
HOW LANGUAGE MAKES INTELLIGENCE GENERAL

PREPRINT

Ean Huddleston 

Independent Researcher
ean@eanhuddleston.com

October 31, 2026

ABSTRACT

Large language models (LLMs) exhibit general-purpose reasoning even when trained on text alone. This raises a fundamental question: could language be not merely a tool for expressing preexisting thought, but a constitutive component of general intelligence? Here I address this question by analyzing what open-ended causal modeling demands of any finite system. These demands expose a core architectural tension: maintaining an adaptive world model requires graded, continuous plasticity, whereas constructing novel causal chains requires systematic, sequential commitment. From these competing constraints, I derive Emergent Symbolic Cognition (ESC), a computational theory proposing that general intelligence arises when a continuous, adaptive substrate—whether biological or artificial—internalizes the relational structure of a symbolic framework like language. This internalization serves a dual function: it organizes the substrate’s continuous geometry into an abstract world model, and it supplies the composable elements for iterative symbolic generation. As each predicted symbol feeds back to reshape the internal state, it acts as a discrete ratchet that converts parallel associative processing into the sequential, step-by-step construction required to solve novel, dependency-structured problems. Because ESC is derived from substrate-agnostic constraints, a single architectural principle accounts for both the unique generality of human intelligence and the emergent capabilities of LLMs. From this architecture, I derive falsifiable predictions (including a sharp decay in the neural or hidden-state decodability of downstream symbols) and evaluate the theory against converging evidence from cognitive science and artificial intelligence—explaining why consolidated reasoning can operate without active language use, and showing how symbolic input alone can ground world models without direct sensory contact. Ultimately, ESC reframes language as a culturally inherited architecture for general-purpose cognition.

Keywords large language models, language and thought, general intelligence, emergent symbolic cognition, chain-of-thought, cognitive architecture, causal modeling, world models

“A long and complex train of thought can no more be carried on without the aid of words, whether spoken or silent, than a long calculation without the use of figures or algebra.”

— Charles Darwin, *The Descent of Man*, 1871

1 Introduction

Homo sapiens has modeled the invisible forces of the atom, calculated the trajectory to the moon, and deciphered the code of life. These accomplishments reflect a *general* form of intelligence: the ability to solve novel problems across a wide and potentially unbounded range of domains. This capacity is particularly remarkable given that only ≈ 5 to 8 million years separate us from our last common ancestor with chimpanzees (e.g., Wood, 2010). What allowed such a leap in cognitive capability to occur within such a short evolutionary window? Was it simply a matter of a larger brain, or did something qualitatively new emerge in the human mind?

Language is a prominent candidate. With its recursive and infinitely generative structure, it is a uniquely human faculty (e.g., Berwick & Chomsky, 2016; Hauser et al., 2002; Pinker, 1994). But whether language merely communicates thoughts that already exist (e.g., Fedorenko et al., 2024; Fodor, 1975; Pinker, 1994) or actively reshapes how the mind thinks (e.g., Bickerton, 1990; Dennett, 1991; Vygotsky, 1934/1986) remains unresolved. The definitive experiment—withholding language from a developing mind—is ethically impossible, leaving us reliant on rare cases of extreme isolation that yield confounded data and contentious interpretations (e.g., Curtiss, 1977; Skuse, 1984). We are left in a scientific bind: the very strength of the conviction that language is essential to a developing mind makes the definitive test of that conviction impossible to conduct.

However, over the past several years, large language models (LLMs)—systems built by learning to predict human language—have begun solving novel problems across mathematics, physics, and software engineering. They engage in multi-step reasoning and construct causal explanations of phenomena they have never observed. A purely statistical system now exhibits structured reasoning capacities that theories of language and cognition did not anticipate.

These systems share little in common with the human brain. They lack evolutionary history, embodiment, sensory contact with the world, and innate cognitive endowments. Their general-purpose Transformer architectures (Vaswani et al., 2017) are domain-agnostic sequence processors that apply the same statistical operations to text as to amino acid sequences (Ferruz et al., 2022) or robotic control signals (Chen et al., 2021). They were not designed to reason; they were trained to predict the next word. That two systems this radically different achieve the capacity for general reasoning raises the question of whether there exists an explanation for this capacity at a higher level of abstraction.

In this paper, I argue that there is, and that the critical variable is language. I introduce *Emergent Symbolic Cognition (ESC)*, a computational theory proposing that symbolic reasoning—and with it, general intelligence—emerges when a continuous, adaptive learning system internalizes the structure of an external symbolic framework, of which human language is the primary instance. This internalization equips the system with a core mechanism I call iterative symbolic generation, which transforms what would otherwise be purely associative prediction into the structured, sequential thought required to solve novel dependency-structured problems. Language, on this account, is not something that greater intelligence gives rise to, but a constitutive component of general-purpose reasoning itself.

ESC is motivated not by the particulars of brains or Transformers, but by an analysis of the computational demands facing any finite general intelligence. This is what distinguishes it from prior accounts of the importance of language in thought: because it is based on the demands of the problem, it applies equally to any finite system.

The argument proceeds as follows. I first derive a set of minimal architectural requirements from the computational demands that general intelligence places on a finite system and apply them to the symbolic and connectionist paradigms, revealing why each fell short and how LLMs have disrupted the resulting impasse. From the same requirements I derive the ESC architecture. I then test the theory’s falsifiable predictions against converging findings and key challenges from human cognition and artificial intelligence, situate it relative to existing theories, and draw out its implications for meaning, understanding, and grounding.

2 Constraining the Architecture of General Intelligence

Many nonhuman organisms, such as corvids, demonstrate impressive problem-solving skills highly adapted to their ecological niches. Similarly, machine learning systems exhibit superhuman performance in specialized tasks like Go (Silver et al., 2016) or protein folding (Jumper et al., 2021). These capabilities, however, represent *domain-specific intelligence*. By contrast, the present inquiry focuses on the distinct challenge of *general intelligence*—solving *novel* problems across an open-ended number of domains.

Rather than cataloging the full range of capabilities a general intelligence must possess, this section approaches the problem from the other direction: it seeks to identify a single diagnostic capability whose computational demands are stringent enough to constrain the space of architectures that could, in principle, support it. The resulting requirements will serve as the evaluative standard against which existing approaches are measured (Section 2.2) and from which the ESC architecture is derived (Section 3).

The capability I isolate is causal modeling. This capacity to reason about interventions (“*what if I do...*?”) and counterfactuals (“*what if I had done...*?”) represents the upper rungs of Judea Pearl’s “Ladder of Causation” (Pearl & Mackenzie, 2018). While many tasks can be solved through associative pattern completion, a truly general intelligence must engage with the causal structure of the world, the dependencies governing how states and events relate to one another.

Crucially, this modeling must operate beyond the immediate sensory environment. A corvid manipulating a tool to solve a physical puzzle engages in a form of local causal reasoning with perceptually available variables. General intelligence, by contrast, requires modeling causes and effects that are decoupled from the here and now. This necessitates a system that can manipulate abstract proxies for reality rather than reality itself.

Even elementary instances of such decoupled reasoning reveal the severity of this demand. Consider a simple case: an organism eats something poisonous, but the illness does not manifest until more than 24 hours later. Despite the obvious survival advantage of identifying the cause, and despite hundreds of millions of years of animal evolution, no nonhuman species is known to be capable of making this connection.

What makes even this basic problem intractable is the search it demands. Within a single 24-hour window, an organism encounters a vast number of candidate causes—every item ingested, every location visited, every substance contacted. Without a mechanism for constructing and evaluating causal hypotheses, the correct link between ingestion and illness cannot be reliably identified from among these alternatives. Even conditioned taste aversion (the associative mechanism natural selection has directly optimized for this problem) reaches its limit at roughly this interval (e.g., Etscorn & Stephens, 1973; Garcia et al., 1966).

And this is merely a one-step inference over a single day. General intelligence requires causal modeling that extends far beyond the immediate environment, spanning vast distances in time and space, or operating at scales imperceptible to biological senses. Once the search space expands to include the deep past, the deep future, and the invisible, any given state has a multitude of potential antecedents, and any action a multitude of potential consequences. The number of possible contingent pathways explodes combinatorially. Navigating this enormous space makes it computationally infeasible for a finite system to rely on prestored answers; a general intelligence must therefore be a constructive system, capable of generating novel solutions on the fly.

This enormous scale might suggest that the solution lies in brute-force parallelism—evaluating many possible causal chains simultaneously. But while potential next steps or even entire chains can be explored in parallel, the construction of any single chain is inherently sequential, as each step is contingent on the one before it (e.g., Newell & Simon, 1976; Pearl, 2009).

2.1 Five Requirements for Causal Modeling and World Construction

Together, the demands of sequential causal construction impose five stringent architectural requirements. These divide into two distinct categories: the mechanics of the constructive process itself, and the nature of the internal world model¹ that makes such processing viable.

To model causal pathways for novel problems, a system must be able to build explanations and plans (chains of dependencies linking a hypothesized cause to an observed effect, or a current state to a desired outcome). This constructive process imposes the first three requirements:

1. **Generative Composition:** The architecture must construct novel causal chains through the systematic combination and recombination of its internal representations. This allows it to form new hypotheses or plans by assembling known cause-and-effect relationships in new ways (e.g., Fodor & Pylyshyn, 1988; Marcus, 2001; Newell & Simon, 1976).
2. **Sequential Processing:** The architecture must support serial, step-by-step operations. This is the necessary mechanism for constructing causal chains, where each event is contingent on the one preceding it and solutions must be constructed in sequence (Elman, 1990; Newell & Simon, 1976; Pearl, 2009).
3. **Goal-Directedness:** To transform an unconstrained generative process into effective causal modeling, the architecture must include a mechanism for representing internal states or criteria (hereafter, Goals) that provide the necessary constraints to guide search and construction (e.g., Newell & Simon, 1972; Simon & Newell, 1971).

To support the demands of generative, sequential causal modeling, the internal world model itself must meet two further requirements:

4. **Continuous Adaptation:** Because real-world evidence is noisy, incremental, and open-ended, the system’s internal model cannot rely on static primitives or all-or-nothing rule revisions. It must support graded plasticity—continuously adjusting its relational structure in response to statistical evidence so its representations remain viable in a dynamic environment (e.g., Hinton et al., 2006; LeCun et al., 2015).
5. **Abstract Representation:** Because novel problems rarely replicate past experience in their surface details, the system cannot merely store concrete instances. It must compress experience into latent representations organized within a continuous similarity space, encoding generalizable causal principles that can be extended to unfamiliar contexts by degree of structural correspondence (e.g., Hinton et al., 2006; LeCun et al., 2015).

Taken together, these five requirements constitute a minimal but necessary benchmark for any general intelligence—and expose a deep architectural tension. Whereas the constructive process (1–3) demands structured, serial commitment, the world model (4–5) demands graded, continuous plasticity; any theory that cannot offer a coherent account of how a single architecture satisfies all of them is fundamentally incomplete.

2.2 The Historical Stalemate and the LLM Disruption

Measured against this benchmark, the two dominant paradigms in artificial intelligence—the symbolic and the connectionist—have historically failed in complementary ways. The symbolic approach, rooted in logic and early AI, posited that intelligence arises from manipulating discrete symbols according to explicit rules (Newell & Simon, 1976). These systems excelled at Generative Composition and Sequential Processing, but they proved brittle, unable to satisfy the requirement for Continuous Adaptation when faced with messy, real-world data.

Conversely, the connectionist paradigm, inspired by neural biology (McCulloch & Pitts, 1943), was built for Continuous Adaptation. These systems learn by adjusting connection strengths based on statistical

¹The term “world model” is sometimes used to refer narrowly to sensorimotor prediction models (e.g., LeCun, 2022). I use it in its fuller sense: a relational structure encoding causal, temporal, social, and abstract dependencies—not merely physical dynamics.

patterns, allowing them to form nuanced, latent Abstract Representations (e.g., Hinton et al., 2006; LeCun et al., 2015). However, they historically failed to satisfy the requirements for structured construction: lacking any mechanism to commit the system to discrete steps against its default dynamics, purely continuous architectures struggled with the systematic Generative Composition and reliable, multi-step Sequential Processing that were symbolic AI’s strength (Fodor & Pylyshyn, 1988; Marcus, 2001).

This created a theoretical stalemate, with each paradigm mastering precisely the capabilities the other lacked. Hybrid models (notably cognitive architectures such as ACT-R and SOAR) sought to bridge this divide (e.g., Anderson et al., 2004; Laird, 2012; Lake et al., 2017), and formal arguments were advanced that compositionality could emerge within neural computation itself (Smolensky, 1988), yet a unifying architecture remained elusive.

The advent of LLMs has profoundly disrupted this landscape. Built on purely connectionist foundations, these models would seem to represent a straightforward triumph of the Continuous Adaptation approach. Yet they paradoxically exhibit a robust capacity for systematic Generative Composition and Sequential Processing: they follow complex instructions, perform few-shot learning, engage in step-by-step reasoning, and generate functional code—behaviors demanding compositional competence traditionally thought to be outside the reach of purely statistical systems. By showing that a purely connectionist system can support systematic symbolic behavior, LLMs have broken the classical stalemate, raising the question of how a single architecture can reconcile these seemingly incompatible demands.

3 Emergent Symbolic Cognition (ESC)

I argue that the requirements of Section 2 rule out any architecture that keeps symbolic and continuous components separate. A system that relies on built-in symbolic primitives—whether engineered or innate—cannot fully satisfy Continuous Adaptation, because static primitives cannot be modified by experience when novel domains demand it. For an intelligence to be truly general, its symbols must be indefinitely malleable, adjusting and adapting as needed. This leaves only one architectural possibility: symbolic capabilities cannot be built in separately, but must emerge as a learned property of the continuous substrate itself.

Emergent Symbolic Cognition (ESC) specifies how this occurs. Its central claim is that compositionality emerges from statistical learning under a specific condition: when an adaptive learning substrate internalizes the structure of an external symbolic framework. On this account, symbolic structure is neither architecturally given nor an illusion of subsymbolic processing; it is absent prior to exposure to the framework and functionally real afterward. Through this process, the agent builds a relational world model capable of the sequential, generative operations that causal modeling demands. I first describe its core components and then the dynamic mechanism—iterative symbolic generation—that drives the process. Notably, while this machinery is derived from the rigorous demands of causal modeling, it will be argued to be effectively domain-general, providing the fundamental basis for open-ended reasoning.

3.1 The Architecture for Causal Modeling

The architecture couples two complementary elements—an adaptive Substrate and a Structured Symbolic Framework—whose interaction is detailed in turn below.

3.1.1 The Adaptive Substrate

The first component is the core learning mechanism required to satisfy the world-model constraints of Section 2.1: a *Functionally Continuous Adaptive Substrate* (hereafter, Substrate). Functionally, the Substrate is any adaptive dynamical system that embeds the statistical regularities of incoming sequences into a continuous, high-dimensional state space to drive prediction. Biological neural networks and engineered architectures such as Transformers represent two physically distinct instantiations of this computational class, differing in implementation while sharing the invariant properties required for graded learning and

abstraction. The Substrate provides the structural mechanism for learning rather than innate symbolic content. In biological agents, it arrives already equipped with innate, domain-specific capacities (e.g., Spelke & Kinzler, 2007); the present theory concerns how such a system acquires the additional capacity for domain-general reasoning.

The Substrate’s defining feature is its graded plasticity. Operating in a continuous state space, it incrementally adjusts its internal transition geometry to capture statistical regularities, dynamically refining its relational structure in response to new evidence (thereby satisfying the requirement for Continuous Adaptation outlined in Section 2.1).² In doing so, it compresses the structure of its input into an elaborate, high-dimensional world model that encodes generalizable principles as geometric relationships within a continuous similarity space, satisfying the corresponding requirement for Abstract Representation.

The Substrate’s expressive power is rooted in the combinatorial scale of this continuous state space. To represent complex, overlapping relational structures without interference, the state space must be high-dimensional. In a contemporary LLM with a 20,000-dimensional internal state, for example, even under conservative assumptions about numerical precision, the number of meaningfully distinct states easily exceeds $10^{6,000}$. This number dwarfs physical benchmarks, such as the 10^{80} atoms in the visible universe, illustrating the vast representational capacity available for modeling the world.

This expressiveness, however, does not answer the question of how the system navigates this space. In typical operation, external input largely drives state transitions—sensory data in a brain, or the input sequence in an LLM. Left to its own internal dynamics, a continuous statistical system will relax toward its strongest attractors, the configurations most reinforced by prior experience. But as established in Section 2, causal modeling requires the sequential construction of novel causal chains decoupled from immediate input—chains whose correct solution depends on a precise arrangement of elements that the Substrate, left to its own dynamics, would never arrive at. The Substrate provides the representational capacity; what it lacks is a mechanism to guide its own traversal, step by step, under internal control alone.

3.1.2 The Structured Symbolic Framework

The *Structured Symbolic Framework* (hereafter, Symbolic Framework) provides this missing mechanism. It is a system of discrete symbols and rules for their combination that, when internalized by the Substrate, imposes the operational structure needed for systematic, step-by-step reasoning. But the Symbolic Framework does more than scaffold the process; it simultaneously supplies content for the world model itself, encoding the causal and relational structure of what it describes. For the task of modeling an open-ended, dynamic world, this framework requires two key features:

1. **Discrete Symbols:** Symbols (e.g., words, morphemes, or the signs of a sign language) act as distinct units whose meaning arises from their use in structured sequences. Their meanings must be flexible, allowing new symbols to be introduced and old ones adapted, with references spanning from concrete objects to abstract ideas.³
2. **Rules for Combination:** Constraints (such as syntax) govern how symbols are combined into meaningful sequences. These rules are not explicitly codified but are internalized by the Substrate through exposure to patterns in symbolic sequences.

Together, these two features define the serial medium. But is it expressive enough to guide the Substrate through its vast state space without becoming a functional bottleneck? Shannon’s analysis of English entropy suggests even a short 20-word sequence still permits approximately $2^{90} \approx 1.24 \times 10^{27}$ distinct, plausible

²While biological substrates adapt continuously across timescales, current pre-trained LLMs separate this plasticity into a training phase (where parameters are updated) and an inference phase (where weights remain static and adaptation is restricted to context-conditioned state transitions).

³Formal systems like logic and mathematics differ in this respect: their symbols have stipulated, invariant meanings and explicitly codified rules—powerful within fixed domains, but not open-endedly extensible, and semantically parasitic on the broader Symbolic Framework.

variations.⁴ Approximately 5×10^{19} words have been spoken in all of human history,⁵ meaning that fewer than 5×10^{19} unique 20-word sequences could ever have been uttered, still over seven orders of magnitude short of the combinatorial space available. All of human speech has barely touched that space. The Symbolic Framework is therefore an expressive medium whose combinatorial depth is sufficient to construct arbitrarily complex lines of reasoning.

The interaction between the Symbolic Framework and the Substrate satisfies the requirements for causal modeling by enabling Generative Composition. This capacity allows the system to construct novel causal chains by systematically combining and recombining abstract representations. This symbolic machinery is what enables the explicit construction of causal chains across temporal horizons, transforming a Substrate like the human brain from a reactive system into a proactive simulator (e.g., Bickerton, 1990; Hockett, 1960; Suddendorf & Corballis, 2007). It allows for retrospective analysis, enabling the agent to trace events backward to understand a cause (*“Why did the clay pot crack in the fire?”*) or to construct counterfactuals to explore unchosen paths. Conversely, it enables prospective planning, allowing the agent to simulate future interventions to achieve a goal (*“How can I build a shelter that withstands the wind?”*) before lifting a single stone (Dennett, 1996).

3.2 Searching the State Space

Having introduced the two core components of the ESC architecture, I now turn to the dynamic process that brings them to life and satisfies the remaining requirements for causal modeling established in Section 2.

The core mechanism is an iterative feedback loop: the Substrate, conditioned by a goal and the preceding sequence, generates a symbol.⁶ This symbol is not merely an output; it is fed back as input, modifying the Substrate’s internal state to guide the generation of the next symbol. This loop transforms the Substrate’s natively parallel processing into a sequential, step-by-step construction, effectively serializing the Substrate’s vast associative power.^{7,8} This serialization is required by the causal modeling constraint established in Section 2. But it simultaneously resolves a representational challenge: encoding open-ended meanings through a finite symbol set is itself a combinatorial problem that serial generation renders tractable.

Generating a discrete symbol acts as a ratchet against the Substrate’s tendency to relax toward prior attractors (Section 3.1.1): it commits the system to a specific point in its state space, forcing the next computation to proceed from that position rather than from wherever the default gradients would have carried it. It is the causal regularities internalized from the symbolic input that structure the landscape through which this ratcheting proceeds.

This iterative process unlocks a search space of astronomical scale. For any nontrivial problem, the number of possible symbolic paths explodes combinatorially. Navigating this space requires more than a simple greedy search (always selecting the most locally probable next symbol) and instead demands a true goal-directed heuristic search.⁹ This involves not only generating plausible next steps but also the critical

⁴Based on Shannon’s estimate of roughly 1 bit of information per letter (Shannon, 1951). Assuming an average of 4.5 letters per word yields 4.5 bits per word.

⁵This estimate assumes 117 billion humans ever lived (Haub & Kaneda, 2022) $\times$ 73 years (modern life expectancy) $\times$ 365 days $\times$ 16,000 words/day (Mehl et al., 2007) $\approx 5 \times 10^{19}$, a deliberate overestimate that assumes speech from birth and uses modern rather than historical life expectancy.

⁶While symbols are the primary carriers of meaning in the Symbolic Framework, the particular granularity of what is generated during each step of iterative generation will vary depending on the Substrate (e.g., phonemes, graphemes, or in LLMs, tokens).

⁷While the Transformer architecture is designed to process sequences, it does so in a single, parallel pass over the input. The serial, step-by-step reasoning discussed here is distinct, emerging from the autoregressive loop where the model’s output is sequentially fed back as its own input.

⁸In humans, ESC makes no claim about what proportion of cognition relies on this explicit process versus continuous dynamics such as direct association or mental imagery. The claim is architectural: this capacity must be available when continuous processing alone proves insufficient.

⁹This aligns with Simon’s conception of problem solving as a search through possible symbolic sequences using heuristics (Simon & Newell, 1971), though it differs in the nature of the symbols and heuristics.

functions of evaluation, revision, and backtracking. These functions arise from the Substrate’s capacity to detect prediction error (the dissonance between a generated symbol’s implications and the constraints of the internal world model or sensory input). When a path leads to a dead end, the system must be able to return to a prior step and explore an alternative branch. By supplying Sequential Processing and Goal-Directedness to guide the Generative Composition, Continuous Adaptation, and Abstract Representation established in Section 3.1, this iterative loop completes the five architectural requirements of Section 2.1.

3.3 From Causal Modeling to General-Purpose Computation

We set out to derive the machinery needed for a single capability: causal modeling. But now that the mechanism is in hand, it is worth stepping back to ask what, exactly, we have built. What happens if we strip away the specifics of physical cause and effect—the cracked pot, the wind-resistant shelter? What remains is a simple loop: assess the current state, generate a symbol, use it to condition the next step.

Nothing in this loop is specific to causation. Causal modeling demands that the system serially construct dependency chains through local, conditional operations (“if A , then B , then C . . .”). But any problem whose solution consists of a sequence of steps where each step is contingent on the one before it reduces to the same cycle. Consider multiplying two large numbers by hand. The solver cannot arrive at the final product without first computing each partial result; each step depends on the output of the one before it, and skipping ahead produces nonsense. The mechanism is identical to causal reasoning: iterative symbolic generation proceeding step by step through a dependency structure—in this case, the learned rules of arithmetic rather than physical relationships. The same holds for constructing a logical proof, debugging an algorithm, or planning a route through unfamiliar terrain. All reduce to the same dependency-governed cycle. The mechanism is domain-general.

A system capable of performing arbitrary such cycles over a sufficiently expressive symbol set is Turing-complete—capable, in principle, of simulating any computation that a universal Turing machine can perform (Church, 1936; Turing, 1937). This convergence is fitting rather than coincidental: Turing’s own formalization was derived from a conceptual analysis of a human carrying out a mechanical procedure step by step with symbols on paper—the same core cycle of iterative symbol generation that ESC describes.

The parallel extends beyond the generative cycle. In a Turing machine, the tape is what allows a finite device to solve problems whose complexity exceeds its finite internal memory. In an ESC system, iterative symbolic generation serves precisely the same function.¹⁰ It decouples the complexity of a problem from the fixed dimensionality of the Substrate by externalizing intermediate results into the symbolic sequence. Without this externalization, the system is limited to whatever its continuous internal dynamics can resolve without symbolic commitment—a constraint that, as Section 4.2 will show, accounts for both the characteristic failures and the dramatic performance gains observed in LLMs.

That the mechanism is Turing-complete, however, is a low bar. Even simple cellular automata such as Rule 110 (Cook, 2004) and Conway’s Game of Life (Rendell, 2002) satisfy the criterion. The significance is not the formal label but what it registers: the Substrate has crossed into an open-ended computational regime.

The resemblance invites an obvious question: is the ESC architecture simply a Turing machine by another name? If iterative symbolic generation reduces to the same cycle of read-transform-write, what, if anything, is new?

A classical Turing machine operates via a static transition table and a fixed symbol set, both specified in advance by its designer. An ESC system’s transitions, by contrast, are derived from the continuous,

¹⁰Merrill and Sabharwal (2024) prove that intermediate token generation fundamentally expands the class of problems a fixed-depth Transformer can solve beyond the limits imposed by its architecture alone. Z. Li et al. (2024) establish the precise characterization: without intermediate steps, constant-depth Transformers are confined to shallow, constant-depth circuits; with T steps, they can simulate any Boolean circuit of size T . Pfau et al. (2024) further demonstrate that this expansion depends on the generated tokens carrying distinct symbolic information rather than merely adding forward passes, as a polynomial number of blank filler tokens fails to lift the model out of the constant-depth regime.

learned geometry of the Substrate, making them probabilistic and context-dependent, and its representational repertoire is continuously expanded and refined through the symbolic input it processes. A fixed symbol set and deterministic transitions are what the ESC architecture looks like when its degrees of freedom are maximally constrained. Put differently, the Turing formalism captures what a system can compute *given* a program, but is silent on where programs come from. ESC models precisely this missing process: the derivation of new programs from the learned geometry of the Substrate.

We set out, in Section 2, to identify minimal architectural requirements that any general intelligence must satisfy. Causal modeling served as the diagnostic capability whose demands were stringent enough to constrain the solution space. The result is stronger than anticipated: an architecture that meets those demands turns out to be a general-purpose computational system. Its representational and computational capacity extends, in principle, to anything that can be expressed in symbols. Its power resides not in a fixed program but in the learned structure of its adaptive Substrate.

4 Empirical Predictions and Evidence

The argument so far has been theoretical. This section tests it against the empirical record. ESC does not predict that reasoning, planning, or problem solving require language: nonhuman animals, domain-specific machine learning systems, and nonlinguistic representations plainly support such capacities. Its narrower claim concerns novel, dependency-structured reasoning that is decoupled from immediate perception and extends across domains. This yields two broad empirical predictions: developmentally, internalizing a Structured Symbolic Framework should make possible cognitive capacities that would not otherwise emerge (Section 4.1); online, solving novel problems whose dependencies cannot be resolved directly from the Substrate’s existing associative geometry should require iterative symbolic generation (Section 4.2), a mechanism whose temporal dynamics can be tested most directly via the sharp-decay prediction (Section 4.3). No single finding establishes ESC on its own; rather, each section below pairs converging evidence from human cognition and AI with the prospective conditions under which the theory would fail.

4.1 Without the Symbolic Framework, Domain-General Reasoning Should Not Develop

ESC makes a strong developmental prediction: an adaptive system that never internalizes a Structured Symbolic Framework should not develop open-ended, domain-general, dependency-structured reasoning. The definitive human test cannot be conducted cleanly, because language cannot ethically be withheld during development. The strongest available evidence therefore comes from narrower natural experiments in which particular symbolic resources are absent or acquired late while other cognitive capacities remain broadly intact.

False-belief reasoning provides one such case. Deaf individuals who grow up without access to a conventional language with recursive structure show severe deficits in representing another person’s false belief, even on nonverbal tasks requiring only a pointing response (Gagne & Coppola, 2017; Pyers & Senghas, 2009). When acquisition of a shared sign language is delayed, false-belief reasoning is correspondingly delayed despite broadly typical nonverbal development (Schick et al., 2007). More specifically, mastery of sentential complements—structures such as “Sally thinks that [X]”—predicts subsequent false-belief success (de Villiers & de Villiers, 2025). The relevant symbolic structure appears to make possible the simultaneous representation of a proposition as true within another agent’s model and false within one’s own.

Exact quantity provides a parallel case. Pirahã speakers, whose language lacks exact number terms, retain approximate magnitude discrimination and can match quantities while the relevant sets remain perceptually available. But when exact cardinality must be preserved across time, space, or modality, performance becomes approximate (Frank et al., 2008; Gordon, 2004). In both cases, the underlying nonlinguistic capacity remains; what is missing is a symbolic structure capable of stabilizing a distinction once perceptual support is removed. These cases therefore provide local instances of the broader constitutive relationship predicted

by ESC: particular symbolic structures make possible cognitive operations that the Substrate alone does not reliably sustain.

Machine learning systems provide a complementary, broader comparison. No system trained without human symbolic input has developed comparable open-ended reasoning across unrelated domains. LLMs, trained on symbolic sequences alone, spontaneously construct internal representations of spatial, temporal, and abstract relational structure, along with compositional machinery that was not explicitly engineered (Gurnee & Tegmark, 2024; K. Li et al., 2023; Modell et al., 2025; Tehenan et al., 2025; Yang et al., 2025). These findings support ESC’s central proposal: an adaptive Substrate can acquire substantial relational and compositional structure through symbolic experience alone, without direct sensory access to much of what that experience describes.

Apparent counterexamples in nonlinguistic animals help sharpen this boundary. Scrub jays cache food for future consumption, and great apes select and transport tools for later use (e.g., Clayton & Dickinson, 1998; Mulcahy & Call, 2006). Like conditioned taste aversion, these behaviors involve temporal decoupling, but they remain bound to narrow, species-typical domains as evolutionarily canalized routines or direct responses to current contextual cues—not the open-ended construction of dependency chains across unrelated domains.

This boundary defines a clear condition for falsification. Any adaptive system that develops genuine open-ended, cross-domain, dependency-structured reasoning without having internalized a Symbolic Framework would falsify this central prediction of the theory.

4.2 Novel Dependency-Structured Reasoning Should Require Iterative Symbolic Generation

Whereas the first prediction concerns what a system must internalize over development, the second concerns what it must do online: when a problem cannot be resolved through the Substrate’s continuous dynamics alone, constructing its solution should require iterative symbolic generation. Not every cognitive operation demands this serial loop. Once the Symbolic Framework has shaped the Substrate’s world model, its relational abstractions remain available to the system’s continuous dynamics—supporting direct associative resolution as well as temporally extended, nonlinguistic trajectories (such as mental imagery or physical simulation in humans), even when no symbols are being produced. Iterative generation should become necessary specifically when a problem requires constructing novel dependency structures that outstrip what continuous dynamics can sustain on their own—when later states cannot be reliably determined until intermediate states have first been discretized and stabilized.

In humans, this boundary is visible in both when inner speech is recruited and when its disruption hurts performance. Routine cognition often relies on condensed or fragmentary inner speech, whereas unfamiliar or demanding problems evoke expanded, syntactically structured forms (Alderson-Day & Fernyhough, 2015; Fernyhough & Borghi, 2023). Conversely, blocking this loop through articulatory suppression—repeating an irrelevant syllable during a task—disrupts the acquisition of new rules and the solution of novel problems far more than the execution of well-practiced routines (van ’t Wout & Jarrold, 2022). In both cases, explicit serialization comes online precisely when an unfamiliar dependency structure outstrips what the Substrate’s learned geometry can resolve directly.

Language models exhibit the computational counterpart of this dynamic. Chain-of-thought prompting and related forms of inference-time computation substantially expand performance on multi-step reasoning tasks (Brown et al., 2025; DeepSeek-AI et al., 2025; OpenAI, 2024; Snell et al., 2024; Wei et al., 2022; Yao et al., 2023; Zelikman et al., 2024) by turning generated intermediate states into inputs for subsequent passes—allowing later steps to condition on earlier results and supporting evaluation, revision, and backtracking. Strikingly, the dynamics of this constructive loop parallel those in humans: reasoning-token use in LLMs covaries with human reaction times on abductive problems, and model error profiles both track human difficulty and become more human-like under stochastic rather than greedy decoding (Arthur, 2026).

This distinction between novel construction and direct associative resolution also explains why explicit reasoning is not universally beneficial. Forced chain-of-thought can degrade model performance, just as deliberate verbalization can impair human performance on tasks better served by automatic or perceptual processing (e.g., Liu et al., 2025), and post hoc verbal chains need not faithfully reflect the computation that produced an answer (e.g., Turpin et al., 2023). Under ESC, iteration is not a blanket requirement for difficult tasks, but a specific mechanism for constructing intermediate states that the Substrate cannot resolve directly; when the relevant relational structure is already encoded in the Substrate, or poorly captured by an imposed verbal sequence, forced serialization adds noise rather than reach. It follows that as a domain’s relational structure is progressively consolidated into the Substrate, disrupting iterative processing should become less costly and additional inference-time generation less beneficial.

Acquired aphasia presents the strongest apparent human counterexample. Adults with profound language impairment can retain capacities for causal reasoning, theory of mind, mathematics, and formal logic, while language-selective cortical regions are often largely inactive during such tasks (Fedorenko et al., 2024; Kean et al., 2026; Varley & Siegal, 2000; Varley et al., 2005). However, because such patients acquired language normally before injury, preservation after language loss shows that some capacities, including domain-specific formal systems, remain executable once constructed—not that they could have developed without symbolic structure (Section 4.1). Moreover, the tasks these patients can still perform typically rely on relational abstractions already encoded in the Substrate or provide perceptual support (e.g., written numerals or visual arrays) that stabilizes intermediate states externally. When nonverbal tasks instead require discovering and switching novel rules or integrating multiple relations, performance drops in proportion to language impairment (Baldo et al., 2005, 2010, 2015). ESC therefore predicts that within acquired aphasia, impairment should increase as familiar abstractions and external support become insufficient, requiring the system to internally stabilize novel intermediate states across multi-step dependencies.

A deeper challenge concerns whether this sequential recurrence must be *symbolic* at all. Replacing chain-of-thought with meaningless filler tokens (e.g., “.”) shows that extra forward passes help only on parallelizable sub-problems, failing whenever later steps depend on earlier intermediate results (Pfau et al., 2024). Hao et al.’s Coconut (Chain of Continuous Thought) (2025) tests the complementary case, feeding continuous hidden states directly back as input without projecting them to discrete tokens. Because projecting an output state to a token and re-embedding it acts as a quantization bottleneck—collapsing competing activations onto a single, canonical vector—bypassing it allows Coconut to maintain multiple candidate paths in superposition. This aids breadth-first search over relationships already specified in the input, but trails conventional chain-of-thought on mathematical problems requiring the construction of novel intermediate quantities. This dissociation aligns directly with ESC: without a discrete bottleneck to purge competing activations and lock in an exact intermediate result, errors compound across steps. Purely continuous recurrence, despite its advantages for parallel search, therefore degrades on deep, dependency-structured problems that require stabilizing newly constructed states—precisely where discrete symbolic commitment provides the necessary ratchet.

Together, these constraints define a clear condition for falsification. ESC’s account of online reasoning would be falsified if a system reliably solved genuinely novel, dependency-structured problems without iterative symbolic generation.

4.3 The Sharp-Decay Prediction

The iterative mechanism of Section 3 yields a direct mechanistic prediction about the temporal dynamics of reasoning. Existing accounts in which structured thoughts are assembled before being expressed in language (e.g., Fedorenko et al., 2024; Fodor, 1975; Pinker, 1994) rarely specify how complex thought unfolds over time or when its internal structure is computed relative to linguistic production. Yet to the extent that such accounts posit that a structured thought is formed prior to its serialization into words, the resolved

content of downstream elements must already be present in the system’s internal state before the sequence is generated.

Under ESC, a system cannot precompile a multi-step symbolic sequence in its internal state. At step n , a complex subsymbolic configuration—encoding the overarching goal, activated concepts, and their interrelations—may persist across the generation of many symbols. However, this configuration only constrains the landscape through which the path is traced; it is not itself a preassembled symbolic sequence. Because each generated symbol acts as an operator that transforms the Substrate to enable the next prediction, the information that determines a distant symbol resides in the *transition dynamics* rather than in a static snapshot of any single state.¹¹

Consequently, when the system’s internal state at step n is probed (via hidden-state vectors or neural activity), predictive fidelity for a specific downstream symbol should decay sharply with distance. While the symbol at $n + 1$ is directly resolvable from the state at n , a symbol at $n + 3$ should be far less reliably recoverable, because the unique internal state required to resolve the latter has not yet been constructed.

Existing evidence from LLMs is consistent with this profile. Pal et al. (2023) probed the hidden states of an LLM and found that future-token decodability drops sharply beyond the immediate next step, even on stimuli drawn from the model’s own training corpus—the conditions under which precompiled or memorized sequences are most plausible. In human language production, an analogous neural test has yet to be conducted, and its prediction cuts against introspective experience: because broad conceptual intentions are present early, it feels as though we form a complete thought and then express it in words, rather than resolving the symbolic sequence progressively.

This prediction would be falsified if distal symbols in a novel sequence could be decoded from the state at step n with high accuracy, indicating retrieval of a preformed symbolic representation. A still stronger causal test would identify a distal symbol’s representation at step n , perturb it, and selectively alter the corresponding output several steps later while leaving the intervening sequence intact.

5 Theoretical Context

With the theory’s boundary conditions sharpened, it becomes possible to identify precisely what it adds to the intellectual tradition linking language to thought, and to the computational formalisms and architectural theories that currently organize cognitive science.

5.1 Intellectual Precedents

Plato asked whether thought and speech are “the same, except that what we call thought is speech that occurs without the voice, inside the soul in conversation with itself” (Plato, 1997). Hobbes argued that children “are not endued with Reason at all, till they have attained the use of Speech” (Hobbes, 1651/1996). Several authors since have maintained that language does not merely express certain forms of thought but, once internalized, brings them into existence. Among those who developed the idea most fully are Vygotsky, Bickerton, and Dennett.¹²

Vygotsky observed that “thought is not merely expressed in words; it comes into existence through them” (Vygotsky, 1934/1986). Bickerton identified the precise ability unlocked by language, arguing that “only language could have broken through the prison of immediate experience in which every other creature is locked, releasing us into infinite freedoms of space and time” (Bickerton, 1990). Dennett framed this

¹¹This “next symbol” orientation does not preclude highly localized planning, such as a model preselecting a rhyming word before composing a line of poetry (Lindsey et al., 2025). Here, the system generates toward a constrained, near-term target, with intermediate words still produced sequentially. This differs from preplanning multiple independent symbols far ahead in an unconstrained sequence.

¹²ESC holds that language is constitutive of certain forms of thought, not merely a modulator of them; the antecedents discussed here are those that share that commitment. For accounts on which language influences or scaffolds cognition without being constitutive of it, see, e.g., Boroditsky (2001); Clark (2008); Lupyan (2012); Lupyan and Bergen (2016); Lupyan et al. (2026).

transformation architecturally, arguing that the cultural acquisition of language and self-talk installs a serial “virtual machine running on the brain’s parallel hardware” (Dennett, 1991).

Yet none of these accounts isolated the mechanism precisely enough to predict—or even reduce surprise at—what has since occurred with LLMs.¹³ What I attempt to add to this lineage is a derivation from first principles. The argument begins not with language but with the structure of causality itself: each link in a causal chain is contingent on the one before it, so any system that models causation must be able to construct such chains step by step (Section 2). Because this derivation proceeds from substrate-agnostic computational constraints rather than observations of the human case, its predictions extend to any system that exhibits general intelligence—including the language models that appear to independently instantiate the mechanism.

5.2 The Missing Mechanism

Cognitive science has acquired rigorous formalisms for what optimal reasoning should look like, and broad principles for how the brain operates. What remains absent is the specific mechanism by which a finite agent transitions from statistical regularities in its input to the structured, compositional representations these formalisms take as given.

Bayesian models of cognition formalize belief revision and hypothesis evaluation (e.g., Tenenbaum et al., 2011), and likely describe the computational ideal that an ESC system approximates. Pearl’s Structural Causal Models formalize the next layer: causal inference over directed graphs, including reasoning about interventions and counterfactuals (Pearl, 2009; Pearl & Mackenzie, 2018). Both formalisms, however, presuppose the structured objects on which they operate: the compositional hypothesis space in the Bayesian case, the causal graph in Pearl’s. Neither addresses how a finite agent acquires the representational building blocks of structured thought, nor the mechanism by which it assembles them into novel constructions. ESC addresses this prior step: the internalization of a Symbolic Framework furnishes the substrate with discrete, combinable elements, and iterative symbolic generation is the process by which candidate structures are assembled, evaluated, and revised in real time.

The Predictive Processing (PP) framework describes the substrate’s general operating principle: the brain minimizes prediction error across hierarchical levels of abstraction (Clark, 2013; Friston, 2010). PP and ESC are compatible but address different problems. PP characterizes the dynamics of error minimization but is agnostic about the representational format of the upper conceptual levels and the mechanism by which hypotheses at those levels are generated. ESC specifies both, and entails a claim PP alone does not make: without internalizing a Symbolic Framework, even a sophisticated predictive substrate cannot extend its upper hierarchy to support reasoning about the absent, the abstract, and the counterfactual.

5.3 Related Frameworks

Where the preceding accounts leave the mechanism unaddressed, several influential theories engage the relationship between architecture and cognition more directly—some proposing structural accounts, others offering functional characterizations or identifying language’s distinctive cognitive role. Each captures a feature of the landscape, but leaves open a critical degree of freedom that ESC serves to pin down.

On the specific question of how compositionality arises within continuous computation, the most closely related program is Smolensky’s neurocompositional computing (e.g., Smolensky, 1988, 1990; Smolensky et al., 2022). Both hold that compositionality must be realized within the continuous substrate, not in a separate architectural component. However, the theories diverge on how compositional structure arises. Smolensky et al. acknowledge that Transformers “can encode general, abstract structure, including webs of inter-related concepts and facts,” but diagnose a specific deficiency: each layer’s attention graph—its implicit relational structure—is consumed in computation and then discarded, never becoming an activation vector

¹³Dennett, for instance, acknowledged as much: “The large language models, like ChatGPT—the generative pre-trained transformers—are largely unanticipated, certainly by me, but also by many people in the field” (Dennett, 2023).

that “encodes the graph’s abstract, compositionally-structured internal information” (Smolensky et al., 2022). The prescribed remedy is to engineer this structure directly into the activation vector through tensor product representations, resolving more compositional structure within each forward pass. This deficiency Smolensky et al. identify in Transformers entails that, in these systems, observed compositionality is realized instead through the trajectory of symbols generated over time. Under ESC, this is not a deficiency to be repaired but the crux of the mechanism (Section 3.2). The two positions accordingly make opposing empirical predictions. If compositional structure is resolved within a single generative state, as the neurocompositional account of human cognition holds, distal symbols should be recoverable from neural activity during language production, contrary to the sharp decay profile predicted by ESC (Section 4.3).

Global Workspace Theory (GWT) proposes that information becomes conscious when specialized processes gain access to a limited-capacity global workspace, making their content available to the brain’s other processes simultaneously, and that this conscious access in turn supports flexible cognition (Baars, 1988; Dehaene & Changeux, 2011). GWT provides a compelling account of how information becomes globally available within the brain, but it is format-agnostic, assigning no special role to *what* is broadcast. Both Baars and Dehaene extend the workspace to non-human animals (Dehaene, 2014), yet no non-human species exhibits the open-ended reasoning, decoupled from immediate perception, that ESC seeks to explain—suggesting that global availability alone is insufficient and that representational format matters. ESC pins down this degree of freedom: general-purpose reasoning depends on the internalization of a combinatorial Symbolic Framework whose generated symbols feed back as computational operators, driving the iterative loop on which such reasoning depends. Whether conscious access operates through a global broadcast is a question on which ESC takes no position; the two theories address different levels of the problem.

Sensorimotor-primacy accounts locate the primary architectural constraint not in representational format but in the perceptuo-motor substrate. Brooks (1991) argued that perception and real-time interaction with a dynamic environment constitute the hard problem, with higher cognition falling into place once that foundation is mature. LeCun (2022) advances a structurally similar position, treating language as a thin layer atop deeper sensorimotor understanding. These capacities are genuinely necessary for navigating the physical world, but the phylogenetic record suggests they are not sufficient: sophisticated sensorimotor systems span the animal kingdom and have been refined over hundreds of millions of years, yet general intelligence appeared exactly once, in the lineage that developed language. LLMs press the point from the opposite direction: structured reasoning emerges in a system that has no perceptuo-motor substrate at all. ESC locates the missing ingredient not in a more sophisticated sensorimotor foundation but in the internalization of a Symbolic Framework whose combinatorial structure furnishes the mechanism for dependency-structured reasoning decoupled from immediate perception.

Finally, accounts of recurrent symbol use and cognitive niche construction anticipate ESC’s iterative loop while stopping short of its architectural necessity. Rumelhart et al. (1986) proposed that a connectionist network can construct sequential thought by generating and reacting to its own internalized linguistic output, and Clark (2008) extended this dynamic to argue that language acts as a cognitive scaffold for reflection and abstract thought. Yet both treat linguistic recurrence as one available route to sequential processing among others—alongside mental imagery, diagrams, and external notation—rather than as the necessary foundation of general intelligence. ESC locates the difference in the architectural requirements established in Section 2: only a Symbolic Framework combines the discrete commitment and open-ended compositional structure required to both install a cross-domain world model and ratchet the Substrate through arbitrary dependency chains. Nonlinguistic simulations and external tools scaffold cognition within specific domains; iterative symbolic generation is what equips the Substrate for open-ended, domain-general reasoning.

6 Implications for Meaning, Understanding, and Grounding

The foregoing account locates general-purpose reasoning in the dynamic interaction between the Substrate and the Symbolic Framework, and establishes that such reasoning cannot remain tethered to the immediate sensory environment. These two commitments have implications for several concepts central to the study of mind, which I explore below.

6.1 Meaning and Understanding Without Phenomenology

In the iterative loop of Section 3.2, each generated symbol serves a dual function: it commits the system to a specific configuration in the Substrate’s high-dimensional world model, and in doing so, reshapes the landscape from which the next symbol emerges. But if a symbol’s identity is fully determined by its relational position and its transformative effect, then the theoretical account contains an implicit account of what meaning *is*. A symbol’s meaning is not a static pointer to an external referent but its location within the learned geometry of the Substrate’s world model: the full pattern of relationships it bears to every other symbol the system has internalized. The meaning of a *sequence*, in turn, is the trajectory it traces through state space—the cumulative transformation it performs on the Substrate.

This account extends naturally to understanding. For a symbolic sequence to perform its intended work—to drive the Substrate through the state transitions the sequence demands—the Substrate must possess geometry rich enough to support those transitions. If a region is richly structured by prior symbolic input, the transition lands in a configuration that opens the correct landscape for the next symbol; if the geometry is sparse, the transition misfires or produces no coherent downstream trajectory. Understanding is therefore not a threshold but a continuum: the more structurally developed the relevant geometry, the more faithfully the Substrate can track the trajectory a given sequence demands. Communication is the same process at a different scale: it succeeds to the degree that a speaker’s sequence drives a listener’s Substrate into an approximately isomorphic arrangement.

This functional account is specified entirely within the theory’s computational vocabulary, without invoking subjective experience. A natural objection is that this omits something essential: the *felt sense* of comprehension. Searle’s “Chinese Room” argument (Searle, 1980) represents the canonical form of this objection, but its force is intuitive rather than empirical, rooted in the difficulty of conceiving of understanding absent its phenomenology. In humans, successful understanding is accompanied by a distinctive affective signature, but this feeling is dissociable from accuracy, arising with equal intensity for correct and incorrect beliefs (Burton, 2008). LLMs now confront this intuition with concrete cases of high-level semantic behavior in entities lacking obvious subjective experience. The theory thus provides principled grounds for demarcating what understanding *is* from what it *feels like*, and for redirecting the binary question “does the system *really* understand?” toward the graded and tractable “how structurally accurate is the system’s world model in the relevant domain?”

6.2 Grounding Without Sensory Contact

The preceding section characterized meaning and understanding as properties of the Substrate’s relational geometry. This leaves open a prior question: how does that geometry acquire fidelity to the world it purports to model?

The standard answer is sensory contact. Harnad (1990) argued that symbols in classical systems are arbitrary tokens whose meaning is entirely injected by their designers, and proposed that grounding requires each agent’s own perceptual experience of the world. Related accounts share this commitment (Deacon, 1997; Pezzulo et al., 2024). Yet all such accounts ultimately depend on internal models rather than ongoing contact, and it is not obvious that sensory experience is the only route by which such models can acquire their fidelity.

Under ESC, language is the output of adaptive Substrates that have internalized the causal and relational structure of the world. Its statistical regularities are shaped by the structure of what it describes. To predict

the next token, a system must reconstruct from within the generative process that produced the sequence, building internal geometries that approximate the world-modeling structures of the agents who spoke and wrote it (Section 3.1.2). Human language was produced by embodied agents under selection pressure: those whose internal models failed to capture causal structure did not survive to shape the Symbolic Framework the next generation internalizes. What a new agent internalizes is therefore not an ungrounded formal system of the kind Harnad critiqued, but the distilled structure of the most successful world models.¹⁴

The productive question, then, is how much world-model fidelity symbolic input alone can transmit. In humans, symbolic learning operates alongside innate priors (Spelke & Kinzler, 2007) and rich sensory experience, making isolation difficult—but cases of deprivation suggest the answer is far more than its typical co-occurrence with sensory experience would imply. Congenitally blind individuals acquire rich, accurate representations of visual properties, such as which animals share appearance features (Kim et al., 2019). Children with severe spinal muscular atrophy type 1, profoundly deprived of motor interaction from infancy, show high-range intelligence and normal language comprehension (Zappa et al., 2021). What survives in both cases is symbolic input. Text-only LLMs sharpen the point further: trained on nothing but symbolic sequences, they construct world models robust enough to predict human judgments across diverse domains (Marjeh et al., 2024).

This inherited fidelity, however, is not uniform. Multimodal models align more closely with human conceptual representations than text-only models for certain categories (Xu et al., 2025), suggesting that some regions of an agent’s world model are shaped by its particular sensory experience in ways that symbolic structure alone does not fully transmit.

7 Limitations and Future Directions

As a high-level theoretical model, ESC defines the necessary functional architecture for general intelligence. However, several critical boundaries and open questions remain regarding the control, integration, and biological implementation of this architecture.

First, the theory’s treatment of Goals remains underspecified. While I posit that goals condition the generative probability distribution, the current model collapses distinct forms of control that likely function very differently in practice. In LLMs, goals are partly exogenous (supplied by user prompts) and partly structural, encoded during training as persistent properties of the learned weights. In humans, socialization and education shape the Substrate in analogous ways, but these acquired orientations coexist and compete with endogenous biological drives and dispositions. How high-level objectives—whatever their source—are converted into the granular symbolic constraints necessary to initiate and guide the heuristic search remains an open and critical question for research on both biological and nonbiological systems.

Second, in humans, this architecture is not implemented on a blank slate; it operates atop a primate brain already rich with intuitive physics, spatial navigation, and basic social dynamics—embodied capacities distributed across specialized cortical regions. While LLMs demonstrate the capabilities of the symbolic component in isolation, human intelligence is defined by the tight coupling of this acquired linguistic mechanism with these preexisting representations. Future work must investigate the neural protocols of this interaction, explaining how the conceptual network built by language constrains—and is constrained by—the brain’s intuitive, nonsymbolic systems.

Finally, further development of the theory requires bridging the gap to Marr’s *implementation* level (Marr, 1982). Having defined the *computational* constraint (sequential causal modeling) and the *algorithmic* mechanism (iterative symbolic generation), the remaining frontier is characterizing the representational geometries that physically instantiate the symbolic loop and the learning dynamics governing their emergence from statistical input. While this is challenging in the brain, it is becoming more tractable in LLMs, where

¹⁴Even the most successful historical models were imperfect approximations; the Substrate internalizes their errors and superstitions by the same mechanism that transmits genuine causal structure.

weights and activations are accessible for direct inspection. Recent work has begun to do so (e.g., Goldstein et al., 2025; Gurnee et al., 2026; K. Li et al., 2023), but empirically grounding the theory at this level remains an open problem.

8 Conclusion

This paper began by asking what allowed general intelligence to emerge within the narrow evolutionary window separating us from our last common ancestor with chimpanzees. On the account developed here, the answer is that, beyond the genetic endowment that led to a Substrate capable of internalizing language, the critical variable is not further genetic change. Whatever fine-tuning has occurred since humans first etched patterns into ochre some 100,000 years ago (Henshilwood et al., 2009), it is unlikely by itself to account for a species that now derives the laws of thermodynamics. The gap is better explained by accumulated cognitive software—the internalized structure of language—transmitted across generations. By this logic, LLMs are an unintended test of this portability: trained on nothing but the symbolic trace of that cumulative process, they reconstruct world models rich enough to support structured reasoning on a substrate entirely unlike the one that produced it.

The introduction identified a scientific bind: we cannot ethically deprive a developing human mind of language to see what remains without it. If ESC is correct, it suggests what such an experiment would reveal: a mind perhaps more capable than any other species’, yet one whose cognition remains mostly bound to the local and the immediate. What the deprived mind lacks is not the words for its thoughts but the machinery of the thinking itself.

This distinction points to why the debate over whether language is necessary for thought has proved so difficult to settle: it treats *thought* as a unitary capacity. Some computations can be resolved directly within a system’s associative geometry and only subsequently expressed in words; others—specifically novel, dependency-structured problems—require symbolic generation to construct an answer at all. The traditional question therefore collapses distinct computational regimes into a single category. The more tractable question is which forms of cognition can occur without symbolic mediation, and which come into existence only through it.

Moreover, the boundary between these regimes shifts with experience. For any *specific* class of problems, the Substrate’s associative geometry can eventually encode the relevant relational structure and resolve the computation directly—this is what expertise is. But the symbolic trace cannot be eliminated *in general*, because problems whose dependencies outstrip the Substrate’s existing abstractions still require step-by-step construction. This carries a practical consequence beyond basic science: in AI systems, because open-ended reasoning must externalize its intermediate steps, the process remains in principle inspectable, providing a basis for auditing and accountability.

By proposing a common, falsifiable principle at work in both humans and machines, ESC lays the groundwork for unifying the study of general cognition. In doing so, it gives computational form to the insight Darwin expressed in the epigraph: words are not the transcription of structured thought but the medium of its construction. The inner monologue we experience as our most private, solitary self is in fact an ancient, ever-changing architecture installed before we had a mind capable of consenting. The generality of human intelligence is, ultimately, the product of two inheritances: at conception, our genes, and thereafter, the symbols of our culture.

References

- Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141(5), 931–965. <https://doi.org/10.1037/bul0000021>
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060. <https://doi.org/10.1037/0033-295X.111.4.1036>

- Arthur, H. (2026). Thinking effort aligns between humans and reasoning models in abductive reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2026*. <https://doi.org/10.48550/arXiv.2609.01867>
- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
- Baldo, J. V., Bunge, S. A., Wilson, S. M., & Dronkers, N. F. (2010). Is relational reasoning dependent on language? A voxel-based lesion symptom mapping study. *Brain and Language*, 113(2), 59–64. <https://doi.org/10.1016/j.bandl.2010.01.004>
- Baldo, J. V., Dronkers, N. F., Wilkins, D., Ludy, C., Raskin, P., & Kim, J. (2005). Is problem solving dependent on language? *Brain and Language*, 92(3), 240–250. <https://doi.org/10.1016/j.bandl.2004.06.103>
- Baldo, J. V., Paulraj, S. R., Curran, B. C., & Dronkers, N. F. (2015). Impaired reasoning and problem-solving in individuals with language impairment due to aphasia or language delay. *Frontiers in Psychology*, 6, 1523. <https://doi.org/10.3389/fpsyg.2015.01523>
- Berwick, R. C., & Chomsky, N. (2016). *Why only us: Language and evolution*. MIT Press. <https://doi.org/10.7551/mitpress/9780262034241.001.0001>
- Bickerton, D. (1990). *Language and species*. University of Chicago Press.
- Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers’ conceptions of time. *Cognitive Psychology*, 43(1), 1–22. <https://doi.org/10.1006/cogp.2001.0748>
- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1–3), 139–159. [https://doi.org/10.1016/0004-3702\(91\)90053-M](https://doi.org/10.1016/0004-3702(91)90053-M)
- Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., & Mirhoseini, A. (2025). Large language mon-keys: Scaling inference compute with repeated sampling. *Proceedings of the Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=0xUEBQV54B>
- Burton, R. A. (2008). *On being certain: Believing you are right even when you’re not*. St. Martin’s Press.
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 34, 15084–15097. https://proceedings.neurips.cc/paper_files/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf
- Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58(2), 345–363. <https://doi.org/10.2307/2371045>
- Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195333213.001.0001>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, 395(6699), 272–274. <https://doi.org/10.1038/26216>
- Cook, M. (2004). Universality in elementary cellular automata. *Complex Systems*, 15(1), 1–40. <https://doi.org/10.25088/ComplexSystems.15.1.1>
- Curtiss, S. (1977). *Genie: A psycholinguistic study of a modern-day “wild child”*. Academic Press.
- Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. W. W. Norton & Company.
- DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., ... Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081), 633–638. <https://doi.org/10.1038/s41586-025-09422-z>
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking Press.
- Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227. <https://doi.org/10.1016/j.neuron.2011.03.018>

- Dennett, D. C. (1991). *Consciousness explained*. Little, Brown & Company.
- Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.
- Dennett, D. C. (2023, October). Daniel Dennett’s been thinking about thinking—and AI [Interview by Taylor McNeil, *Tufts Now*]. <https://now.tufts.edu/2023/10/02/daniel-dennetts-been-thinking-about-thinking-and-ai>
- de Villiers, J. G., & de Villiers, P. (2025). The contributions of language and inhibitory control to false belief reasoning over time. *Frontiers in Psychology*, 15, 1455941. <https://doi.org/10.3389/fpsyg.2024.1455941>
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1
- Etscorn, F., & Stephens, R. (1973). Establishment of conditioned taste aversions with a 24-hour CS-US interval. *Physiological Psychology*, 1(3), 251–253. <https://doi.org/10.3758/BF03326916>
- Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. F. (2024). Language is primarily a tool for communication rather than thought. *Nature*, 630(8017), 575–586. <https://doi.org/10.1038/s41586-024-07522-w>
- Fernyhough, C., & Borghi, A. M. (2023). Inner speech as language process and cognitive tool. *Trends in Cognitive Sciences*, 27(12), 1180–1193. <https://doi.org/10.1016/j.tics.2023.08.014>
- Ferruz, N., Schmidt, S., & Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1), 4348. <https://doi.org/10.1038/s41467-022-32007-7>
- Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3–71. [https://doi.org/10.1016/0010-0277\(88\)90031-5](https://doi.org/10.1016/0010-0277(88)90031-5)
- Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, 108(3), 819–824. <https://doi.org/10.1016/j.cognition.2008.04.007>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Gagne, D. L., & Coppola, M. (2017). Visible social interactions do not support the development of false belief understanding in the absence of linguistic input: Evidence from deaf adult homesigners. *Frontiers in Psychology*, 8, 837. <https://doi.org/10.3389/fpsyg.2017.00837>
- Garcia, J., Ervin, F. R., & Koelling, R. A. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science*, 5(3), 121–122. <https://doi.org/10.3758/BF03328311>
- Goldstein, A., Wang, H., Niekerken, L., Schain, M., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S. A., Gazula, H., Singh, A., Rao, A., Choe, G., Kim, C., Doyle, W., Friedman, D., Devore, S., Dugan, P., Hassidim, A., Brenner, M., ... Hasson, U. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, 9, 1041–1055. <https://doi.org/10.1038/s41562-025-02105-9>
- Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, 306(5695), 496–499. <https://doi.org/10.1126/science.1094492>
- Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2026/workspace>
- Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. *Proceedings of the Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=jE8xbmvFin>
- Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., & Tian, Y. (2025). Training large language models to reason in a continuous latent space. *Proceedings of the Second Conference on Language Modeling*. <https://openreview.net/forum?id=tG4SgayTtk>

- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335–346. [https://doi.org/10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6)
- Haub, C., & Kaneda, T. (2022). *How many people have ever lived on Earth?* Population Reference Bureau. <https://www.prb.org/articles/how-many-people-have-ever-lived-on-earth/>
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579. <https://doi.org/10.1126/science.298.5598.1569>
- Henshilwood, C. S., d’Errico, F., & Watts, I. (2009). Engraved ochres from the Middle Stone Age levels at Blombos Cave, South Africa. *Journal of Human Evolution*, 57(1), 27–47. <https://doi.org/10.1016/j.jhevol.2009.01.005>
- Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554. <https://doi.org/10.1162/neco.2006.18.7.1527>
- Hobbes, T. (1651/1996). *Leviathan* (R. Tuck, Ed.). Cambridge University Press.
- Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203(3), 88–96. <https://doi.org/10.1038/scientificamerican0960-88>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kean, H., Fung, A., Jagers, P., Chen, J., Rule, J. S., Benn, Y., Tenenbaum, J. B., Piantadosi, S. T., Varley, R. A., & Fedorenko, E. (2026). Evidence from formal logical reasoning reveals that the language of thought is not natural language. *Proceedings of the National Academy of Sciences*, 123(28), e2520095123. <https://doi.org/10.1073/pnas.2520095123>
- Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, 116(23), 11213–11222. <https://doi.org/10.1073/pnas.1900952116>
- Laird, J. E. (2012). *The Soar cognitive architecture*. MIT Press.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253. <https://doi.org/10.1017/S0140525X16001837>
- LeCun, Y. (2022). *A path towards autonomous machine intelligence* [Preprint]. OpenReview. <https://openreview.net/forum?id=BZ5alr-kVsf>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *Proceedings of the Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=DeG07_TcZvT
- Li, Z., Liu, H., Zhou, D., & Ma, T. (2024). Chain of thought empowers transformers to solve inherently serial problems. *Proceedings of the Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=yf1icZHJk>
- Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., ... Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>
- Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2025). Mind your step (by step): Chain-of-Thought can reduce performance on tasks where thinking makes humans worse. *Proceedings of the 42nd International Conference on Machine Learning*, 267, 38489–38517. <https://proceedings.mlr.press/v267/liu25t.html>

- Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 3, 54. <https://doi.org/10.3389/fpsyg.2012.00054>
- Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in Cognitive Science*, 8(2), 408–424. <https://doi.org/10.1111/tops.12155>
- Lupyan, G., Gentry, H., & Zettersten, M. (2026). How important is language for human-like intelligence? *Perspectives on Psychological Science*, 1–6. <https://doi.org/10.1177/17456916251398539>
- Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press.
- Marjeh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14, 21445. <https://doi.org/10.1038/s41598-024-72071-1>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133. <https://doi.org/10.1007/BF02478259>
- Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317(5834), 82. <https://doi.org/10.1126/science.1139940>
- Merrill, W., & Sabharwal, A. (2024). The expressive power of transformers with chain of thought. *Proceedings of the Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=NjNGIPh8WWh>
- Modell, A., Rubin-Delanchy, P., & Whiteley, N. (2025). *The origins of representation manifolds in large language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2505.18235>
- Mulcahy, N. J., & Call, J. (2006). Apes save tools for future use. *Science*, 312(5776), 1038–1040. <https://doi.org/10.1126/science.1125456>
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113–126. <https://doi.org/10.1145/360018.360022>
- OpenAI. (2024). *OpenAI o1 system card* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2412.16720>
- Pal, K., Sun, J., Yuan, A., Wallace, B. C., & Bau, D. (2023). Future Lens: Anticipating subsequent tokens from a single hidden state. *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 548–560. <https://doi.org/10.18653/v1/2023.conll-1.37>
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Pezzulo, G., Parr, T., Cisek, P., Clark, A., & Friston, K. (2024). Generating meaning: Active inference and the scope and limits of passive AI. *Trends in Cognitive Sciences*, 28(2), 97–112. <https://doi.org/10.1016/j.tics.2023.10.002>
- Pfau, J., Merrill, W., & Bowman, S. R. (2024). *Let's think dot by dot: Hidden computation in transformer language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2404.15758>
- Pinker, S. (1994). *The language instinct*. William Morrow & Company.
- Plato. (1997). *Sophist* (N. P. White, Trans.). In J. M. Cooper & D. S. Hutchinson (Eds.), *Plato: Complete works* (pp. 235–293). Hackett.
- Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science*, 20(7), 805–812. <https://doi.org/10.1111/j.1467-9280.2009.02377.x>
- Rendell, P. (2002). Turing universality of the Game of Life. In A. Adamatzky (Ed.), *Collision-based computing* (pp. 513–539). Springer. https://doi.org/10.1007/978-1-4471-0129-1_18
- Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.),

- Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models* (pp. 7–57). MIT Press.
- Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Child Development*, 78(2), 376–396. <https://doi.org/10.1111/j.1467-8624.2007.01004.x>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1), 50–64. <https://doi.org/10.1002/j.1538-7305.1951.tb01366.x>
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>
- Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, 26(2), 145–159. <https://doi.org/10.1037/h0030806>
- Skuse, D. H. (1984). Extreme deprivation in early childhood—II. Theoretical issues and a comparative review. *Journal of Child Psychology and Psychiatry*, 25(4), 543–572. <https://doi.org/10.1111/j.1469-7610.1984.tb00172.x>
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1), 1–23. <https://doi.org/10.1017/s0140525x00052432>
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2), 159–216. [https://doi.org/10.1016/0004-3702\(90\)90007-M](https://doi.org/10.1016/0004-3702(90)90007-M)
- Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., & Gao, J. (2022). Neurocompositional computing: From the central paradox of cognition to a new generation of AI systems. *AI Magazine*, 43(3), 308–322. <https://doi.org/10.1002/aaai.12065>
- Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). *Scaling LLM test-time compute optimally can be more effective than scaling model parameters* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2408.03314>
- Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1), 89–96. <https://doi.org/10.1111/j.1467-7687.2007.00569.x>
- Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, 30(3), 299–313. <https://doi.org/10.1017/S0140525X07001975>
- Tehenan, M., Moya, C. B., Long, T., & Lin, G. (2025). *Linear spatial world models emerge in large language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2506.02996>
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285. <https://doi.org/10.1126/science.1192788>
- Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1), 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 74952–74965. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html
- van ’t Wout, F., & Jarrold, C. (2022). How does language support the acquisition of novel cognitive tasks? Investigating the role of task complexity and task instructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(3), 416–431. <https://doi.org/10.1037/xlm0001131>

- Varley, R. A., Klessinger, N. J. C., Romanowski, C. A. J., & Siegal, M. (2005). Agrammatic but numerate. *Proceedings of the National Academy of Sciences*, 102(9), 3519–3524. <https://doi.org/10.1073/pnas.0407470102>
- Varley, R. A., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and ‘theory of mind’ in an agrammatic aphasic patient. *Current Biology*, 10(12), 723–726. [https://doi.org/10.1016/S0960-9822\(00\)00538-8](https://doi.org/10.1016/S0960-9822(00)00538-8)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Vygotsky, L. S. (1934/1986). *Thought and language* (A. Kozulin, Ed. & Trans.) [Revised and expanded translation of the original 1934 Russian edition]. MIT Press.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- Wood, B. (2010). Reconstructing human evolution: Achievements, challenges, and opportunities. *Proceedings of the National Academy of Sciences*, 107(Supplement 2), 8902–8909. <https://doi.org/10.1073/pnas.1001649107>
- Xu, Q., Peng, Y., Nastase, S. A., Chodorow, M., Wu, M., & Li, P. (2025). Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature Human Behaviour*, 9, 1871–1886. <https://doi.org/10.1038/s41562-025-02203-8>
- Yang, Y., Campbell, D., Huang, K., Wang, M., Cohen, J., & Webb, T. (2025). Emergent symbolic mechanisms support abstract reasoning in large language models. *Proceedings of the 42nd International Conference on Machine Learning*, 267, 70240–70269. <https://proceedings.mlr.press/v267/yang25am.html>
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 11809–11822. https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html
- Zappa, G., LoMauro, A., Baranello, G., Cavallo, E., Corti, P., Mastella, C., & Costantino, M. A. (2021). Intellectual abilities, language comprehension, speech, and motor function in children with spinal muscular atrophy type 1. *Journal of Neurodevelopmental Disorders*, 13(1), 9. <https://doi.org/10.1186/s11689-021-09355-4>
- Zelikman, E., Harik, G., Shao, Y., Jayasiri, V., Haber, N., & Goodman, N. D. (2024). Quiet-STaR: Language models can teach themselves to think before speaking. *Proceedings of the First Conference on Language Modeling*. <https://openreview.net/forum?id=oRXPiSOGH9>