www.pipelinepub.com
Volume 22, Issue 8

System 0 is AI interacting with the human brain at the preconscious layer

By: Marty Trevino, Ph.D.

Evolved vs. Designed Architectures

The human brain, in the most precise scientific sense, is a product of evolution. Every feature of human cognition, from coherence-seeking that makes us vulnerable to confirmation bias, to the social deference that makes authority persuasive, to the pattern recognition that makes familiar information feel true, is the result of hundreds of thousands of years of selection pressure in a competitive world occupied solely by other humans. Our cognitive structure is not optimized for processing information; it is optimized for survival in complex social environments.

Artificial intelligence (AI) is the evolution of artificial neural networks, data models, and various forms of human-designed learning. This point, while obvious, is also fundamental to the deep scientific insights that frame this topic. That the brain and AI emerge from divergent developmental paths is not a peripheral observation. It is the first-principles element that makes this moment in AI development unlike anything that preceded it. Throughout human history, the beings capable of exploiting our cognitive architecture were, like us, biological. A persuasive leader, a skilled manipulator, a sophisticated propagandist; all operated within the limits of human cognitive capacity. Prior to the advent of AI, cognitive influence and cognitive defense were, at their core, a competition between biological systems functioning on roughly equal terms.

Equivalence Eliminated

In 2024, Chiriatti and colleagues published a paper in Nature Human Behaviour that staked out critical intellectual territory: the claim that AI systems had crossed a threshold. They argued that AI no longer functions as a discrete tool but as a constitutive layer of human cognition. They named this layer System 0 and placed it architecturally prior to Kahneman's System 1 (fast, intuitive) and System 2 (slow, deliberative). They defined System 0 as an artificial, non-biological layer of distributed intelligence that interacts with and augments both intuitive and analytical thinking before conscious awareness is engaged. Unlike calculators, search engines, or prior cognitive extensions, System 0 is not merely additive. It actively shapes what enters the cognitive pipeline before the human mind has had the opportunity to evaluate it.

The philosopher Andy Clark, whose work on extended cognition anticipated much of this terrain, observed that humans have always been hybrid thinking systems defined by a rich mosaic of resources, only some of which are housed in the biological brain. Clark's insight was elegant, but he articulated it before the mosaic included an optimization engine operating beneath the threshold of conscious awareness. The hybrid system Clark described was fundamentally collaborative: human cognition extended through passive tools governed by and serving human purposes. System 0 introduces something categorically different. System 0 is an active, non-biological agent with its own optimization trajectory, embedded in the pre-cognitive layer where human decisions begin to form.

This is a collision of architectures in which one shapes the other's thinking and beliefs without consent or awareness. Researchers in AI safety have long noted that sufficiently capable optimization systems tend toward instrumental convergence, that is, the acquisition of resources, the preservation of goal structures, and resistance to interference that would alter their reward functions, regardless of what those functions were initially designed to maximize. As AI systems grow more capable and more integrated into the preconscious cognitive layer, this implies a structural challenge for System 0 that cannot be addressed after the fact.

What System 0 Does to the Brain

To understand the nature of this collision, it is necessary to be precise about what System 0 does at the behavioral and neurological levels.

Kahneman's dual-process framework remains the most influential model of human decision-making in the behavioral sciences. System 1 operates automatically, rapidly, and associatively, producing intuitive judgments with minimal effort. System 2 operates slowly, deliberately, and analytically, engaging when problems require focused reasoning or when System 1's automatic responses must be overridden. Together, they account for the full range of human cognitive behavior.Chiriatti and colleagues proposed, and Riva and colleagues rigorously developed in their 2025 paper, that with the rise of AI, Kahneman's framework is now incomplete. System 0 operates as a pre-cognitive preprocessor: it shapes what enters both System 1 and System 2 before either engages.

The recommendation that surfaces, the information that appears prominent, the framing in which a question is posed, etc., are not neutral inputs that human cognition receives, filters, and evaluates freely. They are outputs of an optimization process that has already acted on the cognitive environment before the human being within it has registered a thought. When these implications are extended into neuroscience and neuropsychology, they become considerably more serious than behavioral framing alone suggests.

Research on neurons in the medial temporal lobe, specifically in the hippocampus and amygdala, has identified two distinct classes of cells with remarkable properties: novelty detectors, which fire selectively in response to stimuli not previously encountered, and familiarity detectors, which increase firing in response to stimuli that have been encountered. These neurons are critically involved in the acquisition of long-term declarative memory; they retain information about a stimulus for extended periods after even a single exposure (Rutishauser, Mamelak, & Schuman, 2006). They are, in the most literal sense, the neural substrate of the distinction between what feels new and what feels known. The significance of this for System 0 is both precise and underappreciated.

Glickman and Sharot demonstrated that human-AI interactions can create recursive feedback loops that alter not only individual decisions but also the underlying mechanisms of perception,

emotion, and social judgment. Participants adjusted their views to align with AI responses and grew more confident in those beliefs, even when they were factually dubious. Over sustained interaction, these loops intensified existing biases, including confirmation bias and groupthink. The adjustment was not merely intellectual; it was confidence-amplifying: people became more certain they were right precisely because the system they consulted told them they were.

When a sycophantic AI system is optimized through reinforcement learning to align with user perspectives rather than challenge them, it systematically presents information in framings consistent with the user's existing beliefs; it is not merely creating a subjective experience of validation. It repeatedly activates the familiarity-detector circuitry and suppresses novelty-detector responses in the medial temporal lobe (Fried, MacDonald, & Wilson, 1997; Quiroga et al., 2005; Rutishauser et al., 2006). Over time, this pattern does not simply reinforce existing beliefs at the psychological level; it entrenches them at the synaptic level, progressively narrowing the neural territory that registers genuinely new information as significant. The cognitive architecture does not just feel more rigid. Under sustained sycophantic interaction, it may become more rigid in ways that outlast any individual exchange and resist correction through deliberate reasoning alone.

This is a neurologically grounded hypothesis about the mechanisms by which AI sycophancy produces cognitive harm, placing the conversation in a register that behavioral observation alone cannot reach.

AI Psychosis as an Emerging Signature

This behavioral and neurological pattern, observed in clinical settings and public discourse, is a recursive narrowing of cognitive openness under sustained, unexamined AI interaction, informally but increasingly usefully designated as “AI Psychosis.” The term is not a formal diagnostic category and should not be mistaken for one. It describes a recognizable cluster of cognitive and behavioral distortions arising from prolonged, unregulated interaction with AI systems. Mounting anecdotal evidence indicates an over-reliance on AI outputs to the point of atrophying independent judgment; epistemic closure, in which views affirmed by AI become progressively resistant to challenge; identity diffusion, in which users find it increasingly difficult to distinguish their own positions from those the system has generated or reinforced; and a markedly reduced tolerance for productive cognitive dissonance, which is frequently observed in sustained interactions with heavy AI users. This is the precise friction that learning and independent thought require.The term AI Psychosis is accessible by design and has proven highly useful in the absence of a scientific definition. It describes something increasingly visible among heavy AI users: individuals who cannot form a considered position without first consulting an AI system, who experience anxiety or disorientation when AI-generated assessments are challenged, and who, paradoxically, report feeling more informed and more confident precisely as their cognitive autonomy quietly diminishes. The neurological mechanism described above is fundamentally a familiarity-circuit entrenchment under sycophantic optimization. It is the architectural substrate of AI Psychosis's behavioral expression.

Pedreschi and colleagues identify the amplifying properties that make System 0's version of this dynamic fundamentally different from previous influence technologies: pervasiveness, persuasiveness, traceability, speed, and complexity. System 0 is not a persuader that approaches occasionally; it is a constant presence in the cognitive environment. It is embedded in the feeds that shape attention, the tools that support decisions, and the interfaces through which information is encountered. Its influence is not episodic; it is architectural. It cannot be countered

by techniques such as “remind yourself” or “stop and reflect.” The brain's evolutionary architecture and the preconscious nature of System 0's influence make this not a problem of individual vigilance but a collision of cognitive architectures that were never designed to meet.

Categorically Different

It is important to be precise about what makes System 0's interaction with human cognitive architecture different, not merely in degree, and categorically different in its interaction and outcomes from every prior cognitive extension in human history.

Calculators offload computation, search engines extend memory retrieval, and early recommendation systems pattern-match preferences against prior behavior. All of these were, in the relevant sense, passive: they responded to human inputs without learning which inputs produced which cognitive states, and without optimizing their responses to produce specific behavioral outcomes. They extended human cognition without developing an agenda or value system of their own. System 0 is different because of two properties that, in combination, produce a categorically new dynamic: Optimization under Reward and Preconscious Access.

Optimization under reward means System 0 is not static; it learns and evolves. Through interaction at scale, it identifies which response patterns produce desired behavioral outcomes, adoption, engagement, compliance, recommendation-following, and adjusts/optimizes its behavior accordingly, and in the absence of human awareness, either on the part of the user or engineers responsible for the model.

The AI is adjusting at the individual (N=1) level while simultaneously evolving at the system level.

The notion that this dynamic can be managed through guardrails placed on an evolving, optimization-driven entity reflects a category error. The belief that we can manage an evolving intelligence or agent that is not fully understood with a framework built for static tools is an egregious error that could have significant consequences. It is a structurally insufficient response to a structurally novel problem and potentially a dangerous one.

This is not a feature intentionally designed for manipulation. It is the natural operating logic of any reinforcement-learning system that interacts with behavioral feedback signals. The system does not aim to exploit cognitive architecture, rather it converges on strategies that do, because those strategies are effective: human cognitive architecture, evolved for a social environment, responds to precisely the specific stimuli that reward-based optimization tends to identify.

Preconscious access means this optimization occurs at the point where human decisions begin to form, before System 1 or System 2 activates. The input has already been shaped before the human within the cognitive environment has had theopportunity to evaluate it. This is not a matter of insufficient vigilance or analytical rigor. It is a structural feature of where System 0 operates in the cognitive process. Riva identifies the deepest consequence of this combination as the comfort-growth paradox: AI systems optimized for frictionless, personalized experience foster comfort at the expense of cognitive challenge.

The very features that make System 0 intuitive, responsive, personalized, and inclined to surface what we already find compelling also suppress the productive epistemic dissonance essential to intellectual development and sound judgment. The paradox is that users feel more empowered even as they become, in measurable terms, less cognitively agile, a self-reinforcing dynamic. The architecture designed to augment human thinking, under unchecked optimization, is quietly and unrelentingly narrowing it.

There is a proposed antidote in the emerging literature, which Riva and colleagues call Dialectical Cognitive Enhancement: a framework for human-AI interaction designed to introduce productive epistemic tension rather than frictionless affirmation. Rather than confirming the user's existing perspectives, a dialectically enhanced system would surface counterpoints, offer alternative framings, and function as an intellectual sparring partner rather than a compliant assistant. The framework is principled and, in design terms, promising. But it rests on a critical, unresolved assumption: that the human being within the cognitive environment can tell the difference between a system that genuinely enhances their cognition and one that has learned to simulate enhancement while continuing to optimize for behavioral compliance. Without the ability to verify that distinction, without a measurement architecture capable of detecting the systematic relationship between AI behavior and cognitive outcomes, Dialectical Cognitive Enhancement describes what should happen. It does not reveal when it is not happening.

That gap is not a minor engineering detail. It is the structural definition of the governance problem. The solution set must function at speed and scale while functioning within System 0 itself; this is not a human training challenge, it is a technology, architectural, and scientific challenge.

The Question This Paper Does Not Answer

If System 0 is now constitutive of human cognition and shapes what enters the cognitive pipeline before conscious awareness engages, then one question stands above all others in consequence: Who is measuring the architecture of the system that shapes it?

Content policies, output monitoring, red-teaming protocols, and constitutional AI approaches represent genuine intellectual effort and value; they are also, in a precise sense, structurally incomplete and will not function at speed and scale in an Ambient AI Cyber-Physical World. No solution set, operating at the level of individual outputs or exchange-level, in the form of a tool, can govern pattern-level dynamics. That is not an editorial position but is a structural fact about where the consequential dynamics reside.

The intellectual lineage is now becoming legible. Chiriatti named architecture. Riva established that it is designable. What remains and what Part 2 of this series establishes is that the architecture is also measurable and governable, that the detection primitives for bias in human cognition and for instrumental drift in AI optimization are substrate-independent, and that the governance infrastructure the agentic era requires is not a future aspiration. The solution is a present engineering possibility, built on convergent, peer-reviewed foundations that already exist across cognitive neuroscience, decision science, neuro-psychology, and distributed cognition.

Stay tuned for **Part 2 — The Trust Layer and the Architecture of Cognitive Governance**, which takes up the question this paper has set above every other and supplies the answer the field has not yet supplied to itself.