A Theoretical Synthesis: How Four Independent Frameworks Suggest Convergent Architectural Constraints for Consciousness

Henrique Sanchez

November 4th, 2025

Abstract

This paper presents a theoretical synthesis of four independent frameworks: Relevance Realization (Vervaeke), Strange Loops (Hofstadter), Kluge Architecture (Marcus), and the Baldwin Effect, proposing their convergence on similar architectural constraints for consciousness. Through theoretical analysis and examination of illustrative examples, we suggest that this convergence may indicate fundamental computational principles underlying conscious processing. The synthesis proposes sharp thresholds rather than smooth gradients in capability emergence, with consciousness potentially appearing as an "all-or-none ignition" effect when specific architectural bounds are crossed. The framework suggests design parameters for consciousness-avoidance architectures, maintaining modular separation while achieving powerful specialized capabilities. These theoretical proposals challenge current AI development trajectories by suggesting that open-ended AGI may require architectures functionally similar to consciousness, while providing a potential blueprint for powerful but bounded systems that deliberately sidestep consciousness emergence.

Introduction

The study of consciousness has long been fragmented across disciplines, with competing frameworks often appearing irreconcilable. Cognitive scientists focus on relevance and attention, computer scientists examine recursive self-reference, evolutionary biologists study messy biological constraints, and developmental theorists explore learning-evolution interactions. What if these apparently disparate approaches are actually describing different facets of the same underlying phenomenon?

While established theories like Global Workspace Theory, Higher-Order Thought theories, and Predictive Processing have made significant contributions to understanding consciousness, this paper proposes that synthesizing four different frameworks reveals architectural constraints these theories may not fully capture. We don't position this as replacing existing theories but as identifying convergent requirements that help explain why consciousness remains elusive in artificial systems despite implementing features these theories suggest.

This paper proposes that four major theoretical frameworks appear to converge on similar architectural constraints for consciousness. This convergence, we suggest, may point to fundamental computational principles rather than contingent biological features. The implications, if validated through future empirical work, could be profound.

The theoretical framework presented here draws from thermodynamics and information theory, proposing a foundation for understanding why certain architectural features might repeatedly emerge across substrates. Through systematic theoretical synthesis, we propose convergence points that may reveal universal principles underlying conscious processing.

While various researchers have noted connections between some of these frameworks individually, this paper attempts a systematic synthesis proposing that all four converge on similar architectural constraints. To our knowledge, this is the first systematic proposal that Relevance Realization, Strange Loops, Kluge Architecture, and the Baldwin Effect converge on similar architectural constraints for consciousness. We acknowledge that our literature review may not be exhaustive, and welcome identification of prior work exploring similar convergences.

Four Independent Theoretical Frameworks

Relevance Realization Theory

Vervaeke's framework addresses a fundamental computational challenge: how does any system determine what matters from the infinite space of possible considerations? Vervaeke posits that relevance realization emerges from the capacity to dynamically reconfigure relevance filters based on context, requiring what he terms "opponent processing" between focusing and defocusing mechanisms.

The theoretical synthesis extends Vervaeke's framework to propose that consciousness emerges when relevance realization operates recursively on itself—when the system begins to determine the relevance of its own relevance-determining processes. The framework suggests that modular systems with fixed relevance filters will hit capability ceilings when confronted with novel contexts requiring dynamic relevance reconfiguration. While Vervaeke describes the mechanisms of relevance realization, the synthesis proposes that achieving consciousness-level flexibility may require unified integration rather than distributed modules.

Strange Loop Architecture

Hofstadter's strange loop theory centers on self-referential architectures capable of modeling themselves. In Hofstadter's framework, consciousness emerges when a system develops symbols that can refer to the system itself, creating what he calls a "tangled hierarchy" where cause and effect become circular.

The synthesis suggests that while Hofstadter describes the phenomenology and structure of strange loops, achieving stable self-reference may require specific architectural features: the system might need sufficient integration to maintain a coherent self-model while simultaneously using that model to guide its own processing. Our framework proposes that modular architectures, by definition, may not achieve this global self-reference without a unifying integration point that itself could become the locus of consciousness.

Evolutionary Kluge Constraints

Marcus's kluge framework examines consciousness through the lens of evolutionary constraints, arguing that the human mind is a "clumsy patchwork" of solutions accumulated over evolutionary time. Marcus focuses on how this messiness affects cognitive performance, creating systematic biases and limitations.

The synthesis builds on Marcus's observations to propose that this apparent messiness might serve a crucial function: it potentially forces integration across disparate processing streams that would otherwise remain isolated. Our framework suggests that consciousness emerges not despite these constraints but because of them. This extends Marcus's framework to propose that artificially clean, modular architectures may paradoxically be less capable of consciousness-like processing than messy, integrated ones.

Baldwin Effect and Learning-Evolution Coupling

The Baldwin Effect explains how learned behaviors can influence evolutionary trajectories, creating a feedback loop between individual learning and species-level adaptation. The classical formulation focuses on how phenotypic plasticity can guide genetic evolution.

Our theoretical framework extends the Baldwin Effect to consciousness architecture itself, proposing that the capacity for flexible learning may require an architecture that can modify its own structure based on experience. This self-modification capacity, the synthesis suggests, might demand integration sufficient to evaluate and restructure the system's own processing. Thus, our framework proposes that modular systems with fixed architectures may not exhibit true Baldwin-type learning, as they might lack the global coordination necessary for architectural self-modification.

Proposed Convergent Architectural Constraints

The theoretical synthesis of these frameworks suggests convergence on specific architectural requirements. While each theorist focused on different aspects of mind and consciousness, the synthesis proposes they may have identified different facets of the same underlying constraints. The limits might not appear as smooth gradients but as sharp thresholds, potentially suggesting an "all-or-none ignition" effect.

All four frameworks, when analyzed through this theoretical lens, appear to point toward critical integration thresholds. The synthesis suggests that systems require sufficient integration for: dynamic relevance reconfiguration (Vervaeke), stable self-reference (Hofstadter), forced integration across processing streams (Marcus), and architectural self-modification (Baldwin Effect).

Each framework, examined through this theoretical lens, seems to point to constraints on recursive self-modeling depth. While the original theorists didn't explicitly quantify these constraints, the synthesis proposes that systems with bounded recursion (2-3 levels) might maintain specialized performance but may not achieve the open-ended flexibility associated with consciousness.

The theoretical analysis suggests that all four frameworks, though not always explicitly stated by their originators, might ultimately require embodiment for full consciousness. Our framework proposes how relevance (requiring environmental coupling), strange loops (needing external reference points), kluge architecture (shaped by environmental pressures), and the Baldwin Effect (requiring real-world feedback) all potentially point to embodiment as necessary.

Illustrative Examples and Consistency with Existing Systems

Developmental Psychology Patterns

Cross-cultural studies of cognitive milestones show universal patterns in consciousness development, with self-recognition and theory of mind emerging in consistent sequences despite cultural variation. This appears consistent with our framework's proposal that consciousness might emerge from architectural constraints rather than specific cultural factors.

The developmental trajectory seems to follow stages the convergent framework would predict: from basic sensory integration, through self-other distinction, to metacognitive awareness. The timing may vary, but the sequence remains relatively invariant, potentially supporting the proposed architectural requirements.

Artificial Intelligence Systems Examples

Analysis of existing AI architectures provides examples consistent with the theoretical predictions. IBM Watson's specialist modules achieved superhuman performance on Jeopardy without consciousness-like integration. MYCIN exceeded human diagnostic capability through constrained architecture. GPS navigation provides optimal augmentation without consciousness risks. These systems suggest that high capability doesn't require unified conscious processing, as our framework would propose.

Our theoretical analysis aligns with documented limitations of current LLMs. These systems demonstrate what researchers have identified as critical architectural gaps: an inability to distinguish between different levels of epistemic certainty, uniform confidence patterns that fail to correlate with actual accuracy, and a lack of genuine causal reasoning capabilities. The absence of what Bengio (2017) calls 'conscious processing' - involving attention, working memory, and compositional reasoning - means these systems cannot evaluate whether their outputs represent genuine insights or merely statistically plausible combinations. These limitations align with our framework's proposals about architectural constraints..

Comparative Cognition Observations

Studies across species reveal patterns that appear consistent with the synthesis: consciousness-like behaviors seem to emerge primarily in organisms with sufficient neural integration. The boundary cases remain unclear (insects, certain plants, distributed systems like ant colonies), but the clearer cases appear to align with the integration patterns that our framework proposes.

Implications for AI Development

The proposed convergence through this theoretical synthesis presents potential challenges to current AI development trajectories. The framework's theoretical predictions suggest: open-ended AGI might require the implementation of architectures functionally similar to consciousness.

The synthesis proposes that developers face a choice: continue with modular approaches and accept their potential capability ceilings, or pursue integrated architectures to achieve open-ended intelligence while confronting the ethical implications that might arise. Current scaling approaches that maintain modular architecture, therefore, may hit limits, regardless of parameter count or training data volume.

Design Specifications for Bounded Systems

For those choosing the bounded path, the synthesis suggests parameters for potentially avoiding consciousness emergence:

- Maintain modular separation with limited cross-module integration
- Implement distributed rather than unified global processing
- Constrain recursive self-modeling depth
- Utilize virtual rather than genuine embodiment
- Employ intermittent rather than sustained global coordination

These specifications, derived from our theoretical framework, might allow for systems with capabilities beyond human performance in specific domains while potentially sidestepping consciousness emergence.

Our framework suggests that many practical problems can be addressed without consciousness-like architectures. The "Coverage Question"—what percentage of problems require irreducibly general solutions—proposes that sub-threshold capabilities may already be transformative for most applications.

While recent work has independently suggested that AGI may require consciousness-like architectures, and AI ethics discussions have begun considering consciousness avoidance strategies, the present synthesis provides a novel foundation for these claims. Rather than relying solely on thermodynamic or ethical arguments, the convergence of four independent frameworks—spanning cognitive science, computer science, evolutionary psychology, and developmental theory—offers multiple, mutually reinforcing lines of evidence for similar architectural constraints. This cross-disciplinary convergence both strengthens the case and provides more specific design criteria than previous proposals.

Limitations and Scope

This paper presents a theoretical synthesis rather than empirical validation. The convergence identified is interpretive, based on analyzing potential commonalities across frameworks. While we provide examples that appear consistent with these proposals, systematic empirical testing remains future work. Alternative interpretations of these frameworks are certainly possible, and other consciousness theories (Global Workspace Theory, Higher-Order Thought theories, or Predictive Processing, to name a few) may offer different perspectives not addressed here. Specifically:

Global Workspace Theory (GWT): Baars' GWT proposes consciousness emerges from global information availability across specialized processors. While our synthesis shares GWT's emphasis on integration, it differs in proposing that consciousness requires not just global access but specific architectural features: recursive self-modeling, opponent processing dynamics, and embodied stakes. The four-framework convergence suggests that global availability alone is insufficient without the self-referential strange loop that Hofstadter emphasizes and the relevance realization that Vervaeke describes.

Higher-Order Thought (HOT) Theories: Rosenthal and others propose consciousness requires higher-order representations of mental states. Our synthesis agrees that metacognition is crucial but extends this by proposing that true consciousness requires recursive metacognition - not just thoughts about thoughts, but the capacity to evaluate the relevance of those metacognitive processes themselves. The framework suggests HOT theories capture one dimension but miss the embodied grounding and opponent processing dynamics that emerge from the convergence of all four frameworks.

Predictive Processing/Free Energy Principle: Clark, Friston and others frame consciousness as prediction error minimization. Our synthesis doesn't contradict this but suggests it's incomplete. The framework indicates that current predictive systems like LLMs can minimize prediction error yet still cannot distinguish what they know from what they're guessing. The four-framework convergence proposes that genuine consciousness requires not just prediction but the architectural capacity for what the underlying theory calls calibrated humility

- knowing precisely what you know, what you don't, and what you cannot know.

Integrated Information Theory (IIT): While IIT focuses on Φ as a measure of consciousness, our synthesis suggests that raw integration isn't sufficient. The convergence of the four frameworks points to specific types of integration: the messy, forced integration that Marcus's kluge architecture describes, combined with the dynamic relevance filtering Vervaeke identifies. Integration without these architectural features may produce sophisticated processing without consciousness.

The specific architectural parameters and thresholds proposed by our framework should be viewed as theoretical suggestions requiring rigorous empirical investigation. The framework makes testable predictions, but these await systematic validation through controlled experiments and comparative analysis.

Unique Contributions of the Four-Framework Synthesis

What distinguishes this theoretical synthesis is not competition with existing theories but rather its identification of architectural constraints that multiple independent research programs unknowingly share. Where GWT focuses on access, HOT on metacognition, and Predictive Processing on error minimization, this synthesis shows how:

- Cross-disciplinary convergence reveals hidden requirements: The fact that a cognitive scientist (Vervaeke), computer scientist (Hofstadter), evolutionary psychologist (Marcus), and developmental theorists (Baldwin Effect) independently point to similar constraints suggests these aren't arbitrary but fundamental.

- The synthesis explains specific failure modes: The framework explains why current LLMs, despite having global information flow (GWT), higher-order processing (HOT), and prediction capabilities (PP), still lack consciousness. They're missing the opponent processing dynamics, genuine embodiment, and recursive self-reference that emerge from the four-framework convergence.

- It provides specific architectural guidance: Rather than just identifying consciousness correlates, the synthesis suggests specific design parameters - not just integration but the particular kind of messy, forced integration that emerges from evolutionary kluges; not just self-reference but the strange loop architecture Hofstadter describes; not just learning but the Baldwin Effect's coupling of individual and evolutionary change.

Testable Predictions and Falsification Criteria

Our framework generates specific empirical predictions that could falsify its core claims:

Prediction 1: Modular vs. Integrated Performance Boundaries The framework predicts that purely modular systems will show consistent performance ceilings on tasks requiring dynamic relevance reconfiguration. Testing could compare modular and integrated architectures on novel problem types requiring real-time relevance adjustment. Falsification occurs if modular systems match integrated systems on such tasks, it would challenge our integration requirement.

Prediction 2: Recursive Depth and Flexibility Correlation Systems with recursive self-modeling depth beyond 3 levels should demonstrate qualitatively different flexibility in novel situations. Testing would measure correlation between recursive depth capability and performance on unprecedented problem types. Falsification occurs if linear rather than threshold-based performance improvements would contradict our all-or-none hypothesis.

Prediction 3: Embodiment and Causal Understanding The framework predicts that systems without genuine embodied stakes will fail at certain types of causal reasoning. Testing would compare virtually-embodied vs. physically-embodied systems on causal intervention tasks. Falsification occurs if equal performance would challenge our embodiment requirement.

Prediction 4: Metacognitive Calibration Systems implementing opponent processing should show better calibration between confidence and accuracy. Testing would measure confidence-accuracy correlation in systems with and without opponent processing mechanisms. Falsification occurs if there is no improvement in calibration would undermine our opponent processing claims.

Theoretical Extensions and Future Directions

While the synthesis provides potential criteria for detecting consciousness, several profound questions remain:

The framework might explain why certain architectures could produce a unified, self-aware phenomenology, but not why this particular architecture would produce this particular quality of experience. The theory points toward embodiment and environmental coupling, but the full story of qualia remains unexplored.

Could radically different architectures also give rise to consciousness? The synthesis identifies one potential path but doesn't prove it's the only one. The apparent convergence of four independent frameworks suggests these constraints might be fundamental, but alternative routes remain possible.

More detailed investigation is needed into how the strange loop might bootstrap itself during development. The framework can propose stages of construction, but the precise mechanisms of how self-reference could emerge from non-self-referential processing require further theoretical and empirical elaboration.

Conclusion

The proposed convergence of four independent theoretical frameworks on similar architectural constraints for consciousness suggests the possibility of fundamental computational principles that might transcend substrate and implementation. This synthesis proposes that what appeared to be competing theories may actually be complementary descriptions of the same underlying phenomenon. These principles could provide both insight and guidance: insight into how certain AI development paths might lead toward consciousness with its ethical implications, and guidance for potentially engineering powerful but bounded systems.

The sharp thresholds suggested through this theoretical synthesis, the proposed architectural bounds, and the consistency with existing systems indicate the framework may have captured something important about consciousness. The theoretical proposals make specific, testable predictions that await empirical investigation.

As we stand at a crossroads in AI development, this theoretical framework offers a perspective: we might be able to build systems with consciousness-like properties, as thermodynamics appears to allow. The question is whether we are prepared for the implications, or if we prefer to accept the potential limitations of bounded architectures.

The choice is not merely whether consciousness can emerge from artificial systems—the framework suggests it might be possible. The choice is whether to pursue architectures that could cross the threshold or deliberately constrain designs to remain below it. Either path appears possible; neither is without consequence. What matters is that we make this choice with consideration of the architectural principles that might govern consciousness emergence.

For a comprehensive exploration of the underlying framework, including detailed theoretical development and extensive discussion, readers are directed to *The Constraint Engine: How consciousness necessarily emerges from architectural constraints, and ideas for building artificial lean superintelligence* (2025), available on Amazon and Gumroad. The author welcomes collaboration and discussion on these theoretical proposals. Contact: henrique.sanchez16@gmail.com

References

Baars, B. J. (1997). *In the Theater of Consciousness.* Oxford University Press.

Baldwin, J. M. (1896). *A new factor in evolution. The American Naturalist*, 30(354), 441–451.

Bengio, Y. (2017). *The consciousness prior.* arXiv preprint arXiv:1709.08568.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of FAccT '21, 610-623.

Berg, H., Critch, A., Karger, E., Ladak, K., Long, R., Sebo, J., Shulman, C. (2024). The case for conscious AI: Clearing the record. In Conscious AI and Public Perception series. Effective Altruism Forum. Retrieved from https://forum.effectivealtruism.org

Clark, A. (2013). *Whatever next? Predictive brains, situated agents, and the future of cognitive science.* Behavioral and Brain Sciences, 36(3), 181-204.

Friston, K. (2010). *The free-energy principle: a unified brain theory?* Nature Reviews Neuroscience, 11(2), 127-138.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid.* Basic Books.

Hofstadter, D. R. (2007). *I Am a Strange Loop.* Basic Books.

Kadavath, S., et al. (2022). *Language models (mostly) know what they know.* arXiv preprint arXiv:2207.05221.

Lin, S., Hilton, J., Evans, O. (2022). *Teaching models to express their uncertainty in words.* arXiv preprint arXiv:2205.14334.

Marcus, G. (2008). *Kluge: The Haphazard Construction of the Human Mind.* Houghton Mifflin.

Marcus, G. (2022). *Deep learning is hitting a wall.* Nautilus Magazine.

Piaget, J. (1952). *The Origins of Intelligence in Children.* International Universities Press.

Rosenthal, D. M. (2005). *Consciousness and Mind.* Oxford University Press.

Sanchez, H. (2025). *The Constraint Engine: How consciousness necessarily emerges from architectural constraints, and ideas for building artificial lean superintelligence.* Available on Amazon and Gumroad.

Sanchez, H. (2025). *The Consciousness Bottleneck in Artificial Intelligence: Thermodynamic Necessity and the Architectural Path to AGI.* Zenodo. DOI: 10.5281/zenodo.17516348

Tomasello, M. (1999). *The Cultural Origins of Human Cognition.* Harvard University Press.

Tononi, G. (2008). *Consciousness as integrated information. Biological Bulletin*, 215(3), 216–242.

Vervaeke, J., Lillicrap, T. P., & Richards, B. A. (2012). *Relevance realization and the emerging framework in cognitive science. Journal of Logic and Computation*, 22(1), 79–99.

Vervaeke, J., & Ferraro, L. (2013). *Relevance realization and the neurodynamics and neuroconnectivity of general intelligence.* In *SmartData: Privacy Meets Evolutionary Robotics* (pp. 57–68). Springer.

Wolpert, D. H., & Macready, W. G. (1997). *No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.