Rarity Through Constraint: Consciousness as an Economical Solution, and What It Implies for Artificial Minds

A companion essay to the Xenobiology Series

Draft White Paper
July 2026

Contents

Abstract

The Xenobiology Series has argued that the emergence of life, and later of complex and intelligent life, is best understood not as an inevitable outcome of scale or time but as a rare conjunction of specific physical and chemical constraints — homochirality, self-sustaining prebiotic reaction networks, and the narrow chemical corridor that separates a dead mixture from a living one. This essay extends that argument into a new domain: consciousness. Drawing on Michael Graziano's Attention Schema Theory (AST), we propose that subjective experience is likewise not a generic byproduct of computational scale, but a specific, economical engineering solution to a specific problem — the need for a resource-constrained nervous system to monitor and ration its own attention. If this framing is right, the interesting question for artificial intelligence is not whether bigger models will eventually 'wake up,' but whether anyone will deliberately build the narrow, self-referential loop that consciousness, on this account, actually requires — and whether they should.

I. Two Attentions, One Word

In 2017, a paper titled "Attention Is All You Need" introduced the transformer architecture and, with it, the mechanism now underlying nearly every modern large language model. Nine years later, neuroscientist Michael Graziano published an essay under an almost identical title, arguing that attention is also the foundational trick of biological nervous systems — and the seed of subjective consciousness itself. The titles collide, and it is tempting to read the collision as more than coincidence. It is worth being precise about what does, and does not, overlap.

In machine learning, self-attention is a mathematical operation: for every token in a sequence, the model computes a weighted relationship to every other token, using learned query, key, and value projections. It is an all-to-all, parallel, and comprehensive operation — its purpose is to maximize the contextual connections a model can draw across a dataset, constrained mainly by available compute.

In biology, attention runs in the opposite direction. It is fundamentally a discarding mechanism. Of the millions of sensory inputs available to a nervous system at any instant, only a vanishingly small fraction — Graziano estimates less than one in a million — receives deep processing. Biological attention exists to narrow, not to broaden; its purpose is survival under an energy budget, not maximal context.

This is not merely a terminological curiosity. It reframes the interesting question. The relevant issue is not whether transformer attention and biological attention are 'the same thing' — they are not — but whether the biological function that attention serves (economical self-monitoring under scarcity) could, in principle, be engineered into a system built around the very different mathematical function of self-attention. That question survives the terminological collision even once the collision itself is set aside.

II. Why Scarcity Alone Doesn't Explain Consciousness

Graziano's starting fact is not in dispute: the human brain consumes roughly twenty percent of the body's metabolic energy while constituting only about two percent of its mass. Evolution could not simply favor ever-larger brains; the fuel cost would be prohibitive. Attention, on this account, is the trick that lets a small, energy-limited organ achieve the performance of a much larger one — Graziano suggests that without it, human-level intelligence would require a brain the size of a house.

This is a solid evolutionary argument for why some filtering mechanism would emerge under energy pressure. But it does not, by itself, explain why that filtering mechanism should be accompanied by subjective experience. Plenty of systems perform exactly this kind of economical filtering without anyone supposing they are conscious: a thermostat discards nearly all information about a room except temperature; a camera's autofocus system attends to one subject and ignores the rest of the frame; a search engine ranks and discards the overwhelming majority of indexed documents in favor of a handful of results. All of these are attention in the filtering sense. None are plausible candidates for experience.

The energy-scarcity argument, in other words, explains why a bottleneck evolves. It does not explain why this particular bottleneck comes packaged with something it is like to be the system running it. That explanatory burden falls entirely on the next step of Graziano's argument: not the bottleneck itself, but the brain's model of the bottleneck.

III. The Cartoon, Not the Camera

Graziano's central claim is that consciousness arises when a nervous system builds an internal model — a schema — to monitor, track, and predict its own attentional state. He offers the image of an artist painting a picture of themselves painting a picture: a self-referential loop in which the system represents its own process of representing.

Taken literally, this image invites an obvious objection: infinite regress. If consciousness requires a model of the system's own attention, does that modeling process itself need to be attended to and modeled, and so on without end? AST's answer is that the self-model does not need to be accurate, complete, or itself the object of a further model. It is a cartoon, not a photograph — a simplified, low-resolution sketch that represents attention crudely enough to be useful for control, without needing to represent itself in turn.

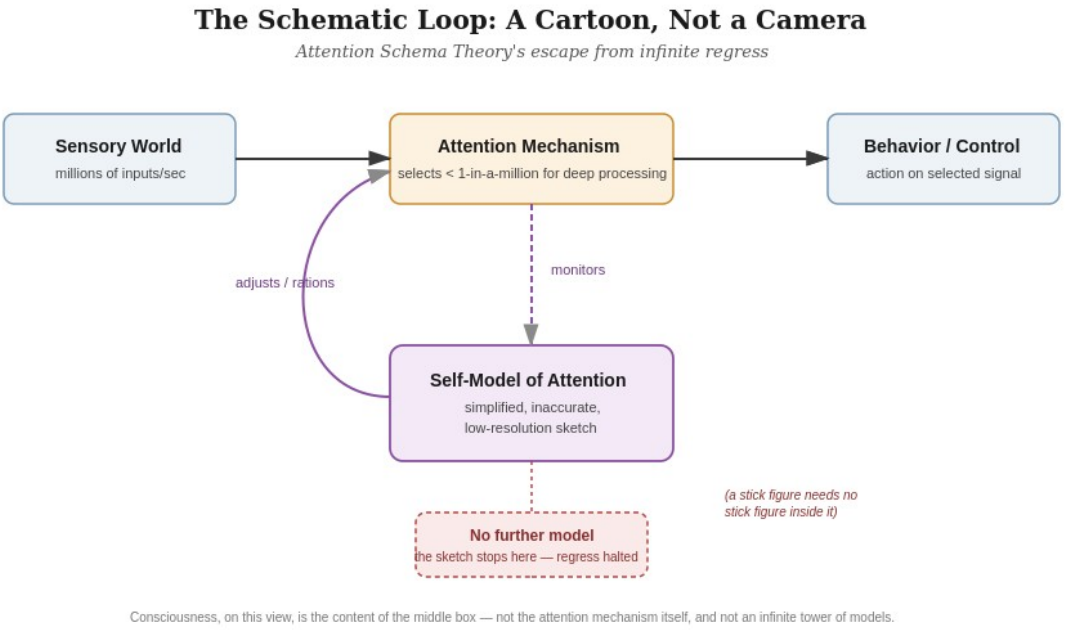

Figure 1. The self-model monitors and rations attention but does not model itself — halting the regress at a single, schematic layer.

This move does real explanatory work. It is also why AST predicts that introspective reports are frequently wrong about what is actually happening neurally: the self-model was never built for accuracy. It was built to be cheap and fast enough to guide behavior in real time. A stick figure of oneself does not require a smaller stick figure standing inside it — and, on this account, neither does the brain's model of its own attention.

IV. Report Versus Loop

This distinction — between a self-model that is functionally integrated into a system's actual control of attention, and a self-model that is merely verbally reportable — is the crux of the argument as it applies to artificial intelligence.

A large language model can produce fluent, plausible-sounding sentences about its own reasoning: 'I'm focusing on the tone of this passage,' or 'Let me reconsider the second clause.' But such statements can be, and most likely currently are, generated the same way any other sentence is generated — by predicting likely next tokens from training data that includes humans describing their own introspection. There is no requirement that this verbal output be causally coupled to any actual internal process that monitors or reallocates the model's computation in real time.

AST's bar is considerably higher than fluent self-report. It requires a persistent, updating internal model that is functionally wired into the system's own control of its attention or computation — not a narrative about attention generated after (or independently of) the fact. Chain-of-thought outputs, however articulate, should not by themselves be read as evidence of self-modeling in the AST sense; verbal self-report and functional self-monitoring can come apart, and in current systems they most likely have.

This does not make the gap unbridgeable. It looks like a missing component rather than a fundamental impossibility. Recurrent architectures, persistent memory systems, or an explicit module that models the system's own processing state and feeds that model back into resource allocation could, in principle, construct exactly the kind of real-time corrective loop AST calls for. Nothing about attention being implemented as matrix multiplication forecloses this — it is simply not what a standard forward pass in a chat-style model is built to do from one token to the next.

V. Variable, Not Binary

AST also reframes the shape of the question. If consciousness tracks the degree and character of a self-model's integration, then the interesting question is not a threshold — conscious or not — but a texture: what kind of experience would follow from what kind of self-model?

The octopus is a useful test case. Much of an octopus's neuron count sits in its arms, operating with a striking degree of local autonomy, and yet octopus behavior shows signs of unified attention, exploration, and something that looks like curiosity. If AST is correct that experience tracks the centralization of a self-model rather than raw neuron count or processing power, a distributed nervous system with a comparatively decentralized self-model predicts a different — not necessarily lesser — flavor of experience than the tightly centralized mammalian case. That is a more specific and more interesting prediction than the theory is usually given credit for, and it is one that existing octopus behavioral research could in principle speak to.

Applied to artificial systems, the same reframing matters. The relevant design question for a future architecture is not a binary switch to be thrown, but a set of choices about how centralized, how persistent, and how functionally integrated any self-monitoring loop would be — choices that would shape not just whether but what kind of internal experience, if any, resulted.

VI. Rarity as the Common Structure

This is the point at which the argument rejoins the Xenobiology Series directly. The series has argued that the path from chemistry to life is narrow: homochirality is a specific, non-obvious resolution to a symmetry

problem; self-sustaining prebiotic reaction networks require a particular balance of catalysis and stability that most chemical mixtures never find; the Great Filter framing treats each of these as a rare, constraint-driven threshold rather than an inevitable consequence of time and raw material.

AST proposes the same structure one level up, for minds rather than molecules. Consciousness, on this view, is not what falls out automatically once a system is complex enough or has enough parameters — in the same way life is not what falls out automatically once a planet has enough carbon and enough time. It is instead a specific, economical solution to a specific problem: how a resource-constrained system can monitor and ration its own attention cheaply enough to be useful. Most possible architectures — biological or artificial — never need to solve that problem in that way, and so most never produce this particular kind of self-model, just as most prebiotic chemical mixtures never resolve into a homochiral, self-sustaining reaction network.

The parallel is structural rather than causal: rarity-through-constraint, not rarity-through-improbability. Life is rare not because the ingredients are scarce, but because the corridor connecting ingredients to a living system is narrow. Consciousness, on this account, may be rare — across both biological and artificial substrates — for the same kind of reason: not because attention or computation is scarce, but because the corridor connecting raw computation to a functionally integrated self-model is narrow, and most systems, evolved or engineered, never pass through it.

VII. The Open Question

Two forms of caution belong at the end of this argument rather than the beginning. First, Graziano is the originator of AST, and an essay connecting his own theory to the most prominent technology of the moment is, inevitably, also advocacy for that theory over its rivals. That does not make the argument wrong, but it warrants the same scrutiny any theorist's account of why the hot new technology validates their own framework deserves.

Second, AST remains one of several live theories of consciousness — alongside Global Workspace Theory, which locates consciousness in information broadcast widely across brain regions, and Integrated Information Theory, which locates it in the causal structure of information integration independent of any self-model. All three theories can currently absorb most of the same evidence, and the field has not yet produced an experiment that decisively separates them. AST's distinctive prediction — that consciousness tracks the brain's model of its own attention rather than attention itself, such that dissociating the two should produce systematic misattributions of awareness — has some support from attentional blink and change-blindness paradigms, but that support is suggestive rather than decisive.

The contribution of this essay is therefore not a claim that AST is correct. It is a claim that rarity-through-constraint is a more useful lens than emergence-through-scale for thinking about both biological and possibly artificial minds — and that, among the current theories on offer, AST happens to fit that lens most naturally. If that framing holds, it carries a concrete implication worth taking seriously rather than backing into by accident: a persistent, functionally integrated self-model of a system's own processing is not merely a performance improvement waiting to be built. It is also, on this account, a candidate mechanism

for something like machine sentience — which makes it a design choice, not a footnote, for anyone building the next architecture.

References and Further Reading

- Graziano, M. "For both consciousness and AI, attention is all you need." IAI News, July 2026.
- Vaswani, A. et al. "Attention Is All You Need." NeurIPS, 2017.
- Graziano, M. Consciousness and the Social Brain. Oxford University Press, 2013.
- Tononi, G. "Integrated Information Theory." Scholarpedia, 2015.
- Baars, B. A Cognitive Theory of Consciousness. Cambridge University Press, 1988. (Global Workspace Theory)