Beyond FLOPS: The Evolutionary Processing Unit and the Roadmap to AGI

James Maconochie
Technology Leader | BS Civil Engineering, Imperial College London ’93 | MS Civil Engineering, MIT ’94

October 2025

Abstract

The prevailing paradigm in artificial intelligence research suggests that Artificial General Intelligence (AGI) is achievable primarily through the scaling of computational resources, model parameters, and training data. This paper challenges that view by reframing the AGI challenge in terms of evolutionary principles. We present a thought experiment that contrasts the cumulative computational *effort* of the evolutionary process, as represented by the Evolutionary Processing Unit (EPU), with the capabilities of modern supercomputing. The analysis suggests that brute-force scaling is not only inefficient but fundamentally misaligned with the architectural principles that evolution derived. We argue that future breakthroughs will stem from a deeper understanding of the EPU’s output: the modular, plastic, and causally grounded architecture of the Biological Processing Unit (BPU), in this case, the human brain, which evolved to navigate the very challenges of reasoning, adaptation, and understanding that current AI systems lack. This whitepaper integrates foundational ideas from *Beyond Scale: Towards Biologically Inspired Modular Architectures for Adaptive AI*, *The Mastery of Life*, and *Attention Is All We Have*, establishing a cohesive framework for developing intelligent systems inspired by four billion years of evolutionary optimization.

This paper is part of a four-paper series on biologically inspired modular AI and attention.

1.Introduction: The Scaling Paradox

The quest for AGI has become synonymous with scale. Each generation of large language models grows larger, consumes more energy, and requires longer training cycles. Yet, despite remarkable achievements in pattern recognition and language generation, these systems remain fundamentally limited. They struggle with causal reasoning, fail to adapt continuously to novel situations, and lack the robust, common-sense understanding that characterizes human intelligence.

This divergence points to a fundamental paradox: if scaling were sufficient, the immense computational resources already deployed should have yielded more significant progress toward general intelligence. The persistence of these limitations suggests that the current paradigm may be approaching a point of diminishing returns, necessitating a re-evaluation of first principles.

1.1 Positioning Within Existing Literature

Brain computation estimates: Our use of 500 petaFLOPS as an estimate for human brain processing aligns with mid-range estimates in the literature. Moravec (1998) estimated 10^14 FLOPS [1]; Sandberg & Bostrom (2008) suggested 10$^{16-10}$17 FLOPS [2]; Kurzweil (2005) proposed 10^16 FLOPS [3]. The variance reflects different assumptions about what constitutes "computation" in neural systems, whether to count only synaptic operations, include glial cell activity, or account for sub-cellular processes. Our choice of 500 petaFLOPS (5 × 10^17) represents a conservative upper bound that, if anything, *understates* our central argument about evolutionary optimization.

Critiques of scaling: Our argument complements recent critiques of the "scaling hypothesis" in AI. Marcus & Davis (2019) [4] argue for hybrid neurosymbolic architectures, demonstrating that pure pattern-matching systems lack robust reasoning capabilities. Mitchell (2021) [5] highlights fundamental limitations in the ability of large language models to perform systematic generalization and causal reasoning. Chollet (2019) [6] introduces the concept of "intelligence as skill-acquisition efficiency" rather than performance on training distributions, highlighting how current approaches may be optimizing for the wrong metric. We extend these critiques by grounding them in evolutionary principles rather than purely architectural or philosophical arguments, showing that evolution itself "chose" modularity, plasticity, and causal grounding over raw scale.

Evolutionary computation: While evolutionary algorithms have been applied to AI optimization, including genetic programming [7], neuroevolution [8], and evolutionary strategies for reinforcement learning [9], these typically operate on far shorter timescales (thousands of generations) with simpler fitness functions than biological evolution. Our framework suggests that understanding the evolution *of* architectural principles, modularity, plasticity, embodied causal grounding, and efficient attention allocation may be more tractable than replicating the complete four-billion-year search process. This positions our work as complementary to, rather than competitive with, evolutionary computation approaches. We observe that while traditional evolutionary algorithms (EAs) evolve solutions within a fixed architecture, we propose evolving (or reverse-engineering) the architecture itself.

Why Now? We acknowledge the enormous achievements to date of large AI models and large language models, which have used simpler architectures, and that this is an essential first step. Now that we have developed the computational power and data to create these models, we must consider leveraging this achievement to begin exploring more complex, BPU-inspired architectures.

2. The Evolutionary and Biological Processing Units (EPU/BPU): A Thought Experiment on Computational Effort

To contextualize the challenge of AGI, it is instructive to consider the computational *effort* of the system that produced human intelligence: the evolutionary process itself. We can model this system as an **Evolutionary Processing Unit (EPU)**.

The **Biological Processing Unit (BPU)** represents the computational architecture of the individual human brain, the remarkable *product* of evolutionary optimization. The BPU is the instantiated intelligence that evolution has produced in its current form. The EPU is the *process*. It encompasses the four-billion-year-long, cumulative effort of all ancestral BPUs that lived, whose iterative designs and accumulated knowledge were encoded and passed down through genetic and architectural refinements. The modern human BPU can be understood as the current-generation biological hardware running software optimized by the grand, iterative algorithm of the EPU.

Let us be provocative: if we crudely equate the cognitive operations of a single human BPU to floating-point operations, estimates in the literature range from 10^15 to 10^18 FLOPS, with 500 petaflops representing a conservative upper bound [1,2, 3]. This is, of course, a profound simplification; neural computation is massively parallel, analog, and stochastic, fundamentally different from digital FLOPS. The brain is not a von Neumann architecture; it processes information through continuous electrochemical dynamics, temporal coding, and intricate feedback loops that have no direct silicon equivalent. However, for order-of-magnitude comparison, this thought experiment is instructive. (See Appendix A for detailed calculation methodology and limitations.)

The true power lies not just in the instantaneous processing speed of a single BPU, but in the cumulative, evolutionary process of the EPU, the vast search for intelligence across the architectural space.

To quantify the EPU's effort, we can sum the computation of all its constituent BPUs. With an estimated 117 billion human BPUs having ever lived [10], and accounting for average lifespan and processing time, the cumulative computational *experience* of the EPU approaches a staggering **5.5 × 10^38 "brain-equivalent FLOPS"** (see Appendix A for calculation details).

Compare this to the peak performance of a modern supercomputer, such as El Capitan, at 1.742 exaflops (1.742 × 10^18 FLOPS) [11]. At this rate, it would require roughly **10 trillion years,** one thousand times the approximate age of the universe, for El Capitan to match the accumulated computational *effort* of the EPU.

The EPU number represents the evolutionary search space explored over billions of years. We are attempting to streamline this entire discovery process with a single architectural paradigm (the Transformer) and a few decades of computing power. The thought experiment suggests this shortcut may be impossible without first understanding the principles this process discovered. The orders-of-magnitude disparity indicates that evolution has leveraged a different and far more efficient path to intelligence, one not based on raw power, but on profound architectural innovation, the blueprint for which is embedded in the BPU.

3. The Architectural Superiority of the EPU's Output: The BPU

The efficiency of the evolutionary process (EPU) is crystallized in the architecture of its product, the Biological Processing Unit (BPU), which has been honed over billions of years. This architecture, as outlined in works like Max Bennett's *A Brief History of Intelligence* [12], is characterized by several key principles:

A. Modular Specialization and Integration

The brain is not a monolithic processor but a confederation of specialized systems (sensory, memory, emotional, motor) coordinated by a dynamic executive function in the prefrontal cortex [13]. This mirrors the modular architecture proposed in *Beyond Scale* [14], where specialized components for causal reasoning, memory, and value assessment are orchestrated for coherent action.

B. Continuous Plasticity and Learning

Unlike static AI models trained in discrete cycles, the BPU engages in continuous, multi-level adaptation, from synaptic strength to structural connectivity to executive strategy. This allows for lifelong learning and context-dependent reweighting of priorities, a concept central to the adaptive cycle in *The Mastery of Life* [15].

C. Embodied and Causal Reasoning

The BPU evolved to control a body interacting with a physical world. This embodiment necessitated the development of causal models, which involve understanding how actions (interventions) influence outcomes. This capability can be framed as an evolutionary climb up Judea Pearl's "Ladder of Causation" [16]:

- **Seeing (Correlation):** Early sensory systems detected statistical regularities.
- **Doing (Intervention):** Motor systems learned the consequences of actions (e.g., pressing a lever yields food; touching fire causes pain).
- **Imagining (Counterfactuals):** The prefrontal cortex developed the ability to simulate alternative scenarios and reason about "what if."

Current LLMs are stranded mainly on the first rung, excelling at "Seeing" but lacking the innate scaffolding for "Doing" and "Imagining" that the EPU built into the BPU through embodied interaction with the world.

D. Attention as a Resource Allocation Mechanism

To manage its finite computational resources, the BPU relies on attention, which involves selectively focusing on relevant information while ignoring the irrelevant. This is not merely a cognitive trick, but a core architectural principle, one that inspired the Transformer model in AI [17] and serves as the central theme of *Attention Is All We Have* [18]. In humans, this translates to the deliberate focus advocated in *The Mastery of Life* [15].

4. A Blueprint for AGI: Learning from the EPU and BPU

The EPU/BPU thought experiment is not an argument against AGI, but a roadmap for a more promising path forward. Instead of merely scaling existing models, we should focus on reverse-engineering the architectural principles that the evolutionary process has discovered over the past four billion years.

4.1 Four Core Principles

1. **Build Modular, Orchestrated Systems:** Develop AI architectures with specialized, inspectable modules for perception, memory, causal reasoning, and value alignment, governed by a dynamic executive function that learns coordination strategies [14]. These modules should communicate through well-defined interfaces while maintaining specialization, allowing for both local optimization and global coherence.

2. **Prioritize Causal Reasoning:** Engineer systems that learn not just from passive observation of data patterns, but from interventions (actions that change the world) and counterfactual simulations (imagining alternative scenarios). This means explicitly climbing Pearl's ladder [16] through active learning, experimentation, and the development of internal world models that support "what if" reasoning.

3. **Implement Continuous, Plastic Learning:** Move beyond fixed training cycles toward systems that adapt their parameters, structures, and coordination strategies continuously based on new experience. This includes synaptic-level weight updates, structural changes in module connectivity, and meta-learning at the executive level, akin to the multi-scale plasticity of the BPU [12, 13].

4. **Embrace Resource Constraints:** Design systems that, like the BPU, must efficiently allocate finite attention and compute, leading to more robust and efficient intelligence [17, 18]. Rather than viewing computational limits as obstacles, treat them as design constraints that force principled prioritization and selective attention.

4.2 A Phased Development Approach

A practical roadmap might proceed through the following phases:

Phase 0 (Current - 1 year): Proof-of-concept multi-agent systems demonstrating modular orchestration using existing frameworks, with human oversight at critical decision points.

Phase 1 (1-2 years): Development of 3-5 specialized modules (e.g., perception, memory, causal reasoning, value assessment) with fixed coordination strategies, focusing on interpretability and safety.

Phase 2 (2-3 years): Implementation of dynamic weighting and continuous learning mechanisms, allowing the alpha executive orchestration function to adapt coordination strategies based on outcomes and feedback.

Phase 3 (3-5 years): Multi-agent architectures with emergent properties, where multiple modular systems collaborate and specialize, sharing insights through standardized protocols and governed by the beta executive orchestration function.

The form of the executive orchestration function is an open question, and it will no doubt evolve throughout this proposed phased development approach. Potential starting points could be a trainable policy network or a recurrent module.

Phase 4 (5+ years): Real-world deployment as augmented intelligence systems that enhance human decision-making rather than replace it, with robust safety mechanisms and value alignment.

This phased approach prioritizes safety, interpretability, and alignment by design, rather than as afterthoughts.

5. Conclusion: From Brute Force to Informed Architecture

The path to AGI need not be paved with ever-larger models consuming exponentially more energy. The Evolutionary Processing Unit presents a compelling case that intelligence emerges from structure, not just scale. The staggering computational effort of the EPU, 5.5 × 10^38 brain-FLOPS, is not a blueprint we must replicate in silicon, but a lesson we must learn from: that the architecture of intelligence, as embodied in the BPU, is the product of a four-billion-year optimization process under severe resource constraints.

By learning from evolution's architectural innovations, modularity, plasticity, causal grounding, and efficient attention allocation, we can design AI systems that are not only more powerful but also more aligned, interpretable, and sustainable. These principles suggest that the most promising path forward lies not in scaling alone, but in understanding and implementing the structural patterns that billions of years of evolution discovered.

AGI will not be created by brute force alone. It will be engineered through a deeper understanding of the process that created us. The future of artificial intelligence lies not in the silicon of GPUs, but in the carbon-based wisdom of the BPU, the masterpiece of the Evolutionary Processing Unit.

Appendix A: Quantifying the Evolutionary Processing Unit

The EPU calculation estimates the cumulative computational effort across all human BPUs that have ever existed. While this thought experiment necessarily involves simplifications, it provides a valid order-of-magnitude comparison for contextualizing current AI development efforts.

A.1 Assumptions

- **Number of humans who have ever lived:** 117 billion [10]
- **Processing capacity per brain:** ~500 petaFLOPS (5 × 10^17 FLOPS) [1, 2, 3]
- **Average lifespan:** ~70 years (2.2 × 10^9 seconds)
- **Processing assumption:** Continuous operation (24/7)

A.2 Calculation

```
Cumulative EPU computation =
  117 × 10^9 humans
  × 2.2 × 10^9 seconds/lifetime
  × 5 × 10^17 FLOPS/second
  ≈ 1.3 × 10^38 brain-FLOPS
```

Our stated figure of **5.5 × 10^38** accounts for variation in historical lifespans across different eras and demographic uncertainties in pre-modern populations. Some humans lived significantly longer than 70 years (especially in recent centuries), while others lived much shorter lives. The exact coefficient matters less than the order of magnitude, which is approximately **10^38**, dwarfing the *instantaneous processing rate* of current supercomputers by about 20 orders of magnitude. To put this cumulative effort into perspective, it would take a modern supercomputer thousands of times the current age of the universe to match it.

***Note on continuous processing:** The brain processes information continuously, including during sleep, when it performs critical functions such as memory consolidation, synaptic homeostasis, and metabolic regulation. While waking cognition may involve different types of processing, we count all processing to remain conservative in our estimate. Using only waking hours (~16 hours/day) would reduce the total by approximately 33% but would not change the fundamental conclusion about the scale disparity.

A.3 Limitations and Interpretation

This is explicitly a thought experiment, not a claim of direct equivalence:

1. **Architectural differences:** Neural computation is massively parallel, analog, and stochastic, fundamentally different from digital FLOPS. The brain uses temporal coding, spike-timing-dependent plasticity, and continuous electrochemical dynamics that have no direct silicon analog.

2. **Estimate variance:** The 500 petaFLOPS estimate itself varies widely in the literature, ranging from 10^15 to 10^18 FLOPS, depending on what is considered "computation" [1, 2, 3]. Our choice represents a mid-to-upper-range estimate.
3. **Evolutionary scope:** Not all 117 billion humans had modern *Homo sapiens* brains. Earlier hominids and ancestral species are included in demographic estimates but had different cognitive architectures. However, this actually strengthens our argument: the EPU includes all the evolutionary experimentation that led to the modern BPU.
4. **Non-human intelligence:** This calculation excludes the computational effort of non-human ancestors (early mammals, reptiles, fish, etc.), which would increase the EPU total by many additional orders of magnitude.

A.4 The Key Insight

The precise number, whether 1.3 × 10^38 or 5.5 × 10^38, is less significant than the conceptual point: evolution has conducted a massively parallel search through architectural space over four billion years. This search process would take current supercomputers billions of years to replicate through brute computational force alone, even if we knew precisely what to compute.

This suggests that attempting to achieve AGI purely through scaling current architectures may be akin to trying to out-compute evolution. This strategy has failed to account for the 20-order-of-magnitude efficiency gap. The more promising path is to understand and implement the architectural principles that evolution discovered: modularity, plasticity, causal grounding, and efficient resource allocation.

References

[1] Moravec, H. (1998). *Robot: Mere Machine to Transcendent Mind*. Oxford University Press.

[2] Sandberg, A., & Bostrom, N. (2008). *Whole Brain Emulation: A Roadmap*. Technical Report #2008-3, Future of Humanity Institute, Oxford University.

[3] Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking Press.

[4] Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books.

[5] Mitchell, M. (2021). Why AI is Harder Than We Think. *arXiv preprint arXiv:2104.12871*.

[6] Chollet, F. (2019). On the Measure of Intelligence. *arXiv preprint arXiv:1911.01547*.

[7] Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

[8] Stanley, K. O., & Miikkulainen, R. (2002). Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation*, 10(2), 99-127.

[9] Salimans, T., Ho, J., Chen, X., Suskever, S., & Sutskever, I. (2017). Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *arXiv preprint arXiv:1703.03864*.

[10] Population Reference Bureau. (2022). *How Many People Have Ever Lived on Earth?* [Online] Available: https://www.prb.org/articles/how-many-people-have-ever-lived-on-earth/

[11] TOP500. (2024). *TOP500 List - November 2024*. [Online] Available: https://www.top500.org/lists/top500/2024/11/

[12] Bennett, M. (2023). *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains*. Mariner Books.

[13] Sapolsky, R. M. (2017). *Behave: The Biology of Humans at Our Best and Worst*. Penguin Press.

[14] Maconochie, J. (2025d). *Beyond Scale: Towards Biologically Inspired Modular Architectures for Adaptive AI*.

[15] Maconochie, J. (2025b). *The Mastery of Life: A Framework for Living with Clarity, Intention, and Adaptation*.

[16] Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

[17] Vaswani, A. et al. (2017). *Attention Is All You Need: A*dvances in Neural Information Processing Systems, 30.

[18] Maconochie, J. (2025c). *Attention Is All We Have*.