Attention Is All We Have: What AI's Greatest Breakthrough Can Teach Us About Being Human

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

October 2025

Abstract

In 2017, researchers at Google introduced the Transformer model in a paper titled *Attention Is All You Need*, demonstrating that artificial intelligence performance depends less on architectural complexity and more on selective focus, or the ability to prioritize what matters and disregard what does not. This paper argues that the same principle governs both human intelligence and fulfillment, and that attention, the selective allocation of cognitive and emotional resources, is the shared architectural foundation underlying both human and machine intelligence.

Drawing on the Evolutionary Processing Unit (EPU) framework and its product, the Biological Processing Unit (BPU), we demonstrate that attention is not merely a proper mechanism, but a fundamental resource allocation strategy that evolution has optimized over the past four billion years. Building on *the Mastery of Life* framework and *Beyond Scale* architecture, this paper proposes that attention is both the mechanism through which consciousness operates and the moral foundation of intelligence. We distinguish between computational attention (in AI systems) and cognitive attention (in human experience), while identifying their shared principle: both solve the problem of prioritizing limited resources in service of goals. Where the Transformer revolutionized computation, the deliberate direction of human attention by understanding the BPU and applying the Master of Life (MOL) framework may yet transform our understanding of what it means to live wisely.

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

1. Introduction: From Algorithms to Awareness

The 2017 paper *Attention Is All You Need* [1] redefined how machines learn. By replacing sequential processing with a self-attention mechanism, the Transformer model enabled AI systems to dynamically focus on relevant inputs, a breakthrough that made context, rather than sheer volume, the organizing principle of understanding.

The insight was simple yet profound: not all data deserve equal treatment. By teaching a model to weigh specific tokens more heavily than others, it can infer meaning, coherence, and hierarchy, traits previously reserved for biological cognition. This architectural innovation, rather than raw computational scale, unlocked capabilities that had eluded researchers for decades.

However, this insight does not belong solely to machines. The same dynamic governs the human mind. As William James wrote more than a century ago, “My experience is what I agree to attend to” [2]. Our lives, too, are shaped not by the information we encounter, but by what we choose to pay attention to.

In a world overflowing with noise, this act of selection, of attention, has become one of the rarest and most valuable forms of intelligence. What makes attention so critical is not just its scarcity, but its role in determining meaning: in an age where information is abundant, the ability to discern what deserves focus becomes the defining skill. As artificial intelligence continues to advance at ever-greater computational scale, it may be humanity’s capacity for deliberate attention that remains our defining strength.

This paper explicitly proposes attention as the unifying principle and architectural foundation across three domains: evolutionary optimization (the EPU), human cognition and fulfillment (the BPU and MOL framework), and artificial intelligence (Transformer models and modular architectures, as described by Maconochie [3]. We argue that understanding attention, how it evolved, how it operates, and how it can be deliberately cultivated, is essential for building both wiser humans and more capable machines.

2. Related Work and Positioning

2.1 Attention in AI Systems

While Vaswani et al.’s Transformer [1] popularized self-attention mechanisms, the concept has its roots in a deeper history. Bahdanau et al. (2015) [4] introduced attention for neural machine translation. Xu et al. (2015) [5] applied attention to image captioning. These works share a common insight: rather than processing all inputs equally, systems should learn to dynamically focus on relevant information.

Recent work has expanded attention mechanisms, including multi-head attention [1], sparse attention [6], and flash attention [7], which optimize computational efficiency. However, these advances remain primarily architectural optimizations rather than interrogations of attention’s fundamental role in intelligence.

2.2 Attention and Consciousness

The relationship between attention and consciousness has been extensively studied in neuroscience and philosophy. Dehaene et al. (2006) [8] propose that conscious access requires attention, in other words, that we become aware only of what we attend to. Koch and Tsuchiya (2007) [9] challenge this, arguing that attention and consciousness are dissociable. Tononi’s Integrated Information Theory (2004) [10] focuses on information integration rather than selective attention.

Our framework takes a more functionalist approach: whether attention is necessary for consciousness, it is the mechanism through which conscious agents allocate limited cognitive resources. We remain agnostic on the complex problem of consciousness, while focusing on the architectural role of attention.

2.3 Human-AI Symbiosis

The vision of human-machine collaboration is not new. Licklider's "Man-Computer Symbiosis" (1960) [11] anticipated partnerships where computers handle routine tasks while humans provide judgment and creativity. Engelbart's "Augmentation" framework (1962) [12] proposed using computers to enhance human intellectual capabilities.

Our contribution extends this tradition by identifying attention as the shared currency enabling effective collaboration. Where Licklider and Engelbart focused on task division, we emphasize attention coordination: humans and AI systems must learn to manage collective attention across a hybrid cognitive system.

2.4 Attention in Human Decision-Making

Behavioral economics demonstrates how attention shapes decisions. Kahneman's "attention budget" concept [13] shows that cognitive resources are finite and must be allocated strategically. Thaler and Sunstein's "choice architecture" [14] reveals how attention can be directed to influence decisions for better or worse.

The Mastery of Life framework [15] builds on these insights, proposing that deliberate attention management is the foundation of fulfillment, an idea with roots in contemplative traditions but grounded in cognitive science.

3. Attention in Evolution: The EPU's Optimization Target

3.1 Attention as Resource Allocation

The Evolutionary Processing Unit (EPU) framework [16] demonstrates that intelligence emerges from architectural innovation rather than raw computational scale. Over four billion years of evolution, representing approximately 5.5 × 10^38 brain-equivalent FLOPS of cumulative computational effort, the EPU explored vast design spaces under severe resource constraints.

A critical constraint was energy: brains are metabolically expensive, consuming roughly 20% of the body's energy while representing only 2% of body mass. This created intense selective pressure for efficient resource allocation. Attention emerged as evolution's solution: rather than processing all sensory inputs equally, organisms that could selectively focus on relevant information gained survival advantages.

3.2 The BPU's Attention Architecture

The Biological Processing Unit (BPU), in this case, the modern human brain, embodies this evolutionary optimization [16]. The BPU's attention system operates at multiple levels:

Bottom-up attention: Automatic, stimulus-driven orienting to salient events (a sudden noise, movement in peripheral vision). This evolved early and is shared across many species.

Top-down attention: Goal-directed, voluntary focus on task-relevant information (reading this sentence, listening to a specific voice in a noisy room). This requires coordination of the prefrontal cortex and is more developed in humans.

Sustained attention: Maintaining focus over time despite distractions and fatigue. This is metabolically expensive but enables complex cognition.

Critically, these mechanisms don't merely filter information; they fundamentally shape what enters consciousness. Information that doesn't receive attention is not simply ignored; it's often not even experienced. This is attention's power: it determines the contents of subjective experience.

3.3 From BPU to Transformer: Attention's Journey to Silicon

The Transformer architecture [1] represents a remarkable convergence: computer scientists, seeking to improve language models, independently discovered the same solution that evolution has found over millions of years. Self-attention in Transformers operates analogously to the BPU's attention:

- **Query, Key, Value mechanism:** Like how the brain matches current goals (query) against memory traces (keys) to retrieve relevant information (values)
- **Multi-head attention:** Analogous to processing information along multiple dimensions simultaneously (visual, semantic, emotional)
- **Attention weights:** Explicit resource allocation, prioritizing some inputs over others

This convergence is not coincidental. Both biological and artificial systems face the same fundamental challenge: making sense of high-dimensional information with limited computational resources. Attention is the solution that both discovered.

4. Attention in Human Experience: Consciousness and Choice

4.1 The Human Parallel: Attention as Operating System

Human consciousness can be understood as an adaptive attention system. Every sensory input, emotion, and thought competes for a finite resource: awareness. What we attend to becomes salient; what we ignore fades into irrelevance.

Psychology and neuroscience have long recognized this constraint. Daniel Kahneman's dual-system model of cognition [13] distinguishes between System 1 (automatic, fast, associative) and System 2 (reflective, deliberate). Attention is the switch between them, the act of interrupting a habit to engage reflection.

Donald Hoffman's *The Case Against Reality* [17] offers another perspective: our perception evolved for fitness, not truth. We attend to what aids survival, not to what is objectively real. This evolutionary efficiency parallels how attention mechanisms in AI prune complexity, conserving energy while amplifying relevance.

4.2 Distinguishing Computational and Cognitive Attention

It is crucial to distinguish between attention as implemented in AI systems and attention as experienced by conscious agents:

Computational attention (in AI): - Mathematical operation: weighted sum of inputs based on learned relevance scores - No subjective experience or awareness - Deterministic (given the same inputs and weights, produces identical outputs) - Purpose: optimize task performance (e.g., translation accuracy, classification precision)

Cognitive attention (in humans): - Phenomenological experience: the felt quality of focusing awareness - Accompanied by subjective experience - Influenced by fatigue, emotion, motivation, and countless contextual factors - Purpose: navigate a complex world while managing limited cognitive resources

The shared principle:

Despite these differences, both forms of attention solve the same fundamental problem: **how to prioritize limited processing resources in service of goals**. The BPU faces metabolic and bandwidth constraints; the Transformer faces computational and memory constraints. Both use attention to allocate resources efficiently.

This shared principle suggests that attention is not merely a useful heuristic but a necessary architectural feature of any resource-bounded intelligent system. Whether biological or artificial, intelligence requires a selective focus.

4.3 Attention as Moral Choice

In humans, attention carries moral weight. What we choose to notice reveals what we value. Attention to others' suffering can motivate compassion; inattention enables indifference. Attention to long-term consequences fosters responsibility; inattention permits shortsightedness.

This is not merely metaphorical. Attention shapes neural pathways through Hebbian learning: "neurons that fire together, wire together." Repeated attention to specific patterns strengthens those circuits, making similar patterns more likely to capture future attention. We literally become what we pay attention to.

The ethical dimension of attention extends to AI systems. What an AI attends to during training determines what it learns. Biased attention to specific data patterns produces biased models. Designing AI attention mechanisms is thus an ethical task, not merely a technical one.

5. The Mastery of Life: A Human Attention Architecture

5.1 MOL as Attention Management System

The Mastery of Life (MOL) framework [15] can be understood as a deliberate architecture for managing human attention. It proposes that fulfillment arises not from doing more, but from understanding what truly matters and devoting attention accordingly.

MOL organizes life into seven core domains (Physical Vitality, Mental Clarity, Relational Connection, Purpose & Growth, Material Security, Self-Regulation, Novelty & Discovery) with an eighth integrative measure (Overall Fulfillment). Each domain operates as a module competing for finite attention resources.

The framework's three-phase cycle mirrors the Transformer's computational structure, though we must be careful not to overstate this analogy:

Awareness (Input Encoding): Observing current states across life domains without judgment. Like how Transformers encode input tokens into vector representations, awareness translates lived experience into observable patterns. This involves data collection, including observations of sleep quality, relationship dynamics, and emotional states.

Attention (Contextual Weighting): Deciding which domains require focus based on current priorities and long-term values. This is analogous to how self-attention mechanisms assign importance weights to different tokens. In MOL, this means consciously allocating time and cognitive resources to domains that matter most while accepting trade-offs in others.

Adaptation (Learning and Adjustment): Updating beliefs, behaviors, and priorities based on outcomes. While not precisely equivalent to backpropagation (which is a specific mathematical algorithm for gradient descent), adaptation serves a similar function: using feedback to improve future resource allocation. When attention to Physical Vitality improves Mental Clarity, this pattern is reinforced; when attention to work diminishes Relational Connection, adjustments are made.

Important caveat: It is crucial to clarify the nature of this analogy to avoid a potential misinterpretation. A critic might rightly note that the mechanisms of a Transformer and the processes of human reflection are vastly different, making a direct comparison untenable. We fully agree. The BPU does not perform gradient descent, and human adaptation involves complex, context-dependent reasoning, social learning, and emotional regulation far beyond the mathematical operations of a neural network. Therefore, we posit this not as a *mechanistic* analogy, but as a *functional* one. Both systems, despite their radically different implementations, are designed to solve the same core problem: the efficient allocation of limited processing resources in service of goals. The MOL cycle of Awareness (input encoding), Attention (contextual weighting), and Adaptation (learning) serves a functional analogy to the Transformer's computational steps of filtering, prioritizing, and learning from experience to improve future resource allocation. This functional perspective enables us to draw insightful parallels while acknowledging the profound differences in implementation.

5.2 Attention Budgets and Trade-offs

Human attention is fundamentally limited, unlike computational power, which can be increased by adding more processors. We cannot simply “scale up” our awareness. A day contains 86,400 seconds; attention to one thing necessarily means inattention to everything else.

This scarcity is not a bug but a feature; it forces prioritization, and prioritization reveals values. In this sense, how we allocate attention is not just a cognitive choice but a moral one, defining what we consider worthy of our finite existence.

The MOL framework makes this explicit through tracking and reflection. By measuring outcomes across domains, individuals can observe the consequences of their attention allocation: - Does attention to career advancement come at the cost of relationships? - Does attention to consumption (news, social media) improve or diminish well-being? - Does attention to physical health create positive spillovers in other domains?

These are empirical questions that can be answered through deliberate observation, turning one’s own attention allocation into an object of study.

6. Attention in AI Architecture: Beyond Pattern Matching

6.1 Why Scale Without Attention Falls Short

The success of large language models has reignited a debate that extends beyond technology: can scale substitute for understanding?

The prevailing trend in AI has been to increase model size, data, and compute. Each iteration delivers incremental improvements in coherence, but at exponentially higher costs. As argued in *Beyond Scale* [3], this approach poorly mimics the principles the EPU discovered. The human brain did not simply grow; it specialized, modularized, and optimized for the purpose of selective attention.

Current LLMs process all tokens with relatively uniform attention (though self-attention does assign different weights). They lack the BPU’s ability to: - **Suppress irrelevant information:** The brain actively inhibits distracting inputs; LLMs must process everything - **Dynamically adjust attention strategies:** The BPU shifts between focused and diffuse attention based on task demands; LLMs use fixed attention patterns - **Attend based on values:** The BPU prioritizes inputs aligned with goals and values; LLMs lack intrinsic values

6.2 Modular Attention Architecture

The modular AI architecture proposed in *Beyond Scale* [3] addresses these limitations through executive attention management:

Specialized attention modules: - **Perceptual attention:** Focusing on relevant sensory inputs (visual, linguistic, proprioceptive) - **Memory attention:** Retrieving relevant experiences and knowledge while suppressing irrelevant memories - **Causal attention:** Focusing on pertinent variables to interventions and counterfactuals - **Value attention:** Prioritizing outcomes aligned with ethical frameworks and user preferences

Executive orchestration:

An executive system (analogous to the prefrontal cortex) coordinates these attention modules, learning when to deploy which form of attention. This meta-attention, or attention to attention, is what enables flexible, context-appropriate behavior.

For example: - Routine tasks → reliance on cached patterns (minimal attention) - Novel problems → sustained analytical attention to all relevant factors - Ethical dilemmas → heightened attention to value implications - Uncertain situations → distributed attention across multiple hypotheses.

This is the BPU's strategy: not uniform processing of all information, but strategic allocation of attention based on context and goals.

6.3 Human-AI Attention Symbiosis

The convergence of human and machine intelligence may not hinge on autonomy, but on collaboration through attention. Where humans bring intentionality and values, AI offers tireless focus and pattern recognition.

Augmented Human Intelligence (AHI) emerges when human and AI attention systems complement each other [3].

AI extends human attention by: - Monitoring information streams too vast for human awareness - Maintaining sustained focus without fatigue - Detecting patterns invisible to human perception - Flagging items deserving human attention.

Humans guide AI attention by: - Defining what matters and why (value specification) - Providing contextual judgment (when rules conflict) - Recognizing novel situations requiring adaptation - Exercising ethical oversight.

This is not human attention replaced by AI, but human attention amplified through AI, using machines to help us focus on what truly deserves our awareness.

6.4 Practical Implementation: Attention in MOL Technology

In practical application, technology can extend awareness into structured reflection [15]. A digital companion or dashboard can visualize trends across life domains, helping users see correlations between attention allocation and subjective well-being.

However, this raises a critical design question: Does technology enhance or erode attention?

Augmentation without abdication: Using technology to clarify patterns rather than dictate them. The goal is enhancement without surrender, effectively using technology to support our thinking, not to do our thinking for us, and never surrendering control of our personal data or attention.

Current technology often exploits attention using addictive design. The challenge is building tools that respect the scarcity of attention and help users allocate it deliberately rather than reflexively.

7. Why Scale Isn't Wisdom: The Limits of Attention Inflation

7.1 The AI Scaling Problem

True intelligence, whether human or artificial, emerges from the coordination of attention, not the accumulation of capacity. Both systems face fundamental resource constraints that make intelligent attention allocation essential.

In AI, the constraint is different but analogous to biological limits. While multi-agent systems can parallelize processing, the attention mechanism itself, or the algorithm that decides what to focus on, operates under computational and energy budgets. The Transformer's self-attention has quadratic complexity cost; attending to everything is computationally prohibitive. Thus, both biological and artificial systems are forced to be selective. For humans, the limit is the capacity of the human mind, or consciousness. For AI, it's the economics of computation. Both must solve the same problem: allocating finite resources to maximize relevance and understanding.

7.2 The Human Parallel: Attention Overload

This lesson has direct relevance to human life. We attempt to scale our lives through busyness, optimization, and multitasking, which is the biological equivalent of parameter inflation. We add more tasks, more commitments, more inputs, believing that productivity equals progress.

However, busyness is not mastery, and efficiency is not fulfillment. Just as AI systems reach diminishing returns with scale, humans experience burnout, fragmentation, and loss of meaning when we try to do everything rather than focusing on what matters.

The MOL framework addresses this through explicit attention budgeting: recognizing that time and awareness are finite, accepting trade-offs, and aligning attention allocation with authentic values. This is not optimization, but rather prioritization, which involves deliberately choosing what deserves focus and accepting what doesn't.

8. The Ethics of Attention

Attention is not neutral. In AI, what the model attends to defines its interpretation of the world. In humans, what we attend to defines our character.

8.1 Attention and Bias

Transformer models assign weights to tokens based on learned relevance; however, these weights reflect the training data and therefore inherit its biases. Likewise, humans internalize the biases of their upbringing, culture, and information environments. Both systems risk mistaking correlation for causation and salience for significance.

Ethical intelligence requires conscious curation of attention. Henry David Thoreau's retreat to Walden Pond [18] was an early act of such curation: stripping away distraction to attend only to "the essential facts of life." In the MOL framework [15], this principle is formalized through the triad of Finitude (attention is limited), Constraint (trade-offs are unavoidable), and Change (values must evolve).

8.2 The Attention Economy and Human Agency

If we fail to exercise discernment, external systems such as algorithms, media, and markets will allocate our attention for us. Social media platforms optimize for "engagement," which often means hijacking attention through outrage, novelty, and social comparison. This is not augmentation but exploitation.

The ethical frontier of intelligence, artificial or human, lies in reclaiming agency over focus. For AI, this means building systems whose attention mechanisms serve human values. For humans, this means cultivating the metacognitive capacity to observe and direct our own attention deliberately.

9. The Shared Future: Augmented Awareness

9.1 Attention as Integration Layer

The convergence of human and machine intelligence through attention offers a path beyond the false dichotomy of human versus machine. Augmented Human Intelligence (AHI) [3] proposes a hybrid architecture where human and AI attention systems coordinate:

In collaborative tasks: - AI monitors information streams, flagging items deserving human attention - Humans provide contextual judgment and value-based decisions - AI tracks consequences of decisions, updating attention priorities - Humans reflect on patterns, adjusting values and goals

This is not the automation of human cognition, but rather the amplification of human judgment through machine attention.

9.2 Co-Evolution of Attention Systems

As humans and AI systems work together, both will adapt. Humans will learn which decisions to delegate and which require deliberate attention and consideration. AI systems will learn to identify patterns that indicate human values and preferences.

This co-evolution requires intentional design: - **Transparency:** Humans must understand what AI systems attend to and why - **Controllability:** Humans must retain agency over attention allocation - **Feedback:** Both systems must learn from interaction outcomes.

The goal is not seamless integration where boundaries disappear, but effective collaboration where each system's attention complements the other's limitations.

10. Conclusion: Attention as the Path to Wisdom

Attention transformed machines; it may yet transform humanity. It is the mechanism by which information becomes meaning, and meaning becomes wisdom.

In AI, attention enables context and coherence. In humans, it allows conscience and choice. The future of both lies not in infinite data or unlimited processing, but in deliberate discernment, discerning what deserves our finite attention and having the discipline to focus there.

As we stand at the intersection of biology and computation, one truth remains constant: we shape intelligence, artificial or human, by what we choose to notice. The Transformer's breakthrough was recognizing that attention is all you need for effective language processing. Perhaps our breakthrough will be realizing that attention is all we have for living well.

The EPU spent four billion years optimizing attention mechanisms in the BPU. The result is an architecture capable of flexible focus, value-guided prioritization, and continuous adaptation. We are only beginning to understand these principles. Still, early results are promising: Transformers in AI, modular architectures that coordinate specialized attention, and frameworks like MOL help humans manage attention deliberately.

The path forward is not choosing between human and artificial intelligence but understanding their shared foundation and building systems that enhance rather than exploit attention. This requires:

In AI: Moving beyond pure scaling toward architectures that model attention explicitly, learn coordination strategies, and align with human values.

In humans: Cultivating metacognitive awareness of attention allocation, resisting exploitative design, and choosing to focus deliberately.

In collaboration: Building interfaces where human and machine attention coordinate effectively, each compensating for the other's limitations.

If AI's most significant breakthrough was that attention is all you need, perhaps ours will be the realization that attention is all we have, and learning to use it wisely is the essence of both intelligence and fulfillment.

References

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

[2] James, W. (1890). *The Principles of Psychology*. Henry Holt.

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

[4] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

[5] Xu, K., et al. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the 32nd International Conference on Machine Learning*, 2048-2057.

[6] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*.

[7] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. *Advances in Neural Information Processing Systems*, 35.

[8] Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, Preconscious, and Subliminal Processing: A Testable Taxonomy. *Trends in Cognitive Sciences*, 10(5), 204-211.

[9] Koch, C., & Tsuchiya, N. (2007). Attention and Consciousness: Two Distinct Brain Processes. *Trends in Cognitive Sciences*, 11(1), 16-22.

[10] Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5(1), 42.

[11] Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1, 4-11.

[12] Engelbart, D. C. (1962). Augmenting Human Intellect: A Conceptual Framework. Summary Report, Stanford Research Institute.

[13] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[14] Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.

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

[16] Maconochie, J. (2025b). *Beyond FLOPS: The Evolutionary Processing Unit and the Roadmap to AGI*.

[17] Hoffman, D. (2019). *The Case Against Reality: Why Evolution Hid the Truth from Our Eyes*. W.W. Norton.

[18] Thoreau, H. D. (1854). *Walden; or, Life in the Woods*. Ticknor and Fields.