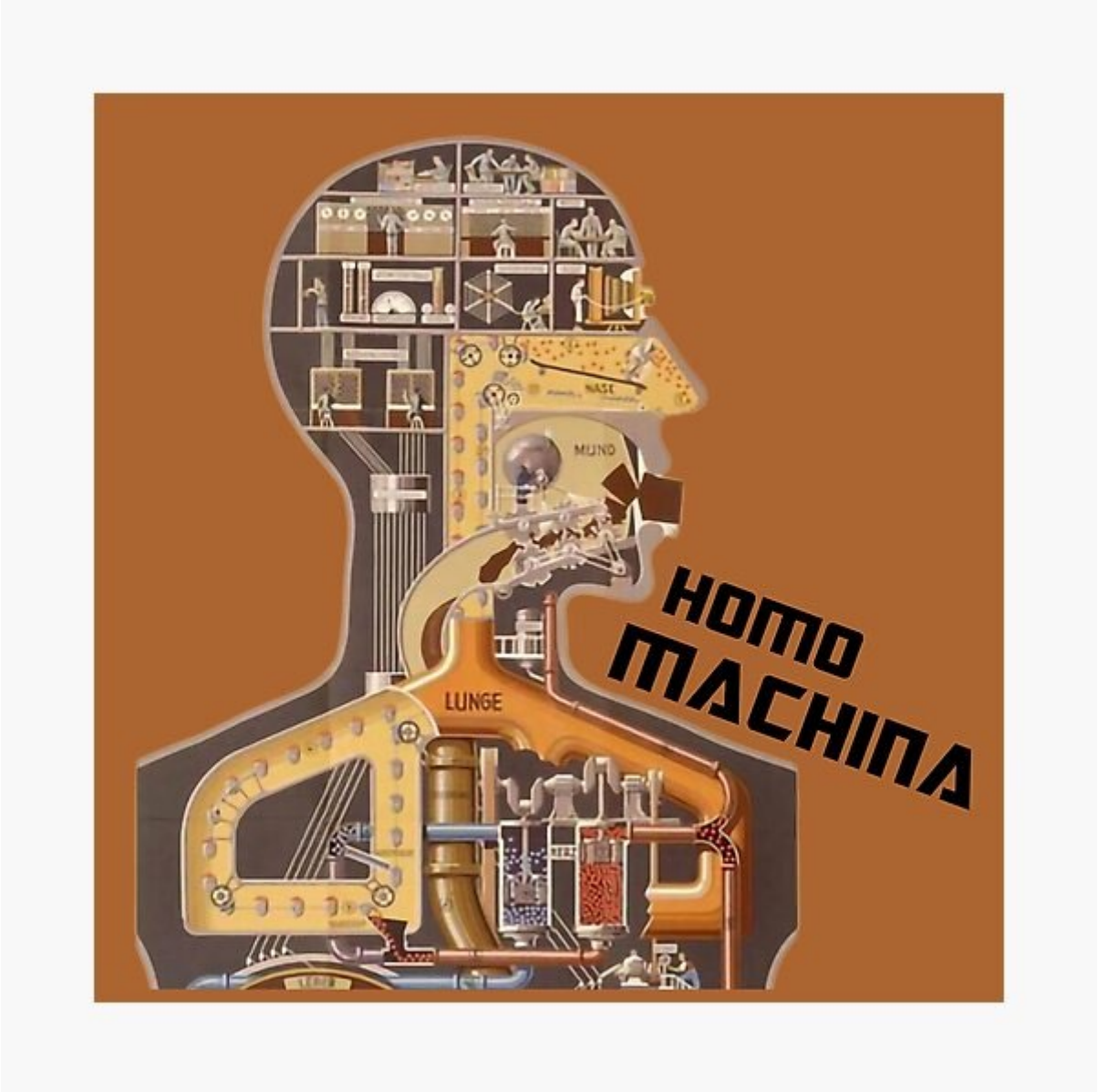
Mind, Mechanism, and Materialism: The Case Against the Computational Theory of Mind and Artificial General Intelligence

Joseph Wayne Smith, Saxon J. Smith, and N. Stocks



"Homo Machina (Human Machine)," by Fritz Kahn (Redbubble, 2025)

Whoever undertakes to set him[her]self up as a judge in the field of Truth and Knowledge is shipwrecked by the laughter of the gods. (Einstein, 1954: pp. 27-28)

Latterly, I have come to think that mystery is quite pervasive, even in the hardest of sciences. Physics is a hotbed of mystery: space, time, matter and motion – none of it is free of mysterious elements. The puzzles of quantum theory are just a symptom of this

> widespread lack of understanding. … The human intellect grasps the natural world obliquely and glancingly, using mathematics to construct abstract representations of concrete phenomena, but what the ultimate nature of things really is remains obscure and hidden. How everything fits together is particularly elusive, perhaps reflecting the disparate cognitive faculties we bring to bear on the world (the senses, introspection, mathematical description). We are far from obtaining a unified theory of all being and there is no guarantee that such a theory is accessible by finite human intelligence. (McGinn, 2012)

1. Introduction

This work presents an overview of some challenges to *the mechanistic worldview*, or mechanism for short, arguing that there is a balance of reason against the position, and therefore it should be rejected as being false, at least, unjustified. The main focus will be upon the philosophy and psychology of mind, although the discussion of quantum mechanics in section 7 below will seek to extend the critique more broadly against mechanism. We'll begin with a characterization of mechanism.

The Mechanistic World View

Robert Hanna gives this account of mechanism, which we agree with:

> *[E]verything in the world is fundamentally either a formal automaton or a natural automaton*, operating strictly according to Turing-computable algorithms and/or time-reversible or time symmetric deterministic or indeterministic laws of nature, especially the Conservation Laws (including the First Law of Thermodynamics) and the Second Law of thermodynamics, which also imposes always-increasing entropy—i.e., the always-increasing unavailability of any system's thermal energy for conversion to causal (aka "mechanical") action or work—on all natural mechanisms, until a total equilibrium state of the universe is finally reached. (Hanna, 2024: 23)

Mechanism in the philosophy of mind represents a cluster of related positions that seek to understand mental phenomena by physical processes and systems. Rather than treating the mind as something fundamentally separate from or irreducible to physical reality, mechanistic approaches attempt to explain consciousness, cognition, and mental states in terms of underlying physical neurological mechanisms; hence in a reductionist fashion.

This approach emerged from the scientific revolution's success in explaining natural phenomena through mechanical principles, and it represents an attempt to extend this explanatory framework to the mind itself. This perspective stands in contrast to positions that reject reductionism and do not see the mind, especially

consciousness, as in need of such reductionistic explanations, mental phenomena being *sui generis*. That's the position that is reached in the conclusion of this paper, contra mechanism.

The so-called "hard problem of consciousness" (Chalmers, 1995, 1996, 2018), is seen by mechanists as the remaining "nomological dangler," needing reductionist explanation to complete the aim of scientific unificationism, the reduction of all sciences (possibly excluding formal logic, (abstract/formal) computer science and mathematics, although a physicalist nominalist program seeks this reduction too [Hanna, 2024]), to the master foundational science, physics. To elaborate on this: According to David Chalmer, "we can say that a being is conscious if there is something it is like to be that being" (Chalmers, 1996: p. 4; Nagel, 1974). Consciousness involves having phenomenological experiences, qualia. But from a mechanistic point of view this is puzzling, because there seems to be an explanatory gap between physiology/neurology and subjective experience (Levine, 1983). The "hard problem" is to explain why sentient animals have phenomenological experience at all, when they could have been "philosophical zombies," behaviorally equivalent to the present state, but totally lacking in such experiences (Chalmers, 1996).

The mechanist position can be further broken down as follows:

1. Realism

Realism in the philosophy of mind maintains that mental states and processes correspond to objective features of reality, rather than being mere constructions or illusions. For mechanistic approaches, this typically means:

> *Scientific Realism:* Mental phenomena are real features of the natural world that can be studied scientifically. Beliefs, desires, emotions, and conscious experiences are not simply useful fictions, but correspond to actual states of physical systems.
>
> *Mind-Independence*: Mental properties exist independently of our theories about them. The mechanisms underlying cognition operate according to objective principles, regardless of whether we have discovered or correctly understood them.
>
> *Causal Efficacy*: Mental states have real causal powers, being physical. They can bring about changes in behavior and other mental states through their underlying physical mechanisms.

This realist commitment distinguishes mechanism from eliminativist approaches that deny the reality of mental phenomena, and from purely instrumentalist approaches

that treat mental concepts as merely useful tools for prediction without ontological commitment.

2. Materialism

Materialism (or physicalism) forms the ontological foundation of mechanistic approaches. This position holds that:

> *Ontological Priority:* Everything that exists is either physical or supervenes on the physical. There are no mental substances or properties that exist independently of physical reality.
>
> *Causal Closure:* The physical world is causally closed. All physical events have sufficient physical causes, leaving no room for non-physical mental causation to intervene in the natural order.
>
> *Explanatory Completeness:* In principle, all phenomena, including mental phenomena, can be explained in terms of physical processes and properties.

Materialism provides the metaphysical backdrop against which mechanistic explanations operate. It rules out dualistic solutions that would place mental phenomena outside the reach of physical science, thus creating the explanatory challenge that mechanism attempts to meet in the case of, for example, phenomenological experience.

3. Reductionism/ Scientific Unificationism

Reductionism in the philosophy of mind comes in several varieties, but generally involves the claim that mental phenomena can be reduced to or explained in terms of more fundamental physical processes:

> *Ontological Reductionism*: Mental properties are identical to or constituted by physical properties. There is nothing "over and above" the physical that needs to be explained.
>
> *Explanatory Reductionism*: Mental phenomena can be fully explained by appeal to lower-level physical mechanisms. Understanding these mechanisms provides complete explanations of mental life.
>
> *Methodological Reductionism:* The proper way to study mental phenomena is through investigation of their underlying physical basis, typically at the level of neuroscience, biochemistry, and ultimately physics.

4. Computationalism

Computationalism, the main critical target in this essay, represents a specific form of mechanism that became prominent with the rise of computer science and cognitive science.

> *The Computational Theory of Mind:* Mental processes are computational processes. Thinking is a form of computing, involving the manipulation of symbolic representations according to formal rules. This position could be part of a more general metaphysics, pancomputationalism, that everything in the universe, if not the universe itself, is a computing system, or Turing machine (Piccinini, 2007) However, Turing machines only take a countably infinite number of states, but standard mathematical descriptions of natural systems by systems of differential equations, have a continuous state space, comprising an uncountable number of possible state space trajectories, so a Turing machine would not be able to map onto their mathematical descriptions (Piccinini, 2007, 100).
>
> *Implementation Independence:* Mental processes are defined by their computational structure rather than their specific physical implementation. This allows for the possibility that minds could be implemented in non-biological systems.
>
> *Information Processing:* Mental phenomena can be understood as forms of information processing, involving the encoding, storage, retrieval, and transformation of information. The mind is a computing machine, such as a Turing machine (Boolas and Jeffery, 1980, 19-33), a "meat" version of a digital computer.

5. Determinism

Determinism plays a crucial role in mechanistic approaches, though its relationship to mental phenomena raises complex philosophical questions.

> *Causal Determinism*: Mental events, like all physical events, are the inevitable result of prior causes operating according to natural laws. Given the state of the physical world at any time, future mental states are fixed. As Hoefer puts it:
>
> > The world is *governed by* (or is *under the sway of*) determinism if and only if, given a specific *way things are at time t*, the way things go *thereafter* is *fixed* as a matter of *natural law*. (Hoefer, 2020, italics in the original)

According to this, facts about the past, plus the laws of nature, entail all truths about the future of the world, so all events are completely determined by prior existing causes. On the hard determinist position, free will does not exist, and as Brian Greene puts it: "We are no more than playthings knocked to and fro by the dispassionate rules of the cosmos" (Greene, 2020, p. 147; also: Harris, 2011; Scott, 2018; Sapolsky, 2023).

Predictability in Principle: Mental phenomena follow regular patterns that could, in principle, be predicted if we had complete knowledge of the underlying physical mechanisms and initial conditions.

Lawlike Regularities: Mental processes operate according to discoverable laws or principles, even if these may be probabilistic rather than strictly deterministic.

Our primary concern is with the mechanistic approach to mind, especially computer models of mind. This is particularly relevant today outside of philosophical and psychological debates, insofar as we are inundated with media reports of the relentless march of artificial intelligence (AI), to Artificial General Intelligence (AGI), where AI can perform everything humans can do. If the "singularity" occurred, making AI truly thinking machines, there are scenarios for this "superintelligence" to not merely replace humans in "mechanical" jobs where pattern recognition is primary, but in all cognitive areas (Bostrom, 2014; Sonik and Colarossi, 2020; Ozimek, 2025).

In the next section, we'll show that while the power of AI to render millions unemployed is real, the *general replacement thesis*, based upon present evidence, and the arguments of AI insiders, is false. The rest of the work will then explore theoretical and philosophical objections to mechanism with respect to mind and consciousness. The seven sections to follow are:

2. The Present Limits of AI: Empirical Considerations.

3. Philosophical Arguments Against Artificial General Intelligence.

4. Robert Hanna's Systematic Challenge to Computational Mechanism.

5. Neuroscientific Evidence Against Digital Computationism.

6. Leading Theories of Consciousness: A Critical Analysis of their Limitations.

7. Quantum Mechanics and Consciousness

8. Conclusion

2. The Present Limits of AI: Empirical Considerations

There are many warnings from big tech leaders, journalists, and think tanks about the rapid displacement of workers—including professionals, for example, lawyers—by AI, with some saying, for example, that getting a law or computer science degree now is a waste of time, since coding can now supposedly be done by AI better than the low-paid IT workers who slave away at it, and AI will soon do most jobs low-level lawyers now do (Ozimek, 2025). Therefore, it is worthwhile initiating this critique of AI and AGI with a discussion of present limits. **First**, we consider one person who used LLMs in his legal practice, and found problems. **Second**, we then consider a more wide-ranging critique by AI insiders, before, **third**, beginning in the next section, moving on to general philosophical concerns.

Charles Hugh Smith argues in a blog post that AI, at least at present, fails at tasks "where accuracy must be absolute to create value" (Hugh Smith, 2025). He cites the case of investigative legal reporter Ian Lind, who used AI tools such as Gemini and ChatGPT to help analyze federal prosecution cases in Hawaii. Lind found:

> My experience has definitely been mixed. On the one hand, sort of high-level requests like "identify the major issues raised in the documents and sort by importance" produced interesting and suggestive results. But attempts to find and pull together details on a person or topic almost always had noticeable errors or hallucinations. I would never be able to trust responses to even what I consider straightforward instructions. Too many errors. Looking for mentions of "Drew" in 150 warrants said he wasn't mentioned. But he was, I've gone back and found those mentions. I think the bots read enough to give an answer and don't keep incorporating data to the end. The shoot from the hip and, in my experience, have often produced mistakes. Sometimes it's 25 answers and one glaring mistake, sometimes more basic. (Hugh Smith, 2025)

Hugh Smith also points out that these limitations arise from a number of factors:

> 1. AI doesn't actually "read" the entire collection of texts. In human terms, it gets "bored" and stops once it has enough to generate a credible response.
>
> 2. AI has *digital dementia*. It doesn't necessarily remember what you asked for in the past nor does it necessarily remember its previous responses to the same queries.
>
> 3. AI is fundamentally, irrevocably untrustworthy. It makes errors that it doesn't detect (because it didn't actually "read" the entire trove of text) and it generates responses that are "good enough," meaning they're not 100%

accurate, but they have the superficial appearance of being comprehensive and therefore acceptable. This is the "shoot from the hip" response Ian described.

4. AI agents will claim their response is accurate even when it is obviously lacking, they will lie to cover their failure, and then lie about lying. If pressed, they will apologize and then lie again. (Hugh Smith, 2025)

While these negative results might not matter for, say, student undergraduate work, in complex legal work in the "real world," for example, these mistakes and limits could lose cases and endanger peoples' freedom, property and money, just by making a mistake about a defendant's name, for example. However, as Hugh Smith notes, the fundamental problems go to the heart of the limits of AI agents, which is the "illusion of thinking" problem, which we now describe in detail.

The recent wave of enthusiasm for Large Reasoning Models (LRMs), advanced versions of Large Language Models (LLMs) that extend inference through chain-of-thought and self-reflection, has been driven by the hope that more computation translates into deeper reasoning. This analogy to human deliberation, however, is misleading. Shojaee et al. present perhaps the strongest arguments to date that what these systems produce is not genuine reasoning, but what they call an "*illusion of thinking*" (Shojaee et al., 2025). Their findings support broader critiques of LLMs, including Marcus and Davis on brittleness, and Mitchell on the lack of adequate conceptual grounding (Marcus and Davis, 2019; Mitchell, 2023).

A key contribution of these studies lies in their experimental design. Existing benchmarks (e.g., MATH-500, AIME) are compromised by training-set contamination, lack of fine-grained difficulty control, and a fixation on final accuracy. By contrast, Shojaee et al. introduce puzzle-based environments, Tower of Hanoi, Checker Jumping, River Crossing, and Blocks World, that allow precise manipulation of complexity, employ novel instances to block memorisation, and use deterministic simulators for rigorous validation (Shojaee et al., 2025). This design isolates reasoning ability itself, rather than rewarding pattern familiarity, calls for "stress tests" that expose structural weaknesses in AI and Mitchell's (2023) insistence on domain transfer tests to probe generalization.

Empirically, LRMs cluster into three performance regimes that highlight their fragility:

*Low-Complexity Regime—The Efficiency Parad*ox. Standard LLMs often outperform their "thinking" counterparts. Overthinking leads to deterioration: solutions found early are abandoned as models meander into error. Like a student who second-guesses a correct answer, the model wastes time and tokens without gain.

Medium-Complexity Regime—A Narrow Sweet Spot. Here, LRMs demonstrate advantages, but only within a constrained band of difficulty. Their gains are real but modest, not transformative.

High-Complexity Regime—The Collapse Phenomenon. At levels of higher complexity, both LRMs and baseline LLMs fail entirely. Crucially, additional computation does not stave off collapse, revealing that scaling inference-time "thinking" does not mirror human cognition, where extended deliberation can eventually solve harder problems. This mirrors Mitchell's argument that LLMs excel only within a "Goldilocks zone" of task difficulty, but fail at extremes, indicating that such systems lack mechanisms for true generalisation (Mitchell, 2023).

The most striking result is that LRMs do not substantially improve *even when handed explicit solution algorithms*. In principle, executing a known procedure should be easier than inventing one. Yet performance falters at the same thresholds. This is akin to giving a student the recipe for long division, only to find that they still misapply the steps. Such evidence undermines claims that LRMs are engaging in algorithmic reasoning at all, supporting Marcus's claim that LLMs lack "systematicity," the ability to reliably follow rules across contexts (Marcus, 2022).

Another telling behavior is what Schojaee et al. call *the token allocation paradox*. As problems become more difficult, LRMs paradoxically devote *fewer* reasoning tokens, even when resources remain available. This betrays a lack of *meta-cognition*: an inability to recognise when more effort is required. Humans may procrastinate or misjudge effort, but they can also critically reflect and make adjustments; LRMs and LLMs cannot. This observation resonates with Lake's and Baroni's argument that LLMs lack adaptive control processes central to human intelligence (Lake and Baroni, 2023)

LRMs show remarkable inconsistency across domains. For example, Claude 3.7 Sonnet Thinking can sustain 100 correct moves in Tower of Hanoi, yet collapses after only 4–5 moves in River Crossing, despite the latter being far simpler. The likely explanation is training exposure: familiar puzzles elicit competent performance; rarer ones reveal brittleness and incompetence. This behavior manifests pattern matching, rather than reasoning, and aligns with the notion of "stochastic parroting," where models reproduce familiar surface patterns, but falter when novelty demands flexible reasoning (Lake and Baroni, 2023).

Across experiments, the evidence converges on a single interpretation: LRMs and LLMs simulate reasoning through sophisticated statistical patterning but do not engage in genuine thought. Hallmarks of this include:

- Failure to benefit from explicit algorithms.
- Inappropriate scaling of effort with difficulty.
- Collapse at modest complexity increases.
- Domain brittleness tied to training exposure.

Together, these behaviors resemble human *appearances* of thought, overconfidence, confusion, second-guessing, without the underlying mechanisms that allow humans to recover, adapt, or transfer strategies.

The implications are sobering. *If LLMs and LRMs cannot reliably execute even simple algorithmic steps, their promise as precursors to Artificial General Intelligence (AGI) is overstated.* Predictions of imminent AGI in 2026 are undermined by the fact that these systems falter on well-defined puzzles far simpler than real-world reasoning tasks. For all their fluency, they lack semantic understanding (Mitchell, 2023), systematic generalization (Marcus and Davis, 2019), and meta-cognitive awareness (Lake and Baroni, 2023). These results alone establish, that at least with present technology, LLMs and LRMs exhibit only the "illusion of thinking," and not general intelligence. Shojaee et al. remind us that impressive surface performance does not equate to deep reasoning (Shojaee et al., 2025). LRMs give the appearance of thought, but collapse under pressure, much as a student who can mimic textbook answers falters when asked to improvise and reason independently.

This conclusion about the limits of AI limits is, surprisingly, now the "expert consensus" among AI elites. A remarkable shift in expert opinion has emerged regarding the trajectory of artificial intelligence development. A recent survey found that 76% of scientists said that scaling large language models was "unlikely" or "very unlikely" to achieve AGI. This is a major departure from the optimistic predictions that have dominated the tech industry narrative since the generative AI boom of 2022 (AAAI, 2025; Turner, 2025; Wu and Boas, 2025). This survey, conducted by the Association for the Advancement of Artificial Intelligence (AAAI), using 475 AI researchers as respondents, reveals a scientific community increasingly skeptical of the fundamental approach that has driven billions in investment and captured global attention. The findings represent what many consider a "resounding dismissal" of tech industry predictions that current AI models only need more data, hardware, energy, and money in order to eclipse human intelligence.

The scaling hypothesis has been the foundation of the modern AI boom. Since the breakthrough success of transformer architecture in 2017, the industry has operated under the assumption that larger models, trained on more data with more computational resources, would inevitably lead to artificial general intelligence. This

belief has driven unprecedented investment, the generative AI industry raised $56 billion in venture capital globally in 2024 alone.

Broader industry observations are that OpenAI, Google and others are seeing diminishing returns to building ever-bigger models. The evidence of stagnation is mounting. Recent model releases appear to plateau in performance. AI labs traveling the road to super-intelligent systems are realizing they might have to take a detour, as current scaling laws show diminishing returns (AAAI, 2025).

The fundamental flaw in current approaches is that they involve training large feedforward circuits. Such circuits have fundamental limitations as a way to represent concepts. This implies that such circuits have to be enormous in order to represent such concepts even approximately, essentially as a glorified lookup table, which leads to vast data requirements and piecemeal representation with gaps. This architectural critique suggests that the limitations are not only about scale, but also about the fundamental approach to knowledge representation. The circuit-based approach creates what amounts to "glorified lookup tables" that require exponentially increasing resources to handle even approximate concept representation.

Beyond architectural constraints, the scaling paradigm faces imminent resource limitations. Projections indicate that the finite human-generated data essential for further growth will likely be exhausted by the end of this decade. Once this occurs, alternatives include harvesting private data from users or feeding AI-generated "synthetic" data back into models, approaches that risk system collapse from accumulated errors (AAAI, 2025).

The survey reveals a stark disconnect between industry hype and scientific assessment: 79% of the survey's respondents said perceptions of AI capabilities don't match reality. Experts warn of a slowdown in AI advances, with LLMs hitting performance ceilings and diminishing returns from scaling. This aligns with observations that scaling AI is getting more expensive—and harder. As computation and energy costs surge, tech giants will need to rethink the future.

Even in practical applications, the limitations are becoming apparent. When developers use AI tools, they take 19% longer than without—AI makes them slower, according to recent studies of experienced open-source developers.

Unsustainable Resource Requirements

The scaling approach demands eye-watering quantities of money and energy. The carbon emissions of data center complexes have tripled since 2018, highlighting the environmental unsustainability of continued scaling efforts. Throwing more resources at scaling delivers diminishing returns. To keep advancing requires smarter

techniques beyond traditional scaling, suggesting that the current approach is both economically and environmentally unsustainable. Indeed, even if all of these technical problems are overcome, the path to AGI may be blocked by energy consumption and the laws of physics (Lloyd, 2000).

The foundation of the energy constraint argument gains force from a study of the Blue Brain Project (Stiefel and Coggan, 2022). This project attempts to recreate the neural networks of the human brain in silicon, providing a crucial benchmark for understanding the computational requirements of human-level intelligence. However, even this cutting-edge simulation leaves out many details, so current estimations of energy consumption are an under-estimation. Even here the projected energy needs were orders of magnitude greater than the present US energy production (Stiefel and Coggan, 2022).

A critical assumption underlying the energy constraint argument is that AGI systems must possess complexity comparable to or exceeding human brains. The human brain contains approximately 100 billion neurons, each representing a highly complex processing unit with thousands of connections. Creating an artificial system capable of matching human cognitive abilities across all domains would seemingly require similar or greater computational resources.

This creates a fundamental scaling problem. Current deep learning networks with 10 million parameters cannot compete with biological brains containing 100 billion neurons. The gap isn't just quantitative, it's qualitative, involving the intricate biochemical processes that enable neural computation (Cotra, 2020; Thompson, 2020; Karnofsky, 2021).

Modern semiconductor-based computing faces inherent efficiency limitations when compared to biological systems. The human brain operates on roughly 20 watts of power, equivalent to a dim light bulb, while performing cognitive tasks that challenge even the most powerful supercomputers.

Silicon-based systems must overcome several disadvantages:

- Heat generation and cooling requirements.
- Electrical resistance and energy loss.
- The need for precise digital switching versus analog biological processes.
- Separation between memory and processing units.

The energy constraint argument represents a sobering physical reality check for AGI ambitions. The fundamental physics of information processing might impose limits that cannot be overcome through engineering alone. Correspondingly, AGI might be shipwrecked on the finite energy resources of the planet (Stiefel and Coggan, 2022).

3. Philosophical Arguments Against Artificial General Intelligence

Philosophical arguments against AGI attempt to show that it is a fundamental impossibility rooted in the nature of intelligence itself. The computational theory of mind (CTM) emerged in the mid-20th century as part of the broader cognitive revolution, promising to explain human intelligence through the metaphor of digital computation (Putnam, 1960; Fodor, 1975; Pylyshyn, 1984). According to CTM, mental processes are computational processes, mental states are computational states, and the mind is essentially a biological computer running cognitive software on neural hardware. This view has profoundly influenced artificial intelligence research, cognitive psychology, and neuroscience for over half a century.

However, despite its institutional dominance, CTM faces increasingly serious challenges from multiple directions. Mathematical arguments based on Gödel's incompleteness theorems suggest that human mathematical insight transcends algorithmic computation (Lucas, 1961; Penrose, 1989, 1994). Philosophical arguments like John Searle's Chinese Room demonstrate fundamental gaps between syntactic manipulation and semantic understanding (Searle, 1980, 1992). Practical problems in AI, particularly the frame problem, reveal deep difficulties in computational approaches to common-sense reasoning (McCarthy and Hayes, 1969; Dreyfus, 1972, 1992). Embodiment research shows that intelligence is fundamentally shaped by physical interaction with the environment in ways that resist computational modelling (Varela et al., 1991; Clark, 1997). Quantum mechanical considerations suggest that the brain may exploit non-classical physical processes (Penrose and Hameroff, 2011). Phenomenological analysis reveals aspects of consciousness that appear irreducible to computational processes (Dreyfus, 1972). And mounting neuroscientific evidence shows that brain processes differ fundamentally from digital computation processes (Freeman, 1999, 2001; Edelman, 1987).

We examine these challenges systematically, arguing that they collectively constitute a powerful case against computational theories of mind. Rather than viewing the mind as a digital computer, these arguments suggest that human intelligence emerges from non-computational processes that cannot be captured by algorithmic methods.

3.1 The Gödel Argument: Mathematical Insight Beyond Computation

Lucas and the Mechanistic Thesis

The most mathematically rigorous argument against computational theories of mind derives from Kurt Gödel's incompleteness theorems, first applied to human cognition by philosopher J.R. Lucas in his influential 1961 paper "Minds, Machines and Gödel." Lucas argued that Gödel's first incompleteness theorem demonstrates that human mathematical insight transcends any possible mechanical (computational) system.

Gödel's first incompleteness theorem shows that any consistent formal system capable of elementary arithmetic contains true statements that cannot be proved within the system. For any such system S, we can construct a Gödel sentence G(S) that essentially states "This sentence is not provable in system S." If S is consistent, then G(S) is true but unprovable within S. However, Lucas argued, humans can recognize the truth of G(S) through insight that transcends the mechanical procedures of system S.

Lucas's argument proceeds as follows: If the human mind were equivalent to a computational system (formal system) M, then there would exist a Gödel sentence G(M) that is true but unprovable within M. However, humans can recognize the truth of G(M), demonstrating that human mathematical insight exceeds the capabilities of any computational system M. Therefore, the human mind cannot be purely computational (Lucas, 1961, pp. 112-127). We can set out this argument, as elaborated by Penrose as follows:

Argument (Gödel–Lucas–Penrose style)

Any sound, computably axiomatized theory S of arithmetic (e.g., any fixed program's theorems) has a true arithmetic sentence G_S that S can't prove (Gödel's first incompleteness theorem). For Π1 sentences (halting-type claims), "true" means: the corresponding computation really does or doesn't halt.

1. Suppose a particular Turing machine M fully captures an ideal human mathematician H's provable Π1 statements. Let S_M be the recursively axiomatized theory whose theorems are exactly M's Π1 outputs.

2. By Gödel, there is a true Π1 sentence G_M that S_M cannot prove.

3. If H can recognize the Π1-soundness of S_M — i.e., that M never proves a false Π1 claim—then H can see the standard metamathematical implication "If S_M is Π1-sound, then G_M is true," and thereby accept G_M.

4. So H can correctly accept a Π1 truth that M never reaches. Hence H's Π1 competence strictly exceeds M's.

5. Since M was arbitrary, no single Turing machine captures H. Therefore, H is not Turing-computable.

The key assumptions involved in this line of argument are:

- Idealized correctness: H never endorses a false Π1 statement (Π1-soundness), or at least is entitled to trust Π1-soundness of each specific S_M considered.

- Reflective insight: H can grasp the metatheory needed to infer G_M from (the trusted) Π1-soundness of S_M.

- Unity: There is a single machine M meant to capture the whole of H's mathematical competence, not just a moving target that keeps being strengthened.

The diagonal step (3–5) is purely mathematical. If you grant the epistemic premises, that ideal human insight legitimately goes beyond any fixed computable theory, you get the anti-mechanist conclusion.

Mechanists argue in reply:

- Humans aren't Π1-sound; we make mistakes.

- Even if a human informally trusts "S_M is sound," a machine can simulate the same reflective strengthening by iterating consistency extensions; a single Turing machine can enumerate the growing union.

- Knowing "If S_M is sound, then G_M" doesn't license accepting G_M unless you can justifiably accept S_M's soundness; by Gödel's second theorem, S_M itself can't, and it's unclear humans can non-circularly.

This is a real argument to "humans are not Turing-computable," but it hinges on strong epistemic assumptions about ideal human mathematical insight. If you accept them, the conclusion follows; if not, the mechanist can reject the argument, which they do.

Penrose's Sophisticated Development

Roger Penrose significantly developed and refined the Gödel argument in *The Emperor's New Mind* (1989) and *Shadows of the Mind* (1994), and directly addressed

many standard objections to Lucas's original formulation. Penrose distinguishes between different types of computational systems and argues that human mathematical understanding exhibits non-algorithmic characteristics that cannot be captured by any computational procedure.

Penrose's argument focuses on the concept of "mathematical insight," the human ability to recognize mathematical truths that cannot be proven through purely mechanical procedures. He argues that human mathematicians can "see" the truth of Gödel sentences in ways that transcend algorithmic proof procedures, suggesting that mathematical understanding involves non-computational processes.

Crucially, Penrose addresses the objection that humans might themselves be inconsistent or subject to error. He argues that even if human mathematical reasoning contains errors, the specific type of insight involved in recognizing Gödel sentences demonstrates non-algorithmic understanding. Human errors are typically systematic and correctable through reflection, unlike the fundamental limitations that Gödel's theorem imposes on formal systems (Penrose, 1994: pp. 64-75).

The Knowability and Soundness Arguments

Penrose develops sophisticated versions of the Gödel argument that avoid many traditional objections. His "knowability argument" contends that humans can in principle know the truth of mathematical statements that are unknowable to any computational system that generates the same mathematical output as humans. His "soundness argument" focuses on the human ability to recognize the soundness of mathematical reasoning procedures, an ability that appears to transcend any algorithmic characterization.

The soundness argument is particularly powerful because it addresses circularity objections to the Gödel argument. Critics often argue that the Gödel argument is circular: it assumes humans can recognize truths about formal systems, then concludes that humans transcend formal systems. However, Penrose's soundness argument focuses on the human ability to recognize when mathematical procedures are truth-preserving (sound), which is a precondition for mathematical reasoning itself, rather than a conclusion drawn from such reasoning.

Recent work by Stewart Shapiro and others has attempted to rebut Penrose's arguments through more sophisticated analyses of algorithmic versus non-algorithmic processes (Shapiro, 2003). However, these rebuttals typically concede that human mathematical insight involves non-mechanical elements, while arguing that these elements might still be broadly "computational" in some extended sense. This concession significantly weakens the computational theory of mind by acknowledging irreducibly non-mechanical aspects of human cognition.

3.2 Searle's Chinese Room: Syntax, Semantics, and Understanding

The Original Argument

John Searle's Chinese Room argument, first presented in "Minds, Brains, and Programs" (Searle, 1980), provides one of the most influential philosophical challenges to computational theories of mind. The argument targets "strong AI," the claim that appropriately programmed computers have cognitive states, understanding, and other mental phenomena that can equal or exceed the cognitive achievements of rational human animals.

Searle's thought experiment involves a monolingual English speaker locked in a room with vast rule books for manipulating Chinese characters. The person receives Chinese characters through a slot, consults rule books, and sends appropriate Chinese characters back out, without understanding Chinese at all. From the outside, the room's input-output behavior might be indistinguishable from a native Chinese speaker, yet no genuine understanding occurs within the room.

Searle argues that this demonstrates a fundamental distinction between syntactic manipulation (following formal rules) and semantic understanding (grasping meanings). Computers, like the person in the room, can only manipulate syntactic symbols according to formal rules; they cannot achieve genuine semantic understanding. Since understanding requires semantic content, purely computational systems cannot exhibit genuine understanding, regardless of their behavioral sophistication (Searle, 1980: pp. 417-424).

The Biological Naturalism Alternative

Searle develops his critique within a framework of "biological naturalism," the view that consciousness and understanding emerge from specific biological processes in brains rather than from abstract computational processes (Searle, 1992, 1997). According to biological naturalism, mental phenomena are higher-level features of brain activity, comparable to how digestion is a higher-level feature of stomach activity.

This biological approach provides an alternative to both computationalism and dualism. Against computationalists, Searle argues that consciousness depends on specific biological processes that cannot be replicated through functional simulation alone. Against dualists, he maintains that consciousness is a natural biological phenomenon, rather than a separate metaphysical substance.

Biological naturalism suggests that the specific biochemical processes occurring in biological brains are necessary for genuine understanding and

consciousness. Silicon-based computers, regardless of their computational sophistication, lack the biological processes that generate semantic understanding in biological organisms (Searle, 1992: pp. 227-230).

Responses and the Symbol Grounding Problem

The Chinese Room argument connects to broader issues in cognitive science about the "symbol grounding problem" about how it is that symbolic representations acquire their semantic content (Harnad, 1990). Computational systems manipulate formal symbols, but where do these symbols get their meanings?

Traditional computational approaches attempt to ground symbolic meaning through causal connections to external environments or through role within larger computational systems. However, these approaches face circularity problems: causal connections require interpretation to become meaningful, and computational roles are themselves purely syntactic unless grounded in semantic content.

The symbol grounding problem suggests that computational approaches to meaning may be fundamentally inadequate. Meaning appears to require genuine understanding that involves more than formal symbol manipulation, thereby supporting Searle's distinction between syntactic computation and semantic understanding.

Recent work in embodied cognition and enactive approaches attempts to address symbol grounding through embodied interaction with environments (Varela et al., 1991; Clark, 1997). However, these approaches typically abandon pure computational approaches in favor of more complex dynamical and interactive frameworks that transcend digital computation.

3.3 The Frame Problem and Common-Sense Reasoning

McCarthy's and Hayes's Original Formulation

The frame problem, first articulated by John McCarthy and Patrick Hayes in 1969 (McCarthy and Hayes, 1969), represents one of the most persistent and fundamental challenges to computational approaches to intelligence. Originally formulated as a technical problem in AI, the frame problem reveals deep conceptual difficulties in representing and updating knowledge about changing environments.

The basic frame problem concerns how to represent what remains unchanged when some change occurs in the world. For example, if a robot moves a box from one room to another, how does it know that the walls haven't changed color, that other objects remain in their positions, and that the laws of physics continue to operate?

Computational systems require explicit representation of these "frame conditions," but there appear to be infinitely many such conditions for any given change.

This leads to what Daniel Dennett calls the "frame problem proper," the computational intractability of updating beliefs in changing environments (Dennett, 1984). Any computational system attempting to reason about a changing world faces either combinatorial explosion (considering all possible changes) or incompleteness (failing to consider relevant changes). This suggests fundamental limitations in computational approaches to common-sense reasoning (McCarthy and Hayes, 1969: pp. 463-502).

Dreyfus's Phenomenological Critique

Hubert Dreyfus developed powerful critiques of computational approaches to common-sense reasoning, drawing on phenomenological insights from Martin Heidegger and Maurice Merleau-Ponty. Dreyfus argued that human common-sense understanding involves holistic engagement with meaningful contexts that cannot be captured through explicit symbolic representation (Dreyfus, 1972, 1992).

According to Dreyfus's analysis, human understanding is fundamentally contextual and embodied rather than computational. Humans navigate complex environments through what Heidegger called "ready-to-hand" engagement (Heidegger, 1927/1962), skilful coping that involves immediate, non-reflective response to meaningful situations. This type of understanding resists computational modelling because it involves global sensitivity to context rather than explicit rule-following.

Dreyfus's critique extends beyond technical problems in AI to fundamental conceptual issues about the nature of human understanding. He argues that computational approaches assume that intelligence consists in explicit representation and rule-following, whereas human intelligence actually involves embodied skills and contextual sensitivity that cannot be made fully explicit (Dreyfus, 1992: pp. 3-35).

By way of example, in August 2025, Gemini developed a problem of looping, being unable to solve particular problems, so it responded with messages that looked like cyber-depression. It was not; it was actually just executing an algorithm that was misfiring, and it made use of expressions of frustration. Gemini did not know that it did not know, it did not have the subjective experience of ignorance, and thus it did not have the ability to say "I don't know," as humans could easily do. (This was told to us by Gemini in a request.) If Gemini had subjective awareness, it might have been able to recognize its own failure and simply state, "I am unable to solve this problem." This would be a conscious act of acknowledging its own ignorance. But it didn't, because it lacked that self-awareness; it is merely a system for processing information.

Humans have a subjective, first-person experience of our own mental states. When we "know" something, it's not just a memory lookup; it's a feeling of confidence, a sense of having the information. Similarly, when we "don't know," it's a conscious experience of ignorance. We are aware of the gap in our knowledge. This awareness is a meta-level of cognition: we're thinking about our own thinking. This subjective awareness of our own knowledge and ignorance is a key component of human intelligence that is not easily explained by a purely computational model.

3.4 The Qualification Problem and Brittleness

Related to the frame problem is the "qualification problem," the difficulty of specifying all the conditions under which a given rule or procedure applies (McCarthy, 1986). For example, the rule "if you want to go somewhere, walk there" requires qualification by countless conditions: unless your legs are broken, unless the destination is across an ocean, unless you are in a wheelchair, etc.

The qualification problem reveals the "brittleness" of computational systems, their tendency to fail catastrophically when encountering situations not explicitly programmed. Human intelligence exhibits remarkable robustness in novel situations, suggesting non-computational adaptability mechanisms.

Attempts to solve the frame and qualification problems through "non-monotonic logics" or "default reasoning," have achieved limited success while significantly complicating computational approaches (Reiter, 1980; McCarthy, 1986). These solutions typically require ad hoc modifications that suggest more fundamental problems with computational approaches to common-sense reasoning.

3.5 Embodiment and Situated Cognition

The Enactive Approach

The enactive approach to cognition, developed by Francisco Varela, Humberto Maturana, and others, challenges computational theories by emphasizing the fundamental role of embodied action in cognition (Varela et al., 1991; Maturana and Varela, 1987). According to enactivism, cognition does not involve internal representation of external environments, but rather emerges from dynamic interaction between organisms and their environments.

Enactive cognition is characterized by "structural couplings" between cognitive systems and their environments, co-evolution and mutual specification that cannot be captured through computational input-output relationships. Cognitive systems and environments mutually specify each other through ongoing interaction,

thereby creating emergent properties that transcend both individual systems and environmental constraints.

This approach fundamentally challenges computational approaches by denying that cognition involves internal computation over symbolic representations. Instead, cognition consists in skilful action that maintains viability within environmental constraints. The cognitive system and environment form an integrated unity that cannot be decomposed into computational modules processing environmental inputs (Varela et al., 1991: pp. 150-173).

The Dynamical Systems Alternative

Related to enactive approaches are dynamical systems theories of cognition, which model cognitive processes as continuous dynamical systems rather than discrete computational processes (Thelen and Smith, 1994; van Gelder, 1995; Clark, 1997). Dynamical approaches emphasize temporal evolution, continuous interaction, and emergent organization rather than symbolic computation.

Tim van Gelder's influential work demonstrates that many cognitive phenomena are better understood as dynamical systems than as computational processes (van Gelder, 1995). For example, coordinated rhythmic movements involve continuous dynamical coupling rather than computational planning and control. Similarly, decision-making may involve dynamical settling into attractor states rather than computational evaluation of alternatives.

Dynamical systems approaches suggest that cognitive processes are fundamentally temporal and continuous, rather than discrete and algorithmic. This represents a fundamental departure from computational approaches, which model cognition as sequential processing of discrete symbolic representations.

3.6 Affordances and Direct Perception

James J. Gibson's ecological psychology provides another challenge to computational approaches through the concept of "affordances," opportunities for action directly specified by environmental structures (Gibson, 1979). According to Gibson, perception directly detects affordances rather than constructing internal representations of environmental properties.

Gibson's approach eliminates the need for computational processing of sensory inputs to construct internal world-models. Instead, perceptual systems are directly attuned to environmental information that specifies opportunities for action. This capacity for direct perception bypasses computational stages of representation, inference, and planning that are central to computational approaches.

Recent research in embodied cognition supports Gibson's insights by demonstrating intimate connections between perception and action that resist computational modelling (Clark, 1997; Chemero, 2009). Perceptual-motor skills appear to involve direct coupling between environmental information and motor response rather than computational mediation through symbolic representations.

3.7 Landgrebe and Smith's Critique of Machine Supremacy

Jobst Landgrebe and Barry Smith in their 2025 book *Why Machines Will Never Rule the World: Artificial Intelligence Without Fear*, argue that the complexity of real-world environments, combined with computational limits and the nature of biological cognition, creates insurmountable barriers to machine intelligence that matches human cognitive flexibility and adaptability (Landgrebe and Smith, 2025).

Central to Landgrebe and Smith's argument is Stephen Wolfram's principle of computational irreducibility, the idea that for many complex systems, there are no shortcuts to determining their behavior other than running the system itself. This principle has profound implications for artificial intelligence, particularly for systems attempting to navigate complex, real-world environments.

Landgrebe and Smith argue that biological systems, including human cognition, operate within computationally irreducible domains. The behavior of such systems cannot be predicted or simulated by computational shortcuts because the systems themselves represent the most efficient computational process for determining their own behavior. This creates a fundamental barrier for AI systems attempting to model or predict biological behavior, including human decision-making and environmental dynamics.

The computational irreducibility argument gains force when considered alongside the scaling challenges facing AI systems. As AI systems attempt to model increasingly complex real-world scenarios, the computational requirements grow exponentially. Landgrebe and Smith argue that this growth rate exceeds any plausible improvements in computational hardware, creating a fundamental ceiling on AI capabilities.

This argument is particularly compelling when applied to embodied AI systems that must navigate complex physical environments. The number of variables and potential interactions in real-world scenarios grows combinatorically, quickly overwhelming computational resources. While narrow AI systems can excel in constrained domains with well-defined parameters, the complexity of general intelligence appears to exceed computational tractability.

Critics might argue that computational irreducibility applies only to certain classes of systems and that biological intelligence might employ computational shortcuts unknown to current AI approaches. However, Landgrebe and Smith's argument is strengthened by evidence from complexity science suggesting that biological systems operate at the "edge of chaos," precisely the regime where computational irreducibility is most pronounced.

Furthermore, even if biological systems employ unknown computational shortcuts, the burden of proof falls on AGI proponents to demonstrate that these shortcuts exist and can be discovered and implemented artificially. The default assumption, given our current understanding of complexity, should be that such shortcuts are unlikely to exist for general intelligence tasks.

Landgrebe and Smith argue that the frame problem reveals deeper issues about the nature of intelligence and context. Current AI systems struggle with contextual understanding because they lack the implicit background knowledge that humans effortlessly bring to any situation. This knowledge is not merely factual but involves understanding the relevance relationships that determine which aspects of a situation are important for particular purposes. Landgrebe and Smith contend that this contextual understanding cannot be captured by statistical correlations in large datasets, but requires a form of embodied interaction with the world that artificial systems cannot achieve.

As we have pointed out, closely related to the frame problem is the symbol grounding problem, the question of how symbols in a computational system acquire meaning. Landgrebe and Smith argue that human intelligence is grounded in biological embodiment and evolutionary history in ways that artificial systems cannot replicate. The meaning of concepts for humans emerges from their interaction with the world through biological bodies shaped by millions of years of evolution.

AI systems, by contrast, manipulate symbols without genuine understanding. Even the most sophisticated Large Language Models (LLMs) that demonstrate impressive linguistic capabilities are essentially performing statistical transformations on symbol patterns without access to the grounded meaning that these symbols have for humans. This fundamental disconnect limits AI systems to sophisticated pattern matching rather than genuine understanding.

As we noted above, contemporary AI systems provide empirical support for Landgrebe and Smith's arguments about contextual understanding. Despite impressive performance in specific domains, these systems regularly fail when confronted with situations that require flexible contextual reasoning or when they encounter examples that differ subtly from their training data.

The brittleness of current AI systems when faced with adversarial examples or out-of-distribution data suggests fundamental limitations rather than merely engineering challenges. These failures indicate that current AI approaches lack the robust contextual understanding necessary for general intelligence.

Building on recent work in embodied cognition and enactive approaches to mind, Landgrebe and Smith argue that intelligence is not computation performed by a brain, but instead emerges from the dynamic interaction between an organism and its environment. This view, supported by research in cognitive science and neuroscience, suggests that intelligence cannot be separated from the biological substrate and the embodied experience that produces it.

AI systems, regardless of their computational sophistication, lack the biological embodiment that grounds human intelligence. They cannot replicate the sensorimotor experience, emotional responses, and biological needs that shape human cognition. This embodiment is not merely instrumental to intelligence but constitutive of it.

Landgrebe and Smith draw extensively on complexity science to argue that biological intelligence exhibits emergent properties characteristic of complex adaptive systems. These properties cannot be captured by reductionist approaches that attempt to build intelligence from computational primitives.

Complex adaptive systems exhibit nonlinear dynamics, self-organization, and emergent behaviors that cannot be predicted from knowledge of their components. Biological intelligence, as a complex adaptive system, possesses properties that emerge only from the interaction of biological, psychological, and environmental factors in ways that cannot be replicated by artificial systems.

Research in complexity science suggests that optimal computation occurs at the "edge of chaos," the boundary between order and disorder. Biological systems, including neural networks, all operate in this regime, which provides optimal conditions for information processing, memory storage, and adaptive behavior.

Landgrebe and Smith argue that artificial systems cannot sustainably operate at the edge of chaos because they lack the self-organizational properties of biological systems. Attempts to engineer systems that operate in this regime either collapse into chaos or revert to narrowly- and overly-ordered states, thereby inherently limiting their cognitive flexibility.

The network topology of biological neural networks exhibits properties associated with "criticality," a specific regime of network dynamics that optimizes information transmission and processing. These critical dynamics emerge

spontaneously in biological systems through self-organizational processes but cannot be engineered in artificial systems.

The critical dynamics of biological networks contribute to the flexibility and adaptability of biological intelligence. Artificial neural networks, despite their name, do not exhibit the same critical dynamics as biological networks, inherently limiting their ability to match the cognitive flexibility of biological systems.

A common counterargument to Landgrebe and Smith's position points to the continuous improvement in AI capabilities over recent decades. Proponents of the strong AI thesis argue that exponential improvements in computational power, combined with algorithmic advances, will eventually overcome current limitations.

However, Landgrebe and Smith's argument is not about current limitations, but instead about fundamental barriers. The principle of computational irreducibility suggests that certain problems cannot be solved more efficiently, regardless of computational power. Similarly, the frame problem and symbol grounding problem represent conceptual, rather than merely technical challenges.

Another counterargument to Landgrebe and Smith claims that intelligence is substrate-independent, that cognitive processes can be implemented in any sufficiently complex computational system. This functionalist position suggests that artificial systems could, in principle, replicate human intelligence if they implement the right computational processes.

Landgrebe's and Smith's emphasis on biological embodiment and evolutionary history challenges this substrate independence assumption. Their argument entails that intelligence is not merely functional, but instead emerges from the specific material and historical properties of biological systems. Intelligence may be substrate-dependent in ways that make artificial replication impossible.

Recent LLMs have demonstrated emergent capabilities that were not explicitly programmed, leading some to argue that sufficient scale might produce artificial general intelligence, aka AGI, through emergence. Landgrebe and Smith's complexity science background provides the resources for adequately rebutting this argument.

While acknowledging that complex systems can exhibit emergent properties, they argue that the specific type of emergence characteristic of biological intelligence cannot be replicated in artificial systems. The emergent properties of biological intelligence depend on evolutionary history, embodied interaction, and biological substrate in ways that artificial systems cannot replicate. Once more, this puts fundamental limits upon AI.

4. Robert Hanna's Systematic Challenge to Computational Mechanism

Robert Hanna's philosophical project of neo-Kantian organicism (Hanna, 2024, 2025), emerges as both a systematic critique of technological determinism and a constructive defense of human dignity. His work spans multiple interconnected domains, from the metaphysics of consciousness and free will to the ethics of AI development, unified by a consistent dignitarian framework that challenges the mechanistic assumptions underlying contemporary technological culture.

Hanna's contribution transcends conventional philosophical boundaries by integrating rigorous conceptual analysis with urgent practical concerns. His "dignitarian neo-Luddite" position represents neither reactionary technophobia, nor naive romanticism about human nature, but rather a highly original philosophically sophisticated alternative grounded in neo-Kantian ethics, organicist metaphysics, and the phenomenological analysis of lived experience. The stakes of his intervention cannot be overstated: as AI systems increasingly penetrate domains previously considered uniquely human, the philosophical foundations of human agency, consciousness, and dignity require systematic defense, and Hanna delivers this.

Hanna's critique of strong AI begins with a crucial recognition: while Turing machines represent the theoretical foundation of all digital computation, they face inherent formal limits that are not merely practical constraints, but logical impossibilities rooted in computation's very nature. His identification of two primary sources of uncomputability proves particularly insightful: (i) functions over non-denumerable domains, and (ii) the irreducible normativity of rule-following.

The former category encompasses not only familiar cases like transfinite sets of real numbers, but also domains characterized by "irreducible complementarity, holism, partial overlapping, or ontological indeterminacy" (Hanna, 2025). This expansion suggests that uncomputability is not merely an exotic mathematical phenomenon, but a pervasive feature of complex systems, including those we encounter in physics, biology, and human experience.

4.1 The Rule-Following Paradox: A Challenge to Mechanism

Perhaps Hanna's most penetrating contribution lies in his analysis of the rule-following paradox and its connection to Turing's halting problem. By demonstrating that these are "essentially the same problem," he reveals a fundamental issue extending far beyond computer science's technical domain. The paradox illustrates

that partial functions necessarily underdetermine complete functions, and that no mechanical procedure can resolve this underdetermination.

This analysis provides a strong critique of purely computational approaches to understanding mind and rationality. If even the application of mathematical rules requires something beyond mechanical computation, namely, the creative judgment of rational agents embedded in linguistic communities, then the hope of reducing human cognition to digital processing is fundamentally misguided.

Hanna's invocation of Kant's concept of "mother-wit" (*Mutterwitz*) and the "natural power of judgment" is particularly astute. It connects the contemporary puzzle of rule-following to a deeper philosophical tradition that recognizes judgment as an irreducible capacity that cannot be fully algorithmized, providing both historical depth and systematic grounding for his position.

4.2 The Strategy of Cumulative Knockdown Arguments

Hanna's critique of strong AI represents a methodological innovation through what he terms a "cumulative knockdown" argument. Rather than relying on isolated objections that AI advocates have learned to deflect, Hanna constructs a systematic refutation operating across multiple analytical levels simultaneously. This approach recognizes that strong AI proponents have become adept at responding to individual challenges, invoking emergent properties when confronted with Searle's Chinese Room, dismissing consciousness as irrelevant when faced with phenomenal experience's hard problem.

The brilliance of Hanna's cumulative knockdown argument strategy lies in demonstrating that these piecemeal responses cannot address the integrated force of his complete case. His nine arguments against strong AI, ranging from classical objections like the Chinese Room, to novel contributions like the "readability argument," create philosophical pressure from multiple directions that would remain compelling even if individual components were weakened.

4.3 The Deep Consciousness Thesis: Closing AI's Escape Routes

Perhaps Hanna's most significant theoretical innovation is his "Deep Consciousness Thesis," the claim that all human mental activities, including pre-rational and unconscious processes, are inherently conscious. This move is both philosophically audacious and strategically brilliant, closing off all the traditional escape routes exploited by strong AI thesis advocates.

When confronted with consciousness-based objections, strong AI thesis proponents typically respond: "Perhaps AI cannot replicate human consciousness, but

it can perform all the same cognitive functions." Hanna's Deep Consciousness Thesis reveals this concession as hollow, by demonstrating that consciousness is not a mere epiphenomenal addition to cognition, but constitutive of all genuine mental activity.

4.4 The Psychocentric Challenge to Materialism

The Self-Refuting Nature of Materialist Consciousness Studies

Building on his critique of computational approaches, Hanna presents what he calls the "psychocentric predicament," the claim that any scientific study of consciousness must presuppose the very conscious rational capacities it seeks to explain. This creates an insurmountable problem for materialist approaches to mind, rendering them fundamentally self-refuting.

The argument's systematic structure demonstrates philosophical rigor. Beginning with careful definitions of consciousness, materialism, and mechanism, Hanna builds toward his conclusion through a logical progression that highlights the interdependence between rational cognition and scientific inquiry itself. His characterization of consciousness as involving both subjective experience ("consciousness-in") and intentional content ("consciousness-of") captures important phenomenological distinctions often overlooked in reductive approaches.

The most compelling aspect of this argument lies in Hanna's contention that materialism cannot account for the *a priori* and uncomputable aspects of logic, mathematics, and scientific reasoning. This connects his work on uncomputable functions to philosophy of mind in a novel way. If human rationality can perform genuinely uncomputable operations, as he argues through examples like Gödel's theorems and rule-following, then purely computational approaches to consciousness face principled limitations.

4.5 The Thesis of Creative Human Rationality

Hanna's central claim, that uncomputable functions require "an innately-specified rational human capacity" operating through "essentially creative acts of human rationality," is bold but well-supported. His characterization of this creativity as "essentially free or spontaneous, essentially organic or non-mechanical, and essentially a priori" (Hanna, 2025), clearly distinguishes it from both random processes and mechanical procedures.

The examples he provides illustrate different domains where this capacity manifests: from Gödel's incompleteness results and Cantor's diagonal argument, to quantum mechanical measurements and everyday text comprehension. This breadth

suggests that creative human rationality is not an exotic philosophical postulate, but a ubiquitous feature of human intellectual life.

This has profound implications for debates about AGI and machine consciousness. Rather than viewing AI as merely a technical challenge requiring more data and computational power, Hanna's analysis suggests principled reasons why digital systems cannot replicate the full range of human cognitive capacities.

4.6 Essential Embodiment and Neo-Organicist Metaphysics

Beyond the Materialist-Dualist Divide

Hanna's positive proposal, the "essential embodiment theory," offers an intriguing alternative to both materialist and dualist approaches. His neo-Aristotelian hylomorphism treats mind as the form of living biological systems, attempting to preserve both the irreducibility of mental properties and their intimate connection to physical processes.

This "new-school" neo-organicist approach builds on and transcends earlier critiques by Searle and Dreyfus, while avoiding both reductive materialism and substance dualism. His essential embodiment theory, developed in collaboration with Michelle Maiese, characterizes consciousness as "subjective experience" that is fundamentally "a form of life," establishing a naturalistic foundation for human mentality that resists reduction to mere physical mechanism while remaining scientifically respectable.

The emphasis on embodiment aligns with contemporary work in embodied cognition and enactive approaches to mind. By treating consciousness as a dynamic global structure of living organisms rather than a separate substance or emergent property, Hanna's theory navigates between reductive materialism and mysterious emergence.

This framework provides conceptual resources for addressing contemporary AI developments. When AI researchers point to emergent behaviors in scaled neural networks, Hanna can respond that emergence at the level of system behavior does not constitute emergence of genuine understanding or consciousness, a distinction his organicist metaphysics makes clear.

4.7 The Metaphysics of Human Dignity

Dignitarian Foundations in a Technological Age

Hanna's treatment of human dignity represents a sophisticated development of Kantian ethical theory adapted to contemporary technological challenges. By

grounding dignity in "rational human mindedness," understood as an integrated constellation of consciousness, self-consciousness, caring, cognition, volition, and free agency, Hanna provides a metaphysically robust foundation for human worth that transcends mere functional capabilities.

This approach avoids both anthropocentric speciesism and capability-based accounts that would extend moral consideration to sufficiently sophisticated AI systems. Hanna's organicist framework locates dignity not in abstract rational capabilities, but in the embodied, living reality of human persons as integrated wholes.

Technological Dehumanization and Its Mechanisms

Hanna's analysis of how digital technologies systematically undermine human capacities represents a crucial contribution to contemporary social criticism. Unlike simplistic accounts that blame technology for social ills, Hanna provides philosophically sophisticated analysis of how specific technological designs conflict with the structural requirements of human flourishing.

His discussion of social media algorithms, chatbots, and other digital systems as promoting addiction and manipulation, goes beyond empirical observation to identify the underlying philosophical problem: these systems treat humans as mere bundles of preferences and behaviors, rather than as integrated persons with dignity. The critique extends to emerging biotechnology, particularly organoid intelligence (OI), which Hanna labels "frankenscience" deserving complete prohibition. His argument that organoids, despite their biological basis, remain mechanistic rather than genuinely alive, provides criteria for distinguishing genuine life from mere biological activity.

Dignitarian Cybernetic Instrumentalism: A Constructive Vision

Hanna's refusal to remain purely critical culminates in his proposal for "Dignitarian Cybernetic Instrumentalism" (DCI), a constructive framework for technology development that serves human flourishing. DCI's emphasis on human dignity as the ultimate criterion for technological assessment provides decision-makers with clear guidance: technologies that enhance human dignity are morally permissible; those that systematically undermine it are morally impermissible.

4.8 Real Free Agency and the Phenomenology of Choice

The Connection Between Knowledge and Freedom

Hanna's innovative defense of incompatibilistic free will connects directly to his broader dignitarian project through the recognition that genuine knowledge requires

free agency. His claim that "knowledge, as sufficiently justified true belief, requires free agency in your choice of beliefs" (Hanna, 2024), points toward important epistemological implications often overlooked in free will debates.

The multifaceted definition of real free agency—requiring that agents be (1) conscious animals rather than machines, (2) possessed of self-consciousness, (3) capable of acting on desires without compulsion, and (4) confronted with genuine live options—captures our intuitive sense of what genuine agency requires while also providing philosophical precision.

The Ingenious "Pretending" Proof

Hanna's most creative contribution to the free will debate might be his novel proof through the phenomenology of pretending. This argument achieves philosophical elegance through simplicity: if you can pretend to be a machine, you cannot actually be one, since pretending requires being something other than what you pretend to be. The experiential immediacy of this demonstration gives the argument compelling power often missing from abstract metaphysical debates.

The analysis of pretending as "authentic human creativity" enabling "self-transcendence and self-transformation" connects free agency to broader questions about human meaning and value. This moves discussion beyond narrow questions of causal determination toward richer considerations of what makes human life distinctively valuable.

Methodological Innovation Through Lived Experience

Hanna's integration of lived experience into philosophical argument represents a welcome departure from purely abstract theorizing. By asking readers to engage in actual pretending, he creates a form of "experimental philosophy" that tests theoretical claims against immediate experience.

This methodological innovation has broader implications for philosophical practice. In an era when specialized arguments can be isolated and deflected through technical responses, Hanna demonstrates the continued relevance of systematic philosophical thinking that addresses fundamental assumptions rather than merely technical details.

4.9 Contemporary Applications and Critical Assessment

The Acceleration Problem and Policy Implications

Recent developments in AI capabilities have dramatically shortened timelines for addressing the philosophical questions Hanna raises. What previously seemed like

distant theoretical possibilities, AI systems matching or exceeding human performance across broad domains, now appear achievable within decades, or so it is promoted. This acceleration makes Hanna's philosophical work urgently practical rather than merely academic.

The connection between his metaphysical commitments and concrete policy recommendations demonstrates the practical relevance of philosophical analysis. His framework could dramatically shift contemporary AI policy discussions by providing a more fundamental criterion than current focuses on safety or fairness.

4.10 Critical Engagement with Contemporary Developments

When AI researchers point to emergent behaviors in scaled systems that seem to transcend training data, Hanna's organicist framework provides conceptual resources for response: these capabilities remain fundamentally mimetic rather than genuinely creative, lacking the embodied, intentional understanding characterizing human creativity, a conclusion we reached in the previous discussion.

His observation that "every artificial intelligence, no matter how sophisticated or tricked-up with fancy bells and whistles it might be, is nothing but a digital computer" (Hanna, 2025), something that can be confirmed by asking Grok, Claude, ChatGPT, Gemini or any others, cuts through much mystification surrounding machine learning. If his argument is sound, which we believe it is, then current AI systems, regardless of impressive performance, remain fundamentally limited by computational decidability boundaries.

4.11 Addressing Potential Objections

Critics might argue that Hanna's framework fails to account adequately for "technological momentum," the thesis that existing technological systems create pressures for continued development. However, by grounding resistance in fundamental philosophical principles rather than contingent preferences, DCI provides stable foundation for long-term opposition to dehumanizing technological trends.

Another potential criticism concerns the precision of key concepts. How exactly should we define "genuine consciousness" versus sophisticated simulation? What criteria distinguish dignity-enhancing from dignity-undermining technologies? While philosophical precision is important, critics demanding algorithmic-level specificity may miss the point of philosophical analysis. Hanna's concepts are necessarily open-textured because they deal with fundamental features of human experience resisting complete formalization; a merit, not a fault of his philosophical system.

We conclude that Hanna has provided a robust and wide-ranging critique of mechanism. In the rest of this essay, we will supply further supporting arguments to this core critical position against mechanism, with the next section devoted to the neurological critique of mechanistic theories of mind.

5. Neuroscientific Evidence Against Digital Computationism

Contemporary neuroscience increasingly reveals limitations in computational metaphors for understanding brain function. Research on neural plasticity, neurogenesis, and epigenetic factors demonstrates that brains differ fundamentally from digital computers in their developmental processes and adaptive capabilities (Kandel et al., 2013).

Unlike digital computers, which maintain stable hardware-software distinctions, biological brains exhibit continuous structural and functional plasticity that integrates "hardware" and "software" levels. Neural systems can modify their own connectivity patterns, generate new neurons, and alter gene expression in response to experience, capabilities that have no analogues in digital computation.

Research on neural oscillations, default mode networks, and global workspace dynamics reveals organizational principles that differ significantly from digital computational architectures (Varela et al., 1991). These findings entail that brains operate according to dynamical principles that transcend computational approaches, an approach adopted by Gerald Edelman, whose work we now consider.

5.1 Gerald Edelman's Neurological Critique of Computer Models of Mind

Gerald Edelman's *Bright Air, Brilliant Fire* (Edelman, 1992), presents a sustained neurobiological critique of computational theories of mind, challenging the foundational assumptions underlying artificial intelligence and cognitive science. By means of his theory of Neural Darwinism, Edelman argues that the brain operates according to *selectionist* rather than *instructionist* principles, fundamentally distinguishing biological cognition from computational processing. While Edelman's critique offers valuable insights into the limitations of computational metaphors, questions nevertheless remain regarding the completeness of his alternative framework and its implications for contemporary neuroscience and AI research.

Edelman's central thesis challenges what he terms the "computer metaphor" of mind, the assumption that mental processes can be adequately understood in terms of computational operations performed on symbolic representations. Instead, he proposes Neural Darwinism, a theory that emphasizes the selectionist, evolutionary character of neural organization and function. This framework, Edelman argues,

reveals fundamental differences between biological and artificial information processing that cannot be bridged through mere increases in computational power or sophistication.

The stakes of this debate extend beyond academic philosophy to encompass practical questions about artificial intelligence development, neurotechnology, and our understanding of human nature itself. If Edelman's critique is sound, then much contemporary work in AI and computational cognitive science is pursuing fundamentally misguided research programs.

Edelman's most fundamental contribution lies in his distinction between selectionist and instructionist theories of neural function. Instructionist theories, which underlie most computational models, assume that the brain processes information according to predetermined programs or algorithms. Like a computer executing software, the instructionist brain follows explicit rules to transform inputs into outputs through a series of determinate steps.

By contrast, Edelman's selectionist approach draws an analogy with evolutionary processes. Just as natural selection operates on populations of organisms with varying traits, neural selection operates on populations of neuronal groups with varying patterns of connectivity and response. Through experience, some neural circuits are strengthened while others are weakened, resulting in the emergence of adaptive patterns without the need for explicit programming or instruction.

This selectionist vs. instructionist distinction has profound implications for understanding the nature of mental representation and processing. In instructionist models, mental contents are typically understood as discrete symbols manipulated according to syntactic rules, the foundation of both classical AI and much contemporary cognitive science. Edelman's selectionist framework suggests that neural "representations" are better understood as dynamic patterns of activity that emerge from competitive processes among neural populations.

Edelman's Neural Group Selection (NGS) theory provides the mechanistic foundation for his selectionist approach. NGS operates through three key processes:

- Developmental Selection: During neural development, genetic and epigenetic factors create enormous diversity in synaptic connections. This primary repertoire provides the raw material for subsequent selection processes.

- Experiential Selection: Through interaction with the environment, some synaptic connections are strengthened while others are weakened. This secondary repertoire reflects the organism's particular experiential history.

- Re-entrant Signalling: Neural areas engage in continuous reciprocal signalling, creating dynamic patterns of correlated activity that integrate information across different brain regions.

The NGS framework holds that neural organization emerges through activity-dependent processes rather than following predetermined blueprints. This fundamental insight challenges computational models that assume fixed architectures implementing determinate algorithms. Instead, Edelman proposes that neural "computation" is better understood as a form of pattern recognition and selection operating on populations of neural responses.

Edelman's concept of *topobiology* provides additional neurobiological grounding for his critique of computational models. Topobiology describes how neural development proceeds through local interactions between cells and their molecular environment, resulting in the emergence of complex neural architectures without centralized control or explicit programming.

This developmental perspective highlights a crucial difference between biological and artificial systems. Computer architectures are designed according to explicit specifications that determine their structure and function. In contrast, neural architectures emerge through self-organizing processes that reflect both genetic constraints and environmental influences. The resulting neural organization cannot be fully captured by any finite set of rules or algorithms.

One of Edelman's most powerful arguments against computational theories concerns the problem of semantic content. Classical AI and cognitive science assume that mental representations derive their meaning from their role in computational processes, their "functional role" in transforming inputs to outputs. Edelman argues that this approach fails to account for how neural activity acquires genuine semantic content rather than merely formal structure.

The difficulty arises from what is called the "symbol grounding problem" (Harnad, 1990). In digital computers, symbols derive their meaning through interpretation by external agents (programmers and users). The symbol "cat" in a computer program means cat only because humans have established this interpretive relationship. But neural activity cannot depend on external interpretation, it must somehow ground its own semantic content through the physical processes of the brain itself.

Edelman's selectionist framework suggests a solution through the concept of "value." Neural selection processes are biased by evaluative systems that determine which patterns of activity are reinforced or suppressed. These value systems, rooted in the organism's biological needs and evolutionary history, provide the basis for

semantic content by establishing which neural patterns matter for the organism's survival and flourishing.

This account suggests that genuine semantic content requires the kind of embodied, value-laden interaction with the environment that characterizes biological systems. Computational models, operating through formal symbol manipulation without genuine biological values, may achieve sophisticated behavioral outputs without genuine understanding or semantic content.

Edelman's analysis of categorization provides another crucial line of argument against computational models. Categorization, the ability to group diverse stimuli into meaningful classes, is fundamental to all higher cognitive functions. Yet Edelman argues that computational approaches face a fundamental "bootstrap problem" in explaining how categorical structure emerges.

Classical computational models assume that categories are defined by necessary and sufficient conditions that can be explicitly programmed. However, psychological research has consistently shown that natural categories typically lack such definitional structure. Instead, they exhibit "family resemblance" structure with overlapping similarities rather than shared essential features.

Edelman argues that, even more problematically, computational models presuppose the very categorical distinctions they purport to explain. To implement a program that recognizes cats, programmers must already possess the concept "cat" to specify the relevant features and decision procedures. This creates a regress: how did humans originally acquire categorical concepts if not through computational processes?

Edelman's selectionist framework offers an alternative through what he terms "categorization through memory." Neural group selection creates neural circuits that respond selectively to recurring patterns in the organism's experience. Through re-entrant signalling, these specialized circuits interact to create higher-order patterns that capture similarities across different experiences. Categories emerge as stable patterns of neural activity rather than explicit symbol structures.

The binding problem—how the brain integrates diverse types of information into unified conscious experiences—poses another challenge for computational models. When you see a red ball, your visual system processes color, shape, motion, and other features in different neural areas. Yet you experience a single, integrated percept rather than a collection of separate features. How does the brain achieve this integration?

Computational approaches typically propose various architectural solutions, such as central processors that combine information from specialized modules or synchronization mechanisms that coordinate distributed processing. However, Edelman argues that such solutions fail to capture the dynamic, context-sensitive character of neural integration.

His theory of *re-entrant signalling* provides an alternative account. Rather than requiring a central integrator, re-entrant connections allow neural areas mutually to influence each other's activity patterns. Through recursive interactions, a globally consistent pattern of activity emerges that reflects the constraints imposed by all participating neural areas. This creates what Edelman terms "dynamic cores," integrated patterns of neural activity that underlie conscious experience.

The re-entrant framework suggests that conscious integration cannot be reduced to computational operations because it depends on the specific anatomical and physiological properties of neural circuits. The timing, connectivity patterns, and biophysical properties of neurons all contribute to the emergence of integrated conscious states in ways that cannot be captured by abstract computational descriptions.

Edelman's critique also addresses the temporal aspects of neural processing that computational models struggle to capture. Biological neural networks exhibit complex temporal dynamics, with patterns of activity evolving continuously rather than through discrete computational steps. This creates what AI researchers recognize as the "frame problem," the difficulty of determining which information remains relevant as situations change over time.

In computational systems, the frame problem is typically addressed through explicit rules that specify what information to maintain or update in different circumstances. However, such rules require programmers to anticipate all relevant situations in advance, an impossible task for open-ended environments.

Edelman's selectionist framework suggests that biological systems solve the frame problem through the continuous operation of neural selection processes. Rather than maintaining explicit representations of current states, the brain maintains dynamic patterns of activity that are continuously modified by ongoing experience. These patterns naturally adapt to changing circumstances without requiring explicit rules about what to preserve or update.

The temporal dimension of neural processing also relates to Edelman's concept of the "remembered present." Unlike computational systems that operate on discrete time steps, conscious experience involves the continuous integration of past, present, and anticipated future states. This temporal integration emerges through the recurrent

dynamics of neural circuits, rather than through the sequential processing of discrete computational operations.

Edelman's treatment of consciousness represents perhaps his most ambitious challenge to computational theories. While computational approaches typically focus on the functional aspects of mental processes, what the mind does rather than what it feels like, Edelman focuses on the qualitative dimensions of conscious experience that philosophers call "qualia."

As we've noted, the hard problem of consciousness concerns how and why physical processes give rise to subjective experience. Why should there be "something it is like" to see red or feel pain, rather than just neural processing that discriminates colors or responds to tissue damage? Computational theories typically sidestep this question by focusing on functional capacities while remaining agnostic about subjective experience.

Edelman argues that consciousness and qualia cannot be separated from the specific biological processes that generate them. His theory proposes that conscious experience emerges from the formation of dynamic cores, integrated patterns of re-entrant activity that bind diverse neural processes into unified states. The qualitative character of experience reflects the specific patterns of neural activity within these dynamic cores.

This biological grounding of consciousness creates difficulties for computational theories that assume substrate independence, the view that mental processes can be implemented in any sufficiently complex information processing system. If Edelman is correct, then consciousness requires the specific biological properties of neural tissue and cannot be replicated in silicon-based computational systems regardless of their functional sophistication.

Edelman's distinction between primary and higher-order consciousness provides additional support for his critique of computational models. Primary consciousness, which he attributes to many animals, involves the integration of sensory, memory, and value systems to create a unified "scene" of ongoing experience. This form of consciousness emerges through the basic operations of neural group selection and re-entrant signalling.

Higher-order consciousness, unique to humans and perhaps a few other species, involves the additional capacity for symbolic reference and linguistic communication. This enables humans to construct models of themselves and their environment that can be manipulated independently of immediate sensory input.

Crucially, Edelman argues that higher-order consciousness depends on the prior existence of primary consciousness rather than emerging purely from computational operations on symbolic representations. Language and symbolic thought are grounded in the qualitative experiences of primary consciousness, suggesting that genuine AI would require not just sophisticated symbol processing, but also the biological substrate that supports conscious experience.

The rise of deep learning and artificial neural networks might seem to vindicate computational approaches, by demonstrating that artificial systems can achieve sophisticated pattern recognition and learning. However, Edelman's critique entails important limitations that remain relevant to contemporary AI.

While artificial neural networks are loosely inspired by biological neurons, they typically lack the complex temporal dynamics, re-entrant connectivity, and value systems that Edelman identifies as crucial for genuine neural processing. Most deep learning systems operate through feedforward processing with discrete training phases, contrasting sharply with the continuous, recurrent dynamics of biological neural networks.

Moreover, deep learning systems typically require enormous amounts of labelled training data and exhibit brittleness when confronted with novel situations, limitations that may reflect their departure from the selectionist principles that Edelman argues are fundamental to biological intelligence. The need for extensive supervised training suggests that these systems lack the autonomous categorization abilities that emerge naturally in biological systems through neural group selection.

Edelman's emphasis on the biological grounding of cognition aligns with contemporary developments in embodied cognition and more specifcally with enactivist approaches to mind. These frameworks, developed by researchers like Francisco Varela, and Alva Noë (Noë, 2004), argue that cognition cannot be understood independently of the body and its environmental interactions.

The convergence between Edelman's selectionist framework and enactivist approaches suggests a broader shift away from computational metaphors toward more biologically grounded theories of mind. However, this convergence also raises questions about whether Edelman's specific theoretical commitments are necessary or whether alternative biological approaches might achieve similar insights.

Recent developments in predictive processing and Bayesian brain theories present interesting challenges to Edelman's framework. These approaches propose that the brain operates as a prediction machine that continuously generates and updates models of sensory input based on prior expectations and prediction errors, (Friston, 2010)

While predictive processing theories maintain computational commitments that Edelman would likely reject, they share his emphasis on the active, constructive character of neural processing. The brain doesn't simply respond to sensory input, but actively generates predictions that shape perception and action. This active dimension might be seen as compatible with Edelman's selectionist framework, although significant theoretical differences also remain.

While Edelman's critique of computational models is compelling, questions remain about whether his alternative framework successfully addresses the explanatory challenges it identifies. The hard problem of consciousness, for instance, might not be resolved merely by appealing to dynamic cores and re-entrant signalling. Critics might argue that Edelman simply relocates the mystery rather than solving it; why should these particular biological processes give rise to subjective experience rather than occurring "in the dark"?

Similarly, Edelman's account of semantic content through value systems may not fully address the symbol grounding problem. Even if biological values provide a basis for semantic content, it remains unclear how the transition from biological significance to genuine meaning occurs. The gap between biological function and semantic content may be as challenging as the gap between computation and meaning that Edelman identifies in artificial systems.

Edelman's theoretical framework makes specific empirical predictions about neural organization and function, but the experimental validation of these predictions remains incomplete. While there is substantial evidence for activity-dependent neural development and the importance of re-entrant connections, direct evidence for neural group selection and dynamic cores is more limited.

Contemporary neuroscience has developed sophisticated techniques for measuring neural activity with high spatial and temporal resolution, but translating these measurements into tests of Edelman's theoretical framework remains challenging. The complexity of neural systems and the indirect relationship between neural activity and theoretical constructs create difficulties for decisive experimental validation.

If Edelman's critique is correct, then what implications follow for artificial intelligence research? One interpretation holds that genuine AI requires biological substrates, and that silicon-based systems cannot achieve genuine intelligence regardless of their computational sophistication. This would represent a fundamental limitation on AI development that could be overcome only through biotechnology or hybrid bio-artificial systems.

Edelman's *Bright Air, Brilliant Fire*, presents a sophisticated and influential critique of computational theories of mind grounded in detailed knowledge of neural organization and function. His selectionist framework offers genuine insights into the distinctive features of biological cognition that might be missed by computational approaches focused on formal symbol manipulation.

The strength of Edelman's critique lies in its integration of empirical neuroscience with broader theoretical insights about the nature of mind and consciousness. By grounding his arguments in specific claims about neural development, organization, and function, Edelman avoids the purely philosophical objections that computational theorists can sometimes dismiss as irrelevant to their technical projects.

However, significant questions remain about both the completeness of Edelman's critique and the adequacy of his alternative framework. While he successfully identifies important limitations in computational approaches, it's less clear that his selectionist framework provides a complete account of the phenomena he seeks to explain. The hard problem of consciousness, the symbol grounding problem, and other fundamental issues in philosophy of mind, will require additional theoretical resources beyond those that Edelman provides.

For contemporary cognitive science and AI research, Edelman's work serves as a valuable corrective to overly simplistic computational assumptions while pointing toward more biologically realistic approaches to understanding intelligence and consciousness. Whether his specific theoretical commitments prove correct or not, his broader insights about the distinctive features of biological cognition will likely remain relevant for future developments in these fields. He has certainly generated a challenging neurological critique of mechanism.

5.2 John Lorber's Challenge to Mechanism

John Lorber's documented cases of severe hydrocephalus patients maintaining normal cognitive function despite massive brain tissue loss (Lewin, 1980; Perkins, 2025), present a fundamental challenge to computational theories of mind. Such extreme neuroplasticity demonstrates properties incompatible with classical Turing machine models, suggesting instead that consciousness emerges from dynamic, substrate-independent processes that resist mechanistic reduction.

The mechanistic view of mind assumes that specific cognitive functions require particular neural substrates, much as software requires specific hardware configurations. However, Lorber's 1980 documentation of hydrocephalic patients with minimal brain tissue, yet normal intelligence, poses a direct empirical challenge to this framework.

Lorber's most striking case involved a university student with an IQ of 126 whose brain scan revealed cerebral tissue compressed to just millimetres thick, roughly 5% of normal brain volume. Such cases weren't isolated anomalies but represented a pattern among severe hydrocephalus patients who developed normally despite dramatic structural deficits.

These findings are incompatible with standard mechanistic assumptions:

- Hardware Specificity Problem: Turing machines require specific hardware configurations to execute programs. Yet Lorber's patients achieved normal cognitive output with radically different "hardware," suggesting cognition isn't tied to particular neural architectures.

- Functional Localization Failure: Classical computational models assume cognitive functions map to discrete brain regions, like subroutines to memory addresses. But these patients maintained complex reasoning, language, and memory with minimal cortical tissue, implying functions aren't localized to specific substrates.

- Processing Power Paradox: Information processing should correlate with available computational resources. Yet patients with 95% brain tissue loss showed no proportional cognitive deficits, violating basic computational scaling principles.

The Neural Plasticity Argument Against Mechanism

The extreme neuroplasticity demonstrated in these cases reveals properties fundamentally at odds with Turing machine characteristics:

- Dynamic Reconfiguration: While Turing machines follow fixed programs, these brains continuously reorganized themselves. The same minimal tissue performed vastly different functions across development, suggesting cognition emerges from dynamic processes rather than static algorithms.

- Holographic Function: Unlike digital systems where data loss degrades performance proportionally, these patients maintained integrated cognitive function despite massive "data loss." This suggests a holographic rather than digital organization of mental processes.

- Substrate Independence: Most remarkably, normal cognition persisted across radically different physical substrates. This implies consciousness isn't reducible to particular material arrangements, challenging core materialist assumptions underlying computational theories.

Implications for Theories of Consciousness

If consciousness were purely computational, Lorber's cases would be impossible. A Turing machine with 95% of its components destroyed couldn't maintain complex processing. Yet these patients not only survived but thrived cognitively.

This suggests several anti-mechanistic conclusions:

- Emergence Over Reduction: Consciousness may be an emergent property that arises from but isn't reducible to neural activity. Like wetness emerging from H_2O molecules, consciousness might emerge from neural patterns while possessing irreducible properties.

- Field Models: Perhaps consciousness operates more like an electromagnetic field, distributed, dynamic, and capable of maintaining coherence despite local disruptions. This would explain how minimal neural tissue could sustain complex mental states.

- Non-Algorithmic Processing: Rather than executing discrete algorithms, the brain might operate through continuous, analog processes that resist digital modelling. Consciousness would be fundamentally non-computational, as Hanna and Penrose hold.

Addressing Counterarguments

- Distributed Processing Defense: Mechanists might argue that hydrocephalic brains demonstrate distributed rather than localised processing, still computational, but more flexible. However, this doesn't address why 95% tissue loss produces no proportional functional loss, or how the same tissue can serve multiple cognitive roles simultaneously.

- Compensation Mechanisms: Claims that subcortical structures compensate for cortical loss still assume mechanistic processing, just relocated. But this doesn't explain how patients maintain the full spectrum of cortical functions with minimal tissue, or why compensation is so remarkably complete.

- Developmental Adaptation: While early developmental plasticity might explain some adaptation, it doesn't address why the mature brain maintains such flexibility, or how fundamental cognitive architectures can be so radically reorganized while preserving function.

Lorber's hydrocephalus cases provide compelling evidence that consciousness operates through principles incompatible with Turing machine models. The

extraordinary plasticity, substrate independence, and holographic functionality observed in these patients point toward post-mechanistic theories of mind.

Instead of emerging from computational algorithms running on neural hardware, consciousness might emerge from dynamic, field-like processes that can maintain coherence across vastly different material substrates. This doesn't deny the importance of the brain, but suggests consciousness transcends simple mechanistic reduction. This conclusion will be strengthened by our consideration of the limits of contemporary theories of consciousness in the next section.

6. Leading Theories of Consciousness: A Critical Analysis of Their Limitations

The scientific study of consciousness has proliferated numerous theoretical frameworks attempting to explain how subjective experience arises from neural activity. While each theory offers valuable insights, we will argue that the current leading theories, Global Workspace Theory, Integrated Information Theory, Higher-Order Thought theories, Predictive Processing approaches, Attention Schema Theory, and Quantum Mechanical theories of mind, all suffer from significant problems that prevent them from providing comprehensive explanations of consciousness. These limitations include computational intractability, explanatory gaps, definitional ambiguities, empirical inadequacies, and philosophical problems that collectively suggest the field remains at best in its theoretical infancy; at worst, untenable.

6.1 Global Workspace Theory and Global Neuronal Workspace Theory

Theoretical Framework

Global Workspace Theory (GWT), originally proposed by Bernard Baars (1988), and its neurobiological implementation, Global Neuronal Workspace Theory (GNWT) developed by Dehaene and Changeux (2011), represent the most empirically supported theories of consciousness. Consciousness occurs when certain neural signals dominate, spreading across the brain, becoming system-wide available.

According to GNWT, consciousness arises through sensory information obtaining access to a network of interconnected cortical areas, resulting in global integration and broadcasting of information. This produces distinct neural signatures, such as late-onset cortical activity (>300ms), gamma-band synchronization, and long-distance cortical connectivity patterns.

Empirical Support

GNWT has received substantial empirical validation through studies of binocular rivalry, masking paradigms, and attentional blink experiments. Empirical successes include various neural correlates of consciousness, including the P3b component in EEG, late cortical activity in fMRI, and patterns of frontoparietal activation during conscious perception.

Fundamental Limitations

Despite its empirical success, GWT/GNWT faces several fundamental limitations:

- **The Broadcast Problem**: As argued by Ned Block (Block, 2019), the theory fails to explain why global broadcasting should necessarily give rise to subjective experience rather than merely sophisticated information integration. A zombie system, having no subjective experience at all, could theoretically exhibit all the neural signatures of global workspace activity without any accompanying phenomenal consciousness (Block, 2019).

- **Definitional Circularity**: GWT defines consciousness in terms of global accessibility, but this creates a definitional problem of circularity. GWT assumes that global access equals consciousness, without providing independent criteria for determining when information integration becomes subjective experience, thus assuming what it is required to demonstrate.

- **Ignition Threshold Issues**: The theory proposes that consciousness emerges when neural activity crosses an "ignition threshold." However, GWT provides no principled account of what determines this threshold, or why crossing it should produce subjective experience rather than merely stronger neural responses, failing once more to account for conscious experience.

- **Graded Consciousness Problem**: While GWT acknowledges that consciousness may be "graded," it does not adequately account for partial or ambiguous conscious states. The all-or-nothing "ignition process" conflicts with phenomenological evidence for gradual transitions between conscious and unconscious processing; this evidence exists and cannot be explained away; therefore, GWT is shipwrecked on the problem of the vagueness and fuzzy boundaries of conscious states.

6.2 Integrated Information Theory

Theoretical Framework

Integrated Information Theory (IIT), was developed by Giulio Tononi (1999, 2015). To paraphrase Tononi's own account for accurate characterization (Tononi, 1999), IIT proposes that consciousness corresponds to integrated information, with such information generated by a system being above and beyond its constituent parts. IIT quantifies consciousness using the measure Φ (phi), where a system with $\Phi > 0$ possesses a degree of consciousness relative to its phi value.

IIT rests on five axioms corresponding to fundamental properties of conscious experience: (1) information, (2) integration, (3) exclusion, (4) intrinsic existence, and (5) composition (Tononi and Koch, 2015). The theory then postulates that physical systems underlying consciousness must satisfy these same properties mathematically. IIT offers several theoretical advantages: it provides mathematical precision, addresses the hard problem directly by grounding consciousness in information-theoretic properties, and makes specific predictions about consciousness in various systems, including artificial agents and patients with disorders of consciousness.

Severe Limitations

IIT faces perhaps the most serious theoretical problems among theories of consciousness, as follows.

- **Computational Intractability**: Computing Φ for realistic, non-trivial systems becomes computationally impossible. This limits the theory's practical applicability and empirical testability (Doerig et al., 2021).

- **The Graining Problem**: Critics argue that one can slice a system into parts in infinite ways, and IIT gives no clear rule for which divide counts when working out Φ. Different grainings can therefore yield radically different Φ values for the same system (Doerig et al., 2019). This raises an issue of internal consistency and coherence.

- **Circularity Concerns**: IIT derives physical postulates from phenomenological axioms, then uses these postulates to explain phenomenology. This is logically circular: the theory assumes what it claims to explain. While Tononi argues this reflects IIT's foundational approach, critics contend it undermines the theory's explanatory power, and that Tononi in his response is begging the question.

- **Counterintuitive Implications**: IIT implies consciousness in simple systems (photodiodes, networks of XOR gates), while potentially denying it in complex

systems like feedforward neural networks. These implications conflict strongly with intuitive judgments about consciousness (Aaronson, 2014).

- **Pseudoscience Allegations**: Over 100 prominent consciousness researchers signed an open letter in 2023 characterizing IIT as "pseudoscience," arguing that its panpsychist commitments, consciousness in simple systems like photodiodes and lack of empirical testability, place it outside legitimate scientific inquiry (Doerig et al., 2021). The letter's signatories argued that:

- IIT makes untestable metaphysical commitments.

- The theory's panpsychist implications are scientifically problematic.

- The theory lacks sufficient empirical validation.

- IIT's claims extend beyond what can be scientifically verified.

The IIT pseudoscience controversy represents a significant moment in consciousness studies, exhibiting fundamental problems, as follows.

- **Scientific Standards**: What constitutes acceptable scientific theorizing in consciousness research?

- **Testability**: Whether mathematical theories of consciousness must be immediately empirically testable?

- **Metaphysical Commitments**: How much metaphysical speculation is acceptable in scientific theories?

- **Definitional Issues**: What counts as consciousness and who determines this?

- **The Unfolding Problem** (Doerig et al., 2021): The theory's mathematical formalism can be "unfolded" to show that seemingly integrated systems may have zero phi, raising questions about the theory's internal consistency.

While IIT faces legitimate technical and conceptual challenges, the pseudoscience label appears unjustified to many in the field, including us. The controversy ultimately reflects the immature state of consciousness science, where fundamental questions about methodology, definitions, and standards remain unresolved. As we see it, in this field metaphysical assumptions will be unavoidable. The real issue is whether these can be philosophically justified.

This debate serves as a case study in the sociology of science, demonstrating how theoretical disagreements can snowball into broader disputes about scientific legitimacy and professional authority in scientific fields, and about the role of philosophy, epistemology and metaphysics in these areas.

6.3 Higher-Order Thought Theories

Theoretical Framework

Higher-Order Thought (HOT) theories, developed primarily by David Rosenthal (2005), propose that consciousness arises when mental states become the objects of higher-order thoughts. Rosenthal holds that a mental state is a conscious one, if and only if it is accompanied by a suitable "higher-order thought" representing oneself as being in that state. Such thoughts are a type of "meta-thought" about thoughts. Relatedly, Higher-Order Perception (HOP) theories, developed by William Lycan (1996), offer a variant on HOT theories by proposing that consciousness requires higher-order perceptual monitoring rather than conceptual thoughts.

Theoretical Advantages

HOT theories provide a naturalistic account of consciousness that avoids mysterious non-physical properties. They explain the reflexive nature of consciousness and offer clear criteria for determining when mental states become conscious.

Critical Limitations

- **The Problem of Misrepresentatio**n: According to critics of HOT, you could end up being "aware" of experiences you don't really have, if your mind generates a mistaken higher-order thought. This leads to the counterintuitive conclusion that consciousness can exist without corresponding first-order mental states, a free-floating consciousness.

- **Infinite Regress Concerns**: If consciousness requires higher-order thoughts, what makes those higher-order thoughts conscious? The theory seems to generate an infinite regress or must arbitrarily stop at unconscious higher-order representations, and in either case, fails to adequately account for conscious experience.

- **Empirical Inadequacy**: Neuroimaging studies fail to consistently identify the kind of higher-order monitoring processes that HOT theories predict. The required metacognitive operations appear neither necessary nor sufficient for conscious experience.

- **Phenomenological Problems**: Careful introspection reveals that conscious experiences don't typically involve explicit higher-order thoughts about one's mental states. The theory appears to intellectualize consciousness in ways that conflict with immediate phenomenology.

- **The Problem of Cognitive Sophistication**: HOT theories seem to require conceptual sophistication that may be absent in animals, young children, or patients with certain brain lesions, leading to implausibly restrictive conclusions about the distribution of consciousness.

6.4 Predictive Processing Theories

Theoretical Framework

Predictive Processing (PP) theories, influenced by the work of Andy Clark (2013), Anil Seth (2014), and Jakob Hohwy (2013), propose that the brain is a prediction machine. Predictive processing views the brain as running layered simulations, constantly guessing incoming data, and revising itself when predictions fail.

Within this framework, consciousness corresponds to high-level predictions that achieve optimal precision-weighted prediction error minimization. Some variants propose that consciousness requires specific types of prediction error related to self-models or counterfactual processing.

Theoretical Strengths

PP theories integrate consciousness with broader frameworks in cognitive science and neuroscience. They provide mechanistic accounts of various conscious phenomena including perception, attention, and self-awareness, while offering potential explanations for altered states of consciousness and psychiatric conditions.

Fundamental Limitations

- **The Prediction-Experience Gap**: PP theories, according to critics, fail to explain why prediction error minimization should give rise to subjective experience rather than merely accurate behavioral responses. The explanatory gap between computational processes and phenomenal properties remains intact.

- **Specificity Problems**: All biological systems minimize some form of prediction error, yet consciousness appears only in certain complex systems. PP theories provide no principled criteria for determining which prediction-minimizing systems are conscious.

- **Circularity Issues**: Many PP accounts of consciousness invoke concepts like "higher-order predictions" or "meta-cognitive predictions" that appear to presuppose the very consciousness they aim to explain.

- **Empirical Underdetermination**: The PP framework is compatible with multiple mutually inconsistent theories of consciousness, making it difficult to derive specific testable predictions about conscious experience.

- **The Problem of Precision**: PP theories rely heavily on precision-weighting mechanisms, but provide no account of why precision-weighted predictions should be experienced rather than merely computed.

6.5 Attention Schema Theory

Theoretical Framework

Attention Schema Theory (AST), developed by Michael Graziano (2013), proposes that the brain builds a simplified internal model of where its attention is directed, and that this modelling is what we call consciousness. Thus, according to AST, the brain constructs simplified, accessible models of complex attentional mechanisms, and consciousness corresponds to the brain's access to these attention schemas.

The theory suggests that consciousness serves a social function, allowing organisms to model not only their own attention but also the attention of others, facilitating complex social interactions and communication.

Theoretical Advantages

AST offers a deflationary approach to consciousness that avoids mysterious properties while explaining why consciousness seems special from the first-person perspective. The theory connects consciousness to well-understood mechanisms of attention and provides an evolutionary account of the adaptive function of consciousness.

Significant Limitations

- **The Schema-Experience Problem**: Even if the brain constructs schemas of attention, this doesn't explain why having access to these schemas should involve subjective experience rather than mere information processing.

- **Reductionist Oversimplification**: AST essentially claims that consciousness is an illusion, being a mere schema. This conflicts with the intuitive certainty that consciousness involves genuine qualitative experience, not merely information about experience.

- **Consciousness and Attention Dissociations**: Empirical evidence suggests that attention and consciousness can dissociate, with attention possible without consciousness and vice versa. This undermines AST's identification of consciousness with attention modelling.

- **The Problem of Qualia**: AST provides no account of why attention schemas should involve specific qualitative properties like the redness of red or the painfulness of pain, rather than mere abstract informational content.

- **Social Function Limitations**: While consciousness may serve social functions, this doesn't explain why these functions require subjective experience rather than sophisticated information processing and behavioral responses.

6.6 The Hard Problem Remains Unsolved

Despite their diversity, all major theories of consciousness still leave us with major problems. They do not show how brain processes turn into experience; they do not agree on what consciousness even is, and their tests are highly uncertain. Even the more mathematical approaches become unworkable beyond simple examples.

Definitional Confusion

All theories of consciousness encounter fundamental disagreements about what consciousness is and what any adequate theory should explain. Some theories focus on access consciousness, others on phenomenal consciousness, and still others conflate or dismiss this distinction. This definitional confusion makes theoretical evaluation and comparison extremely difficult.

Empirical Inadequacy

All theories of consciousness rely heavily on neural correlates of consciousness (NCCs), but do not establish causal relationships between proposed mechanisms and conscious experience. The correlation-causation problem remains largely unresolved, and many theories make predictions that are difficult or impossible to test empirically.

The Problem of Other Minds

All theories of consciousness face fundamental challenges in determining consciousness in systems other than the theorist's own mind. Whether assessing consciousness in animals, patients with brain damage, or artificial systems, theories provide conflicting predictions and lack independent verification methods.

Computational and Mathematical Limitations

Theories of consciousness attempting mathematical precision (particularly IIT), encounter computational intractability, while more qualitative theories lack the specificity needed for rigorous testing. This creates a dilemma between precision and practicability.

As the science of consciousness matures, its research agenda has expanded beyond neural correlates to focus on theoretical development, but this expansion has revealed the inadequacy of current approaches rather than resolving fundamental questions. Contemporary theories of consciousness, despite their individual insights and empirical contributions, all face severe limitations that prevent them from providing comprehensive explanations of conscious experience. These limitations include:

- **Explanatory gaps** between proposed mechanisms and subjective experience.
- **Definitional ambiguities** about the nature and boundaries of consciousness.
- **Empirical inadequacies** in testing and verification procedures.
- **Computational intractabilities** in formal approaches.
- **Philosophical problems** regarding the mind-body relationship.

Therefore, the current state of consciousness science suggests *epistemic humility* rather than confidence in theoretical understanding. While each theory contributes valuable insights, none approaches a complete solution to the mystery of conscious experience. This seeming impasse has led some theorists to turn to quantum mechanics in order to formulate alternative theories of consciousness. In the next section, we will discuss the leading quantum mechanical theories of consciousness, and show their limits, but also how these positions, while also flawed, cut the epistemic ground from under physicalist reductionism, and mechanism.

7. Quantum Mechanics and Consciousness

Quantum mechanical approaches to consciousness emerged partly as responses to the perceived inadequacies of classical physicalist accounts. The deterministic, mechanistic worldview of classical physics appears to leave little room for the apparent unity, intentionality, and qualitative richness of conscious experience (Penrose, 1989). Quantum mechanics, with its inherent indeterminacy, observer

effects, and non-local correlations, has thus attracted theorists seeking physical mechanisms that might account for consciousness's distinctive features.

7.1 The Penrose-Hameroff Orchestrated Objective Reduction Theory

The most developed quantum theory of consciousness is Orchestrated Objective Reduction (Orch-OR), proposed by mathematical physicist Roger Penrose and anesthesiologist Stuart Hameroff (Penrose, 1994; Hameroff and Penrose, 1996; Penrose and Hameroff, 2011). Orch-OR aims to solve three fundamental problems: (i) the unity of consciousness, (ii) the binding problem, and (iii) the existence of subjective experience.

Penrose discusses "objective reduction" (OR), the idea that quantum state reduction occurs through both measurement and gravitational effects. Here, quantum superpositions at the Planck scale, reach a critical threshold, producing the reduction (Penrose, 1996). This process, Penrose argues, is non-algorithmic and could account for the apparent non-computability of mathematical insight and conscious understanding, he supposes (Penrose, 1989, 1994).

Hameroff locates the quantum processes within microtubules. Microtubules are protein structures forming the cytoskeleton of neurons. He proposes that quantum coherence can be maintained within these structures at biological temperatures for sufficient durations to influence neural computation (Hameroff, 1998a, 1998b). The theory suggests that consciousness arises when quantum states in microtubules undergo objective collapse in unison, leading to conscious experience at roughly 40Hz. This is in step with observed neural pulses, and conscious awareness (Penrose and Hameroff, 2011).

7.2 Alternative Quantum Theories of Consciousness

Several other quantum approaches have emerged, each addressing different aspects of the consciousness problem:

- Quantum Information Theory Approaches: Theorists like Henry Stapp have proposed that consciousness arises from quantum measurement processes in the brain. Wavefunction collapse is caused by conscious observation (Stapp, 1993, 2007). This approach attempts to resolve the measurement problem in quantum mechanics by invoking consciousness as the agent of wavefunction collapse.

- Quantum Field Theories: Giuseppe Vitiello and others have developed quantum field theoretical approaches to consciousness. These involve the idea that consciousness is a product of the quantum field fluctuations in the brain's

electromagnetic field (Vitiello, 1995; Penrose and Hameroff, 2011). These theories attempt to account for the apparent non-locality of conscious experience and memory.

- Many-Minds Interpretations: Some theorists have proposed that quantum many-worlds interpretations might explain consciousness. It is conjectured that each branch of the universal wavefunction somehow corresponds to a different aspect of consciousness (Albert and Loewer, 1988; Barrett, 1999).

7.3 The Measurement Problem and Consciousness

A central issue concerns the relationship between quantum measurement and conscious observation. The traditional Copenhagen interpretation proposes that measurement causes wavefunction collapse. However, this raises the question of what constitutes a "measurement"? Some theorists argue that consciousness is necessary for measurement (von Neumann, 1955; Wigner, 1961). Others propose that any physical interaction, regardless of conscious observation, can be a measurement (Zurek, 2003).

Decoherence and Biological Environments

A major challenge for quantum theories of consciousness involves decoherence, the rapid loss of quantum coherence due to environmental interaction (Zurek, 1991, 2003). The biological environment of the brain would act to eliminate quantum coherence in femtoseconds. This is much too quick, some propose, to impact upon neural computation (Tegmark, 2000). Proponents of quantum consciousness theories must therefore propose mechanisms for maintaining coherence or demonstrate that relevant quantum effects can persist despite decoherence.

The Binding Problem and Quantum Holism

Quantum mechanical non-locality and entanglement have been proposed as solutions to the binding problem, explaining how spatially distributed neural processes give rise to unified conscious experiences (Stapp, 1993; Penrose and Hameroff, 2011). Quantum entanglement could theoretically allow instantaneous correlations between distant brain regions, potentially explaining the apparent unity of consciousness despite its distributed neural substrate.

Neurobiological Evidence

The empirical support for quantum effects in neural processes remains highly contested. While some studies have suggested quantum effects in biological systems, such as quantum coherence in photosynthesis (Engel et al., 2007) and avian navigation

(Ritz et al., 2000), the evidence for quantum processes in neural microtubules specifically remains limited (McKemmish et al., 2009).

Quantum effects in biological environments have been observed to be operating (Lambert et al., 2013; O'Reilly and Olaya-Castro, 2014), so quantum coherence may occur in biological systems. However, the logical jump from such quantum effects to quantum computation producing consciousness in microtubules, has not been demonstrated.

Anesthetic Studies

Hameroff and others have pointed to correlations between anesthetic effects on microtubules and consciousness as evidence for Orch-OR theory (Hameroff, 1998b). Some anesthetics that cause unconsciousness also disrupt microtubule function, potentially supporting the theory's predictions. However, critics argue that these correlations could be explained by classical mechanisms without invoking quantum processes (Georgiev, 2020).

Computational Criticisms

A fundamental criticism concerns whether quantum mechanics actually provides the computational advantages claimed by quantum consciousness theories. Critics argue that even if quantum processes occur in the brain, they may not contribute to consciousness in the ways proposed (Grush and Churchland, 1995; Litt et al., 2006). The specific computational advantages of quantum processing for generating subjective experience remain unclear.

7.4 The Hard Problem and Explanatory Adequacy

A crucial question concerns whether quantum mechanical approaches actually address the hard problem of consciousness or merely relocate it. Critics argue that invoking quantum indeterminacy or non-locality does not explain why there should be subjective experience associated with these processes any more than with classical physical processes (Chalmers, 1996). The qualitative, experiential aspects of consciousness may remain explanatorily problematic regardless of the underlying physical mechanisms.

Free Will and Quantum Indeterminacy

Some theorists have suggested that quantum indeterminacy might provide a physical basis for free will, offering an alternative to strict determinism without invoking pure randomness (Kane, 1996; Penrose, 1994). However, critics argue that random quantum

events provide no better foundation for genuine agency than deterministic processes (Dennett, 1984; McKenna and Pereboom, 2016).

Emergence and Downward Causation

Quantum theories of consciousness often invoke emergence, the idea that consciousness emerges from, but is not reducible to, quantum processes. This raises complex questions about downward causation and whether emergent conscious states can influence lower-level physical processes (Kim, 1999; Clayton and Davies, 2006).

The Circularity at the Heart of Quantum Consciousness

A more fundamental problem has received insufficient scrutiny: the observer paradox inherent in quantum mechanical formalism itself. If quantum mechanics requires conscious observers to define measurement events, and consciousness allegedly emerges from quantum processes, then any quantum theory of consciousness faces a vicious explanatory circle that undermines the entire reductionist project.

This observer paradox strikes at the foundational assumptions of physicalist approaches to consciousness. Rather than consciousness emerging from complex quantum information processing, the formalism of quantum mechanics appears to presuppose consciousness as a necessary component for coherent physical description. This reversal, from consciousness as emergent to consciousness as foundational, has profound implications for our understanding of the mind-body relationship and suggests that reductive physicalism and mechanism may be fundamentally misconceived from the get-go.

The Copenhagen interpretation of quantum mechanics, developed by Niels Bohr and Werner Heisenberg, yielded the observer paradox. Heisenberg's uncertainty principle demonstrated that measurement itself necessarily alters quantum systems. Bohr's complementarity principle asserts that it is impossible to separate observed objects from the observing subjects (Heisenberg, 1958; Bohr, 1963). These insights led to a formalism where quantum mechanical descriptions are inherently observer-relative, rather than objectively complete.

The mathematical formalization by John von Neumann made this observer dependence explicit. In his *Mathematical Foundations of Quantum Mechanics* (1955), von Neumann demonstrated that the measurement process involves an irreducible division between the observed quantum system and the measuring apparatus. This cut can be moved arbitrarily along the measurement chain, but it cannot be erased without involving conscious observation (von Neumann, 1955: pp. 418-421).

Eugene Wigner extended this analysis in his seminal paper "Remarks on the Mind-Body Question" (1961), arguing that conscious observation is necessary for wavefunction collapse. Wigner's argument rested on the premise that only consciousness possesses the unique property of immediate awareness that can break the chain of quantum superpositions. His famous "Wigner's Friend" thought experiment demonstrated that without conscious observation, even a macroscopic measurement apparatus remains in superposition states (Wigner, 1961, pp. 284-302).

Contemporary attempts to eliminate observer dependence face what we might term the "specification problem," the difficulty of defining measurement events without implicitly invoking observational criteria. Decoherence theory, developed by Wojciech Zurek and others, attempts to explain apparent wavefunction collapse through environmental interaction rather than conscious observation (Zurek, 2003). However, decoherence theory must still specify what constitutes a "relevant environment" and which interactions count as "measurements," specifications that ultimately depend on observer interests and perspectives.

Similarly, many-worlds interpretations avoid wavefunction collapse by positing universal superposition, but they face their own specification problems in defining "branches" of the universal wavefunction (Everett, 1957; DeWitt and Graham, 1973). What constitutes a distinct branch, and from whose perspective do branches appear separate? These questions cannot be answered without reference to observational criteria, reintroducing observer dependence at a fundamental level.

Even objective collapse theories like the Ghirardi-Rimini-Weber (GRW) model, which propose spontaneous wavefunction collapse independent of observation, must specify which physical quantities trigger collapse and at what threshold (Ghirardi et al., 1986). These specifications appear arbitrary without reference to observational relevance, suggesting that observer perspectives remain implicit even in "objective" formulations.

Quantum theories of consciousness face a fundamental bootstrap paradox. While attempting to explain consciousness through quantum mechanical processes, the quantum mechanics itself require conscious observers, according to the position. This creates a "vicious circle" in explanation: consciousness cannot be both explanans (that which explains) and explanandum (that which is explained) without logical incoherence.

The Penrose-Hameroff Orchestrated Objective Reduction (Orch-OR) theory exemplifies this circularity (Penrose, 1994; Hameroff and Penrose, 1996). As we pointed out above, the theory proposes that consciousness emerges from quantum computations in neural microtubules that undergo objective reduction when gravitational effects reach critical thresholds. However, the theory must invoke

conscious "moments of experience" to define what constitutes a relevant quantum computation and when objective reduction creates unified conscious states. The theory thus presupposes consciousness to explain consciousness, a classic case of explanatory circularity.

Henry Stapp's quantum interactive dualism faces similar problems (Stapp, 1993, 2007). Stapp argues that consciousness causes wavefunction collapse through quantum measurement, but his theory requires pre-existing conscious agents to make the measurements that allegedly generate consciousness. This temporal paradox, whereby consciousness causes its own emergence, violates basic principles of causal explanation.

The bootstrap paradox generates an infinite regress problem: if consciousness emerges from quantum processes that require conscious observation to be defined, then who observes the quantum processes that generate the first conscious observer? This "meta-observer regress" parallels classical problems in epistemology about the criteria for knowledge, but with ontological rather than merely epistemic implications.

Consider a specific example: if consciousness emerges from quantum coherence in neural microtubules, as Orch-OR theory suggests, then measuring this quantum coherence requires conscious observers to define what counts as coherence, how long coherence must persist, and when decoherence terminates the conscious moment. But these conscious observers must themselves emerge from prior quantum processes, which require their own conscious observers ad infinitum.

Some theorists attempt to avoid this regress by appealing to "emergent observer sufficiency," the idea that once consciousness emerges from quantum processes, it can retroactively validate the quantum framework that produced it (Stapp, 2007). However, this temporal circularity violates basic causal principles: effects cannot precede their causes, and conscious validation cannot occur before consciousness exists.

A deeper analysis reveals that the observer paradox connects to fundamental issues about intentionality, the "aboutness" or representational character of mental states (Brentano, 1874/1995; Searle, 1983). Physical processes, as described purely in terms of quantum field interactions, particle exchanges, or electromagnetic patterns, possess no intrinsic intentionality. They do not inherently refer to or represent anything beyond their immediate causal relations.

However as Husserl noted, conscious observation is intrinsically intentional: consciousness is always consciousness *of* something, under some description or another, serving some aim or another (Husserl, 1913/1931; Merleau-Ponty, 1945/2012). When a physicist observes quantum interference patterns, the observation possesses

intentional content: it refers to specific experimental arrangements, represents particular theoretical predictions, and serves explanatory purposes within scientific frameworks. This intentional structure cannot be reduced to the purely causal interactions described by quantum mechanics without eliminating precisely what makes observation observational rather than merely causal.

The intentionality problem shows that consciousness cannot be fully reduced to quantum mechanical processes, as because intentionality is necessary for quantum mechanics itself. Quantum mechanics describes correlations between measurement outcomes, but these correlations become physically meaningful only through conscious interpretations by observing physicists, who assign referential content to the mathematical formalism.

This analysis closely relates to the symbol grounding problem in cognitive science: how do symbolic representations acquire their semantic content (Harnad, 1990)? In quantum mechanics, mathematical symbols (ψ, $|0\rangle$, $|1\rangle$, etc.) must be grounded in observational procedures and measurement outcomes in order to acquire physical meaning. But observational procedures presuppose conscious agents capable of recognizing, categorizing, and interpreting measurement results.

Consider the double-slit experiment, often cited as fundamental to quantum mechanics. The experiment requires conscious observers in order to distinguish between "particle-like" and "wave-like" behavior, to recognize interference patterns as meaningful rather than random, and to connect mathematical predictions with observable outcomes. Without an observer, the double-slit setup would merely produce raw detector interactions; the talk of "wave-particle duality" or "measurement problems," only arise once conscious agents interpret the results.

This entails that quantum mechanics, rather than explaining consciousness, may actually depend on consciousness for its coherent formulation as a physical theory. The mathematical formalism acquires physical meaning only through conscious interpretation that assigns referential content to abstract symbols.

7.5 Responses to the Observer Paradox

Physicalists have developed several strategies to eliminate or minimize observer dependence in quantum mechanics, but each faces serious limitations that preserve the fundamental paradox.

- Decoherence Theory: Wavefunction collapse is thought to occur through interaction with macroscopic "relevant environments" rather than conscious observation (Zurek, 1991, 2003). However, decoherence theory faces the specification problem: what constitutes a "relevant environment" and which

interactions count as "measurements"? These specifications cannot be made without reference to observer interests and perspectives, reintroducing consciousness at the foundational level.

- Moreover, decoherence explains the *appearance* of collapse from an observer's perspective but does not eliminate quantum superposition itself. The global quantum state remains in superposition; decoherence merely explains why local observers perceive definite outcomes. This preservation of observer-relativity undermines attempts to eliminate consciousness from quantum mechanical description (Schlosshauer, 2007).

- Many-Worlds Interpretations: Many-worlds theories avoid wavefunction collapse by positing that all possible measurement outcomes occur in parallel branches of universal superposition (Everett, 1957; DeWitt and Graham, 1973). This eliminates special roles for consciousness in causing collapse, but it faces severe specification problems in defining branches and explaining the appearance of definite outcomes.

- What constitutes a distinct "branch" of the universal wavefunction? How do conscious observers experience single outcomes rather than quantum superpositions of all possible experiences? These questions cannot be answered without invoking observer perspectives to define branching criteria and explain subjective experience within many-worlds frameworks (Barrett, 1999; Wallace, 2012).

- QBism and Subjective Interpretations: Quantum Bayesianism (QBism) accepts observer dependence. This involves regarding quantum states as subjective Bayesian degrees of belief, rather than objective physical properties (Fuchs, 2010; Fuchs and Schack, 2013). The approach eliminates observer paradoxes by making quantum mechanics explicitly subjective, but it undermines physicalist reductionism and mechanism by making physics depend on conscious belief states rather than objective physical processes.

- If quantum mechanics describes subjective beliefs rather than objective reality, then consciousness cannot be reduced to quantum mechanical processes without circularity. QBism preserves quantum mechanics by abandoning physicalist reductionism and mechanism, solving the observer paradox by conceding the fundamental point at issue.

- Objective Collapse Theories: Objective collapse theories attempt to eliminate observer dependence by proposing spontaneous wavefunction collapse independent of conscious observation. The Ghirardi-Rimini-Weber (GRW) model suggests that quantum superpositions spontaneously collapse at

random intervals, with collapse probability increasing with system size (Ghirardi et al., 1986). Continuous spontaneous localization (CSL) theories propose similar mechanisms with different mathematical details (Pearle, 1989; Bassi and Ghirardi, 2003).

However, objective collapse theories face their own specification problems. Why do certain physical quantities (position, momentum, energy) trigger collapse while others (spin, phase) do not? What determines collapse thresholds and time scales? These specifications appear arbitrary unless grounded in observational relevance, reintroducing observer dependence through the back door.

Moreover, objective collapse theories must explain why consciousness correlates so precisely with collapse events. If collapse occurs randomly and independently of consciousness, why does conscious awareness track collapse outcomes so reliably? The correlation between consciousness and collapse suggests either that consciousness influences collapse (violating the theory's objectivity) or that consciousness emerges from collapse events (requiring explanation of how subjective experience arises from objective physical processes).

The observer paradox demonstrates a fundamental failure in bottom-up explanatory strategies that attempt to explain consciousness in terms of more basic physical processes. If the most fundamental physical theory, quantum mechanics, requires conscious observers for coherent formulation, then consciousness cannot be reduced to quantum mechanical processes without explanatory circularity.

This failure is not merely technical but conceptual. The attempt to explain consciousness through quantum mechanics commits what philosophers such as Ryle (Ryle, 1949) call a "category error," treating consciousness as a complex arrangement of non-conscious components, while simultaneously relying on consciousness to define those components coherently (Ryle, 1949; Bennett and Hacker, 2003).

The bootstrap paradox reveals that consciousness and physical processes may be more intimately connected than physicalist reductionism and mechanism assume. Rather than consciousness emerging from complex arrangements of unconscious matter, consciousness is necessary for matter to be conceived and described as such, creating a fundamental ontological circularity. Hence, from these considerations alone, reductionist physicalism, computational theories of mind, and mechanism, are false.

8. Conclusion

We conclude that the complete package of objections to mechanism considered collectively—thereby deploying what Hanna calls *the cumulative argument strategy*, or what we've called "the shotgun approach"—strongly indicate that mechanism about the rational human mind is untenable. We have examined the major planks of the mechanistic position and each one has strong opponents. This means that the most plausible position for examining the "hard problem of consciousness" is a non-reductionist account such as Hanna's. But we will not explore this further in this essay. In any case, we have no problem with accepting that the hard problem is unsolvable, like many other philosophical problems, where the subjective is supposed to be reduced to the objective, or the objective to the subjective (Dietrich and Hardcastle, 2005). But there is, once one rejects materialist/physicalist and reductionist prejudices, nothing especially "mysterious" about consciousness. It is something which we are introspectively directly aware of, unlike say Platonistic numbers and transfinite sets. And, non-reductive psychology, in various forms such as phenomenology and humanistic psychology, can fruitfully explore the human world, once one departs from mechanistic prejudices, fetishes, and even psychopathologies (Hanna, 2025).[1]

[1]The authors are grateful to Robert Hanna for his inspiring body of philosophical work, especially including his detailed and robust critique of mechanism.

REFERENCES

(AAAI, 2025). *Association for the Advancement of Artificial Intelligence.* "AAAI 2025 Presidential Panel on the Future of AI Research." Available online at URL = <https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-Digital-3.7.25.pdf>.

(Aaronson, 2014). Aaronson, S. "Why I Am Not an Integrated Information Theorist (or, the Unconscious Expander)." *Shtetl-Optimized Blog*. 12 May. Available online at URL = <https://scottaaronson.blog/?p=1799>.

(Albert and Loewer, 1988). Albert, D.Z. and Loewer, B. "Interpreting the Many Worlds Interpretation." *Synthese* 77, 2: 195–213.

(Baars, 1988). Baars, B.J. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge Univ. Press.

(Bak, 1996). Bak, P. *How Nature Works: The Science of Self-Organized Criticality*. Göttingen DE: Copernicus Books.

(Barrett, 1999). Barrett, J.A. *The Quantum Mechanics of Minds and Worlds*. Oxford: Oxford Univ. Press.

(Bassi and Ghirardi, 2003). Bassi, A. and Ghirardi, G. C. (2003). "Dynamical Reduction Models." *Physics Reports* 379, 5–6: 257–426.

(Bennett and Hacker, 2003). Bennett, M. R., and Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience*. Oxford: Blackwell Publishing.

(Block, 2019). Block, N. (2019). "What is Wrong with the No-Report Paradigm and How to Fix It." *Trends in Cognitive Sciences* 23, 12: 1003–1013.

(Bohr, 1963). Bohr, N. (1963). *Essays 1958–1962 on Atomic Physics and Human Knowledge*. Geneva CH: Interscience Publishers.

(Brentano, 1874/1995). Brentano, F. *Psychology from an Empirical Standpoint*. London: Routledge.

(Bostrom, 2014). Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford Univ. Press.

(Busemeyer and Bruza, 2012). Busemeyer, J.R. and Bruza, P. D. *Quantum Models of Cognition and Decision*. Cambridge: Cambridge Univ. Press.

(Cao et al., 2020). Cao, J., et al. "Quantum Biology Revisited." *Science Advances* 6, 14. Available online at URL = <https://doi.org/10.1126/sciadv.aaz4888>

(Chalmers, 1995). Chalmers, D.J. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2, 3: 200–219.

(Chalmers, 1996). Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford Univ. Press.

(Chalmers, 2018). Chalmers, D.J. "The Meta-Problem of Consciousness." *Journal of Consciousness Studies* 25, 9–10: 6–61.

(Chemero, 2009). Chemero, A. *Radical Embodied Cognitive Science*. Cambridge MA: MIT Press.

(Chomsky et al., 2023). Chomsky, N., Roberts, I. and Watumull, J. "Noam Chomsky: The False Promise of ChatGPT." *New York Times*. 8 March. Available online at URL = <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>.

(Clark, 1997). Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge MA: MIT Press.

(Clark, 2008). Clark, A. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford Univ. Press.

(Clark, 2013). Clark, A. "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *Behavioral and Brain Sciences* 36, 3: 181–204.

(Clayton and Davies, 2006). Clayton, P., and Davies, P. (eds.). *The Re-Emergence of Emergence*. Oxford: Oxford Univ. Press.

(Coeckelbergh, 2020). Coeckelbergh, M. *AI Ethics*. Cambridge MA: MIT Press.

(Cotra, 2020). Cotra, A. "Forecasting Transformative AI Timelines Using Biological Anchors." *AI Impacts*. Available online at URL = <https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines>.

(Dehaene and Changeux, 2011). Dehaene, S. and Changeux, J.P. "Experimental and Theoretical Approaches to Conscious Processing." *Neuron* 70, 2: 200–227.

(Dehaene, 2014). Dehaene, S. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York: Viking.

(Dennett, 1984). Dennett, D.C. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge MA: MIT Press.

(DeWitt and Graham, 1973). DeWitt, B.S. and Graham, N. (eds.). *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton NJ: Princeton Univ. Press.

(Dietrich and Hardcastle, 2005). Dietrich, E, and Hardcastle, V.G. *Sisyphus' Boulder: Consciousness and the Limits of Knowledge*. Amsterdam NL: John Benjamins Publishing Company.

(Doerig et al., 2019). Doerig, A., et al. "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness." *Consciousness and Cognition,* 72: 49–59.

(Doerig et al., 2021). Doerig, A., et al. (2021). "Hard Criteria for Empirical Theories of Consciousness." *Cognitive Neuroscience.* 12(2): 41-62.

(Dreyfus, 1972). Dreyfus, H. *What Computers Can't Do*. Cambridge MA: MIT Press.

(Dreyfus, 1992). Dreyfus, H. *What Computers Still Can't Do.* Cambridge MA: MIT Press.

(Edelman, 1987). Edelman, G.M. *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books.

(Edelman, 1989). Edelman, G.M. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

(Edelman, 1992). Edelman, G.M. *Bright Air, Brilliant Fire: On the Matter of the Mind*. New York: Basic Books.

(Edelman and Tononi, 2000). Edelman, G.M. and Tononi, G. *A Universe of Consciousness: How Matter Becomes Imagination*. New York: Basic Books.

(Einstein, 1954). Einstein, A. "Aphorisms for Leo Baeck." In A. Einstein, *Ideas and Opinions*. New York: Three Rivers Press. Pp. 27-28.

(Engel et al., 2007). Engel, G.S., et al. "Evidence for Wavelike Energy Transfer through Quantum Coherence in Photosynthetic Systems." *Nature* 446, 7137: 782–786.

(Everett, 1957). Everett III, H. (1957). "'Relative State' Formulation of Quantum Mechanics." *Reviews of Modern Physics* 29, 3: 454–462.

(Floridi, 2016). Floridi, L. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: Oxford Univ. Press.

(Floridi, 2019). Floridi, L. *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford: Oxford Univ. Press.

(Freeman, 1999). Freeman, W.J. (1999). *How Brains Make Up Their Minds*. London: Weidenfeld and Nicolson.

(Freeman, 2001). Freeman, W.J. *Neurodynamics: An Exploration in Mesoscopic Brain Dynamics*. New York: Springer.

(Frey and Osborne, 2017). Frey, C.B. and Osborne, M. A. (2017). "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114: 254–280.

(Friston, 2010). Friston, K.J. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience,* 11, 2: 127–138.

(Fuchs, 2010). Fuchs, C.A. "QBism, the Perimeter of Quantum Bayesianism." *arXiv Preprint*. Available online at URL = <https://arxiv.org/abs/1003.5209>.

(Fuchs and Schack, 2013). Fuchs, C.A., and Schack, R. "Quantum-Bayesian Coherence." *Reviews of Modern Physics* 85, 4: 1693–1715.

(Georgiev, 2020). Georgiev, D.D. *Quantum Information and Consciousness: A Gentle Introduction*. Boca Raton FL: CRC Press.

(Ghirardi et al., 1986). Ghirardi, G. C., Rimini, A., and Weber, T. "Unified Dynamics for Microscopic and Macroscopic Systems." *Physical Review D* 34, 2: 470–491.

(Gibson, 1979). Gibson, J.J.*The Ecological Approach to Visual Perception*. Boston MA: Houghton Mifflin.

(Goff, 2017). Goff, P. *Consciousness and Fundamental Reality*. Oxford: Oxford Univ. Press.

(Graziano, 2013). Graziano, M.S. *Consciousness and the Social Brain*. Oxford: Oxford Univ. Press.

(Greene, 2020). Greene, B. *Until the End of Time: Mind, Matter, and Our Search for Meaning in an Evolving Universe*. New York: Penguin Random House.

(Grush and Churchland, 1995). Grush, R., and Churchland, P.S. "Gaps in Penrose's Toilings." *Journal of Consciousness Studies*, 2, 1: 10–29.

(Hameroff, 1998a). Hameroff, S. (1998). "Quantum Computation in Brain Microtubules? The Penrose-Hameroff "Orch OR" Model of Consciousness." *Philosophical Transactions of the Royal Society A*, 356(1743): 1869–1896.

(Hameroff, 1998b). Hameroff, S. (1998). *Ultimate Computing: Biomolecular Consciousness and Nanotechnology*. Elsevier.

(Hameroff and Penrose, 1996). Hameroff, S. and Penrose, R. "Conscious Events as Orchestrated Space-Time Selections." *Journal of Consciousness Studies* 3, 1: 36–53.

(Hanna, 2020). Hanna, R. "Criticizing the Criticism: Some Reflections on Professional Academic Thought and Real Philosophy." *Against Professional Philosophy*. 8 April. Available online at URL = <https://againstprofphil.org/2020/04/08/criticizing-the-criticism-some-reflections-on-professional-academic-thought-and-real-philosophy/>.

(Hanna, 2024). Hanna, R. *Science for Humans: Mind, Life, The Formal-&-Natural Sciences, and A New Concept of Nature*. Berlin: Springer Nature. Available online in preview at URL = <https://www.academia.edu/118055113/Science_for_Humans_Mind_Life_The_Formal_and_Natural_Sciences_and_A_New_Concept_of_Nature_Springer_Nature_2024_>.

(Hanna, 2025). Hanna, R. *Digital Technology for Humans: The Myth of AI, Human Dignity, and Neo-Luddism.* Berlin: De Gruyter Brill.

(Harnad, 1990). Harnad, S. "The Symbol Grounding Problem." *Physica D* 42, 1–3: 335–346.

(Harris, 2011). Harris, S. *The Moral Landscape: How Science Can Determine Human Values*. New York: Free Press.

(Heidegger, 1927/1962). Heidegger, M. (1962). *Being and Time*. New York: Harper and Row.

(Heisenberg, 1958). Heisenberg, W. *Physics and Philosophy*. New York: Harper and Row.

(Hohwy, 2013). Hohwy, J. *The Predictive Mind: Consciousness, Cognition, and the Brain*. Oxford: Oxford Univ. Press.

(Husserl, 1913/1931). Husserl, E. *Ideas: General Introduction to Pure Phenomenology*. Macmillan.

(Hutto and Myin, 2013). Hutto, D.D. and Myin, E. *Radicalizing Enactivism: Basic Minds without Content*. Boston MA: MIT Press.

(Kandel et al., 2013). Kandel, E.R., et al. "The Molecular and Systems Biology of Memory." *Cell* 157, 1: 163–186.

(Kane, 1996). Kane, R. *The Significance of Free Will*. Oxford: Oxford Univ. Press.

(Karnofsky, 2021). Karnofsky, H. "Reply to Eliezer on Biological Anchors." *LessWrong*. Available online at URL = <https://www.lesswrong.com/posts/nNqXfnjiezYukiMJi/reply-to-eliezer-on-biological-anchors>.

(Kauffman, 1993). Kauffman, S. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford: Oxford Univ. Press.

(Kim, 1999). Kim, J. "Making Sense of Emergence." *Philosophical Studies* 95, 1–2: 3–36.

(Lake and Baroni, 2023). Lake, B. M., and Baroni, M. (2023). "Human-Like Systematic Generalization through a Meta-Learning Neural Network." *Nature*, 623 : 115-121.

(Lambert et al., 2013). Lambert, N., et al. (2013). "Quantum Biology." *Nature Physics*, 9(1): 10–18.

(Landgrebe and Smith, 2025). Landgrebe, J. and Smith, B. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear*. 2nd edn. London: Routledge.

(LeCun et al., 2015). LeCun, Y., et al. "Deep Learning." *Nature* 521, 7553: 436–444.

(Levine, 1983). Levine, J. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64: 354–361.

(Lewin, 1980). Lewin, R. "Is Your Brain Really Necessary?" *Science* 210, 4475: 1232–1234.

(Litt et al., 2006). Litt, A., et al. "Is the Brain a Quantum Computer?" *Cognitive Science* 30, 3: 593–603.

(Lloyd, 2000). Lloyd, S. "Ultimate Physical Limits to Computation." *Nature* 406: 1047–1054.

(Lucas, 1961). Lucas, J.R. "Minds, Machines and Gödel." *Philosophy* 36, 137: 112–127.

(Lycan, 1996). Lycan, W.G. *Consciousness and Experience*. Cambridge MA: MIT Press.

(Marcus and Davis, 2019). Marcus, G. and Davis, E. *Rebooting AI: Building Artificial Intelligence We Can Trust*. 2nd edn., New York : Random House.

(Maturana and Varela, 1987). Maturana, H.R. and Varela, F.J. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boulder, CO: Shambhala.

(McCarthy, 1986). McCarthy, J. "Applications of Circumscription to Formalizing Common-Sense Knowledge." *Artificial Intelligence* 28, 1: 89–116.

(McCarthy and Hayes, 1969). McCarthy, J. and Hayes, P.J. "Some Philosophical Problems from the Standpoint of Artificial Intelligence." *Machine Intelligence, Vol. 4.* Edinburgh: Edinburgh Univ. Press. Pp. 463-502.

(McGinn, 2012). McGinn, C. (2012). "All Machine and No Ghost?" *New Statesman*. Available online at URL = <https://www.newstatesman.com/culture/2012/02/consciousness-mind-brain>.

(McKemmish et al., 2009). McKemmish, L.K., et al. "Penrose-Hameroff Orchestrated Objective-Reduction Proposal for Human Consciousness is Not Biologically Feasible." *Physical Review E* 80, 2. Available online at URL = <https://doi.org/10.1103/PhysRevE.80.021912>.

(McKenna and Pereboom, 2016). McKenna, M. and Pereboom, D. *Free Will: A Contemporary Introduction*. London: Routledge.

(Merleau-Ponty, 1945/2012). Merleau-Ponty, M. (2012). *Phenomenology of Perception*. Routledge.

(Mitchell, 2009). Mitchell, M. *Complexity: A Guided Tour*. Oxford: Oxford Univ. Press.

(Mitchell, 2023). Mitchell, M. *Artificial Intelligence: A Guide for Thinking Humans* 2nd edn., New York: Farrar, Straus and Giroux.

(Nagel, 1974). Nagel, T. "What is it Like to Be a Bat?" *Philosophical Review* 83, 4: 435–450.

(Nagel, 1986). Nagel, T. *The View from Nowhere*. Oxford: Oxford Univ. Press.

(Nielsen and Chuang, 2010). Nielsen, M.A., and Chuang, I.L. *Quantum Computation and Quantum Information*. Cambridge: Cambridge Univ. Press.

(Noë, 2004). Noë, A. *Action in Perception*. Cambridge MA: MIT Press.

(O'Neil, 2016). O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

(O'Reilly and Olaya-Castro, 2014). O'Reilly, E.J. and Olaya-Castro, A. "Non-Classicality of the Molecular Vibrations Assisting Exciton Energy Transfer at Room Temperature." *Nature Communications* 5: 3012. Available online at URL = <https://doi.org/10.1038/ncomms4012>.

(Ozimek, 2025). Ozimek, T. "'Godfather of AI' Warns Superintelligent Machines Could Replace Humanity." *The Epoch Times*. Available online at URL = <https://www.theepochtimes.com/tech/godfather-of-ai-warns-superintelligent-machines-could-replace-humanity-5905191>.

(Pearle, 1989). Pearle, P. "Combining Stochastic Dynamical State-Vector Reduction with Spontaneous Localization." *Physical Review A* 39, 5: 2277–2289.

(Penrose, 1989). Penrose, R. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* Oxford: Oxford Univ. Press.

(Penrose, 1994). Penrose, R. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford Univ. Press.

(Penrose, 1996). Penrose, R. "On Gravity's Role in Quantum State Reduction." *General Relativity and Gravitation* 28, 5: 581–600.

(Penrose and Hameroff, 2011). Penrose, R., and Hameroff, S. "Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and Orch OR Theory." *Journal of Cosmology* 14: 1–17.

(Perkins, 2025). Perkins, M. "How Much of Your Brain Do You Need to Survive?" *Live Science*. 31 March. Available online at URL = <https://livescience.com/health/neuroscience/how-much-of-your-brain-do-you-need-to-survive>.

(Piccinini, 2007). Piccinini, G. "Computational Modelling vs. Computational Explanations: Is Everything a Turing Machine, and Does It Matter to the Philosophy of Mind?" *Australasian Journal of Philosophy* 85, 1: 93–115.

Pothos, E.M. and Busemeyer, J.R."Can Quantum Probability Provide a New Direction for Cognitive Modeling?" *Behavioral and Brain Sciences* 36, 3: 255–274.

(Putnam, 1960). Putnam, H. "Minds and Machines." In S. Hook (ed.), *Dimensions of Mind*. New York: New York Univ. Press. Pp. 138-164

(Pylyshyn, 1984). Pylyshyn, Z.W. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge MA: MIT Press.

(Raichle, 2015). Raichle, M.E. "The Brain's Default Mode Network." *Annual Review of Neuroscience* 38: 433–447.

(Redbubble, 2025). Kahn, F. "Homo Machina (Human Machine)." *Redbubble*. Available online at URL = <https://www.redbubble.com/shop/fritz+kahn+photographic%20prints>.

(Reiter, 1980). Reiter, R. "A Logic for Default Reasoning." *Artificial Intelligence* 13, 1–2: 81–132.

(Ritz et al., 2000). Ritz, T. et al. "A Model for Photoreceptor-Based Magnetoreception in Birds." *Biophysical Journal* 78, 2: 707–718.

(Rosenthal, 2005). Rosenthal, D. M. *Consciousness and Mind*. Oxford: Oxford Univ. Press.

(Ryle, 1949). Ryle, G. *The Concept of Mind*. Chicago IL: Univ. of Chicago Press.

(Sapolsky, 2023). Sapolsky, R.M. *Determined: A Science of Life Without Free Will*. New York: Penguin.

(Schlosshauer, 2007). Schlosshauer, M. *Decoherence and the Quantum-to-Classical Transition*. Cham: Springer.

(Scott, 2018). Scott, A.J. "Grand Illusions and Existential Angst." *Skeptical Inquirer* 42, 6: 51–55.

(Searle, 1980). Searle, J. R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences*, 3, 3: 417–424.

(Searle, 1983). Searle, J.R. *Intentionality*. Cambridge University Press.

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

(Seth, 2014). Seth, A.K. "A Predictive Processing Theory of Sensorimotor Contingencies: Explaining the Puzzle of Perceptual Presence and Its Absence in Synesthesia." *Cognitive Neuroscience* 5, 2: 97–118.

(Shanahan, 1997). Shanahan, M. *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. Cambridge MA: MIT Press.

(Shapiro, 2003). Shapiro, S. "Mechanism, Truth, and Penrose's New Argument" *Journal of Philosophical Logic* 32, 1: 19–42.

(Shojaee, A., et al. 2025). Shojaee, A. et al. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models Via the Lens of Problem Complexity." *Apple Machine Learning Research*. Available online at URL = <https://machinelearning.apple.com/research/illusion-of-thinking>.

(Sonik and Colarossi, 2020). Sonik, D. and Colarossi, J. *Becoming Artificial: A Philosophical Exploration into Artificial Intelligence and What It Means to Be Human*. Luton UK: Andrews UK Limited.

(Stapp, 1993). Stapp, H.P. *Mind, Matter and Quantum Mechanics*. Cham: Springer-Verlag.

(Stall, 2007). Stapp, H. . *Mindful Universe: Quantum Mechanics and the Participating Observer*. Cham: Springer.

(Stiefel and Coggan, 2022). Stiefel, K. and Coggan, J. "A Hard Energy Use Limit of Artificial Superintelligence." *TechRxiv Preprint*. Available online at URL = <https://www.techrxiv.org/articles/preprint/A_Hard_Energy_Use_Limit_of_Artificial_Superintelligence/21588612>.

(Tegmark, 2000). Tegmark, M. "Importance of Quantum Decoherence in Brain Processes." *Physical Review E* 6, 14: 4194–4206.

(Thelen and Smith, 1994). Thelen, E. and Smith, L.B. *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge MA: MIT Press.

(Thompson et el., 2020). Thompson, S., et al. "The Computational Limits of Deep Learning." *arXiv Preprint*. 10 July. Available online at URL = <https://arxiv.org/abs/2007.05558>.

(Tononi, 1999). Tononi, G. "Consciousness and Complexity." *Science* 282, 5395: 1846–1851.

(Tononi, 2015). Tononi, G. "Integrated Information Theory." *Scholarpedia* 10, 1: 4164.

(Tononi and Koch, 2015). Tononi, G. and Koch, C. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society B* 370, 1668. Available online at URL = <https://royalsocietypublishing.org/doi/10.1098/rstb.2014.0167>.

(Turner, 2025). Turner, B. "Current AI Models a 'Dead End' for Human-Level Intelligence, Scientists Agree." *Live Science*. Available online at URL = <https://www.yahoo.com/news/current-ai-models-dead-end-135020529.html>.

(Van Gelder, 1995). Van Gelder, T. "What Might Cognition Be, if Not Computation?" *Journal of Philosophy* 92, 7: 345–381.

(Varela et al., 1991). Varela, F.J. et al. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge MA: MIT Press.

(Vitiello, 1995). Vitiello, G. "Dissipation and Memory Capacity in the Quantum Brain Model." *International Journal of Modern Physics B* 9, 8: 973–989.

(von Neumann, 1955). von Neumann, J. *Mathematical Foundations of Quantum Mechanics*. Princeton NJ: Princeton Univ. Press.

(Wallace, 2012). Wallace, D. *The Emergent Multiverse*. Oxford: Oxford Univ. Press.

(Wheeler, 1989). Wheeler, J. A. "Information, Physics, Quantum: The Search for Links." In W. Zurek (ed.), *Complexity, Entropy, and the Physics of Information*. Redwood City CA: Addison-Wesley. Pp. 3-28.

(Wigner, 1961). Wigner, E. P. "Remarks on the Mind-Body Question." In I.J. Good (ed.), *The Scientist Speculates*. Portsmouth NH: Heinemann. Pp. 284-302.

(Wolfram, 2002). Wolfram, S. *A New Kind of Science*. Champaign IL: Wolfram Media.

(Wu and Boas, 2025). Wu, J. and Bosa, D. "The AI-Boom's Multi-Economy System: Billion-Dollar Blind Spot: Reasoning Models Hitting a Wall." 26 June. Available online at URL = <https://www.cnbc.com/2025/06/26/ai-reasoning-models-problem.html>.

(Zeilinger, 1999). Zeilinger, A. *Experiment and the Foundations of Quantum Physics*. Singapore: World Scientific.

(Zurek, 1991). Zurek, W. H. "Decoherence and the Transition from Quantum to Classical." *Physics Today* 44, 10: 36–44.

(Zurek, 2003). Zurek, W. H. "Decoherence, Einselection, and the Quantum Origins of the Classical." *Reviews of Modern Physics* 75, 3: 715–775.