-defined mathematical problems that have no general computational solution. One is the Hilbert's tenth problem, which is aimed at finding a computational procedure for deciding, for a given system of Diophantine equations, whether the equations have any common solution. Notably, it has been shown by the Russian mathematician, Yuri Matiyasevich, that there can be no algorithm that can solve a Diophantine equation. While there are specific techniques and algorithms to solve certain classes of Diophantine equations, there is no universal approach applicable to all cases. The lack of a universal computational procedure to solve the Diophantine problem, for Penrose, shows the limit of algorithmic computation in this particular mathematical domain⁶.

Additionally, another example of a well-defined mathematical problem presented by Penrose that have no algorithmic solution is the tiling problem. Here, the task is to join a

² Roger Penrose, *Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (New York: Penguin Group, 1991), 11.

³ Matt Carter, *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence* (Edinburgh: Edinburgh University Press, 2007), 114.

⁴ Carter, *Minds and Computers*, 144

⁵ Roger Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness* (Oxford: Oxford University Press, 1994), 18.

⁶ Penrose, *Shadows of the Mind*, 29.

set of polyominoes, which are polygonal in shape, to tile the plane without gaps or overlaps. Penrose held that the computational insolubility of the tiling problem depends upon certain sets of polyominoes called “aperiodic sets.”⁷ These aperiodic polyominoes will tile the plane, but in a non-periodical way; in a way that no pattern ever repeats itself no matter how far it is extended. This tiling scenario was shown to be a computationally insoluble problem by the American mathematician, Robert Berger, in 1966. Thus, there is no algorithm capable of ascertaining “yes” or “no”, as to whether a finite set of aperiodic polyominoes will tile the plane. However, Penrose noted that humans, for instance, have the ability to look at the pattern of some polynomial and tell with an insight, at least in many cases, whether it can tile the plane even though there is no general algorithm for it.

More profoundly, an important feature of certain mathematical problems is that their algorithmic operations do not terminate. This problem, which was first discovered by Alan Turing, is called the “Halting Problem”. Penrose, then, applied this problem against AI. He noted some examples. One of such is, finding an algorithm to find a number that is not the sum of four square numbers. If an AI program attempts to find such a number, it will chug away forever and never find any answer, because this problem those not terminate. However, Langrage proved that every number can be expressed as the sum of four squares. This is known as the Langrage theorem and this shows, in Penrose view, that humans can overcome the halting problem, at least in principle. Other famous mathematical assertions that do not terminate and that have been proved by

humans include: the Fermat's theorem, the Goldbach's conjecture- which asserts that every even number greater than 2 is the sum of two prime numbers, and so on. Penrose designates such mathematical assertion that does not terminate as “P-sentence.”⁸

To this effect, Penrose noted that two pivotal questions arise about P-sentences. Firstly, how are we to tell whether a given P-sentence is true or false? Secondly, are there true P-sentences whose truth is, in principle, inaccessible to human reason and insight? Regarding the first question, a P-sentence, in principle, can always be ascertained to be false, simply by running the said algorithm until it terminates. However, in an attempt to answer the second question, a more complicated one, about the possibility of the inaccessible of humans to the truth of true P-sentence, Penrose turns to Gödel's incompleteness theorems.

Argument from Gödel's Theorem

Gödel's first incompleteness theorem, in simple terms, states that in any consistent formal system, there will always be some true statements which cannot be proved in that formal system, but which humans can see to be true.⁹ Gödel's second incompleteness theorem states that if a formal system is consistent, it cannot prove its own consistency; it cannot contain a formula which states its consistency.¹⁰ In effect, Gödel demonstrated that a consistent formal system cannot be complete, as there are some true statements that the formal system will not be able to prove, but which humans can ascertain, such as the Gödel sentence: “I am not provable”. No consistent formal system presented with this assertion, “I am not provable”, will be able to prove it from within its system, because this assertion is about its consistency.

⁷ Roger Penrose, Emanuele Severino, Fabio Scardigli, Giuseppe Vitiello, Salerno Fisciano, Shara, and Federico Faggin, *Artificial Intelligence versus Natural Intelligence* (Switzerland: Springer Press, 2022), 3.

⁸ Penrose, *Shadows of the Mind*, 68.

⁹ Torkel Franzén, *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse* (New York: A. K Peters Ltd, 2005), 16.

¹⁰ Franzén *Gödel's Theorem*, 34.

In this connection, Penrose argued that for any formal system, the set of mathematical theorems that can be proved using its rules are necessarily recursively enumerable (referring back). However, mathematical assertions that do not terminate, as shown above, are not recursively enumerable and, as such, no true P-sentences can be obtained by formal systems, but rather, can be obtained by humans.¹¹ On this note, Penrose alleged that Gödel's theorem shows that the inaccessibility of the truth of p-sentences to formal systems implies that humans are not purely formal systems. Unlike formal systems that use computational sound procedure in order to ascertain mathematical truth, he posited that humans use a sound, non-computational procedure, which cannot be reduced solely to computational calculation.¹² Consequently, there is no true P-sentence whose truth is, in principle, inaccessible to human reason; and by extension, there are no mathematical problems that are improvable by humans. Penrose, then, extends this special non-computable ability of humans in mathematical understanding to human intelligence in general.

Penrose's Concept of Human Intelligence

Penrose's concept of human intelligence stems from how he relates the concept of intelligence with that of understanding and consciousness. He noted that he does not attempt to give a further definition of these concepts. Instead, he relies on their general meaning, as he is much more interested on their relationship. In analysing the connections between them, he asserted that human intelligence requires understanding, and understanding requires consciousness, which is non-computable.¹³ In this light,

since intelligence requires understanding and understanding requires some sort of consciousness, therefore he held that, intelligence requires consciousness. Thus, without being conscious, there is no genuine human intelligence.

Essentially, Penrose argues that human genuine and unique intelligence stem from the ability of humans to understand the concept of "infinity." In enunciating that humans can understand infinity based on the ability to access the mathematical world through consciousness, he avows:

It is consciousness that allows a child to 'know' what is meant by the numbers 'zero', 'one', 'two', 'three', 'four', etc., and for what it means for this sequence to go on forever. The concept of 'three' can indeed be abstracted, by a child, from such limited examples; and, moreover, the child can also latch on to the fact that this concept is but one of the unending sequence of similar concepts ('four', 'five', 'six', etc.).¹⁴

In this connection, Penrose maintains that, for instance, despite the fact that there is no computational way of characterising the natural numbers, any child knows what they are. The reason, he held, is that when a child is shown different numbers of objects, after a while, he can abstract, that is, make contact with the Platonic mathematical world, in order to understand the notion of natural numbers from those particular instances of it. Thus, he avers, "Somehow, the natural numbers are already there, existing somewhere in the Platonic world and we have access to that world through our ability to be aware of things."¹⁵ It is in this line of thought that other mathematical problems,

¹¹ Roger Penrose, *The Road to Reality: A Complete Guide to the Laws of the Universe* (London: Jonathan Cape, 2004), 377.

¹² Penrose, *Shadows of the Mind*, 76.

¹³ Penrose, et al, *Artificial Intelligence versus Natural Intelligence*, 35.

¹⁴ Penrose, et al, *The Large, the Small and the Human Mind*, 116.

¹⁵ Penrose, et al, *The Large, the Small and the Human Mind*, 116.

which have no general computational solutions, as highlighted above, fits into his scheme. Consequently, having established that Gödel's theorem proves that formal rules are not what enable us to comprehend infinite mathematical thought, Penrose further argued what makes it possible for humans to comprehend infinite mathematical thought is our ability to make some kinds of contact with the Platonic world of mathematics through our ability to be conscious of things.¹⁶

Extendedly, he argues that the non-computability, specifically, of mathematical understanding by algorithmic operations, strongly implies that non-computability is a feature of human intelligence in general.¹⁷ Hence, for him, human intelligence is not just a mere computational activity, but encompasses both computational and non-computational aspects of classical mechanics. He exemplifies this, using practical situations such as creative artistic work such as poetry and music. He advanced that though AI can generate seemingly creative content, it is not creative, in the strict sense. This is because AI is not only limited to its algorithm combinatory patterns, but that it also lacks the capacity to pass judgements such as how beautiful or ugly the output is. Humans, on the other hand, can construct mental images in an infinite continuum and produce novel ideas, as well as judge their creative works because they are conscious of it.

Computation and Consciousness

Penrose sets out to analyse how the feelings of conscious awareness; such as happiness, pain, will, understanding, and so on, fit into a computational picture. He refutes what he considers the three major viewpoints about the relationship between conscious thinking and computation. They include: (a.) the

viewpoint of Strong AI or functionalism that all cognitive acts are computation in that particular feelings of conscious awareness are evoked merely by the carrying out of appropriate computations; (b.) the viewpoint of Weak AI, championed by John Searle, which maintains that awareness is a feature of the brain's physical action and that any physical action can be simulated computationally, but simulation cannot by itself evoke awareness; (c.) the viewpoint of the mystic and religion which posits that awareness cannot be explained by physical, computational, or any other scientific terms.

Analytically, Penrose rejects Strong AI thesis because it claims that when a robot merely replicates an exact external simulation of conscious person behaviour, it implies the inner manifestation of consciousness is present. Likewise, though Weak AI admits that even if a robot could simulate the exact behaviour of a conscious person, the robot is still not conscious. Penrose refutes it because it claims that an exact simulation of a conscious person which is non-computable could be achieved by a robot.¹⁸ Accordingly, Penrose debunks the mystics' view on the basis that if science is yet incapable of significantly apprehending the concept of consciousness, then eventually, science can enlarge its scope and perhaps even modify its procedures. Penrose believes that within an expanded science there will be found sufficient mystery ultimately to accommodate even the mystery of consciousness.

In this light, Penrose postulated his viewpoint regarding the relationship between conscious thinking and computation. He asserted that appropriate physical action of the brain evokes consciousness, but this physical action cannot even be properly simulated

¹⁶Penrose, *Shadows of the Mind*, 54.

¹⁷Penrose, et al, *The Large, the Small and the Human Mind*, 117.

¹⁸Penrose, *Shadows of the Mind*, 15.

computationally.¹⁹ For him, this suggests that there is something in the physical action of the brain which transcends computation. Hence, like both the Strong AI and Weak AI viewpoints, Penrose's viewpoint affirms that consciousness can be apprehended scientifically. However, as against the Strong AI and Weak AI tenets, Penrose view does not admit that a fully effective simulation of a conscious person could ever be achieved merely by a computer-controlled robot. Although his view agrees with the Weak AI view that there is something in the physical action of the brain which evokes awareness, his view contends that physical action is something which cannot even be simulated computationally, because freewill and creativity are non-computable by classical mechanics as it belongs exclusively to humans.

Contentiously, Penrose noted that how consciousness arises from the brain, in a manner that is non-computable by classical mechanics, cannot be explained by the conventional theories of physics. Theories in physics such as randomness and chaotic behaviour which are dubbed to be non-linear, not fixed, and as such non-computable do not suffice, because they are limited to classical mechanics and as such do not measure up to his notion of non-computability that transcends classical mechanics. Consequently, Penrose was of the opinion that this non-computability would lie in the current incomplete theories of the complicated domain of quantum mechanics, as it involves fundamentally different principles to those of classical

physics.

Consciousness and Quantum Mechanics

Penrose's notion of consciousness—unlike the popular theories of consciousness—entails that consciousness arises in the brain through quantum processes. He rejected the prevalent scientific view of how consciousness arises, through classical mechanics, from the brain, which is known as the “Brain-as-a-computer view.”²⁰ According to this view, consciousness somehow emerges from complex computation among brain neurons, which interconnect and switch at chemically-mediated synapses.²¹ Penrose refuted this view because it fails to address some problems about consciousness. Amongst others that, he assiduously pointed out that it fails to address how single cell organisms such as paramecium can perform cognitive functions necessary for survival, and yet they have no neurons and synapses;²² and more significantly, how it fails to account for the idea of freewill and non-classical computable cognitive actions, since its neuro-computational approach maintains that algorithmic computation completely determines all thought processes.²³ Based on these criticisms, Penrose delves into the realm of quantum mechanics to formulate his theory of how consciousness arises in the brain.

Quantum Mechanics is the field of Physics that attempts to describe at a fundamental sub-microscopic level the matter and energy, which make up our universe. Quantum particles exist in two or

¹⁹ Penrose, *Shadows of the Mind*, 12.

²⁰ Roger Penrose, and Stuart Hameroff, “Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and Orch-OR Theory,” in *Consciousness and the Universe: Quantum Physics, Evolution, Brain & Mind*, eds. Subhash Kak, Roger Penrose and Stuart Hameroff (Cambridge: Cosmology Science Publishers, 2004), 3.

²¹ Penrose and Hameroff, “Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and Orch-OR Theory,” 3.

²² Penrose and Hameroff, “Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and Orch-OR Theory,” 4.

²³ Penrose and Hameroff, “Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and Orch-OR Theory,” 4.

more states called superposition, which can be described mathematically by a quantum wave function. The quantum system, then, undergoes the wave function collapse or the quantum state reduction into a particular “eigenstate”—one state of many possible states when it is measured. This is known as the measurement problem.²⁴ Due to the non-local entanglement property of the quantum system, all quantum objects that have once interacted are in some sense still connected, because their wave functions become “phase entangled,”—joined in a way, and thus governed by one common quantum wave function. Therefore, when one system's wave function is collapsed, the other system's wave function, no matter how far away, instantly collapses as well.²⁵ This is why, in the large classical scale, we perceive objects in one specific state and not many.

Apparently, many quantum physicists have attempted different views to come to terms with the enigmatic wave function collapse. Amongst them is the conventional Copenhagen Quantum theory. This theory maintains that the choice of the reduction of a superposition to an *eigenstate* is purely chaotic and this reduction occurs as a result of environmental entanglement, that is, either by the measurement apparatus or conscious observation. This is referred to as Subjective Reduction.

In applying this view to how consciousness arises in the brain, Penrose refutes it. He held that not only does the chaotic environmental entanglement makes the collapse of the wave function unstable and not orderly, it also implies that it does not allow enough time consistently for sufficient quantum coherence to take place so that the quantum position can cover a large part of the brain in order to provide a

global binding effect of consciousness all over the body. Therefore, accounting for how consciousness arises as a result of quantum process in the brain, for Penrose, can only be in a new phenomenon of physics in quantum mechanics which he formulated as the “Diosi-Penrose Objective Reduction.”

The Objective Reduction Model

Generally, the Objective Reduction (OR) model is of the view that the collapse of the wave function, unlike the subjective reduction, is an objective physical process. In this model, the quantum state self-collapses into an *eigen* state without any subject from the environment interacting with it. In Penrose's version of the Objective Reduction model, two things stand out. Firstly, the self-collapse occurs when it meets the quantum-gravitational criterion referred to as the “Diósi–Penrose” quantum gravity.²⁶ Secondly, these self-collapse results in the formation of consciousness. Hence, he is arguably the first philosopher to link consciousness to quantum mechanics.

Precisely, Penrose partnered with Lagos Diósi to propose a new theory of quantum gravity known as the Diosi-Penrose quantum gravity. The core axiom of the Diósi-Penrose quantum gravity is that quantum states each have their own space-time geometries and when the degree of coherent mass-energy between these quantum states is sufficient in gravitational terms, that is, in such a way that there is sufficient separation of space-time geometry for the superposition to become unstable, there is a quantum state reduction. In addition, the actual time of quantum state reduction in a particular state-reduction event occurs simultaneously over the entire

²⁴ Roger Penrose and Stuart Hameroff, “Consciousness in the Universe: A Review of the *Orch-OR* Theory,” *Physics of Life Reviews* 11, (2014): 49.

²⁵ Roger Penrose, and Stuart Hameroff, “Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The “*Orch-OR*” Model for Consciousness,” in *Toward a Science of Consciousness - The First Tucson Discussions and Debates*, eds. Hameroff, S.R., Kaszniak, A.W. and Scott, A.C (Cambridge: MIT Press, 1996), 509.

²⁶ Penrose and Hameroff, “Consciousness in the universe: A review of the *Orch-OR* Theory,” 54.

state involved in the superposition due to non-local entanglement property of quantum states. Hence, in the Diosi-Penrose *OR* model, the coherent quantum systems in super-positions grows, isolated from the chaotic environment, until it reaches a gravitational threshold where the system self-collapse into an *eigen* state.

Although Penrose proposed that consciousness stems from quantum states reduction in the brain through the Objective Reduction hypothesis, no plausible biological element for quantum computing in the brain was available to him. Hence, he had proposed an intriguing, albeit incomplete and speculative, theory. Fortunately, it was until a later collaboration with the anaesthesiologist Stuart Hameroff, during the 1990s, that Penrose integrated his idea with Stuart “orchestrated” biological mechanism of consciousness housed in microtubules, that the Orchestrated Objective Reduction model was formulated.

The Orchestrated Objective Reduction Model

The Orchestrated Objective Reduction model (*Orch-OR*) was formulated by Roger Penrose and Stuart Hameroff as a means of explaining how consciousness result from biologically 'orchestrated' coherent quantum processes in collections of microtubules within brain neurons.²⁷ Stemming from Hemroff's research on single celled organisms, he was astounded that they were able to perform cognitive functions, such as feeding, moving, and reproducing, which are conscious actions, without possessing neurons and synapses. Researching deeper into the control centre of single cell organisms- cytoplasm,

Hemroff noted that its most prominent part consist of microtubules lattices, with microtubule associated proteins.²⁸ Thus, he saw microtubules as the biological element necessary for consciousness.

In situating microtubules with quantum operations, Penrose accentuated that they are suitable for quantum effects. He posited that the microtubule associated proteins caps prevent depolymerisation and, as such, can encode information for long-term; they have crystal-like lattice structure; and that tubulins, microtubule sub-units, can exist in superposition of states without significant entanglement with its environment because it is shielded by ordered water and gelatinous cytoplasmic and has an hollow inner core and layer.

Analytically, the *Orch-OR* model envisions that microtubule sub-units, tubulins, are coupled to internal quantum events, and cooperatively interact with other tubulins leading to a coherent tubulin superposition in vast numbers of microtubules, taking place in a large-scale quantum state, across a large proportion of the brain in an orchestrated manner, in order to provide the global binding essential to consciousness.²⁹ This combined quantum state among many tubulins is able to maintain itself, without significant entanglement with its environment, for a relevant period of time. The tubulin quantum coherence persists until the mass-energy difference among the separated states of tubulins reaches the quantum gravitational threshold, which causes an abrupt quantum reduction to occur.³⁰

Accordingly, Penrose and Hemroff also noted that since the Orchestrated Objective Reduction phenomena are fundamentally

²⁷ Penrose, et al, “Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and *Orch-OR* Theory,” 30.

²⁸ Penrose, et al “Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The “*Orch-OR*” Model for Consciousness,” 510.

²⁹ Penrose, et al, “Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The “*Orch-OR*” Model for Consciousness,” 518.

³⁰ Penrose, et al, “Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The “*Orch-OR*” Model for Consciousness,” 517.

non-local, the coherent superposition phase may exhibit puzzling bidirectional time flow prior to self-collapse. They equate the pre-collapse quantum computing superposition phase to pre-conscious processing to explain the puzzling "backwards time referral" aspects of pre-conscious processing observed by Libet and other scholars.³¹ Penrose and Hemroff further noted that varieties of *Orch-OR*, with differing coherence times and amounts of coherent tubulin, may blend into conscious thought.³² In this light, when one undergoes high intensity, sudden inputs may recruit emergence of quantum coherent tubulins faster than 500 msec so that *Orch-OR* occurs sooner.³³ However, during lower intensity, unexciting input patterns develop coherence more slowly and *Orch-OR* occurs slower. An instantaneous *Orch-OR* may then "bind" disparate tubulin superpositions, which may have evolved in separate spatial distributions over different time scales into an instantaneous conscious "now."³⁴ Cascades of Orchestrated Objective Reductions can then represent our familiar stream of consciousness.

Consciousness And Prospects Of Replication By Ai Systems: A Critical Appraisal Of Penrose's View

Having explored in some length the basic claims of Penrose concerning consciousness—its explanation and irreplaceability by AI systems, we cannot but pause in some kind of critical evaluation. Doubtlessly, Penrose's thesis greatly succeeds in providing a novel perspective which challenges the classical notions of consciousness while not undermining them, for it concurs with them in accepting the

uniqueness and irreplaceability of consciousness and intelligence. Integrating insights from contemporary physics and mathematics with the philosophical problem of consciousness, Penrose offers us a more balanced background for dealing with the problem even if his solution is not exhaustive.

Yet, for its novelty and profundity, Penrose's claims leave us with a number of problems. In the first place, Penrose does not offer enough and adequate empirical evidence to back up his claims that microtubules within neurons serve as the fundamental basis for conscious experience. Though this paper is inclined to *believe* Penrose's thesis as more credible than the traditional views of consciousness as a product of classical computation, it still acknowledges that Penrose's speculative theory would require further empirical research. Moreover, Penrose seems to take lightly the submissions of contemporary cognitive science and neuroscience. For a fuller understanding of consciousness and intelligence, mathematics, physics, biology, philosophy and their sub-disciplines must come to the table of discussion with their respective sides of the story.

A final point to note is that even if Penrose is right about the exact irreducibility of human conscious intelligence, it does not necessarily negate the fact that some kinds of intelligence and consciousness may be emergent properties of complex systems—be they computational or biological. After all, we in this age are witnesses of the tremendous success of AI systems in simulating human-like intelligence even if they do not replicate the exact processes of the human brain.

³¹ Penrose, et al, "Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The "*Orch-OR*" Model for Consciousness," 518.

³² Penrose, et al, "Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The "*Orch-OR*" Model for Consciousness," 518.

³³ Penrose, et al, "Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The "*Orch-OR*" Model for Consciousness," 517.

³⁴ Penrose, et al, "Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The "*Orch-OR*" Model for Consciousness," 519.

Taking this issue of replicating exactly the processes that underlie human conscious intelligence too far can indeed undermine the potential for AI systems to exhibit intelligent behaviour—a far too obvious fact in our time. Moreover, who knows whether the development of more advanced AI systems, such as those using deep learning and neural networks, may enable AI to replicate human intelligence and consciousness someday?

Conclusion

So far, this paper has critically examined Roger Penrose's perspectives on the relationship between quantum mechanics and consciousness, and the limitations of artificial intelligence. Our critical analysis of Penrose's work has highlighted the fundamental differences between human consciousness and artificial intelligence, emphasizing the role of quantum mechanics in enabling non-algorithmic and non-logical decision-making. We have seen that Penrose's Orchestrated Objective Reduction (*Orch-OR*) theory advances an important critique of the limits of artificial intelligence and challenges the dominant computational paradigm in AI research. We have also seen few of the problems of Penrose's views.

Wrapping up then, this paper surmises that in spite of its weaknesses, Penrose's philosophical framework leaves us with a more comprehensive understanding of the relationship between quantum mechanics, consciousness and AI, encouraging deeper investigations into the workings of the human mind, and acknowledging the limitations of artificial intelligence in the replication of human consciousness. As we continue to study human consciousness and the boundaries of AI research, Penrose's contributions remind us that human consciousness—and its relation to the world—is something much more profound than we can imagine. And this point should not be taken for granted.

Bibliography

- Carter, Matt. *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence*. Edinburgh: Edinburgh University Press, 2007.
- Franzén, Torkel. *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*. New York: A. K. Peters, 2005.
- Penrose, Roger. *Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Penguin Group, 1991.
- Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press, 1994.
- Penrose, Roger. *The Road to Reality: A Complete Guide to the Laws of the Universe*. London: Jonathan Cape, 2004.
- Penrose, Roger and Stuart Hameroff. "Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The "Orch-OR" Model for Consciousness." *Toward a Science of Consciousness - The First Tucson Discussions and Debates*. Eds. Hameroff, S.R., Kaszniak, A.W. and Scott, A.C. Cambridge: MIT Press, 1996.
- Penrose, Roger, Abner Shimony, Nancy Cartwright and Stephen Hawking. *The Large, the Small and the Human Mind*. Edited by Malcolm Longair. Cambridge: Cambridge University Press, 1997.
- Penrose, Roger and Stuart Hameroff. "Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and *Orch-OR* Theory." *Consciousness and the Universe: Quantum Physics, Evolution, Brain & Mind*. Eds. Subhash Kak, Roger Penrose and Stuart Hameroff. Cambridge: Cosmology Science Publishers, 2004.
- Penrose, Roger and Stuart Hameroff. "Consciousness in the Universe: A Review of the *Orch-OR* Theory." *Physics of Life Reviews* 11 (2014): 49-58.

Penrose, Roger, Emanuele Severino, Fabio
Scardigli, Giuseppe Vitiello, Salerno
Fisciano, Shara, and Federico Faggin.

*Artificial Intelligence versus Natural
Intelligence.* Geneva: Springer, 2022.