Narrative Review Article

Recurrent Integration and the Empirical Grounding of Phenomenal Consciousness in Artificial Intelligence Systems

Akshaj Devireddy

Evergreen Valley High School, 3300 Quimby Rd, San Jose, CA, 95148, United States

ABSTRACT

Artificial intelligence systems continue to increase in sophistication, renewing the questions of what structurally distinguishes conscious experience from computation. This paper develops a unified framework for consciousness by combining Recurrent Processing Theory (RPT) with a weakened form of Integrated Information Theory (IIT). The aim is to articulate a mechanistic account in which recurrent feedback stabilizes perceptual contents and structural integration unifies them into a single, irreducible experiential field, and then to evaluate whether contemporary AI architectures exhibit these features. Using this framework, the analysis examines the consciousness-relevant organization of two major classes of models: Large Language Models (LLMs) and Emergent Models (EMs). The discussion shows that EMs, due to their intrinsically recurrent dynamics and globally interdependent state evolution, more closely approximate the structural conditions identified by the RPT and weak IIT account than do standard feedforward transformer-based LLMs. The paper also reconsiders the debate between phenomenal and access consciousness by providing an RPT and weak IIT interpretation of the Sperling experiment and by showing how EMs offer a way to render the posited structure of phenomenal consciousness empirically tractable.

Keywords: phenomenal consciousness; artificial consciousness; recurrence; integration; Large Language Models; Emergent Models

INTRODUCTION

Phenomenal consciousness refers to the qualitative, subjective character of experience itself, which includes the felt presence of perception, emotion, sensation, imagery, thought, and even dream or hallucinatory states that need not track the external world. It encompasses not only the vividness and unity of waking perception but also the cognitive phenomenology of thinking, the affective tone of anxiety, after-images, and the richly structured scenes of dreaming. A central challenge in philosophy and cognitive science is to explain what physical or computational structures, if any, correlate with this qualitative dimension of mental life. Access consciousness, in contrast, concerns the availability of information for reasoning, report, and control; the two are often intertwined in practice, but conceptually distinct enough to motivate different explanatory approaches.

Corresponding author: Akshaj Devireddy, E-mail: akshaj.devireddy@gmail.com.

Accepted December 1, 2025

https://doi.org/10.70251/HYJR2348.36740751

Several major theories aim to account for consciousness. Higher-Order Thought (HOT) theories attribute consciousness to a system's capacity to form thoughts about its own mental states (1). Attention Schema Theory (AST) proposes that consciousness arises when the brain constructs an internal model of its own attentional processes (2). Predictive Processing accounts characterize conscious experience as the brain's best, integrated prediction of its sensory inputs (3). Functional theories such as Global Workspace Theory emphasize reportability and global availability (4), while structural theories like Recurrent Processing Theory (RPT) (5) and Integrated Information Theory (IIT) (6) target the neural dynamics and organizational features associated with conscious states. Each framework captures something important, but none fully explains why phenomenal consciousness is stable, unified, and structured in the way experience suggests.

The present paper argues that a synthesis of Recurrent Processing Theory and a weakened form of Integrated Information Theory ("RPT + weak IIT") provides the most coherent account of the structural features associated with phenomenal consciousness. This combined framework does not claim to identify the metaphysical origin of experience, nor does it assert that recurrence and integration cause phenomenal consciousness. Instead, it proposes that recurrent stabilization of content and integrated organization are the structural conditions most closely linked to the form phenomenal consciousness appears to take. This approach begins with the phenomenology itself and then identifies the neurocomputational properties that best correspond to it.

Recent advances in artificial intelligence (AI) make the underlying question more urgent: whether any artificial system could represent the organizational conditions associated with phenomenal consciousness. As Large Language Models (LLMs) display increasingly robust behavior and new architectures such as Emergent Models (EMs) exhibit more recurrent and self-organizing dynamics, it becomes natural to ask whether such systems approximate these structural conditions. This paper develops the RPT + weak IIT framework, applies it to contemporary AI architectures, and argues that EMs more closely instantiate the structural features linked to phenomenal consciousness. It concludes by revisiting the phenomenal-access distinction, analyzing the Sperling experiment through this lens, and clarifying how EMs make the scientific study of artificial phenomenal consciousness a meaningful possibility.

COMBINING RECURRENT PROCESSING THEORY AND WEAK INTEGRATED INFORMATION THEORY

Experience provides a distinctive source of evidence about consciousness. Through a vivid, detailed, and unified perspective, conscious experience constructs the subjective mental life that differs sharply from the fragmented and mechanistic structure of the external world. For example, the world does not tend to appear as a collection of disconnected shapes and colors, showing that consciousness forms a very cohesive scene that includes all the sights, sounds, and colors encountered in everyday life. Additionally, consciousness is never uniform or static, waxing and waning depending on factors like attention, sleep, or anesthesia. These phenomenological features constitute fundamental data that any scientific theory of consciousness must accommodate.

Two scientific frameworks, RPT and a weakened form of Integrated Information Theory (IIT), are well suited to explain these basic characteristics. This study develops a combined RPT and weak IIT framework grounded in these phenomenological and neuroscientific considerations, and uses this unified account to motivate the analyses that follow.

Recurrent Processing Theory (RPT)

RPT was first developed by Victor Lamme and others, and asserts that consciousness comes about through feedback loops within the brain's processing hierarchy. These sorts of feedback loops contrast from the simple forward flow of sensory information (5). Visual neuroscience offers a helpful example. When light strikes the retina, signals travel through multiple layers of the visual cortex in a feedforward sweep. Essentially, the brain's electrical impulses (action potential) are triggered by edges, colors, and motion cues in the external world, which are then extracted and passed up to higher-order regions, creating a process that is extremely quick and characterizes what is known as reflex-like behavior. Critically, these reflexes are without full awareness (such as ducking when something moves rapidly toward the face).

However, feedforward processing does not produce conscious experience. RPT proposes that the true key to consciousness is when later brain areas send information back to earlier ones. Through this loop of recurrent activity, the initial sensory signals are re-amplified and stabilized as they go through multiple passes. These signals are also refined as they loop between earlier and later brain areas. Recent neuroscientific evidence confirms that, when the brain must segment complex visual scenes, recurrent processing is especially necessary (7). For example, when walking through a crowded crosswalk and needing to distinguish a friend's face from a blur of moving bodies, the brain uses recurrent feedback to carve the scene into meaningful objects instead of letting it remain a flat jumble of colors and shapes. RPT also requires that information be "locked in" by closed-loop re-entry, which refers to the way signals loop back and forth between higher brain areas to earlier ones in a continuous loop. The repeated exchange strengthens and stabilizes the content without letting it fade after a single pass. In other words, recurrence prevents perceptual content from dissolving into noise, maintaining it long enough to enter awareness in a steady and coherent way.

Additionally, evidence from neuroscience supports this link. Studies concerning "backward masking" show that a certain stimulus can enter the visual system, but if the recurrent loops are disrupted, it vanishes from awareness (8). A potential concern is that this relationship might be merely correlational, with both recurrence and conscious perception caused by some shared deeper factor. However, the masking procedure is designed to interrupt directly on the recurrence itself, flashing an image quickly and then replacing it with a "masking" pattern that disrupts feedback activity without removing the initial feedforward sweep. The fact that perception fails precisely when recurrence is blocked provides stronger evidence for a causal role, even if it does not eliminate all alternative explanations. Similarly, under anesthesia, recurrent activity is significantly weakened, correlating with a loss of conscious experience (9). These findings suggest that recurrent loops are not just by-products of perception but play a direct role in stabilizing the contents of conscious awareness.

From a phenomenal perspective, the stability explained by RPT corresponds to the lived character of experience. Consciousness is not a rapid series of fleeting impressions but a stable, enduring field of awareness that allows us to track a variety of features across time (objects, colors, movements, sound, pain). Crucially, this occurs without constant dissolution into noise, giving perceptual scenes their characteristic richness and detail. However, if RPT were the entire explanation, phenomenal consciousness might appear distributed across different modules in the brain, with separate percepts "locked in" but not combined into a unified conscious scene. This dis-unification problem is where weak Integrated Information Theory becomes necessary in order to explain the phenomenal evidence that consciousness resembles a seamless field rather than a patchwork of recurrent processes. Essentially, it accounts for the binding and unity that RPT alone leaves unexplained.

Weak Integrated Information Theory (IIT)

Integrated Information Theory (IIT) was first developed by Giulio Tononi, and begins from quite a different intuition. The theory proposes that consciousness, despite being stable, is also unified and irreducible (6). For example, when a subject sees a red apple, the experience does not separate the characteristics of the apple into categories like "red" and "round". Rather, it appears as one coherent perceptual whole. To this end, IIT measures how much the information in the brain is differentiated (each part makes unique contributions) and integrated (no subset of parts can be removed without losing the core structure of the experience).

Furthermore, there are different degrees to which IIT can be implemented, which matters for constructing a synthesis of RPT and IIT. The strongest version of IIT proposes that a mathematical measure Φ (phi) captures the properties of differentiation and integration. Consequently, a system with high Φ would be conscious to a high degree. However, this "strong IIT" view raises a key concern: it implies that many physical systems, including simple circuits, could possess some level of consciousness (10). This implication aligns with panpsychism, the highly controversial view that consciousness is a fundamental property of the physical world. More importantly, strong IIT makes the substantive metaphysical claim that consciousness is identical to a maximally irreducible cause-effect structure, a commitment that goes far beyond what is needed for a structural theory of experience. Because the goal here is to use integration as an empirical indicator rather than assert a metaphysical identity, the stronger version of IIT is set aside.

By contrast, "weak IIT" assesses consciousness through a more minimal and practical lens (11). It doesn't claim that Φ is the literal essence of consciousness. Instead, it proposes that measures of information integration can serve as useful indicators of when consciousness is likely to be present. For example, compared to brain states during deep sleep or coma, brain states during wakefulness show more complex and integrated patterns, which aligns with phenomenal evidence where experience fades or disappears in sleep or anesthesia (12). Hence, the system's information becomes less differentiated and integrated during such states.

Therefore, instead of treating Φ as a metaphysical criterion, weak IIT highlights two structural properties of conscious experience. The first is unity, meaning that all contents of awareness belong to one field rather than a disconnected jumble. The second is irreducibility, meaning that experiences cannot be broken down into independent parts without losing what makes them experiences. Although experiences may be described in terms of "visual" and "auditory" components, weak IIT emphasizes that these aspects do not exist in isolation at the phenomenal level. Removing one stream of information would not leave an autonomous experience of the other; instead, the entire conscious scene would change into something different in kind. An analogy arises from music: an individual with perfect pitch may identify each note in a chord, but the chord cannot be reduced to hearing those notes separately. Removing one note does not yield a "partial chord" but produces an entirely different harmony. Similarly, the phenomenal scene cannot be disassembled without altering the whole. Related to this idea, the concept of computational irreducibility becomes relevant for later sections, especially in relation to artificial systems that may approximate consciousness.

Pairing RPT With Weak IIT

Nonetheless, the two theories of Recurrent Processing Theory and Integrated Information Theory have their own limitations, which leads us to combine them for a truly cohesive picture of consciousness.

The Limitation of Unity with RPT Alone

RPT alone provides a strong account of how specific mental contents enter consciousness. According to the theory, information stabilizes in consciousness when neural signals loop back and influence earlier stages of processing, reinforcing a representation so that it isn't fleeting or "in transit". In other words, stabilization means the neural activity is maintained long enough (and with enough mutual reinforcement across levels) so that it can influence perception and behavior in a durable way rather than fading out. These feedback loops explain why consciousness feels phenomenally rich. When a subject perceives the color red, hears the hum of an air conditioner, or discerns the shape of a cup, the subject is aware of grounded precepts instead of a vague or undifferentiated background.

However, RPT doesn't explain why these contents are unified into a single, unified stream of consciousness. For example, consider a system in which multiple disconnected loops of recurrence remain isolated from each other. One recurrent loop in the visual cortex might stabilize the perception of a red square, while another in the auditory cortex separately stabilizes the sound of a bell. Importantly, this is without any cross-linking integration. Such a system could generate isolated "conscious islands" (awareness of a shape and awareness of a sound) but with no sense that both belong to the same experiential field. The system would produce fragments of content but no structural glue to bind them into "one" consciousness. Therefore, RPT provides a mechanism for the "presence of content" in consciousness (the occurrence of something rather than nothing appearing), but not for the "unity of experience" (the fact that different contents attach together as part of a single scene or perspective). Without an additional structural principle of integration, RPT risks portraying consciousness as a patchwork of disconnected experiences rather than the seamless, unified flow encountered in actual phenomenal life.

The Limitation of Selection with Weak IIT Alone

Here, weak IIT fills in the other half of the story. The theory explains why consciousness feels like a seamless whole rather than many isolated islands, focusing on the degree of information integration across a system. It captures the intuition that experience is irreducibly unified, and that it is resistant to being broken into independent parts. In spite of that, weak IIT alone leaves out the question of exactly which specific contents actually populate consciousness. For instance, a system could have very high integration, but that does not mean the conscious subject in question is aware of any specific thing (whether a color, sound, or thought).

To illustrate this, imagine a subject walking into a café. Inside the café, the brain simultaneously encodes the smell of coffee, the sight of a red mug on the counter, the hum of background conversation, and the feeling of a chair against the body. All of these could contribute to a strongly integrated system, as they are highly interconnected. On the basis of weak IIT alone, the subject has a unified conscious field, though the theory cannot determine which of these signals actually show up in experience. Does the subject consciously notice the red mug, or is attention captured by the smell of the coffee instead? In other words, using weak IIT by itself only predicts unity because the theory cannot discriminate among competing inputs to determine what is consciously selected.

At this point, RPT becomes necessary. Specific neural representations become conscious when feedback loops between higher and lower brain areas stabilize them long enough to influence processing happening in that moment. Returning to the café example, recurrent reinforcement of the visual representation of the red mug could allow it to stabilize as a conscious perception. Meanwhile, the hum of conversation could remain at a preconscious level because it isn't being recurrently amplified. Afterward, weak IIT structurally integrates these stabilized perceptions (the red mug in this case) with other conscious contents being perceived at the same time (such as the smell of coffee), binding them into a single unified experience. In this way, the combination of RPT and weak IIT explains selection, referring to which contents rise into consciousness, and why. It also explains unity, namely why these contents, once selected, belong to one seamless conscious scene.

Overall, the starting point for explaining consciousness should concern the structure of experience itself, in which phenomenal evidence shows that consciousness is both stable and unified. Recurrent Processing Theory and Weak Integrated Information Theory together explain why this is the case. Recurrence supports phenomenal stability and detail, while integrated information supports unity and irreducibility. While other theories may focus on pieces of the puzzle, this combined framework offers the most direct, complete, and phenomenally grounded account of what consciousness is like. Additionally, while this framework does not claim that recurrence and integration produce phenomenal consciousness, it does identify the structural conditions most consistently associated with the form that phenomenal experience takes.

LARGE LANGUAGE MODELS, EMERGENT MODELS, AND THE PROSPECTS FOR ARTIFICIAL PHENOMENAL CONSCIOUSNESS

In the previous section, the synthesis of Recurrent Processing Theory (RPT) and weak Integrated Information Theory (IIT) was established as a theoretical framework for consciousness. By explaining the detailed stability of conscious contents and their unified and irreducible character, these theories together provide a strong foundation for investigating whether existing and emerging artificial intelligence (AI) architectures contain the structural features which are crucial for consciousness.

This section examines two broad categories of AI systems. The first is large language models (LLMs), which represent the dominant paradigm of contemporary machine learning. The second concerns Emergent Models (EMs), a novel family of architectures that differ significantly in design and general dynamics. The analysis evaluates which architecture most plausibly approximates the structures associated with phenomenal consciousness by examining them in relation to RPT and weak IIT. The discussion begins with LLMs, focusing on their current limitations and on recent innovations that gesture toward architectures more compatible with the "RPT + weak IIT" framework.

Large Language Models (LLMs)

LLMs (like the GPT-style transformers) have achieved considerable success in tasks involving text generation and symbolic manipulation. These models can even reason through complex problems. However, their core architecture reveals important limitations. These limitations especially regard the criteria developed from RPT and weak IIT. Introduced by Vaswani and colleagues, the Transformer framework prioritizes self-attention but eliminates recurrence (13). While this enables high parallelism and scalability, traditional LLMs which follow this architecture diverge from a core requirement of RPT. Specifically, RPT requires the presence of intrinsic feedback loops that stabilize and refine content over time (5). Despite this, the LLM design is still computationally efficient, leading to the astounding results observed in current LLMs. Yet, in transformers, once a forward pass is completed, there is no closed-loop recurrence within the same processing episode, which raises doubts about their suitability as candidates for phenomenal consciousness.

However, within the AI community, recurrence is becoming more central and is being reintroduced into transformer-based architectures. Some transformer models have been configured to a specific type of memory (memory tokens) which allows information to be fed back to previous parts of the system. The Recurrent Memory Transformer (RMT) has a form of recurrent design while keeping the overall efficiency of the original transformer model intact (14) (Figure 1). Additionally, the RWKV and Liger models both combine transformer-like training efficiency with recurrent-style inference. The RWKV model mixes the efficiency of transformers with the looping style of traditional recurrent networks (15). The Liger model "recurrentizes" pre-trained transformers weights by adding recurrent elements into otherwise feedforward systems, and it does so without a dramatic increase in parameter count (16). Although recurrence was initially excluded from transformer design, it is now being reconsidered as an important architectural component.

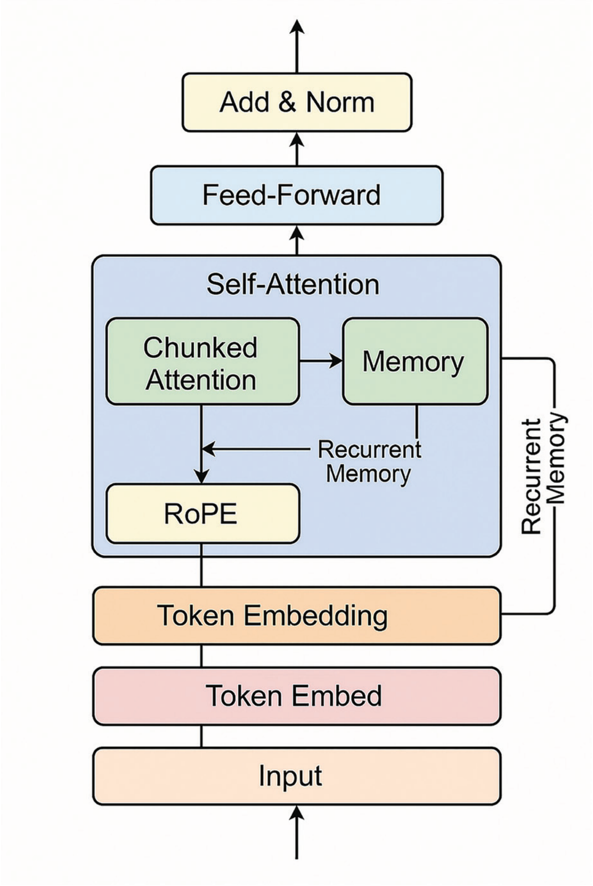


Figure 1. Architecture diagram of a Recurrent Memory Transformer. Adapted from Kashyap, "Recurrent Memory-Augmented Transformers with Chunked Attention for Long-Context Language Modeling" (17). The model is structurally similar to a standard transformer but with an added memory loop. Each input is turned into tokens and then embedded. After, it passes through position encodings (RoPE) and chunked attention. The memory module keeps track of past inputs and sends them back into the system, so the model can hold on to context over time. The recurrent loop lets information carry forward while preserving earlier states, and it makes the model operate more continuously than one-shot feedforward transformers.

Despite these developments, LLMs still face challenges with satisfying weak IIT's criteria of being unified, irreducible, and integrated. Stephen Wolfram has noted that modern machine learning systems may contain localized "pockets" of computational irreducibility (18). This means that the system's internal behavior in some parts of the model cannot be shortcut or reduced without simulation of the entire process. In other words, certain parts of the model are complex and unpredictable, but these parts are mostly separate from each other and don't combine into one fully connected whole. Weak IIT, by contrast, holds that conscious experience requires irreducibility across the whole system, where every component contributes to one inseparable informational state. In this case, the pockets of complexity in LLMs do show local unpredictability and sophistication, but not the kind of whole integration and structural unity that weak IIT associates with unified phenomenal consciousness (19).

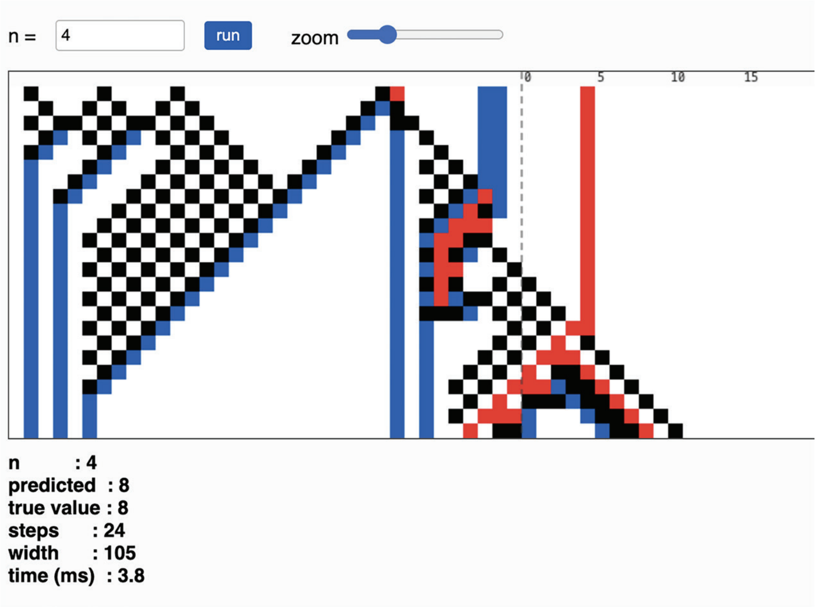
Philosophers of AI consciousness have raised parallel concerns. Susan Schneider argues that conscious AI would require architectures with sophisticated self-modeling capacities and globally integrative information flow, conditions that current transformer systems do not satisfy (20). In addition, Eric Schwitzgebel argues that because different theories of consciousness emphasize different structural or functional features, including temporally extended and causally integrated processing, it remains uncertain how to evaluate AI systems whose architectures diverge sharply from these models. He notes that if theories such as RPT or IIT are correct, architectures like standard LLMs (which lack intrinsic recurrence) would not meet their proposed criteria for unified phenomenal consciousness (21). However, he still stresses that we currently lack the theoretical clarity required to determine which structural features are truly essential. Given the preceding analysis of RPT and weak IIT, Schwitzgebel's caution reinforces the structural limitations already identified in LLMs. If recurrence and integrated causal organization are among the features that matter for phenomenal consciousness (as these theories suggest) then the absence of such properties in standard transformer architectures further underscores their weakness as candidates for unified conscious experience.

David Chalmers likewise argues that current LLMs lack several structural properties emphasized by leading theories of consciousness. He notes that contemporary transformer systems are "almost entirely feedforward" and therefore do not exhibit the recurrent processing that theories such as Lamme's RPT and Tononi's IIT treat central to conscious experience (22). Although Chalmers remains open to the possibility that future "LLM+" architectures incorporating recurrence, memory, multimodal grounding, and unified goal structures could become plausible candidates for consciousness, his analysis implies that present transformer-based LLMs fall short of the relevant structural conditions identified by RPT and weak IIT.

All in all, traditional LLMs demonstrate important limitations regarding recurrence and integration. New LLM architectures are indeed introducing recurrence in transformer-based systems, but the lack of global integration across subsystems is a flaw when it comes to phenomenal consciousness. A new form of AI architecture known as Emergent Models (EMs) provides a more promising direction and offers intrinsic recurrence and irreducible integration, which is examined in the next subsection.

Emergent Models (EMs)

Emergent Models (EMs) are a class of computing systems designed to solve tasks by evolving a simple dynamical rule over time (23). This differs from neural networks, which rely on deep stacks of trained weights in order to produce an output. EMs sit conceptually between two classic models of computation, namely cellular automata and Turing machines. Cellular automata are grids of cells in which each cell follows the same simple local rule (based on its surrounding neighbors) updated in parallel, while Turing machines are step-by-step devices that manipulate symbols on a tape until a halting state is triggered. EMs, in this regard, combine aspects of both systems. From cellular automata, EMs use local updates in parallel across an entire state space, and from Turing machines, they incorporate structured task-solving and have clear readout conditions. In EMs, the input is encoded into the system's initial state, and then a fixed update rule is applied repeatedly across all cells. Finally, the solution "emerges" as the system stabilizes into a recognizable pattern when a halting condition is triggered (i.e. when a majority of cells are in a particular cell state), which is then decoded and interpreted. The update rule itself is simple and fixed, while the "program" lies in the initial conditions and in how the system's own dynamics evolve over multiple timesteps (Figure 2).



```
n          : 4
predicted  : 8
true value : 8
steps      : 24
width      : 105
time (ms)  : 3.8
```

Figure 1. Example of EM-43 performing a doubling task (n = 4 → 8). Adapted from Emergent Models – EM43 Viewer (24). The system begins with a one-dimensional tape initialized as a fixed program (black P cells), a separator (BB), and the encoded input (n + 1 zeroes, followed by a red R marker and trailing 0). While the fixed program provides structure, the true computational rule that dictates the subsequent behavior of the EM is evolved through a genetic algorithm. As the rule iterates across time, local interactions cascade into global patterns. The computation halts once >50% of the tap is filled with B cells, at which point the system decodes n × 2 = 8 by finding the location of the rightmost R cell relative to the 0 point. This illustrates how solutions in EMs emerge from iterative dynamics rather than explicit step-by-step programming.

Figure 2 shows only one task, but the argument does not depend on the doubling problem itself. The example is merely an illustration of the architectural dynamics that distinguish EMs: iterative updating, intrinsic recurrence (recurrence implemented within the architecture's own timestep dynamics rather than across separate forward passes), and globally evolving state patterns. The claims made in this section concern these structural properties, and not the specific computation depicted in the figure. Nothing about the doubling task is meant to suggest that task performance alone has relevance for consciousness; rather, the emphasis lies on the underlying dynamical organization displayed during the computation.

Importantly, no external controller determines each timestep of an EM's evolution. Instead, the answer arises purely from repeated local interactions. This is why they are called "Emergent Models": high-level patterns arise from simple, systematic rules rather than from explicitly programmed behavior. While LLM behavior is also sometimes described as "emergent", that emergence is statistical, encoded in large matrices of learned parameters. EMs differ by repeatedly applying the same simple rule until a stable solution appears. They are temporal, recurrent systems rather than one-shot feedforward ones.

At this point, the connection to the earlier theory of phenomenal consciousness becomes clearer. RPT holds that consciousness requires feedback loops within the brain's processing hierarchy, stabilizing perceptual content. To this end, EMs exhibit a structurally similar strategy. Instead of producing outputs in a single feedforward sweep, EMs refine representations through iterative updates, in which recurrent evolution allows the system to stabilize and maintain information across time (23). This maps onto the stability emphasized by RPT: in EMs, informational states are repeatedly re-amplified until a final stable configuration is reached.

Weak IIT highlights unity and irreducibility as core features of conscious experience, where each component of a state depends on one another and contributes uniquely to the whole. EMs display structural features that resemble these conditions because their state space evolves cohesively: each cell's update depends on its local neighborhood, but the resulting global dynamics cannot be separated into independent parts without losing the overall informational structure. This interdependence is a well-documented feature of cellular automata and recurrent dynamical systems more broadly. Foundational work in cellular automata has shown that simple local update rules can generate globally coordinated patterns whose behavior depends on systemwide interactions rather than independently functioning parts. For example, Mitchell, Crutchfield, and Hraber demonstrate that effective CA-based computation emerges only when large-scale dynamical patterns propagate information across the entire lattice in a unified manner (25).

In the field of dynamical-systems AI research, Beer's analysis of autonomous recurrent agents shows that their behavior arises from the dynamics of the entire coupled system rather than from any separable component, since "properties of the coupled system cannot generally be attributed to either subsystem alone" (26). This demonstrates that recurrent dynamical architectures naturally produce globally integrated, non-decomposable state evolutions, a structural property weak IIT associates with unity and irreducibility. EMs share this feature, since their iterative updates generate systemwide trajectories in which small local changes propagate through the entire state, yielding behavior that depends on the configuration of the whole system.

In summary, Emergent Models appear more naturally aligned with the structural features associated with phenomenal consciousness than standard LLMs. EMs' recurrent, iterative evolution parallels the stability emphasized by RPT, and their globally interdependent state space resembles the unity stressed by weak IIT. By contrast, LLMs struggle to explain why experience feels unified rather than disjointed because of their fragmented pockets of computation. However, none of this is to prove that EMs are conscious. Rather, for these reasons, EMs provide a more promising architectural candidate for modeling the structural conditions that theories like RPT and weak IIT associate with phenomenal experience.

ON PHENOMENAL VS. ACCESS CONSCIOUSNESS

A central implication of the combined RPT and weak IIT framework is that the mechanisms responsible for generating conscious experience may stabilize and integrate more information than what later becomes available for deliberate report or reasoning. This possibility introduces a structural distinction between phenomenal consciousness and access consciousness. Determining whether such a distinction is coherent and empirically grounded is directly relevant to assessing the plausibility of the larger theoretical framework. Accordingly, this section examines whether conscious experience can exceed what is accessed by analyzing the Sperling paradigm and the major competing interpretations it has generated. The subsequent discussion evaluates how RPT and weak IIT characterize the underlying mechanisms of early visual processing, and how Emergent Models (EMs) provide a concrete way to operationalize these distinctions in artificial systems.

The Sperling Experiment

One of the earliest examples that phenomenal consciousness may differ from access consciousness comes from the Sperling experiment (27). In this study, people were shown a grid of twelve letters for only a fraction of a second. When asked to recall the whole grid in a comprehensive report, they could only name three or four letters. However, if a tone is played right after the display, it cued them to recall just one row by being asked to name a single letter from a chosen row. Under this condition, they could report nearly all the letters from that chosen row. Ned Block interprets this as evidence that the subjects phenomenally experienced the entire grid for a brief moment, even though they could only access a smaller part of it for report at once. This suggests that phenomenal experience "overflows" access (27).

Michael Cohen and Daniel Dennett, however, argue that this doesn't prove a real separation between the two kinds of consciousness, instead claiming that the cue merely directs which information becomes accessed and reportable. On their interpretation, there is no need to claim that the entire grid was phenomenally experienced; the phenomenon is instead a demonstration of how access is influenced by attention and memory cues. Calling the unreported letters "experienced but inaccessible" introduces a category that, on their view, lacks independent evidence. Thus, Sperling illustrates how cues influence access, not that experience exceeds access (28).

Although Block's overflow thesis is influential, several competing theories challenge the claim that phenomenal consciousness exceeds access. Attention-based accounts argue that phenomenal and access consciousness coincide because the mechanisms that stabilize representations for report are the same mechanisms that generate conscious experience. Jesse Prinz defends this position by identifying consciousness with attended intermediate-level representations, proposing that without attention a representation is not conscious at all (29). Model-based theories offer a related challenge. Michael Graziano's Attention Schema Theory (AST) holds that consciousness arises when the brain constructs a simplified internal model of its own attentional processes, such that only information represented as accessible within this schema becomes conscious; representations outside of it are classified as unconscious (30). Neurocomputational approaches similarly caution that impressions of rich, pre-access phenomenal fields may arise from post hoc interpretive processes rather than from a genuinely distinct domain of conscious contents. Paul Churchland argues that introspection is an unreliable guide to cognitive architecture and that positing hidden phenomenal states risks making theoretical entities unsupported by empirical evidence real (31).

These alternative interpretations highlight that the Sperling paradigm does not, on its own, resolve whether phenomenal consciousness outstrips access. The theoretical dispute instead turns on how different models characterize the mechanisms that generate conscious experience. Attention-based, model-based, and neurocomputational accounts interpret the uncued rows as unconscious because they lack the functional markers tied to consciousness on those theories. The following account develops how RPT and weak IIT interpret the structure of early visual processing and why, under these frameworks, the experiment is consistent with a transient phenomenal field that exceeds what becomes accessible for report. This analysis does not presuppose Block's overflow thesis; rather, it assesses the Sperling data through the lens of independently motivated mechanistic principles. Whether one ultimately sides with Block or his critics, the value of the RPT and weak IIT account lies in clarifying what these theories predict should occur during brief, high-capacity visual presentations.

The RPT and Weak IIT Account

The combined RPT and weak IIT view explains the Sperling experiment without mystery by spelling out what the "contents" of phenomenal consciousness are and how they behave over periods of time.

When the grid first appears, early visual areas figure out the individual features of each letter like the edges, brightness, and position. These feature patterns loop back through short feedback circuits between visual regions. The recurrent feedback briefly stabilizes many letter-patterns at once even after the image disappears, with each stabilized pattern being a phenomenal content in the simplest sense: what it feels like to see that letter in that location.

After recurrence stabilizes individual letters, integration links those stabilized patterns to each other inside the sensory system itself. This occurs before any of them are sent to higher areas. Horizontal (side-to-side) and feedback (top-down) signals weave the local loops together so that the visual cortex represents the entire grid as one pattern. This is about forming a self-consistent sensory map, not about making the contents available to language or decision. In this self-consistent sensory map, each part depends on its neighbors, so in weak IIT terms, the system's structure at this stage has high internal connectivity. That is, removing any part would change how the rest behaves. At this point, the experience feels like a single visual scene rather than separate flashes, even though none of it has yet been "broadcast" to the higher-order systems that handle report and reasoning.

Only some stabilized, integrated contents are then broadcast to higher-order systems for language and decision. These become access contents. The cue tone works by steering attention toward the still-active portion of the integrated field (the cued row) before these stabilized patterns decay. Because of this, participants can report almost all letters from that row. The remaining rows were part of the unified phenomenal field for a brief moment, but never reached access before fading.

From this perspective, Block's stance is not that there are "mysterious" experiences that can never be studied. Using RPT with weak IIT, Block's notion of "overflow" can be interpreted as the claim that recurrent and structured contents can exist and be integrated before the system uses them. Dennett and Cohen think of this sequence differently, claiming that until a representation is accessed by higher cognitive systems (attention, working memory, or language), the early sensory trace in the Sperling experiment is not conscious at all. On this interpretation, the entire letter grid being "phenomenally experienced" before the cue is an unnecessary postulate.

RPT with weak IIT offers a way to make sense of why that interpretation may be too narrow. Recurrent feedback allows the visual system to keep many letter representations active at once, and structural integration combines these stabilized elements into a single conscious field. This means that the brain temporarily sustains a structure that already has the organization and stability required for consciousness. The structure is not a collection of unconscious pieces, but a tightly knit grid from which some portions are accessed. To this end, the cue does not create the conscious contents or bring them into being for the first time; it selects from a pre-existing, recurrently stabilized, and integrated field corresponding to Block's "overflow" of phenomenal consciousness.

Emergent Models and Making Phenomenology Meaningful

The concept of "phenomenal consciousness", according to Dennett, is scientifically meaningless because it cannot be measured apart from access (32). Emergent Models (EMs) challenge that claim by acting as a rough model to connect structure and experience-like organization. As EMs run, they often develop stable and integrated patterns before any decoding happens or output is produced (Figure 2). These patterns can be tracked over time, along with the dependence of the parts on each other and the degree to which the whole state depends on all its components. EMs therefore provide an operational definition of the very features (recurrence and integration) that RPT with weak IIT link to phenomenal contents, but they do so in a system where every internal state is observable.

This does not demonstrate that EMs are conscious. It does, however, show that the category "integrated, pre-access structure" is coherent and measurable. Once such structures are well-defined in an artificial system, they can be compared with biological data, such as time-varying neural patterns in humans, to determine whether similar structural signatures appear before access and predict what later becomes reportable. This approach does not "measure qualia", nor does it assume that these structures are qualia, cause qualia, or merely correlate with them. It examines a clear, structural hypothesis about the conditions under which experience-like organization emerges, which is enough to make the phenomenal/access distinction scientifically meaningful. If such structures consistently appear prior to access and matter for what later gets accessed, the distinction gains empirical support; if not, it loses support. Either way, it is testable in principle, contrary to Dennett's claim.

What If Phenomenal and Access Consciousness Turn Out to Coincide?

Suppose future work shows that whenever a content is recurrently stabilized and integrated, it is always accessed. In that case, phenomenal and access consciousness coincide in practice. The combined theory of RPT and weak IIT still stands, but its interpretation shifts. The theory would then identify the shared mechanism for both: recurrence creates stable contents, integration unifies them, and broadcast expresses them. In this scenario, the phenomenal/access distinction collapses, yet the core mechanism remains. Any conscious system (biological or artificial) must implement recurrent stabilization within a structurally integrated architecture. The theory therefore continues to provide a precise account of what makes a conscious state possible.

CONCLUSION

This paper has argued that the synthesis of RPT and weak Integrated Information Theory (IIT) offers the most coherent and empirically grounded account of phenomenal consciousness. By linking recurrence to the stabilization of conscious contents and integration to the unity of experience, this combined framework fixes the explanatory gaps that each theory faces on its own. Applying this model to artificial intelligence reveals that LLMs don't have the deeply recurrent and widely integrated architectures that our theory associates with consciousness. Emergent Models, in contrast, more closely approximate these structural conditions through repeating, self-organizing patterns that resemble how the human brain constantly loops and connects information.

The implications of this framework are beyond the philosophy of the mind. There is now a roadmap for future empirical work (both in neuroscience and AI) in identifying measurable structural signatures that could test claims about conscious-like organization. Ethically, as AI systems become more autonomous and complex, understanding which designs might induce experience becomes an increasingly urgent concern. Ultimately, the synthesis of RPT and weak IIT supports a rethinking of consciousness as a structural and temporal phenomenon instead of a functional result. A phenomenon that can, in theory, be studied, modeled, and perhaps even engineered through AI models.

ACKNOWLEDGEMENTS

I would like to sincerely thank my mentor Dr. Alasdair Craig for his constant feedback and encouragement throughout this project. I would also like to thank the Horizon Academic Research Program for offering me this wonderful opportunity to explore my interests.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. Rosenthal DM. 2. Varieties of higher-order theory. In: Higher-Order Theories of Consciousness. Amsterdam: John Benjamins Publishing Company; 2004; p.17–44. https://doi.org/10.1075/aicr.56.04ros
2. Graziano MSA, Kastner S. Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cogn Neurosci [Internet].* 2011; 2 (2): 98–113. https://doi.org/10.1080/17588928.2011.565121
3. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci [Internet].* 2013; 36 (3): 181–204, https://doi.org/10.1017/S0140525X12000477
4. Baars BJ. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Prog Brain Res [Internet].* 2005; 150: 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9
5. Lamme VA, Roelfsema PR. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci [Internet].* 2000; 23 (11): 571–9. https://doi.org/10.1016/s0166-2236(00)01657-x
6. Tononi G. An information integration theory of consciousness. *BMC Neurosci [Internet].* 2004; 5: 42 [cited 2025 Oct 18]. Available from: https://doi.org/10.1186/1471-2202-5-42
7. Seijdel N, Loke J, van de Klundert R, van der Meer M, *et al.* On the necessity of recurrent processing during object recognition: It depends on the need for scene segmentation. *J Neurosci [Internet].* 2021; 41 (29): 6281–9. https://doi.org/10.1523/JNEUROSCI.2851-20.2021
8. Peterson MA, Campbell ES. Backward masking implicates cortico-cortical recurrent processes in convex figure context effects and cortico-thalamic recurrent processes in resolving figure-ground ambiguity. *Front Psychol [Internet].* 2023; 14: 1243405. https://doi.org/10.3389/fpsyg.2023.1243405
9. Mashour GA. Anesthesia and the neurobiology of consciousness. *Neuron [Internet].* 2024; 112 (10): 1553–67. https://doi.org/10.1016/j.neuron.2024.03.002
10. Albantakis L, Tononi G. The intrinsic cause-effect power of discrete dynamical systems—from elementary cellular automata to adapting animats. *Entropy (Basel) [Internet].* 2015; 17 (8): 5472–502, https://doi.org/10.3390/e17085472
11. Butlin P, Long R, Elmoznino E, Bengio Y, *et al.* Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv [cs.AI]. Available from: https://arxiv.org/abs/2308.08708 (accessed on 2025-09-19).
12. Boly M, Moran R, Murphy M, Boveroux P, *et al.* Connectivity changes underlying spectral EEG changes during propofol-induced loss of consciousness. *J Neurosci [Internet].* 2012; 32 (20): 7082–90. https://doi.org/10.1523/JNEUROSCI.3769-11.

2012

13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, *et al*. Attention is all you need [Internet]. arXiv [cs.CL]. 2017. Available from: https://arxiv.org/abs/1706.03762 (accessed on 2025-09-30). https://doi.org/10.65215/ysbyhc05
14. Bulatov A, Kuratov Y, Burtsev MS. Recurrent Memory Transformer [Internet]. arXiv [cs.CL]. 2022. Available from: https://arxiv.org/abs/2207.06881 (accessed on 2025-10-18).
15. Peng B, Alcaide E, Anthony Q, Albalak A, *et al*. RWKV: Reinventing RNNs for the Transformer era [Internet]. arXiv [cs.CL]. Available from: https://arxiv.org/abs/2305.13048 (accessed on 2025-09-30).
16. Lan D, Sun W, Hu J, Du J, Cheng Y. Liger: Linearizing large language models to gated recurrent structures [Internet]. arXiv [cs.CL]. 2025. Available from: https://arxiv.org/abs/2503.01496 (accessed on 2025-10-09).
17. Kashyap A. Recurrent memory-augmented transformers with chunked attention for long-context language modeling [Internet]. arXiv [cs.LG]. 2025. Available from: https://arxiv.org/abs/2507.00453 (accessed on 2025-10-18).
18. Wolfram S. What's really going on in machine learning? Some minimal models. Stephen Wolfram Writings [Internet]. 2024. Available from: https://writings.stephenwolfram.com/2024/08/whats-really-going-on-in-machine-learning-some-minimal-models/ (accessed on 2025-09-10). https://doi.org/10.31855/e0e30753-e3f
19. Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci [Internet]*. 2016; 17 (7): 450–61. https://doi.org/10.1038/nrn.2016.44
20. Schneider S. Artificial You: AI and the Future of Your Mind. Princeton, NJ: Princeton University Press; 2019. https://doi.org/10.2307/j.ctvfjd00r
21. Schwitzgebel E. AI and Consciousness [Internet]. arXiv [cs.AI]. 2025. Available from: https://arxiv.org/abs/2510.09858 (accessed on 2025-11-25).
22. Chalmers DJ. Could a Large Language Model be Conscious? [Internet]. arXiv [cs.AI]. 2023. Available from: https://arxiv.org/abs/2303.07103 (accessed on 2025-11-25).
23. Bocchese G. Emergent models: Machine learning from cellular automata [Internet]. ResearchHub. 2025 [cited 2025 Sep 30]. Available from: https://www.researchhub.com/post/4073/emergent-models-machine-learning-from-cellular-automata (accessed on 2025-09-30). https://doi.org/10.55277/ResearchHub.70e8enig
24. Bocchese G. Emergent model - EM43 - number doubler [Internet]. Github.io. Available from: https://bocchesegiacomo.github.io/em43viewer/ (accessed on 2025-11-28).
25. Mitchell M, Crutchfield JP, Hraber PT. Evolving cellular automata to perform computations: mechanisms and impediments. Physica D [Internet]. 1994; 75 (1–3): 361–91. https://doi.org/10.1016/0167-2789(94)90293-3
26. Beer RD. A dynamical systems perspective on agent-environment interaction. *Artif Intell [Internet]*. 1995; 72 (1–2): 173–215. https://doi.org/10.1016/0004-3702(94)00005-l
27. Block N. Perceptual consciousness overflows cognitive access. *Trends Cogn Sci [Internet]*. 2011; 15 (12): 567–75. https://doi.org/10.1016/j.tics.2011.11.001
28. Cohen MA, Dennett DC. Consciousness cannot be separated from function. *Trends Cogn Sci [Internet]*. 2011; 15 (8): 358–64. https://doi.org/10.1016/j.tics.2011.06.008
29. Prinz J. The conscious brain: How attention engenders experience. Cary, NC: Oxford University Press; 2012.
30. Graziano M. Rethinking consciousness: A scientific theory of subjective experience. New York, NY: WW Norton; 2019.
31. Churchland PM. Matter and consciousness: A contemporary introduction to the philosophy of mind. London, England: MIT Press; 1999.
32. Dennett DC. The fantasy of first-person science. In: The Map and the Territory. Cham: Springer International Publishing; 2018; p.455–73. https://doi.org/10.1007/978-3-319-72478-2_26