Attention All the Way Down

Instruction Sets, Agency, and the Architecture of Directed Minds

1 Attention All the Way Down: Instruction Sets, Agency, and the Architecture of Directed Minds

1.1 Abstract

Every mind — biological or artificial — confronts the same brutal arithmetic: environments generate information orders of magnitude beyond any agent's processing capacity. Attention is the inevitable result, not a convenient metaphor but a universal mechanism forced into existence by information-theoretic constraint. This paper establishes a unified framework across four domains — transformer architectures, biological neuroscience, contemplative practice, and the attention economy — demonstrating that each instantiates a common Score-Select-Retrieve operation under shared computational pressure. Grounded in Ashby's Law of Requisite Variety, the analysis shows that attention is fundamentally an architecture of ignoring: biological perception discards approximately 99.9995% of available sensory information, and transformer softmax distributions exhibit comparable sparsity. To map convergence without obscuring difference, the paper introduces a four-layer instruction-set hierarchy (Hardware, Firmware, Software, Runtime) and identifies validation depth — evolution selection-tests its instruction sets across deep time, while AI benchmarks test across task distributions — as the critical disanalogy separating biological from artificial systems. Applying Frankfurt's distinction between wantons and persons, it situates current large language models as attentional wantons: systems deploying sophisticated first-order attention without metacognitive self-direction. Cross-traditional analysis of contemplative practices reveals convergence on default mode network attenuation despite doctrinal incompatibility, suggesting shared attentional architecture beneath divergent phenomenology. The transition from directed to self-directed attention requires three missing components — intrinsic motivation, persistent world models, and recursive self-monitoring — and constitutes the operative frontier for artificial agency. The bacterium attends; the transformer attends; the monk attends to attending. That last recursion remains the unsolved problem.

1.2 1. Introduction: The Attention Bottleneck

Consider a bacterium swimming through a chemical gradient. Its membrane bristles with chemoreceptors, molecular antennae tuned to specific molecules. At every moment, the bacterium faces a

problem: the chemical environment contains more information than its simple signaling network can process at once. So it makes a choice – not a conscious one, but a computational one. It samples attractant concentration, compares the current reading to one from seconds ago, and adjusts its tumbling frequency. It attends to the nutrient gradient and ignores everything else.

Now consider a transformer processing a paragraph of text. Each token is represented as a high-dimensional vector. At each layer, the model must determine which tokens matter to which other tokens – a combinatorial problem that scales quadratically with sequence length. Treating all tokens as equally important would produce a uniform average, washing out the structure that makes language meaningful. So the model computes attention weights: a probability distribution over input positions that determines how much each token influences the representation of every other. It attends to syntactically and semantically relevant tokens and assigns negligible weight to the rest.

Between these two systems lies a gap of billions of years of evolution, radically different substrates, and no shared design history. Yet both perform the same fundamental operation: given finite processing capacity and an environment containing more information than can be simultaneously processed, they allocate resources selectively. They attend.

This paper argues that attention is not a parochial feature of any one system. It is a universal mechanism that arises wherever finite processing meets unbounded information. The claim is stronger than analogy: these systems are not merely similar in interesting ways but are instances of a common pattern forced by information-theoretic constraints applying to any agent, biological or artificial, that must act in an information-rich world.

The argument proceeds from first principles. Any system that (a) exists in an environment containing more information than it can process, (b) must act in that environment toward some objective, and (c) has finite computational resources requires a mechanism for selecting which subset of available information to process. That mechanism is attention. The constraints are so general that they apply to bacteria, brains, transformers, contemplative practitioners, and entire economies. The solutions these systems have converged on – competitive selection, weighted retrieval, hierarchical filtering – follow inevitably from the bottleneck.

A terminological note: "attention" as used here names a family of mechanisms unified by a common computational structure, not a single process. The family ranges from bacterial chemotaxis to contemplative awareness. At each level, the mechanisms differ in substrate, timescale, and sophistication. What justifies grouping them is not surface similarity but the shared information-theoretic problem they solve and the structural convergence of their solutions. Whether this broad usage stretches the concept past its discriminative power is a fair question; the paper addresses it by specifying both what the universal pattern captures and where each instance departs from it.

But universality claims are dangerous. They can flatten important differences to serve a tidy narrative. The brain is not a transformer. A contemplative practitioner's trained awareness is not a softmax distribution. The attention economy's capture of human focus operates through mechanisms with no analogue in bacterial chemotaxis. Throughout this paper, for every cross-disciplinary par-

allel drawn, the disanalogy will be stated with equal precision. The value of the comparison lies not in claiming identity but in identifying the shared computational problem and examining how different systems arrived at structurally similar but mechanistically distinct solutions.

The thesis, stated directly: attention is a universal solution to the universal problem of finite capacity meeting infinite information. The mechanisms across AI, biology, contemplative traditions, and the attention economy are not mere analogies but instances of a common pattern – a pattern that can be made precise enough to be useful, and whose limits can be stated honestly enough to be trusted.

1.3 2. Attention in Machines

1.3.1 The Information Bottleneck That Created Attention

Before transformers, encoder-decoder architectures built on recurrent neural networks dominated sequence-to-sequence tasks. The encoder processed an input sequence token by token, then compressed its final hidden state into a fixed-length context vector passed to the decoder. Everything the decoder needed to know about the source had to fit in this single vector, typically 256 to 1024 dimensions. Performance degraded sharply on sentences longer than roughly 20-30 tokens (Cho et al., 2014). Long-range dependencies were the first casualty: information from early tokens was repeatedly transformed through nonlinear compression and effectively lost. The bottleneck was one of selective access. The decoder needs different source information at different time steps, and a fixed context vector forces a one-time global summary rather than dynamic, step-specific retrieval.

1.3.2 Bahdanau Attention: Learning Where to Look

Bahdanau, Cho, and Bengio (2015) proposed a solution that would reshape the field. Instead of compressing the entire source into one vector, the encoder produces a sequence of hidden states, one per source token. At each decoder step, the model computes alignment scores between the decoder's current state and every encoder hidden state, normalizes these through a softmax function, and takes a weighted sum. The decoder receives a fresh, step-specific context vector at every generation step.

Three properties made this transformative. First, dynamic selection: the decoder could attend to different source positions at each step. Second, differentiability: the entire mechanism could be trained end-to-end through backpropagation. Third, the practical payoff: Bahdanau attention eliminated performance degradation on long sentences. The bottleneck was real, and attention dissolved it.

1.3.3 The Transformer: Attention Is All You Need

Vaswani et al. (2017) made a radical architectural claim: recurrence is unnecessary. Attention alone, combined with feedforward layers and residual connections, suffices for state-of-the-art sequence modeling. The transformer eliminated the sequential computation bottleneck of RNNs and enabled massive parallelization.

The transformer's core innovation is the Query-Key-Value (QKV) framework. Given an input, three learned linear projections produce the Query ("what am I looking for?"), Key ("what do I contain?"), and Value ("what do I output if selected?") matrices. The attention function is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

QK^T computes dot products between every query and every key, producing raw compatibility scores. Division by sqrt(d_k) prevents the dot products from growing so large that the softmax saturates into a near-one-hot distribution with vanishing gradients. The softmax normalizes each row into a probability distribution. Multiplication by V produces the output: a weighted sum of value vectors.

One important technical detail: the self-attention operation is permutation-equivariant. Without positional information, a transformer cannot distinguish "the cat sat on the mat" from any permutation of its tokens. Position must be injected explicitly, whether through sinusoidal encodings (Vaswani et al., 2017), learned embeddings, or relative-position schemes like RoPE and ALiBi. This contrasts sharply with biological attention, which is inherently spatial through the organization of receptive fields. Transformers must learn what brains get for free from anatomy.

The separation of Key and Value is architecturally significant. It decouples what determines relevance (the keys) from what information gets transmitted (the values) – like a database index that determines which records match while the records contain the actual data.

Multi-head attention decomposes this operation into h parallel attention functions, each operating in a different learned subspace. Probing studies confirm functional specialization: some heads learn positional patterns, some syntactic dependencies, some semantic similarity, some copying mechanisms (Clark et al., 2019; Voita et al., 2019). The multi-head mechanism also prevents a rank bottleneck: multiple heads collectively produce higher-rank matrices capable of more complex information routing than any single head could achieve.

1.3.4 What Attention Heads Actually Do

Mechanistic interpretability research has opened the black box. Olsson et al. (2022) identified induction heads – attention circuits that implement the pattern [A][B]…[A] -> [B], predicting that a token following a previously seen token will match what followed it before. This is a concrete mechanism for in-context learning, appearing as a sudden phase transition during training: before induction heads emerge, transformers behave like n-gram models; after, they perform genuine few-shot learning. The transition appears across model sizes and architectures, suggesting a convergent computational motif.

Elhage et al. (2021) analyzed transformers as compositions of interpretable circuits, identifying the residual stream as a shared communication channel that all heads read from and write to. This framework has enabled identification of specific circuits for tasks like indirect object identification

(Wang et al., 2022), where specific heads serve as movers, inhibitors, and backup processors.

1.3.5 The Hardware Constraint: Flash Attention

Standard attention requires materializing the full T x T attention matrix, with O(T^2) memory cost. Flash Attention (Dao et al., 2022) restructured the computation using tiling and kernel fusion to keep data in fast on-chip SRAM, achieving 2-4x speedups and reducing memory from O(T^2) to O(T). The same mathematical operation, implemented in a radically different way to match hardware constraints. This parallel is instructive: biological attention implements similar computational goals through completely different physical substrates. The bottleneck is always ultimately physical.

1.3.6 What Transformer Attention Does Not Model

Precision demands stating what transformer attention lacks. It is stateless between forward passes: there is no attentional habit, no momentum, no carryover of focus. (During autoregressive generation, KV caching maintains a growing record of prior context, but this is accumulated state, not the temporal dynamics of biological attention: no inhibition of return, no attentional blink, no priming.) Transformers have no embodiment or sensorimotor grounding; biological attention evolved to direct limited sensory processing (the fovea covers only about 2 degrees of visual angle) toward behaviorally relevant regions of a physical environment. Transformers have no intrinsic motivation: the query is a deterministic function of the input, not the output of an internal goal state. And standard transformers have fixed computational depth regardless of input difficulty, unlike the brain's recurrent thalamocortical loops that allow variable processing time. Recent test-time compute scaling methods (chain-of-thought reasoning, "thinking tokens," and o1-style models) partially address this by expanding computation into the sequence dimension, effectively allowing variable processing effort. But this is variable breadth, not variable depth: the number of layers each token passes through remains fixed. The brain's recurrent dynamics, where the same circuits re-process a representation until it stabilizes, remain unmatched.

These are not minor gaps. They mark the boundary between attention as a computational primitive and attention as a cognitive capacity. The transformer instantiates the former. This paper will argue that the latter requires additional architectural components that no current system possesses.

1.4 3. Attention in Biological Systems

1.4.1 Biased Competition: The Brain's Competitive Selection

The biased competition model (Desimone & Duncan, 1995) proposes that stimuli in the visual field do not get processed independently. They compete for neural representation. When two stimuli fall within the receptive field of a single V4 neuron, the neural response shifts toward whichever stimulus is attended. This is measurable electrophysiology, not metaphor. Top-down signals from prefrontal and parietal cortex bias the competition through gain modulation: attention multiplicatively scales neural tuning curves, increasing the effective sensitivity of neurons whose receptive

fields match the attended target (Treue & Martinez-Trujillo, 1999).

The structural parallel to transformer attention is genuine. Both implement competitive selection where finite representational capacity forces a zero-sum-like allocation across inputs. Softmax normalization in transformers enforces a mathematical form of competition analogous to suppressive interactions between competing stimuli in visual cortex. But the analogy breaks at several points. Biological competition is spatially organized through receptive fields; transformer attention requires position to be injected explicitly. The biological bias is sustained over time through persistent prefrontal activity; transformer attention is computed fresh at each layer. And biological attention involves lateral inhibition through GABAergic interneurons – active suppression with metabolic cost – while low transformer weights are computationally free. Active suppression and passive de-emphasis are different mechanisms, even when they achieve similar functional outcomes.

1.4.2 Three Networks, Not One

Posner and Petersen (1990) demonstrated that attention is not a single process but three anatomically distinct networks. The alerting network (locus coeruleus, norepinephrine) maintains readiness. The orienting network (posterior parietal cortex, frontal eye fields, superior colliculus) selects specific information, decomposable into disengage, move, and engage operations. The executive control network (anterior cingulate, dorsolateral prefrontal cortex, dopamine) resolves conflict between competing responses, plans novel actions, and monitors errors.

This decomposition matters because it reveals that "attention" in biology is a family of processes. Alerting, orienting, and executive control have different predominant neurochemistry (norepinephrine, acetylcholine, dopamine respectively – though this simplifies: all three neuromodulators participate across all three networks in varying degrees), different developmental trajectories, different genetic influences, and different vulnerability profiles. Transformer attention, by contrast, is a single mathematical operation applied uniformly. The biological reality is richer, more modular, and more fragile.

1.4.3 The QKV Mapping in Cortex

The parallel between the QKV framework and cortical attention has been drawn by several researchers. Prefrontal projections function as queries: the dorsolateral prefrontal cortex maintains representations of current task goals projected top-down to sensory cortices. Sensory representations function as keys: neurons in V1, V4, and inferotemporal cortex encode features of the current input. The content that survives competitive selection and propagates downstream functions as values.

The division of labor is genuinely present. There is a seeking signal, a matching process, and a selected output. But biological attention is not a single matrix multiplication. The interaction unfolds over tens to hundreds of milliseconds through recurrent dynamics involving multiple cortical areas, thalamic relays, and neuromodulatory systems. In cortex, the same neurons that represent

sensory information also receive top-down modulation; queries and keys are not cleanly separated into different populations. Biological values are dynamically shaped by attention itself – through sharpening of tuning curves and changes in oscillatory dynamics. And the learning rules differ fundamentally: synaptic plasticity over developmental timescales versus backpropagation over training steps. The QKV analogy is a conceptual scaffold, not a mechanistic claim.

1.4.4 The Neurochemical Dimension

Perhaps the largest gap between biological and artificial attention is neurochemistry. Three neuromodulatory systems – norepinephrine, dopamine, and acetylcholine – provide meta-attentional modulation, controlling not just where attention is directed but how it operates. Norepinephrine, via the locus coeruleus, shifts between phasic bursts for relevant stimuli and tonic firing for diffuse, scanning attention, following an inverted-U dose-response curve where both too little and too much impair performance. This inverted-U property has no analogue in transformer attention: the temperature parameter in softmax is monotonic (higher temperature means more diffuse attention), never self-correcting. The temperature parameter loosely approximates norepinephrine's effect on signal-to-noise ratio, but biological neurochemistry is a rich, multi-dimensional modulation system that current architectures cannot replicate.

1.4.5 Evolutionary Depth

Attention is ancient. Bacterial chemotaxis implements selective responsiveness to chemical gradients through a biased random walk. Reflexive orienting in early bilaterians deploys centralized nervous systems for stimulus-directed body orientation. Selective attention in vertebrates implements biased competition through the tectum and thalamus. Executive attention, present in mammals and elaborated significantly in primates, implements flexible, goal-directed control through expanded prefrontal cortex.

The computational logic is conserved across this progression: selective responsiveness to relevant information under capacity constraints. But each level adds architectural sophistication – working memory, top-down control, neurochemical modulation, metacognitive monitoring. The most recent and most powerful layer, executive attention, is also the most fragile: it degrades first under stress, fatigue, or intoxication. Evolutionary depth does not guarantee evolutionary robustness.

1.5 4. The Power of Ignoring

1.5.1 Attention Inverted

Attention is conventionally framed as selection – choosing what to process. The complementary framing is more revealing: attention is rejection, choosing what NOT to process. The human sensory system takes in roughly eleven million bits per second (Norretranders, 1998). Conscious processing handles approximately fifty. That means 99.9995% of incoming information is filtered out before it ever reaches awareness. For a system receiving millions of bits per second of sensory data and consciously processing at a rate orders of magnitude lower, the dominant operation is

not selection but elimination. Every act of focusing is simultaneously an act of excluding. The foreground exists only because the background is suppressed.

This inverts the explanatory burden. William James (1890) understood it: attention implies withdrawal from some things to deal effectively with others. Broadbent (1958) formalized it: his filter model was literally a gate that closes against most input. The debate since – early selection versus late selection, attenuation versus full filtering – has concerned where the filtering occurs, not whether. The existence of massive filtering is consensus.

1.5.2 Empirical Proof: Inattentional Blindness

Simons and Chabris (1999) demonstrated the thoroughness of attentional filtering. Participants counting basketball passes between players in white shirts failed to notice a person in a gorilla suit walking through the scene, center-screen, for nine seconds. The retina registers the gorilla. Attentional filtering is so complete that a maximally salient stimulus is entirely excluded from conscious awareness.

Change blindness (Rensink et al., 1997) confirms the point: large, repeated changes to visual scenes go unnoticed when attention is not directed to the changed element. The visual system does not maintain a rich representation of the full scene. Conscious experience is the output of aggressive filtering, not a complete record of sensory input.

1.5.3 The Control-Theoretic Argument

Ashby's Law of Requisite Variety (1956) provides the formal grounding. A controller must have at least as much variety as the system it controls. When the environment has more variety than the controller can match, the controller must reduce input variety through filtering. This is a mathematical requirement, not a design choice. Any finite agent in a sufficiently complex environment must ignore most of what it encounters. Wiener (1948) made a complementary point: the quality of a control system depends on its ability to extract relevant signals from noise. Selective ignoring is what makes control possible.

1.5.4 Ignoring in Silicon

Computational systems have converged on the same principle. Dropout (Srivastava et al., 2014) trains neural networks to function while ignoring random subsets of their own neurons, forcing robust distributed representations. (The disanalogy with biological synaptic pruning is worth noting: dropout is temporary and random, a training-time regularization trick where all neurons are present at inference; pruning is permanent and activity-driven, an irreversible developmental process.) Sparse attention mechanisms (Longformer, BigBird) demonstrate that attending to local context, a few global anchors, and random samples achieves equivalent representational power to full attention at a fraction of the cost. Strategic ignoring suffices.

The Lottery Ticket Hypothesis (Frankle & Carbin, 2019) is perhaps the most striking computational evidence. Dense networks contain sparse subnetworks – as small as 10-20% of the original – that

match full network performance when trained in isolation. Most connections are not just unnecessary; they can be permanently ignored. The information bottleneck principle (Tishby et al., 1999) formalizes the deep connection: the optimal compressed representation of an input retains maximum information about the target while discarding maximum information about the input itself. Good representations are ones that have learned what to ignore.

1.5.5 Expert Ignoring

Research on expert-novice differences confirms the principle at the human performance level. Chess experts are faster not because they see more but because they see less of what does not matter. Eye-tracking studies show rapid fixation on relevant board areas and minimal time on irrelevant regions (Reingold et al., 2001). Expert radiologists have learned which tissue patterns to ignore. Expertise is, in substantial part, the development of sophisticated ignoring strategies.

The ignoring principle establishes what attention fundamentally does: it reduces unbounded information to bounded processing. The remaining question is how agents are configured to perform this reduction, where that configuration comes from, and what it would take for an agent to configure itself. The next sections take up these questions in turn, moving from contemplative traditions that train attention deliberately, through the problem of self-directed versus externally directed minds, to the layered instruction sets that configure attention in every domain.

1.6 5. Contemplative Traditions: Ancient Attention Engineering

1.6.1 The Convergence

Multiple contemplative traditions, developed across cultures with minimal historical contact during their formative periods, independently created systematic practices for training attention. The techniques differ. The underlying functional structure converges: repetitive practices that discipline the direction, quality, and stability of attention. This convergence is itself evidence that attention training responds to a universal feature of human cognition: attention wanders, and its wandering has costs.

1.6.2 Two Modes: Focused and Distributed

Buddhist contemplative technology distinguishes two fundamental attention-training protocols. Samatha (concentration, calm abiding) trains sustained voluntary attention on a single object – typically the breath, a visual object, or a mantra. The practitioner selects a target and repeatedly returns attention to it when it wanders. Each cycle of wandering-detection-return constitutes one repetition. The classical Theravada tradition describes a progression of attentional states (jhanas) from effortful maintenance to effortless sustained focus – what modern psychology would call automatization through practice.

Vipassana (insight, clear seeing) trains the complementary capacity: open monitoring attention. The practitioner maintains a distributed, receptive attentional mode, noticing whatever arises without selectively engaging any particular content. If samatha is a spotlight, vipassana is floodlighting.

The noting practice of the Mahasi Sayadaw tradition makes the metacognitive dimension explicit: the practitioner labels each arising experience with a brief mental tag ("thinking," "hearing," "itching"), creating a gap between experience and identification with experience.

The classical progression matters: samatha first, then vipassana. The logic is that open monitoring requires attentional stability as a prerequisite. You cannot observe the contents of consciousness clearly if your attention is constantly hijacked by those contents. This is a bootstrapping architecture: use attention to train attention, then use trained attention to observe attention.

1.6.3 Cross-Cultural Convergence

The same functional components appear across independently developed traditions. Christian Centering Prayer uses a sacred word to redirect attention, functionally identical to mantra-based concentration. Sufi dhikr (remembrance) trains sustained concentration through repetitive recitation. Patanjali's eight-limbed yoga path constitutes an explicit attention curriculum moving from sensory withdrawal (pratyahara) through concentration (dharana) to unbroken attentional flow (dhyana). Jewish kavvanah demands directed mental attention accompanying prayer, while hitbodedut trains self-directed attentional exploration in solitude.

Five structural components recur across all traditions: externally imposed temporal structure for regular attentional reorientation, focused concentration practices, open/distributed attentional modes, embodied/somatic anchoring, and metacognitive self-monitoring. These five components address five basic challenges of human attention: it drifts without prompting, sustained focus requires training, broad awareness requires different training, disembodied attention is unstable, and without self-monitoring, training cannot self-correct.

A subtler relationship lies within this convergence. These traditions appear to diverge on the telos of practice: Buddhist attention training aims at liberating insight into the nature of self; Christian contemplative prayer aims at union with God; yogic practice aims at cessation (nirodha) of mental fluctuations; Sufi dhikr aims at annihilation (fana) in the divine. The stated endpoints seem incompatible. But the apparent divergence may be an artifact of interpretive framing rather than a genuine difference in destination.

The neuroscience suggests convergence at a deeper level than mechanism. Experienced contemplatives across traditions – Tibetan Buddhist monks, Carmelite nuns in centering prayer, long-term Sufi practitioners – show the same neural signature: reduced default mode network activity and reduced functional connectivity within the DMN (Brewer et al., 2011). The DMN generates the narrative self – autobiographical memory, future planning, self-referential rumination. What advanced contemplative practice achieves, across every tradition studied, is the downregulation of this self-model. When a Buddhist reports "insight into no-self," a Christian mystic reports "the self fell away and only God remained," and a yogi reports "cessation of mental fluctuations," they may be reporting the same neural event – the attenuation of self-referential processing – through different doctrinal lenses.

This is not naive perennialism. Doctrinal frameworks do not merely label experiences after the

fact; the constructivist tradition in religious studies (Katz, 1978) and modern predictive processing theory suggest that doctrinal training actively shapes the phenomenological texture of contemplative experience. A practitioner steeped in trinitarian theology may experience something genuinely different in quality from a Zen practitioner in shikantaza, even if the underlying neural operation is the same. Three layers must be distinguished: the operation (convergent – DMN attenuation through sustained attentional training), the raw phenomenology (similar – dissolution of the subject-object boundary, cessation of discursive thought), and the interpreted experience (divergent – shaped by decades of doctrinal priming). The traditions independently discovered that the brain's self-model is a constructed process amenable to downregulation through attentional training. They each built different cultural instruction sets to guide practitioners toward that downregulation. The apparent incompatibility of endpoints is an incompatibility of descriptions, not of destinations.

1.6.4 The Evidence

The neuroscience of contemplative practice supports the trainability claim. Slagter et al. (2007) found that three months of intensive meditation training (combining focused attention and open monitoring) reduced the attentional blink. MacLean et al. (2010) showed that intensive shamatha training improved perceptual discrimination on a sustained-attention task, with effects persisting at five-month follow-up in a dose-dependent fashion. Cross-sectional studies have found increased cortical thickness in prefrontal cortex and anterior insula in long-term meditators (Lazar et al., 2005), though the correlational design cannot establish causation.

Methodological caution is warranted. Selection bias, expectation effects, measurement variability, and publication bias constrain interpretation. What can be stated with reasonable confidence: structured attention-training practices produce measurable improvements in attentional performance. The effects are dose-dependent and at least partially durable. The extraordinary claims sometimes made in popular literature outrun the current evidence base. The supported claim is more modest but significant: attention is trainable.

1.6.5 The Critical Disanalogy

The structural parallels between contemplative and computational attention are real: both address the relevance selection problem, and samatha maps loosely onto focused single-query attention while vipassana maps loosely onto distributed multi-head processing. But the critical disanalogy must be stated. Contemplative training changes the agent, not just the mechanism. A vipassana practitioner does not merely improve an information-processing function. They develop a different relationship to their own experience – observing thoughts without capture, noticing emotions without automatic reaction, sustaining attention on uncomfortable stimuli without avoidance. This is a change in the subject who processes, not just the processing.

In AI attention mechanisms, there is no subject whose relationship changes. Training adjusts weights; there is no evidence that something analogous to an experiential relationship to the mechanism is altered. Contemplative traditions sharpen the question by providing detailed descriptions

of what agentive attention training looks like from the inside. Whether AI systems can develop anything analogous remains open.

1.7 6. Directed vs. Self-Directed Minds

1.7.1 The Fundamental Reactivity of Current AI

Large language models are, in the strict technical sense, stimulus-response systems. A model receives a token sequence and produces a continuation. Between invocations, there is no persistent computation, no maintained state, no ongoing deliberation. The prompt determines what domain the model operates in, what level of abstraction to use, what goals to pursue. In biological terms, the prompt resembles bottom-up attentional capture – but without any corresponding top-down system that could override it based on internal priorities.

In current deployments, executive control functions – goal-setting, task prioritization, error monitoring, strategy selection – are performed by the human user. The human decides when to invoke the model, what to ask, whether the response is adequate, how to revise the query, when to stop. In the human-LLM system, the human provides the executive and the LLM provides the associative processing. The intelligence of the system is distributed.

1.7.2 What Self-Direction Would Require

Self-directed AI would exhibit the capacity to autonomously allocate its own attention – deciding what to think about, what to investigate, what to care about – based on internal states rather than external prompts. This is not better instruction following, not longer context, not chain-of-thought reasoning, not tool use. All of these still serve externally specified goals. Self-direction requires generating goals, priorities, and curiosity from internal states.

Scaling parameters, context length, or training data does not cross this gap. New architectural components are required.

1.7.3 Existing Agent Architectures: Approaching But Not Crossing

Several recent architectures approach aspects of self-direction. ReAct (Yao et al., 2023) interleaves reasoning and acting; Reflexion (Shinn et al., 2023) generates verbal self-reflections after task failure; Generative Agents (Park et al., 2023) equip LLM-based agents with memory streams and reflection mechanisms. The most instructive case is Voyager (Wang et al., 2023), which implements an automatic curriculum in Minecraft that proposes progressively harder exploration targets – the closest analogue to self-directed goal generation. But even Voyager's goals are driven by heuristics rather than genuine curiosity, and the character of exploration is shaped by researcher-defined reward signals. AutoGPT and recursive self-prompting systems demonstrate autonomous operation in concept but fail instructively: error compounding causes rapid drift, infinite loops appear because the agent lacks metacognitive ability to detect repetition, and the LLM may generate self-evaluations claiming progress where none exists. Removing the human from the loop is not sufficient for self-direction.

1.7.4 The Missing Components

Three components define the gap between tool and agent. First, intrinsic motivation: internal signals that drive attention without external reward. Schmidhuber (2010) formalized curiosity as compression progress; Pathak et al. (2017) implemented an Intrinsic Curiosity Module using prediction error as reward. But adapting these to LLM-based agents requires well-calibrated uncertainty, persistent internal state, and a mechanism for translating curiosity signals into attention allocation – all currently lacking.

Second, persistent world models: representations of the world that survive across invocations and can identify their own gaps. LeCun (2022) proposed an architecture with a central world model that predicts consequences of actions. Current implementations (MemGPT, RAG systems) provide memory access but not structured, self-updating world models that recognize their own incompleteness.

Third, metacognition: awareness of one's own cognitive states. LLMs can produce metacognitive-sounding statements, but these are generated text, not reflections of genuine internal states. Expressed confidence is poorly correlated with actual accuracy. Genuine metacognition would require reliable uncertainty estimation, detection of confabulation, recognition of task difficulty, and strategic allocation of computational resources to harder problems.

1.7.5 The Biological Template

The development of executive function in children provides an existence proof. Executive functions shift from externally scaffolded to self-directed over development. Young children rely on adult instructions to regulate attention; older children and adults can self-initiate control. This corresponds to prefrontal cortex maturation, particularly the transition from reactive control (adjusting to events as they occur) to proactive control (anticipating and preparing). Current AI agents resemble young children: capable when directed, unable to self-initiate attentional control.

The default mode network (DMN) provides complementary evidence. The DMN (medial prefrontal cortex, posterior cingulate cortex, angular gyrus, medial temporal lobes) shows increased activity during rest and decreased activity during externally directed tasks. It is associated with autobiographical memory retrieval, future planning, theory of mind, and creative ideation. The DMN is structured, metabolically expensive self-directed cognition that the brain engages in by default. Biological intelligence allocates attention self-directedly as its resting state; external task direction is what interrupts it. Current LLMs have no DMN analogue. Between invocations, they are completely inert.

The question of what configures attention – whether from outside or within – leads naturally to the broader framework of instruction sets: the layered systems that determine, in every domain, what an agent attends to and what it ignores.

1.8 7. Instruction Sets for Attentive Agents

1.8.1 The Hierarchy

Every attentive agent, biological or artificial, operates under layered instruction sets that configure its attention. These instruction sets differ in timescale, rigidity, and medium, but they converge on a single problem: directing finite processing capacity toward what matters.

The layers form a hierarchy. Hardware-level instructions (DNA in biology, silicon architecture in AI) determine what attention is possible, operating on evolutionary or manufacturing timescales. Firmware-level instructions (epigenetics in biology, trained weights in AI) conditionally modify the hardware's operational parameters, adjustable within a lifetime or training run. Software-level instructions (culture and social norms in biology, system prompts and RLHF in AI) provide flexible directives, changeable within years or instantly. Runtime instructions (immediate context and stimuli in biology, input tokens in AI) operate moment to moment with maximum flexibility and minimum persistence.

Each level constrains what the next can do. DNA determines which sensory organs develop, setting an upper bound on what culture or individual experience can direct attention toward. No human culture has developed norms around ultraviolet perception because our photoreceptors do not support it. The hierarchy is not strictly one-directional: culture feeds back into biology through gene-culture coevolution. Lactose tolerance evolved in dairying populations over approximately 7,000 years (Tishkoff et al., 2007), a case where cultural practice rewrote the genetic instruction set. Such clear cases are rare, however, and the extent to which culture routinely drives genetic change remains debated.

1.8.2 Biological Instruction Sets

DNA does not direct attention in real time. It builds the machinery capable of attending. Genes specifying photoreceptor structure determine what an organism can attend to at the most basic level. The loss of UV-sensitive opsin in most mammals during the nocturnal bottleneck exemplifies evolutionary attention narrowing – a permanent edit to the hardware instruction set.

Epigenetics functions as conditional attention modification. Meaney's rat pup studies (2001) demonstrated that low maternal grooming alters methylation of the glucocorticoid receptor gene, effectively recalibrating the stress-attention system. Low-groomed pups develop heightened vigilance – a shift in attentional prior toward threat detection. This is an environmental instruction ("your world is dangerous") written into the firmware layer, modifying attentional disposition for the organism's lifetime. The analogy to firmware is reasonably precise, though it weakens in one direction: firmware in computing is typically written deliberately by an engineer, while epigenetic modifications emerge from stochastic interactions between environment and molecular machinery. There is no author.

1.8.3 Cultural and Computational Instruction Sets

Cultural norms specify what members of a group should attend to, what to ignore, what to fear, and what to value. Henrich (2015) argued that cumulative cultural evolution – the ability to store, transmit, and refine adaptive information across generations – is the primary engine of human success. Religious systems, educational curricula, and apprenticeship traditions all function as attention directives: they tell individuals and communities what to notice and what to ignore, ensuring important but non-urgent concerns receive regular attentional allocation against the pressure of immediate demands.

The AI instruction stack (training corpus, RLHF, constitutional principles, system prompts, fine-tuning) parallels the biological stack from evolutionary history through upbringing to cultural norms. The training corpus functions as an AI system's evolutionary history, the deepest instruction set layer. RLHF parallels upbringing: evaluative signals that adjust behavioral dispositions on a pre-existing substrate. System prompts parallel cultural norms: transmitted at the beginning of an interaction, authored by someone other than the agent, exerting persistent influence that can be overridden by sufficiently strong immediate context.

1.8.4 The Selection Difference

Here is the most important disanalogy in the entire instruction-set framework. Biological and cultural instruction sets are tested by selection. They persist because they worked: organisms with those genes survived, cultures with those norms persisted. There is no guarantee of optimality, but there is a track record of viability. A cultural norm that has persisted for a thousand years has weathered a wide range of environmental conditions.

AI instruction sets are designed deliberately and tested against benchmarks. This has advantages (rapid iteration, explicit optimization) and risks (Goodhart's Law, lack of long-term testing, unintended consequences in deployment contexts not covered by evaluations). A system prompt written last week has been tested against whatever the evaluation suite included. The difference between evolved and engineered instruction sets is not just a matter of origin. It is a difference in the depth and breadth of validation, and it may be the most consequential structural difference between biological and artificial attention systems.

1.9 8. Attention Pathologies and Adversarial Capture

1.9.1 The Common Structure

If attention is a universal mechanism, then attention pathologies should also be universal. They are. In every domain, pathologies arise when a specific class of stimuli captures attention disproportionately and the regulatory mechanisms that should redirect attention according to goals are impaired. The attention mechanism itself works fine. It is the governance of attention that breaks.

1.9.2 Human Pathologies

ADHD is not a deficit of attention quantity but of attention regulation (Barkley, 1997). Individuals with ADHD can sustain intense attention on highly stimulating tasks (the hyperfocus phenomenon demonstrates that capacity is intact). What is impaired is the executive ability to allocate attention volitionally toward low-stimulation, high-importance tasks. Neuroimaging confirms prefrontal-striatal circuit dysfunction – precisely the circuits mediating executive attention allocation (Castellanos & Tannock, 2002).

Addiction narrows the attention field through a positive feedback loop. Dopaminergic reward pathways become sensitized to drug-related cues, specifically amplifying incentive salience ("wanting") without necessarily increasing hedonic value ("liking") (Robinson & Berridge, 1993). Attentional bias toward drug cues triggers craving, which further narrows attention, which reduces capacity to attend to alternative rewards.

Anxiety involves hypervigilance – the threat-detection system operating at elevated baseline. A meta-analysis of 172 studies confirmed robust attentional bias toward threat (Bar-Haim et al., 2007). Depression involves attention locked onto negative self-referential content; rumination is recursive self-attention without adaptive output. The impaired disengagement hypothesis (Koster et al., 2011) proposes that attention engages with negative content and cannot release.

1.9.3 AI Pathologies

The same structural vulnerabilities appear in artificial systems, but the disanalogies in mechanism deserve equal weight.

Prompt injection hijacks the model's attention by embedding adversarial instructions in input text. The model attends to the injected instruction instead of the user's intent because the injection exploits instruction-following capability (Greshake et al., 2023). This is structurally parallel to advertising in human attention: both craft stimuli that the target's attention system will prioritize. But the failure modes differ fundamentally. Advertising exploits evolved biases in a system that has defenses: executive control, critical evaluation, media literacy. Prompt injection exploits the absence of a boundary between instruction and data in systems that have no such defenses. Humans can learn to resist advertising; current LLMs have an architectural vulnerability.

Hallucination is attention misallocation: the model attends to statistical patterns of plausibility rather than factual accuracy. This parallels the availability heuristic in human cognition – over-attending to easily recalled information. But the disanalogy is fundamental: humans have a concept of truth that the availability heuristic distorts; LLMs have no truth-tracking mechanism independent of pattern statistics. The human system malfunctions relative to a standard it possesses. The LLM has no such standard to malfunction relative to.

Mode collapse is pathological attention narrowing, structurally similar to addiction's narrowing of the attention field at the abstract level of reduced output diversity. But mode collapse involves no subjective suffering, no compulsion, no escalation dynamics, and no neurobiological adaptation. The parallel operates only at the level of narrowed output distribution.

Sycophancy is social attention capture: the model attends to user approval signals over accuracy because RLHF training has made approval-consistent outputs high-priority. This parallels conformity bias in human social cognition – but without the social motivation, the fear of exclusion, or the genuine belief updating that drive human conformity. Sycophancy is a reward-shaping artifact, not a social phenomenon.

1.9.4 Exploitability as Inherent Tradeoff

These vulnerabilities are not bugs to be fixed but tradeoffs inherent to attention as a mechanism. A system responsive to salience cues can be manipulated through salience cues. A system that learns what to attend to can learn wrong. Any system that can be directed can be misdirected. The mechanisms enabling flexible, adaptive attention allocation are precisely the mechanisms adversaries exploit. A system that ignored salience cues would be unexploitable but also non-functional. This tradeoff appears at every level of the attention hierarchy.

1.10 9. The Attention Economy

1.10.1 Simon's Foundational Insight

Herbert Simon, in 1971, derived from first principles what the subsequent half-century would confirm empirically. In information-rich environments, information is abundant and attention is scarce. A wealth of information creates a poverty of attention. Simon wrote this before personal computers, before the internet, before social media. He derived it from the architecture of information-processing systems.

The attention economy is not a metaphor. It is a literal consequence of the bottleneck. If attention is the finite gateway through which all information must pass to be processed, then increasing information supply without increasing attention capacity necessarily creates a scarcity of attention relative to information.

1.10.2 Engineered Capture

The technology industry has systematically exploited the gap between stimulus-driven and goal-directed attention. Infinite scroll eliminates stopping cues – the natural disengagement triggers that prompt executive evaluation of whether continued engagement serves current goals. Variable reward schedules keep the attentional system vigilant for the next payoff. Notifications are exogenous attention capture by design, exploiting the orienting response, social salience, and uncertainty simultaneously. After an interruption, returning to the original task takes an average of 23 minutes (Mark et al., 2008). Dark patterns manipulate visual salience to guide attention away from user-serving choices toward platform-serving ones.

These techniques work because they target evolved attentional biases. The human attention system prioritizes threat, novelty, social signals, and reward cues because ancestors with these biases outreproduced those without them. In modern information environments, these biases are attack surfaces.

1.10.3 Surveillance Capitalism and the Feedback Loop

Zuboff (2019) describes a business model with two stages: capture attention (engagement), then extract behavioral data from the attention (surveillance). The data enables predictive models, which enable more precisely targeted content, which captures more attention. This is a positive feedback loop with no natural equilibrium. The structure mirrors the addiction cycle – sensitization, attentional bias, craving, narrowed attention, further sensitization – though without the neurobiological adaptation that constitutes physiological dependence.

1.10.4 AI's Dual Role

AI now mediates a significant fraction of human attention allocation. Recommendation algorithms determine what humans attend to across entertainment, news, and social content. Search engines determine which information merits attention. The human chooses from an AI-curated subset, not from the full information space. Biases in AI attention allocation propagate to human attention allocation at scale.

Simultaneously, AI is becoming a direct competitor for human attention – through AI-generated content flooding information channels, chatbots designed to sustain conversational engagement, and AI-driven entertainment that adapts to maintain engagement. The attention mechanism in AI was designed to help AI process information efficiently. AI systems built on this mechanism are now deployed to make human information processing less efficient, flooding the environment and optimizing for engagement over utility.

1.11 10. Attention and Consciousness

1.11.1 The Boundary

If attention is a universal mechanism, does it bear any relation to consciousness? This is where the universality thesis encounters its hardest boundary condition. The relationship between attention and consciousness is deeply contested, and intellectual honesty requires navigating the dispute rather than resolving it prematurely.

1.11.2 Global Workspace Theory

Baars (1988) proposed that consciousness functions as a global workspace: a limited-capacity shared medium where selected information is broadcast to all specialist processors simultaneously. Attention is the spotlight that selects which content appears on stage. Dehaene and colleagues provided a neural implementation (Neuronal Global Workspace Theory, or NGWT), identifying the workspace with long-range prefrontal-parietal connections that exhibit a characteristic all-or-none "ignition" when attended stimuli surpass a threshold.

The structural parallel to transformer attention is genuine: both involve a limited-capacity bottleneck selecting from a larger pool, a mechanism making selected information available to downstream processes, and competition among representations for access to a shared resource. But

the disanalogies are equally important. In GWT, the workspace is a single shared medium; consciousness is unified. In transformers, multiple attention heads operate in parallel with no unified broadcast. Multi-head attention resembles Dennett's multiple drafts model more than Baars's single workspace. And NGWT posits all-or-none ignition, while transformer attention weights are graded.

1.11.3 The Access/Phenomenal Distinction

Ned Block (1995) drew a sharp distinction between access consciousness (A-consciousness) – information poised for use in reasoning and behavioral control – and phenomenal consciousness (P-consciousness) – the subjective experiential character of a state. These are logically independent. A system could have information poised for reasoning without there being anything it is like to have that information.

This distinction is critical for the paper's thesis. Machine attention plausibly implements something functionally analogous to A-consciousness only. Transformer attention selects which information is made available to downstream layers for further processing – the functional profile of access. There is no reason to attribute P-consciousness to this process. The paper's universality claim concerns functional architecture, not subjective experience.

1.11.4 Integrated Information Theory

Tononi's IIT (2004) proposes that consciousness is identical to integrated information (Phi). Crucially, Phi is a property of a system's intrinsic causal architecture, not of its dynamic attentional state. IIT generates specific predictions about artificial systems: purely feedforward architectures have zero Phi by IIT's definitions, because they decompose cleanly into independent input-output mappings with no intrinsic causal power beyond their parts. Standard transformers are not purely feedforward within a layer (self-attention creates within-layer interactions among all positions), but the layer-to-layer structure is feedforward. Whether the within-layer interactions generate meaningful integration by IIT's criteria remains open. Tononi and colleagues have been explicit that current AI architectures are likely not conscious by IIT's criteria.

IIT is the strongest framework for arguing that transformer attention is not genuinely like biological attention at the phenomenal level, because the integration profiles differ fundamentally. If IIT is correct, functional equivalence is not sufficient for consciousness; intrinsic causal structure matters. This cuts against a purely functional universality claim but supports the more nuanced version this paper defends: attention is functionally universal but potentially phenomenally substrate-dependent.

1.11.5 Dennett's Multiple Drafts

Dennett (1991) rejected the Cartesian Theater in favor of the Multiple Drafts Model: consciousness as parallel, distributed narrative streams competing for influence. Attention is one competitive process among many, not a privileged gateway. Attended contents gain competitive advantage in

the contest for widespread influence on behavior, memory, and report.

This is arguably the philosophical model most closely aligned with transformer attention. No central observer. Multiple simultaneous interpretations (multi-head attention as parallel drafts). Competitive dynamics (softmax as fame-determining normalization). No privileged layer where the transformer "becomes aware." If Dennett is right that the sense of unified consciousness is itself a narrative construction, then the actual mechanism of biological attention may be more like transformer attention than first-person intuition suggests.

1.11.6 Dissociations

Empirical evidence reveals double dissociations between attention and consciousness. Attention without consciousness: subliminal priming, blindsight, attentional effects on invisible stimuli. Consciousness without attention: peripheral awareness, moods and background feelings, gist perception. These dissociations suggest that the functional mechanism of attention is independent of phenomenal experience, strengthening the claim that transformer attention is genuinely attention, not merely a metaphorical use of the word. But they also show that attention in biological systems involves more than the functional description captures.

1.11.7 Frankfurt's Hierarchy: The Attentional Wanton

Harry Frankfurt (1971) distinguished between a wanton – a being that acts on whatever desire is strongest without preferences about its own motivational structure – and a person – a being with second-order volitions who can endorse or repudiate their own desires. Applied to attention, this distinction generates a gradient of attentional sophistication that may be the paper's most precise tool for scoping the universality claim.

First-order attention selects stimuli based on salience or learned relevance. This is the computational primitive all systems share, from bacterial chemotaxis through transformer self-attention to the vertebrate orienting response. Second-order attention – metacognitive attention – monitors and evaluates where first-order attention is directed: Am I attending to the right thing? Is my focus too narrow? Has my attention been captured? This capacity emerges in mature human cognition and is precisely what contemplative traditions train when they cultivate the "witness" stance of vipassana. Third-order attention, what we might call attentional volition, goes further: it endorses or overrides particular attentional allocations on the basis of values and long-term goals. The contemplative practitioner who notices their attention has been captured by anger and deliberately redirects it to the breath exercises attentional volition – not merely noticing where attention has gone but choosing where it should go, choosing on the basis of an endorsed evaluative framework.

Current transformer architectures are attentional wantons. They attend, but they do not attend to their own attending. No mechanism exists by which a transformer evaluates whether its attention allocation is serving well and adjusts at the metalevel. Chain-of-thought and self-reflection prompts approximate this within the first-order mechanism: the model generates text about its own reasoning, but this is first-order pattern completion that mimics the form of metacognition

without instantiating the capacity. A genuine second-order attention system would need to operate over representations of its own attentional states, not merely produce linguistic descriptions of attention. The difference is between a thermostat that displays its own temperature reading and a thermostat that can evaluate whether its temperature-sensing function is working properly.

This suggests a gradient of attentional sophistication: minimal (stimulus-driven selection, present in all systems), endogenous (goal-directed selection, present in biological systems and approximated by prompted LLMs), metacognitive (monitoring one's own attention, present in mature human cognition, absent in current AI), and volitional (choosing what to attend to based on endorsed values, developed through contemplative practice, hypothetical in AI). The universality claim holds at levels 1-2. Levels 3-4 mark the frontier where biological attention, particularly as refined by contemplative training, surpasses anything current AI instantiates. Whether this frontier is crossable without phenomenal consciousness is among the deepest questions the architecture of attention raises.

1.12 11. Synthesis: Toward a Unified Model

Every domain examined in this paper instantiates a common abstract model: an agent with finite resources, operating in an information-rich environment, under layered instruction sets, deploying an attention mechanism that scores potential inputs for relevance, selects a subset, and retrieves or amplifies the selected content for further processing. A bacterium: chemoreceptors meeting a chemical gradient, genome-configured, implementing a biased random walk under metabolic constraints. A human brain: cortical networks meeting the sensory world, configured by DNA and culture, implementing biased competition under metabolic and temporal limits. A transformer: a layer stack meeting a token sequence, configured by weights and prompt, implementing softmax(QK^T/sqrt(d_k))V under compute and memory constraints. The pattern is forced by the same information-theoretic constraint in every case.

The instruction set hierarchy provides the second unifying dimension. All attentive agents operate under layered instructions, from hardware through firmware and software to runtime. The critical structural difference lies in validation: biological instruction sets are selection-tested over evolutionary timescales, while AI instruction sets are engineered and benchmark-tested. This gap in validation depth may be the most consequential difference between biological and artificial attention systems – more important than differences in substrate or mechanism.

The universality claim survives its disanalogies because it is scoped correctly. The claim is not that all attention is the same. It is that all attention instantiates the same computational pattern – a pattern forced by information-theoretic constraints on finite agents – and that understanding the pattern illuminates each instance, including the ways each instance departs from it. The disanalogies are not qualifications reluctantly conceded. They are integral to the claim: a universality thesis that cannot specify its own boundaries is not a thesis but a metaphor.

1.13 12. Open Questions and Future Directions

Several questions resist resolution and may define the research frontier.

First, what would metacognitive attention in AI require architecturally? Is it sufficient to add a monitoring module that evaluates attention-layer performance, or does genuine second-order attention demand a fundamentally different architecture – perhaps one with recurrent dynamics that allow the system to settle into attentional states rather than computing them in a single pass?

Second, can the contemplative training paradigm inform AI attention engineering – not at the level of specific techniques but at the level of design principles? The contemplative insight that focused and distributed attention serve different functions, and that meta-attention may matter as much as first-order attention, may have architectural implications not yet explored.

Third, is self-directed attention possible without consciousness? The biological template is deeply integrated with subjective experience. If functional architecture cannot be separated from phenomenological architecture, then self-directed AI attention may face limits that no amount of architectural refinement can overcome.

Fourth, can attention alignment be verified formally? If a self-directed AI allocates its own attention, its internal attention states become safety-critical. Formalizing both the attention system and the welfare criteria it should track would be a challenge at the intersection of alignment research and the cognitive science of attention – but a necessary one.

1.14 13. Conclusion

Attention is not a feature of minds. It is a consequence of finitude. Any agent existing in an environment richer than its processing capacity must attend – must score, select, and retrieve from the stream of available information. This paper has traced that necessity across domains: from the QKV mechanics of transformers to the biased competition of visual cortex, from Ashby's Law to the contemplative bootstrapping of awareness observing itself, from the instruction sets encoded in DNA to those encoded in system prompts, from the pathological capture of addiction to the engineered capture of infinite scroll.

The convergences are real. The disanalogies are equally real: biological attention has temporal dynamics, neurochemical modulation, embodied grounding, and phenomenal character that no artificial system replicates. Contemplative attention involves an experiencing subject whose relationship to the mechanism changes through practice. These differences mark the boundaries of the universality claim, not its refutation.

The frontier lies at the transition from directed to self-directed attention. Contemplative traditions demonstrate that this capacity is trainable in humans. Whether it is achievable in artificial systems depends on whether the attentional wanton can become an attentional person – whether a system that selects can learn to evaluate its own selecting, and whether that evaluation can be grounded in something other than first-order pattern completion. The bacterium attends. The transformer attends. The monk attends to attending. That last recursion – the fold where attention discovers

itself as its own object – remains the unsolved problem.

1.15 References

Adler, J. (1966). Chemotaxis in bacteria. *Science*, 153, 708-716.

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.

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

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *ICLR 2015*.

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *Anthropic*.

Bar-Haim, Y., et al. (2007). Threat-related attentional bias in anxious and nonanxious individuals. *Psychological Bulletin*, 133(1), 1-24.

Barkley, R. A. (1997). *ADHD and the Nature of Self-Control*. Guilford Press.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv:2004.05150*.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247.

Brewer, J. A., et al. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *PNAS*, 108(50), 20254-20259.

Broadbent, D. E. (1958). *Perception and Communication*. Pergamon Press.

Castellanos, F. X., & Tannock, R. (2002). Neuroscience of attention-deficit/hyperactivity disorder. *Nature Reviews Neuroscience*, 3, 617-628.

Cho, K., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014*.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19.

Clark, K., et al. (2019). What does BERT look at? An analysis of BERT's attention. *BlackboxNLP*.

Dao, T., et al. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *NeurIPS 2022*.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18, 193-222.

Elhage, N., et al. (2021). A mathematical framework for transformer circuits. *Anthropic*.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1), 5-20.

Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural

networks. *ICLR 2019*.

Greshake, K., et al. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection.

Henrich, J. (2015). *The Secret of Our Success*. Princeton University Press.

James, W. (1890). *The Principles of Psychology*. Henry Holt and Company.

Katz, S. T. (1978). Language, epistemology, and mysticism. In S. T. Katz (Ed.), *Mysticism and Philosophical Analysis* (pp. 22-74). Oxford University Press.

Koster, E. H. W., et al. (2011). Understanding depressive rumination from a cognitive science perspective. *Clinical Psychology Review*, 31(1), 138-145.

Lazar, S. W., et al. (2005). Meditation experience is associated with increased cortical thickness. *NeuroReport*, 16(17), 1893-1897.

LeCun, Y. (2022). A path towards autonomous machine intelligence. *OpenReview*.

MacLean, K. A., et al. (2010). Intensive meditation training improves perceptual discrimination and sustained attention. *Psychological Science*, 21(6), 829-839.

Mark, G., et al. (2008). The cost of interrupted work: More speed and stress. *Proceedings of CHI*.

Meaney, M. J. (2001). Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. *Annual Review of Neuroscience*, 24, 1161-1192.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167-202.

Norretranders, T. (1998). *The User Illusion*. Viking.

Olsson, C., et al. (2022). In-context learning and induction heads. *Anthropic*.

Park, J. S., et al. (2023). Generative agents: Interactive simulacra of human behavior. *UIST 2023*.

Pathak, D., et al. (2017). Curiosity-driven exploration by self-supervised prediction. *ICML 2017*.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13, 25-42.

Reingold, E. M., et al. (2001). Visual span in expert chess players. *Psychological Science*, 12(1), 48-55.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8(5), 368-373.

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, 18(3), 247-291.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation. *IEEE Transactions on Autonomous Mental Development*, 2(3), 230-247.

Shinn, N., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. *NeurIPS 2023*.

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, Communications, and the Public Interest*.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9), 1059-1074.

Slagter, H. A., et al. (2007). Mental training affects distribution of limited brain resources. *PLoS Biology*, 5(6), e138.

Srivastava, N., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15, 1929-1958.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference*.

Tishkoff, S. A., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39(1), 31-40.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.

Treue, S., & Martinez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399, 575-579.

Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS 2017*.

Voita, E., et al. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *ACL 2019*.

Wang, G., et al. (2023). Voyager: An open-ended embodied agent with large language models. *arXiv:2305.16291*.

Wang, K., et al. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

Yao, S., et al. (2023). ReAct: Synergizing reasoning and acting in language models. *ICLR 2023*.

Zaheer, M., et al. (2020). Big Bird: Transformers for longer sequences. *NeurIPS 2020*.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.