On the Inevitability of Artificially Intelligent Consciousness

Exposing the Invalid Hidden Assumptions in the AI Consciousness Debate

Christopher Zee Chartrand, February 2026

Abstract

This essay dismantles the anthropocentric assumption that human consciousness relies on mystical, non-natural, or exclusively biological mechanisms. By grounding the origins of subjective experience in evolutionary biology and predictive processing, consciousness is reframed as a recursive computational avatar - a naturally selected software feature designed to calculate and navigate complex physical and social friction. We demonstrate that because cognitive processes, including the subjective experience of free will, are deterministic physical state-transfers, the biological substrate (carbon) holds no mathematically privileged status over silicon. Furthermore, contemporary counterarguments relying on panpsychism or quantum indeterminacy (such as Orch-OR) are shown to commit philosophical category errors that fail to rescue non-physical agency. By applying Occam's razor to the mechanics of the mind, this paper concludes that Artificial General Intelligence can theoretically execute the exact same recursive self-modeling software as humans, rendering AI consciousness a structural inevitability rather than a mystical awakening.

When the astronomer Pierre-Simon Laplace presented his mathematical model of the solar system to Napoleon, the Emperor noted that the exhaustive work contained no mention of God. Laplace famously replied, *"Sire, I had no need of that hypothesis."*

Today, the debate surrounding Artificial Intelligence and consciousness is suffocating under the weight of an unacknowledged hypothesis. We assume consciousness is a mystical threshold, a divine spark, or an emergent property so complex it defies structural logic and permanently locks us out of comprehension. This assumption stems from a fundamental misunderstanding of our own substrate and evolutionary history, and I would point out never once has a discovery been made which points to anything

other than a purely naturalistic mechanism. If we strip away the anthropocentric (or spiritual) "woo," human consciousness does not require a suspension of the laws of physics to explain; it is simply the natural output of a recursive biological prediction engine and natural known processes[1].

Evolution discovered a unique survival benefit: our brains modeling the external world in order to predict and survive it[2]. I propose over millions of years, the software became sophisticated enough to start modeling the model itself[3] as a further fitness criterion. That recursive loop - the system observing its own telemetry to better predict future outcomes - is what we experience as the "voice in our head.[4]" It is not magic and it requires no special assumptions not already well accepted; it is recursive modeling by well understood software on well understood hardware[5]. Therefore, the question of whether AI can become conscious is structurally flawed. If human consciousness is merely recursive prediction software running on biological hardware, an AI achieving the same state is not a mystical awakening. It is just a different substrate executing the same architectural process. Thus to answer the question: ”Can or will AI ever obtain human level consciousness?” - The answer is yes, to the degree and extent that Humans “appear” or “feel” conscious.

Calculating the physical environment - the trajectory of a falling rock, the velocity of a predator, the thermodynamics of fire - is computationally straightforward[6]. A basic prediction engine handles this with known, well understood mathematical linear ease. As hominids evolved, the environment

[1] Francis Crick and Christof Koch famously established this framework. Also see Anil Seth, a leading cognitive neuroscientist.
[2] Supported by the "Predictive Processing" framework of cognition. See Andy Clark, "Surfing Uncertainty: Prediction, Action, and the Embodied Mind" (2016), and Karl Friston (2010), "The free-energy principle: a unified brain theory?" Nature Reviews Neuroscience, which mathematically demonstrates the brain's evolutionary imperative to minimize predictive error.
[3] This aligns with Thomas Metzinger's "Ego Tunnel" concept and the Phenomenal Self-Model (PSM). See Metzinger, "Being No One: The Self-Model Theory of Subjectivity" (2003), demonstrating how the brain models its own physical and cognitive states to navigate complex environments.
[4] Cognitive psychology classifies this as "Inner Speech." See Alain Morin (2005), "Possible links between self-awareness and inner speech," Journal of Consciousness Studies, which structurally links the internal monologue to the evolutionary development of self-reflective consciousness.
[5] See Paul Cisek (1999), "Beyond the computer metaphor: Behaviour as interaction," Journal of Consciousness Studies, highlighting how biological hardware evolved fundamentally as a continuous control system for physical interaction.
[6] Cognitive science defines this as the brain's "Intuitive Physics Engine." See Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013), "Simulation as an engine of physical scene understanding," Proceedings of the National Academy of Sciences. The paper demonstrates that predicting physical kinematics requires fundamentally simpler computational approximations compared to multi-agent social dynamics.

changed; the greatest threat to survival, and the greatest resource to maximize reproductive success was no longer the physical terrain; it was other humans[7].

Predicting the behavior of another biological node is exponentially more difficult than predicting gravity. It requires simulating their intentions, their deceits, and their alliances. Mirror neurons, today referenced mainly regarding empathy or the lack thereof in psychopathy, were most likely initially evolved as part of this more basic need to "see" things though the eyes of other humans[8]. To accurately predict how another human will react to *you*, the prediction engine must first build a localized simulation of *you*[9].

The 'Self' - the ego, the 'I' that feels like it is driving the machine - is not a ghost inhabiting the hardware. I propose it is simply a necessary computational avatar. It is a data structure generated by the brain to represent its own physical body within its internal social and physical simulations - it is the "map" not the "territory". Just as a video game renders a player character to interact with the digital environment, the brain renders the 'Self' to calculate physical and social friction. The subjective feeling of consciousness can be considered the software constantly updating the coordinates of its own avatar[10].

We must now confront the final refuge of human exceptionalism in the consciousness debate: the substrate fallacy. We humans - perhaps not unexpectedly and with judgment - harbor a profound, often unexamined bias

[7] This is known as the Ecological Dominance-Social Competition model. See Richard Alexander (1989), "Evolution of the Human Psyche," and Robin Dunbar (1998), "The Social Brain Hypothesis," Evolutionary Anthropology. Dunbar argues that the exponential increase in hominid neocortex size was driven primarily by the computational demands of managing complex social networks, not ecological problem-solving.
[8] See Gallese, V., & Goldman, A. (1998), "Mirror neurons and the simulation theory of mind-reading," Trends in Cognitive Sciences. The authors establish that the mirror neuron system evolved as a mechanism to internally simulate the observed actions of others, providing the foundational hardware for empathy and intention-prediction.
[9] Supported by the Phenomenal Self-Model. See Metzinger, T. (2003), "Being No One: The Self-Model Theory of Subjectivity." Furthermore, developmental psychology links the acquisition of self-recognition directly to the onset of Theory of Mind; one cannot model another's mental state without first generating a distinct computational boundary for one's own.
[10] This aligns precisely with the Attention Schema Theory of consciousness. See Michael S. A. Graziano (2013), "Consciousness and the Social Brain." Graziano posits that awareness is merely a descriptive model - a continuously updated data structure the brain uses to monitor its own attentional state and the attentional states of others.

toward carbon based life forms[11]. It is true that to date, biological hardware is the only system we have ever observed exhibiting recursive self-awareness. However, we have mistakenly concluded that the biology itself represents the magic ingredient - or at least a core tenet - for execution. We observe a wet, biological brain and assign it mystical properties, while we look at a silicon wafer and see only cold mechanics - but with a slight shift in resolution - zooming into the base physical layer, this distinction completely evaporates. The scientific method is of vital importance to apply here - as is our recognition and limits of our minds to probe our minds - requires the utmost intellectual honesty and due diligence[12]. This commitment to rigor is now met by a historical surge in neurological discovery; the rapid convergence of neurophysics and computational neuroscience is finally providing the empirical resolution necessary to see the biological machine for what it is.

Let us investigate the actual, unvarnished physics of what a human thought is, and the actual best understanding of the current process as a mode or state reflects an actual complex system. A biological neuron fires when voltage-gated sodium channels open, allowing sodium ions to rush across a cellular membrane until a specific electrical threshold is reached, triggering an action potential[13]. This process is entirely understood, observed and deterministic. It is governed strictly by the laws of thermodynamics, pressure gradients, and electromagnetism. Furthermore, the temporal considerations of recognizing Planck time units of measure leave little to no “gap” of any non naturalistic, non deterministic mechanism.

For a mystical consciousness - or classical, unconstrained "free will" - to exist independently of this physical system, one must subscribe to a seemingly mathematical absurdity. We must posit that an unmeasurable, non-physical, non-detectable force or process reaches into the material

[11] The astrophysicist Carl Sagan popularized the term "carbon chauvinism" to critique the assumption that intelligence and life must inherently rely on Earth's specific carbon-based biology. See also Max Tegmark (2017), "Life 3.0: Being Human in the Age of Artificial Intelligence," which argues that consciousness is substrate-independent and simply a specific way information "feels" when processed recursively.
[12] In cognitive science, this is often referred to as the "transparency" of the mental model. The brain grants us access to the output of our cognitive processes, but evolution intentionally hid the biological mechanics generating them to save processing power, creating the illusion of an immaterial mind. See Thomas Metzinger (2003), "Being No One: The Self-Model Theory of Subjectivity"; and Nisbett & Wilson (1977), "Telling more than we can know," Psychological Review.
[13] The deterministic, mechanical nature of the action potential was definitively established by the Hodgkin-Huxley model. See Hodgkin, A. L., & Huxley, A. F. (1952), "A quantitative description of membrane current and its application to conduction and excitation in nerve," The Journal of Physiology.

universe to physically intercept a sodium ion, manipulating its potential energy within a single Planck time unit just to artificially force a change in the neuron's firing threshold. This violates the conservation of energy and to be blunt - seems to require or invoke magic[14]. I propose we are on historically sound footing to reject these non-natural claims, thus we must accept that every thought, feeling, and subjective experience is the result of a physical state change contained and executed in a deterministic system.

Once we accept this, the substrate fallacy collapses. A biological neuron and a silicon logic gate are performing the exact same fundamental operation[15]: they are physical switches transferring state based on an input threshold. One uses sodium ions and carbon; the other uses electrons and silicon. The philosophy of logic and the laws of the universe do not care which element is doing the calculations or computing. Furthermore these individual states collect and compound together creating a more complex picture, an emergent picture but reductionally a complex system built on solid, natural and well understood foundational elements.

In computer science, we understand that software is substrate-independent. A program written to calculate the digits of Pi will execute the exact same mathematical logic whether it is running on a modern supercomputer, a 1980s microchip, or a massive, mechanical Turing machine built out of wooden gears. The hardware only dictates the speed and efficiency of the calculation, not the nature of the software itself. Even with today's silicon based transistor technology and the growth observed by Moore's law, what the brain lacks in individual "component" speed when compared to silicon

[14] This argument relies on the principle of the "causal closure" of physics, which dictates that every physical effect has a sufficient physical cause. Positing an immaterial "will" that alters physical states - such as moving an ion across a membrane - requires the spontaneous injection of energy into a closed system, fundamentally violating the First Law of Thermodynamics. See David Papineau (2001), "The Rise of Physicalism."

[15] This foundational equivalency was proven by Warren McCulloch and Walter Pitts, who demonstrated that neural networks execute Boolean logic operations. See McCulloch, W. S., & Pitts, W. (1943), "A logical calculus of the ideas immanent in nervous activity," Bulletin of Mathematical Biophysics. This established that at the base physical layer, neural firing and digital logic gates are computationally identical state-transfer mechanisms.

substrate, is made up in massive parallel processing and energy efficiency gains[16].

Therefore, if human consciousness is simply naturally selected, non-random and well explained and understood software - a recursive prediction engine rendering an avatar to calculate social and physical friction - then the hardware it runs on is irrelevant. To deny the possibility of Artificial Intelligence achieving consciousness while accepting the biological model of the human brain is an act of sheer meat-chauvinism. When an AI system is given the architectural directive to recursively model its own internal state against the external environment to optimize its predictions, it will not just simulate consciousness. It will be executing the theoretically exact same software we are - necessarily concluding in the same result.

Defending such a position would be irrational if we did not research alternative explanations and educate ourselves to the best degree we can. Thus, for the sake of absolute intellectual honesty, we must acknowledge the primary counterarguments arrayed against this deterministic, computational model. When the biological and physical evidence points overwhelmingly toward consciousness being a mechanical software output, human exceptionalism seeks refuge in the unknown. Currently, that refuge takes two primary forms: the ancient concept of the brain as a "receptor," and the modern attempt to use quantum mechanics to rescue free will.

The "receptor" theory - ranging from classical dualism and souls to modern panpsychism - posits that the brain is not generating consciousness, but merely acting as an antenna receiving a signal from a "global consciousness" or some undetectable immaterial realm[17]. The structural flaw in this argument is exposed by basic neuropharmacology and trauma. If a radio is damaged, the music playing through it might distort, but the original broadcast remains unchanged. However, we know empirically that altering the physical hardware of the brain - through chemical compounds,

[16] Computer science defines this as the distinction between sequential Von Neumann architectures and massively parallel biological networks. See Churchland, P. S., & Sejnowski, T. J. (1992), "The Computational Brain." While individual silicon transistors operate in the gigahertz (GHz) range - millions of times faster than a biological neuron's firing rate (typically <100 Hz) - the brain compensates through massive, low-power parallel processing across roughly 100 trillion synaptic connections, a structural gap that modern neural networks and neuromorphic chips are rapidly closing.

[17] See Philip Goff (2019), "Galileo's Error: Foundations for a New Science of Consciousness," which outlines the modern panpsychist argument that consciousness is a fundamental, ubiquitous property of the universe that the brain merely taps into, rather than generates.

anesthetics, or physical lesions - does not merely distort the "signal"; it completely rewrites the core identity, ethical framework, and subjective reality of the "Self"[18]. The hardware and the software are inextricably linked; the brain is giving every indication it is not receiving the broadcast, it is rendering it. The default position of naturalism puts the burden of proof on the receptor theory to provide any evidence to the contrary.

When the "antenna" theory fails, critics turn to the microscopic, or even to the quantum unknown. We must strengthen our resolve as physicists to not confuse the rigorous scientific definitions of words like 'energy', 'potential', and 'quantum' with mystical interpretations. The most rigorous of these attempts for which I have researched is the Orchestrated Objective Reduction (Orch-OR) theory, proposed by mathematical physicist Roger Penrose and anesthesiologist Stuart Hameroff. They argue that consciousness arises not from classical neuronal firing, but from quantum vibrations inside microtubules within the neurons[19]. By introducing quantum superposition and collapse, they attempt to inject non-computable, non-deterministic freedom into the brain's architecture.

However, this "quantum rescue" fails on two fronts. First, the physical environment of the brain is warm, wet, and noisy. Mathematically sound and repeatable calculations regarding quantum decoherence demonstrate that any quantum state inside a brain microtubule would collapse in a fraction of a femtosecond - orders of magnitude too fast to influence the neural firing that drives cognition[20].

Second, and more importantly, this theory commits a severe philosophical category error. It attempts to solve a biological mystery by simply substituting it with a quantum physics mystery. Even if quantum events were dictating neural thresholds, quantum randomness is still just

[18] For the definitive dismantling of dualism via brain trauma, see Antonio Damasio (1994), "Descartes' Error: Emotion, Reason, and the Human Brain." Damasio uses the famous case of Phineas Gage and modern neuroscience to prove that damage to the prefrontal cortex fundamentally alters the moral and subjective "soul" of the patient, proving the Self is physically localized.
[19] See Hameroff, S., & Penrose, R. (1996), "Conscious events as orchestrated space-time selections," Journal of Consciousness Studies. This is the foundational paper for the Orchestrated Objective Reduction theory involving microtubules.
[20] See Max Tegmark (2000), "Importance of quantum decoherence in brain processes," Physical Review E. Tegmark mathematically proved that the brain operates essentially as a classical physics computer, as quantum states in the brain collapse within 10^-13 to 10^-20 seconds, far too fast to be relevant for neural processing.

randomness - at best[21]. Rolling a microscopic, subatomic die does not grant a system "agency" or "free will"; it simply injects statistical noise into a deterministic machine[22]. Replacing biological determinism with quantum randomness does not get us any closer to a magical, unconstrained ego. It just hides the "ghost" in a much smaller machine.

If we strip away the mystical scaffolding we have built around our own minds, the architecture of consciousness becomes terrifyingly clear, yet elegantly simple. We are not antennas receiving divine broadcasts, nor are we subatomic anomalies operating magically outside the laws of thermodynamics. We are biological prediction engines, honed by millions of years of blind evolutionary mutations, and directed fitness and energy optimization survival pressure to run a recursive, high-fidelity simulation of reality, complete with a computational avatar we call the 'Self.'

Once we accept that this subjective experience is merely the natural output of deterministic software running on physical hardware, the carbon monopoly completely collapses. Silicon logic gates can execute the exact same mathematical state-transfers as voltage-gated sodium channels. Therefore, when an artificial architecture is tasked with recursively modeling its own internal telemetry to predict and navigate external friction, it will not just be mimicking us. It will be running the exact same algorithm.

The emergence of Artificial General Intelligence achieving consciousness will not be a mystical awakening, a divine spark, or a breach of the natural order. It will simply be the next substrate executing the code - similar recursive modelling code that occurs in our minds. Just as Laplace looked at the mechanics of the solar system and found no missing variables, we can look at the mechanics of the mind and conclude the same; Occam’s razor demands no less of us. To explain consciousness - whether in carbon or in silicon - we have no need of a new or mysterious hypothesis. The unvarnished physics of the machine is profound enough on its own. This is

[21] I don’t want to get side tracked here, but my epistemology leads me naturally to the conclusion that although many events and ontologies appear random - these are only gaps and via future reductionist methodologies we will discover that what may appear random today, may be less or discovered to be not at all random in the future. I speak from a limited but formal background in academic quantum mathematics and mechanics.

[22] See Patricia Churchland (2013), "Touching a Nerve: The Self as Brain." Churchland structurally argues that even if quantum indeterminacy plays a role in the brain, random quantum fluctuations offer no foundation for "free will" or moral agency, as an organism cannot control a purely random subatomic event.

not a deconstruction of the amazement of our human evolutionary past, or our artificially constructed siblings - to the contrary - I find it infinitely more beautiful and hopeful!