Autocatalytic Constraint Closure as an Organizational Principle for Machine Consciousness

Armando Vieira[1] **and Liane Gabora**[2]

[1]Tartu University, Estonia
[2]University of British Columbia, Canada
armando.vieira@ut.ee, liane.gabora@ubc.ca

Abstract

Current AI systems are largely organized around prediction-error correction and reward optimization. While these frameworks have been highly successful, they do not address how AI systems can develop the kind of integrated, self-maintaining world-model widely regarded as central to consciousness. This paper proposes autocatalytic constraint closure as a necessary organizational principle for machine consciousness. Reflexively Autocatalytic Foodset-derived (RAF) networks provide a general-purpose formal framework for describing and analyzing the emergence of systems whose components catalyze the generation of new components that increase the coherence of the whole. This can result in a phase transition to a self-organizing system with history-dependent dynamics. Applied to cognition, external stimuli and internal representations 'catalyze' mental operations yielding new representations, spurring formation of an integrated representational network and coherent world-model. We show that (i) AI systems exhibit limited, task-bound autocatalytic organization without persistent closure, and (ii) in-context learning may reflect transient RAF formation. Next steps include fostering sensitivity to internal incoherence and question-asking in AI systems with the aim of fostering endogenously driven self-organization, a prerequisite for conscious systems, as well as using RAF algorithms to analyze candidates for machine consciousness.

Introduction

Current AI systems excel at pattern recognition, prediction, and optimization, but are largely organized around externally imposed objectives rather than internally generated structure (Russell and Norvig 2021). Even highly capable systems lack the kind of coherent, self-sustaining internal organization characteristic of conscious minds. This suggests that recent progress, while practically useful, does not by itself advance the goal of identifying organizational principles required for machine consciousness (Dehaene, Lau, and Kouider 2017; Seth and Bayne 2022).

AI frameworks emphasize error correction or reward maximization without specifying how internally integrated, self-maintaining representational structures arise. While predictive processing and reinforcement learning offer powerful accounts of perception and action (Friston 2010; Sutton and Barto 2018), they leave under-specified the mechanisms by which a system's internal components become mutually constraining in ways that stabilize a unified perspective, or world-model. Addressing this requires a shift in focus from performance optimization to internal organization.

This paper explores Autocatalytic Constraint Closure (ACC) as a potential organizational principle for conscious systems. The abstract mathematical framework of autocatalytic networks was introduced by Kauffman to explain how self-sustaining systems could emerge and evolve without external control (Kauffman 1993). The framework was subsequently formalized as RAF (Reflexively Autocatalytic and Foodset-derived) networks by Steel and colleagues (Steel, Hordijk, and Smith 2013; Hordijk, Hein, and Steel 2010), and extended to cognition (Gabora and Steel 2017; Gabora, Beckage, and Steel 2022). RAF networks explicitly distinguish elements present from the start or assuming the same form within the network as outside it (the *foodset*) from elements produced through network interactions (*foodset-derived*). As the density of mutually imposed constraints among foodset-derived elements increases, the system undergoes a phase transition to a coherent, self-organizing system with history-dependent dynamics. When ACC is achieved, the system's internal activity is endogenously driven and can adapt through both assimilation of outside elements and reorganization from within.

Applied to cognition, mental representations play the role of molecules, cognitive operations (e.g., deduction, analogy, concept combination) play the role of reactions, and external stimuli or internally generated thoughts that trigger new insights serve as *catalysts*. A cognitive system achieves ACC when its conceptual network becomes sufficiently rich to sustain extended chains of thought without constant environmental prompting—the essence of self-awareness and creativity (Gabora and Steel 2022), as well as self-directed behavior and self-modeling (Gabora and Bach 2023). Thus, in a cognitive system that exhibits ACC, representations are integrated in the sense that they are mutually accessible and can catalyze mental operations in the absence of external cues. As small self-regenerative units of meaning merge into larger units of mutually consistent meanings, meaningful chains of thought can be endogenously generated and sustained.

We suggest that achieving robust ACC may be a necessary

step toward machine consciousness. This paper explores the extent to which AI systems exhibit partial or transient forms of autocatalytic organization (particularly in within-context learning) or the kind of large-scale persistent ACC across time needed to sustain a coherent world-model. We suggest a concrete direction for extending autocatalytic architectures: enabling systems to register and respond to inconsistencies within their own internal structure, thereby allowing reorganization to be guided by endogenous rather than user-imposed demands.

Learning as Prediction Error Correction Versus Constraint Closure

A key feature of conscious systems is the capacity for learning. Currently, learning is predominantly framed as the minimization of prediction error through Bayesian inference: systems learn by building internal models of the world and continuously updating these models to minimize prediction error (Friston 2010; Clark 2013). This framework, exemplified by the Free Energy Principle and predictive processing theories (Hohwy 2013), posits that the brain is fundamentally an inference engine.

While powerful for explaining alignment with external data, this paradigm faces challenges. Bayesian inference in realistic cognitive architectures is computationally intractable (Harkness and Keshavan 2019). Despite recent efforts to operationalize creativity using free energy (Constant, Friston, and Clark 2024), the framework struggles to explain creative cognition in which genuinely novel ideas arise through synthesis and reorganization of existing concepts (Gabora and Steel 2017). It locates the driving force of learning outside the system—in the discrepancy between prediction and reality—rather than in the system's intrinsic organization.

We propose a complementary framework: *learning as constraint closure*. It posits that a conscious system seeks not just to minimize prediction error but to achieve and maintain ACC—robust internal coherence in the face of new information from outside and endogenously-generated knowledge.

Autocatalytic Networks

Reflexively Autocatalytic and Foodset-derived (RAF) network theory was developed by Mike Steel and colleagues as a formalization of autocatalytic networks (Hordijk, Hein, and Steel 2010; Steel, Hordijk, and Smith 2013). It is a domain-general mathematical framework for describing the emergence, self-maintenance, and perpetuation of systems that are separate from yet interact with their environment. Cognitive networks define mental models that exhibit RAF-like organization and evolution (Gabora and Steel 2017; Gabora, Beckage, and Steel 2022). They replicate in a patchwork manner through transmission of ideas, with evolution evidenced in how technologies and scientific theories build on one another.

In a machine, the foodset consists of input tokens or learned embeddings from training data and prompts, while representations resulting from compositional functions or transformations are foodset-derived items. The critical distinction is that only foodset-derived items come into existence in their current form in the system.

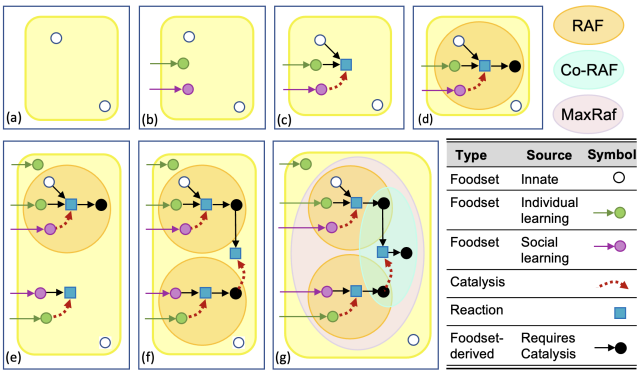

Figure 1: Foodset items in cognitive RAF scenarios can be innate, or acquired through social or individual learning of information already present in the world.

Figure 1 illustrates cognitive catalysis resulting in a RAF, and its merger with other RAFs to form a more extensive RAF.

A key insight is that when a sufficiently diverse collection of interacting elements reaches a critical threshold of catalytic potential, the system undergoes a sudden phase transition. Below this threshold, reactions are isolated and ephemeral, but above it they connect into an integrated network capable of sustaining itself (Hordijk, Steel, and Kauffman 2012).

Mathematical Framework

RAF Network Theory: Formal Definitions

We begin by formalizing the core concepts of RAF theory, adapted from chemical systems to representational systems.

Definition 1. *Reaction Network*
A reaction network is a tuple $\mathcal{N} = (X, R, C, F)$ *where:*

- X *is a finite set of elements (representations, concepts, or states)*
- R *is a set of reactions, where each* $r \in R$ *is a transformation* $r : X^n \to X^m$ *for some* $n, m \geq 1$
- $C : R \to 2^X$ *is a catalysis function mapping each reaction to its set of catalysts*
- $F \subseteq X$ *is the food set—elements available from the environment*

Definition 2 (Reflexively Autocatalytic)**.** *A subset* $\mathcal{R} \subseteq R$ *is reflexively autocatalytic if every reaction* $r \in \mathcal{R}$ *is catalyzed by at least one element that is either in the food set* F *or produced by a reaction in* $\mathcal{R}$ *itself.*
Formally: $\forall r \in \mathcal{R} : C(r) \cap (F \cup products(\mathcal{R})) \neq \emptyset$
where $products(\mathcal{R}) = \bigcup_{r \in \mathcal{R}} output(r)$.

Definition 3 (Foodset-Derived (F-derived))**.** *A subset* $\mathcal{R} \subseteq R$ *is F-derived if all reactants required by reactions in* $\mathcal{R}$ *can be constructed starting from* F *using only reactions in* $\mathcal{R}$.

Let $\text{closure}_{\mathcal{R}}(F)$ denote the set of all elements reachable from F by applying reactions in $\mathcal{R}$. Then $\mathcal{R}$ is F-derived if:

$$\forall r \in \mathcal{R} : \text{reactants}(r) \subseteq \text{closure}_{\mathcal{R}}(F)$$

Definition 4 (RAF Set)**.** *A subset $\mathcal{R} \subseteq R$ is a Reflexively Autocatalytic and Foodset-derived (RAF) set if it is both reflexively autocatalytic and foodset-derived.*

RAF sets are the fundamental self-sustaining structures. They can create all their own catalysts starting from basic inputs and maintain themselves through recursive catalysis.

Definition 5 (MaxRAFs, IrrRAFs, and SubRAF)**.**
- *The maximal RAF (MaxRAF) is the union of all RAF sets in $\mathcal{N}$. It is unique (Steel, Hordijk, and Smith 2013).*
- *An irreducible RAF (IrrRAF) is a RAF that contains no proper subset that is also a RAF.*
- *A subRAF is a RAF that is a subset of another RAF.*

Definition 6 (Co-RAF)**.** *A set of reactions $\mathcal{Q} \subseteq R$ is a co-RAF relative to a RAF $\mathcal{R}$ if $\mathcal{Q}$ is not itself a RAF but $\mathcal{R} \cup \mathcal{Q}$ is a RAF (Steel, Hordijk, and Smith 2013).*

Co-RAFs represent potential growth—structures that can be integrated into an existing RAF to form a larger one.

Constraint Closure in Representational Systems

We now adapt these formal structures to representational learning systems.

Definition 7 (RAF Network)**.** *A representational learning system is a RAF network $\mathcal{N}_{rep} = (X_{rep}, R_{rep}, C_{rep}, F_{rep})$ where:*
- *X_{rep} is a set of representations (internal states encoding information)*
- *R_{rep} is a set of cognitive operations (transformations such as composition or analogy)*
- *C_{rep} maps operations to representational states, cognitive states, or mental state that enable them*
- *F_{rep} is the environmental input (sensory data, training examples, prior knowledge)*

Definition 8 (Constraint Closure)**.** *A representational system achieves constraint closure when it contains a RAF set: a network of representations where each representation helps enable the construction and refinement of others, and all necessary representations can be constructed from environmental inputs.*

A key insight is that constraints in a representational system are not external requirements but internal relationships. A representation constrains and is constrained by other representations. Closure occurs when these mutual constraints form a self-supporting web.

Phase Transitions and Percolation Thresholds

A central result in RAF theory is the existence of abrupt phase transitions.

Theorem 1 (Percolation Threshold)**.** *In random reaction networks with uniformly random catalysis, there exists a phase transition from small, disconnected RAFs to a giant RAF as catalytic density crosses a critical threshold p_c, after which the system undergoes a phase transition to a coherent, self-organizing system with history-dependent dynamics.*

Below p_c, the network is fragmented. Above it, a macroscopic connected component emerges.

Proposition 1 (Learning as Crossing the Threshold)**.** *Learning in a representational system can be modeled as increasing the catalytic density ρ (the average number of representations that can catalyze each operation) until the system crosses a percolation threshold, leading to the sudden emergence of a coherent, self-sustaining conceptual network.*

This proposition explains emergent capabilities in neural networks: they are phase transitions occurring when internal connectivity crosses critical thresholds (Wei et al. 2022).

Constraint Closure vs. Prediction Error Correction

We now formally contrast constraint closure with Bayesian prediction error minimization.

Bayesian Framework: In the Bayesian view, learning optimizes a generative model $p_\theta(x, z)$ over observations x and latent states z to minimize prediction error:

$$\mathcal{L}_{\text{Bayes}}(\theta) = \\ -\mathbb{E}_{p(x)}[\log p_\theta(x)] = \mathbb{E}_{p(x)}[D_{KL}(p(z|x) \| p_\theta(z|x))]$$

The system seeks to match its posterior to the true posterior over latent states (Friston 2010).

Constraint Closure Framework: In the constraint closure view, learning maximizes internal coherence:

$$\mathcal{L}_{\text{closure}}(\mathcal{N}) = \\ f(\text{Integration}(\mathcal{N}), \text{Autocatalysis}(\mathcal{N}), \text{Coherence}(\mathcal{N}))$$

where $\text{Integration}(\mathcal{N})$ measures the size of the largest RAF, $\text{Autocatalysis}(\mathcal{N})$ measures the degree of reflexive catalysis, and $\text{Coherence}(\mathcal{N})$ measures consistency of the representational structure. Note that our framework is more organizational than algorithmic and the function f accounts for the fact that learning is not about fitting data but about becoming a self-sustaining, mutually constraining system.

Coherence can be approximated by the persistence and redundancy of reactions within the MaxRAF across contexts, or by internal surprisal—how surprising a representation is to other representations within the system rather than to external data.

Note that internal coherence refers to the degree to which representations participate in mutually sustaining catalytic constraints that remain dynamically stable across time, rather than to logical consistency or accuracy with respect to the external world.

Theorem 2 (Divergence of Objectives)**.** *For a representational system $\mathcal{N}$ with sufficient complexity (e.g., high-dimensional hypothesis space), minimizing $\mathcal{L}_{Bayes}$ does not necessarily minimize $\mathcal{L}_{ACC}$ (or vice versa). The objectives can be partially orthogonal: improvements in one do not always imply improvements in the other.*

For the Proof of Theorem 2, please see Appendix One.

Note that the autocatalytic framework does not require that learned representations accurately predict the world—only that they form a coherent, self-sustaining whole. Pre-

Algorithm 1: ACC Learning

Initialize food set F from environment (e.g., sensory data, training examples).
Initialize empty reaction set $\mathcal{R} \leftarrow \emptyset$ (reactions represent cognitive operations).
while not converged **do**
  Sample potential reactions $\mathcal{R}_{\text{new}}$ from representational space.
  for all $r \in \mathcal{R}_{\text{new}}$ **do**
    if r is catalyzed by elements in $\text{ACC}_{\mathcal{R}}(F)$ **then**
      Add r to $\mathcal{R}$.
      Update $\text{ACC}_{\mathcal{R}}(F)$ with products of r (new representations).
    end if
  end for
  Compute $\text{MaxRAF}(\mathcal{R})$.
  if $|\text{MaxRAF}(\mathcal{R})|/|X| > \theta$ **then**
    Phase transition: ACC achieved—break loop.
    break
  end if
end while
return $\text{MaxRAF}(\mathcal{R})$.

diction accuracy may emerge as a consequence of coherence, but it is not the fundamental driver. However, in practice, the two can be integrated, with prediction guiding the foodset and ACC organizing it.

This algorithm contrasts with gradient-based optimization in Bayesian learning. The system explores the space of possible cognitive operations, integrating those that can be catalyzed by existing structure, until a self-sustaining network emerges.

The parameter θ is a system-dependent threshold governing emergence of large-scale ACC. Prior work on RAF formation and percolation suggests a practical default of $\theta = 0.5$, i.e., the maxRAF encompasses at least half of all representations (Hordijk, Steel, and Kauffman 2012). At this point the system becomes largely self-sustaining and can exhibit emergent behavior.

Application to Representation Learning in Foundation Models

Transformer Architecture as Autocatalytic System

We show that transformer architectures naturally instantiate autocatalytic structure and dynamics. In current systems, however, these dynamics are organized in service of a user rather than a self: prompts supplied by the user trigger internally self-reinforcing processes that advance the user's goals, not the system's own. For consciousness, autocatalytic dynamics must be organized around the maintenance and growth of a self, rather than externally imposed objectives (Gabora and Bach 2023). Our claim is that transformers instantiate autocatalytic structure as a necessary condition for consciousness, while lacking the endogenous catalytic organization required for self-maintaining internal coherence.

Self-Attention as Mutual Catalysis The core operation in transformers is scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are queries, keys, and values derived from input representations (Vaswani et al. 2017).

Proposition 2 (Attention as Catalytic Reaction)**.** *Each attention operation is a cognitive reaction where:*

- *Reactants: The value vectors V being aggregated*
- *Catalyst: The query vector Q that determines which values to attend to*
- *Product: The updated representation $\sum_i \alpha_i v_i$ where $\alpha_i = \text{softmax}(q^T k_i/\sqrt{d_k})$*

Crucially, attention is *reflexive*: the query that catalyzes the reaction is itself a representation produced by previous reactions. This is the essence of reflexive autocatalysis.

Layer-wise Refinement as RAF Formation A transformer with L layers applies attention iteratively:

$$h^{(l+1)} = \text{Attention}(h^{(l)}) + \text{FFN}(h^{(l)})$$

where $h^{(l)}$ is the representation at layer l.

Proposition 3 (Transformers Build Hierarchical RAFs)**.** *Each layer l constructs representations that are F-derived from layer $l-1$.*

This explains why multi-head attention is more powerful than single-head: it increases the likelihood of discovering and integrating diverse autocatalytic structures.

Self-Supervised Learning as Foodset-Derived Structure Construction

Self-supervised learning objectives do not provide explicit labels but impose constraints that guide the formation of structure.

Masked Language Modeling In BERT-style masked language modeling (MLM), a fraction of input tokens is masked, and the model predicts them from context (Devlin et al. 2019):

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{x\sim D, M\sim\text{Mask}}\left[\sum_{i\in M}\log p_\theta(x_i|x_{\setminus M})\right]$$

Proposition 4 (MLM Drives ACC)**.** *MLM encourages the formation of representations where each token's representation is catalyzed by (and catalyzes) the representations of surrounding tokens. This mutual constraint satisfaction drives the system toward ACC.*

The masking procedure ensures that different tokens are hidden across training examples, forcing the model to develop representations where any token can serve as both a catalyst (providing context for reconstructing others) and a product (being reconstructed from context). This bidirectional dependency increases the catalytic density of the representational network.

Emergent Capabilities as Phase Transitions

Large language models exhibit qualitatively new capabilities at specific model scales, a phenomenon often described as *emergence* (Wei et al. 2022). Tasks such as multi-step arithmetic, compositional reasoning, and in-context learning transition from near-chance performance to reliable execution over relatively narrow scaling ranges.

We argue that emergent capabilities correspond to crossings of thresholds in internal representational networks. As model scale increases, both the number of available representations and the density of catalytic operations increase. When these parameters exceed a critical value, a large reflexively autocatalytic, food-generated structure forms, enabling coordinated execution of operations required for complex capabilities.

Theorem 3 (New Capabilities as RAF Phase Transitions)**.** *Let $\{Q_n = (X_n, R_n, F_n)\}_{n\geq 1}$ be a sequence of representational reaction systems with bounded reaction arity. Assume a random catalysis model in which each pair $(x, r) \in X_n \times R_n$ is independently catalytic with probability p_n, and define the catalytic density $\rho_n = p_n|X_n|$.*

Assume further that the system is F-feasible, in the sense that there exists $\epsilon > 0$ such that for all sufficiently large n, the F_n-generated ACC (ignoring catalysis) contains at least $\epsilon|R_n|$ reactions with high probability.

Then there exists a critical value $\rho_c > 0$ and a constant $\alpha \in (0, 1)$ such that:

1. *If $\limsup \rho_n < \rho_c$, then with high probability, no RAF has size $\geq \alpha|R_n|$.*
2. *If $\liminf \rho_n > \rho_c$, then with high probability, the max RAF has size $\geq \alpha|R_n|$.*
3. *The event that the max RAF has size $\geq \alpha|R_n|$ has a sharp threshold around ρ_c in a window $w_n = o(1)$.*

Consequently, capabilities requiring a large RAF emerge sharply as ρ_n crosses ρ_c.

For the Proof of Theorem 3, please see Appendix Two.

In-Context Learning as Recursive Self-Organization

In-context learning (ICL) refers to the ability of foundation models to perform novel tasks after observing a small number of input–output examples within the prompt, without any update to model parameters (Brown et al. 2020). Unlike conventional learning, which proceeds through gradient-based parameter modification, ICL operates entirely at inference time.

We propose that ICL can be understood as a process of *recursive self-organization* within the model's existing representational dynamics. During pre-training, the model acquires a large, densely connected reflexively autocatalytic network—a global MaxRAF—capable of supporting a wide variety of compositional processes.

Proposition 5 (ICL as Rapid RAF Formation)**.** *In-context learning corresponds to the rapid emergence of a transient, task-specific RAF network induced by a prompt. It temporarily organizes the system's representational dynamics without persisting across contexts. The prompt supplies a representational catalyst that triggers representational change in the model's pre-trained MaxRAF through mutual reinforcement among task-relevant operations. Extending AI systems with mechanisms that detect and respond to internally-generated incoherence might enable ACC to be driven by the system itself rather than external prompts.*

From this perspective, the prompt plays a dual role. First, it introduces *foodset elements*: concrete examples that seed the construction of task-relevant representations. Second, it provides *contextual catalysts*: instructions or formatting regularities that bias which operations become active. Given the model's pre-existing catalytic connectivity, these elements trigger rapid ACC in which representations and operations recursively enable one another, forming a self-sustaining subnetwork. When the prompt changes, the food set and catalytic constraints change, leading to dissolution of the previous subRAF and possible formation of a new one. ICL thus reflects not parameter learning, but the dynamic reconfiguration of an already rich autocatalytic representational system.

Toy Experiments

We present toy experiments illustrating key principles of ACC, noting they are meant as illustrative demonstrations of structural principles, not empirical validations of consciousness.

Experiment 1: Simple Autocatalytic Network

Setup: We simulate a small reaction network with 20 elements and 50 potential reactions. Each reaction has a random probability of being catalyzed by each element. We start with a food set of 3 elements and iteratively apply reactions whose catalysts are available.

Hypothesis: As catalytic density increases, the system will undergo a sharp transition from small, fragmented RAFs to a large, integrated RAF.

Results: Figure 2 shows the size of the MaxRAF as a function of catalytic probability p. Below $p \approx 0.12$, the MaxRAF is small (< 5 reactions). Above $p \approx 0.15$, it grows rapidly to encompass most reactions, confirming a phase transition.

Experiment 2: Phase Transition in Constraint Satisfaction

We construct a random graph with $n = 100$ nodes representing learned representations. An edge between nodes i and j exists with probability p, indicating mutual constraint. We vary $p \in [0, 0.05]$ and measure the fraction of nodes in the largest connected component (LCC), which serves as a proxy for the MaxRAF. For each value of p, we average results over 100 independent graph realizations.

Hypothesis: The system will exhibit a percolation phase transition. Below a critical threshold p_c, the network remains fragmented into small clusters; above it, a giant connected component emerges. Theory predicts $p_c = 1/n = 0.01$ for $n = 100$ (Erdős and Rényi 1960).

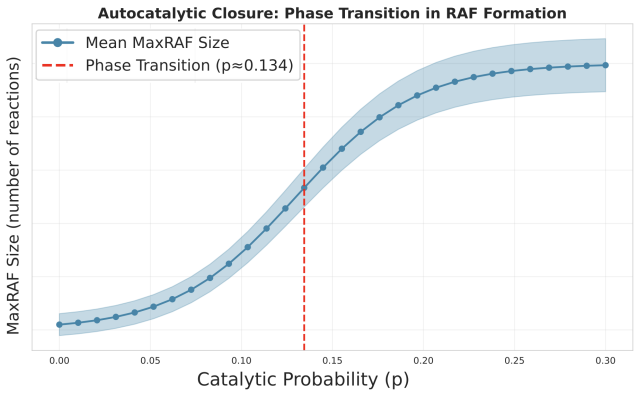

Figure 2: MaxRAF size vs. catalytic probability. A transition occurs near $p = 0.13$.

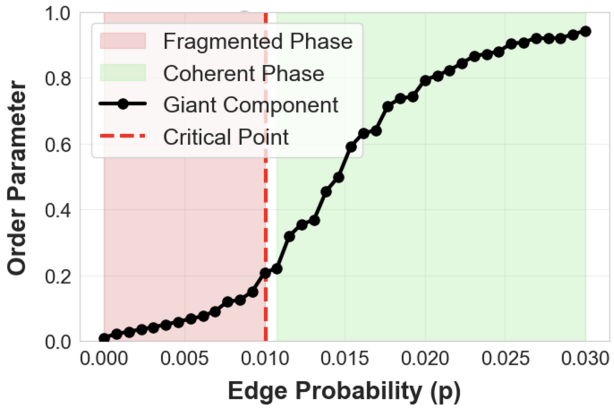

Figure 3: Phase diagram showing percolation transition in a constraint network. The giant component emerges at $p_c \approx 0.01$.

Results: Figure 3 shows the fraction of nodes in the largest component as a function of p. The transition occurs sharply at $p_c \approx 0.01$ (as expected for $N = 100$).

Experiment 3: Transformer-like Model as RAF

Setup: We implement a simplified transformer with 2 layers, 4 attention heads, and embeddings of dimension 32. We train it on a toy language (sequences of 10 tokens from a vocabulary of 50) using masked language modeling. We track the formation of self-attention patterns.

Hypothesis: Initially, attention patterns will be diffuse (low ACC). As training progresses, coherent attention patterns and RAF formation will emerge (high ACC).

Results: Figure 4 shows the evolution of attention entropy (lower entropy = more focused attention = higher catalysis) over training steps.

Interpretation: The model learns to form autocatalytic representations. Early in training, attention is random (no RAF). As training proceeds, structured attention patterns emerge, corresponding to the formation of a RAF where each token's representation is catalyzed by specific others.

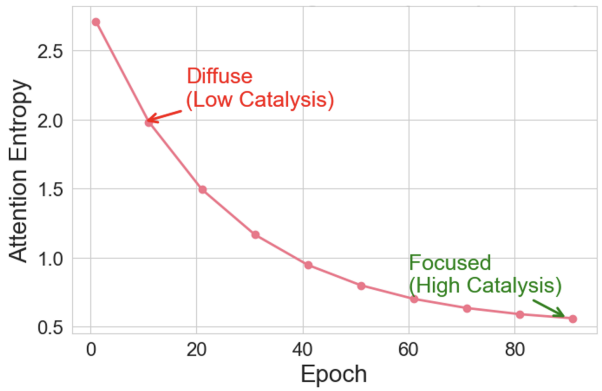

Figure 4: Attention entropy decreases as training progresses, indicating more focused, catalytic attention.

Discussion

The ACC approach frames machine consciousness as an organizational phenomenon grounded in internal dynamics. By modeling cognition in terms of ACC, we have emphasized how internally generated, self-sustaining representational structure arises from mutual enablement among cognitive operations.

We distinguish between static ACC—where a fixed set of representations achieves mutual enablement—and dynamic ACC, wherein the system continuously reconfigures its constraint structure. In the first, the system finds increasingly robust configurations. In the second, new foodset items are continuously incorporated. Human-like cognition operates in the dynamic regime in both respects: it can seek increasingly robust configurations, and each new learning episode potentially perturbs the existing ACC. This creates a tension between integrating novelty and preserving coherence that static ACC models do not capture.

The RAF approach reveals why continuous learning is computationally non-trivial. Each new foodset item introduces potential perturbations to existing catalytic cycles. The system faces a choice: assimilate the new item into existing structure (risking distortion), accommodate by restructuring (risking destabilization), or quarantine (risking fragmentation). We suggest that human-like intelligence excels at navigating this trade-off dynamically—a capacity that may not be fully realized even in AI systems that exhibit forms of ACC. Current LLMs exhibit static ACC: their parameter space represents a fixed closure over a predetermined training corpus. However, they lack the ongoing reorganization characteristic of human cognition. When humans encounter genuinely novel information that conflicts with existing structure, they can undergo conceptual change: a restructuring of the closure itself. Whether LLMs can achieve comparable dynamic closure remains an open question.

In the RAF framework, contemporary machine learning systems exhibit limited, task-bound instances of autocatalytic organization. In particular, in-context learning suggests that transformer architectures can transiently

form reflexively autocatalytic subnetworks that organize task-relevant representations. These structures, however, are short-lived and externally scaffolded: they depend on prompts supplied by the user and dissolve once the task context changes. As such, they lack the persistent ACC required for stable internal world-models.

Because ACC is defined in organizational rather than substrate-specific terms, it provides a framework for exploring consciousness-relevant dynamics across biological, computational, and alternative physical media.

Coherence-First Design Current AI design focuses on loss minimization. This work suggests a *coherence-first design* approach, wherein the goal is to achieve increasingly robust ACC. Practically, this means architectures that support mutual constraint satisfaction (e.g., attention, graph neural networks), enable accumulation of structural change across time, use training objectives that encourage RAF formation (e.g., consistency regularization), and monitor internal coherence metrics alongside loss.

Emergence Engineering If emergent capabilities are phase transitions, we can engineer them:

- *Identify the desired capability:* What RAF structure would support it?
- *Seed the food set:* Provide training data containing necessary foodset items.
- *Increase catalytic density:* Design tasks that encourage necessary catalytic relationships.
- *Monitor for transition:* Track maxRAF size; the capability should emerge when it crosses threshold θ.

Conclusions

This work investigates the relevance of ACC to machine consciousness. In contrast to prediction-error correction, ACC foregrounds internal coherence and self-sustaining organization as primary drivers of the learning process by which conscious systems develop models of their world.

Through toy experiments, we illustrated the core dynamics of ACC: the emergence of coherent structure at critical thresholds, the importance of catalytic density, and the formation of self-sustaining representational loops in transformer-like models. The present work suggests an avenue for designing conscious machines that prioritizes coherence over accuracy.

We suggest that the development and growth of ACC structure may play a key role in the development of conscious machines. This work is nascent and has many limitations. It does not prove that conscious systems necessarily exhibit ACC; it merely makes the case for why we should expect ACC to figure prominently in conscious AI, and shows that existing learning systems exhibit signature features of ACC.

One of the most interesting findings is the relationship between in-context learning and transient RAF formation. Looking ahead, algorithms for RAF-detection and analysis (Hordijk and Steel 2004; Hordijk, Smith, and Steel 2015; Steel, Xavier, and Huson 2020) offer a promising methodological bridge between theoretical accounts of internal organization and empirical analysis of artificial systems. Such analyses would enable systematic comparisons between architectures that exhibit only transient, task-elicited organization and those capable of sustaining internally generated structure over time.

Future work should focus on (1) adapting existing RAF algorithms to compute maxRAFs in high-dimensional representational spaces, (2) assessing whether standard training procedures systematically increase ACC, and (3) enabling networks to generate their own prompts so that learning can follow gradients shaped by the system's internal organization.

Supplementary Material

Appendix One: Proof of Theorem 2

Proof. We construct two counterexamples showing that optima of one objective need not be optima of the other.

Case 1: High Bayes performance, low ACC. Consider an overparameterized lookup-table-like model (e.g., a sufficiently wide neural network in the interpolation regime). Such a system can achieve near-zero training error ($\mathcal{L}_{\text{Bayes}} \approx 0$) by memorizing input-output pairs without building integrated structure. In RAF terms, representations remain fragmented: few reflexive catalytic loops form, MaxRAF size stays small, and $\mathcal{L}_{\text{ACC}}$ (e.g., measured by integration or catalytic density) remains low. Thus, Bayes minimization succeeds while ACC does not.

Case 2: High ACC, low Bayes performance. Consider a tightly constrained but misaligned representational network, such as a small but irreducible RAF with strong mutual constraints (high internal coherence) yet priors badly mismatched to the data distribution. In predictive processing terms, this resembles overly rigid high-level priors leading to persistent prediction errors. Here, $\mathcal{L}_{\text{ACC}}$ is minimized (large, self-sustaining RAF), but $\mathcal{L}_{\text{Bayes}}$ remains high due to systematic mismatch with external data.

Scalar illustration. In high-dimensional spaces, many directions allow minimization of one objective without affecting the other (e.g., adding isolated memorization capacity increases Bayes fit without growing RAF; strengthening internal loops increases ACC without reducing KL divergence). Thus, the gradients $\nabla\mathcal{L}_{\text{Bayes}}$ and $\nabla\mathcal{L}_{\text{ACC}}$ are not necessarily aligned, allowing divergence of optima. □

This distinction is crucial: the autocatalytic framework does not require that learned representations accurately predict the world—only that they form a coherent, self-sustaining whole. Prediction accuracy may emerge as a consequence of coherence, but it is not the fundamental driver. However, in practice, the two can be integrated, with prediction guiding the food set and closure organizing it.

Appendix Two: Proof of Theorem 3

Proof. **Setup and monotonicity.** For each n, the random structure is completely determined by independent Bernoulli(p_n) choices on $X_n \times R_n$. Increasing p_n can only

add catalysis edges, never remove them. Hence the property A_n is *monotone increasing* in p_n (equivalently in ρ_n).

(1) Subcritical regime (no linear-size RAF). A RAF $\mathcal{R} \subseteq R_n$ requires two constraints: (i) *Reflexive autocatalysis* (every $r \in \mathcal{R}$ has at least one catalyst in $F_n \cup \mathrm{Prod}(\mathcal{R})$), and (ii) F*-generation/ACC* (all reactants needed by reactions in $\mathcal{R}$ can be built from F_n using reactions in $\mathcal{R}$).

Under bounded arity, the foodset-derived ACC process can be exposed by iterating: start with $X^{(0)} := F_n$, and at step t add products of reactions whose reactants lie in $X^{(t)}$ and that are catalyzed by some element of $X^{(t)}$. This exploration can be dominated by a Galton–Watson-type branching process whose mean reproduction number is proportional to ρ_n, with constants depending only on the RRS arity bounds. When ρ_n is below a critical value ρ_c (the point at which this effective branching mean drops below 1), the dominated branching process dies out quickly with high probability, implying that the catalyzed F-ACC reaches only $o(|R_n|)$ reactions. Since any RAF must be contained in the catalyzed F-ACC (it must be foodset-derived and internally catalyzed), this rules out RAFs of size $\geq \alpha|R_n|$ w.h.p.

(2) Supercritical regime (existence of a giant RAF). When ρ_n is above the same critical value, the corresponding exploration process becomes supercritical: with probability bounded away from 0, the catalyzed F-ACC expands to include a linear fraction of reactions. Standard results in random catalytic reaction networks show that, above this regime, the ACC contains (and in fact typically equals, up to negligible boundary effects) a large reflexively autocatalytic, food-generated set; taking unions of RAFs and using maximality of MaxRAF yields that a maximal RAF exists and has size at least $\alpha|R_n|$ w.h.p. (Existence/maximality follows because the union of RAFs is again a RAF.)

(3) Sharpness of the transition. Because A_n is a monotone property of independent Bernoulli trials on $X_n \times R_n$, general sharp-threshold theorems for monotone properties (e.g. Friedgut–Kalai type results) imply that the change of $\mathbb{P}(\mathsf{A}_n)$ from near 0 to near 1 occurs in a vanishing window of p_n (equivalently of ρ_n), provided the RRS family is not dominated by a single exceptional coordinate (a mild regularity condition typically satisfied when the RRS is "spread out" and arity is bounded). Converting p_n to $\rho_n = p_n|X_n|$ gives the stated sharp window $w_n = o(1)$.

Capability corollary. Let $\mathcal{C}$ be a capability requiring a set of operations $\mathcal{R}_\mathcal{C} \subseteq R_n$ that must be simultaneously (i) food-generated and (ii) mutually catalytically supported. If $\mathcal{R}_\mathcal{C}$ is contained in some RAF whenever A_n holds, then $\mathbb{P}(\mathcal{C} \text{ is supported}) \geq \mathbb{P}(\mathsf{A}_n)$. By sharpness, this probability rises rapidly from near 0 to near 1 as ρ_n crosses ρ_c, yielding an apparent "sudden" emergence over a narrow scaling window (though the underlying order parameter need not be literally discontinuous). □

Acknowledgments

This research was conducted with funds from grant GR026749 to LG from the Natural Sciences and Engineering Research Council of Canada (NSERC). We thank Mike Steel for comments on the manuscript.

References

Brown, T.; Mann, B.; Ryder, N.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

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

Constant, A.; Friston, K.; and Clark, A. 2024. Cultivating creativity: predictive brains and the enlightened room problem. *Philos Trans R Soc Lond B Biol Sci*, 379(1895):20220415.

Dehaene, S.; Lau, H.; and Kouider, S. 2017. What is consciousness, and could machines have it? *Science*, 358(6362): 486–492.

Devlin, J.; Chang, M.-W.; Lee, K.; et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Erdős, P.; and Rényi, A. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1): 17–60.

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

Gabora, L.; and Bach, J. 2023. A path to generative artificial selves. In *Proc. 22nd EPIA Conf. on Artificial Intelligence*, 15–29. Cham: Springer.

Gabora, L.; Beckage, N.; and Steel, M. 2022. An autocatalytic network model of conceptual change. *Topics in Cognitive Science*, 14(1): 163–188.

Gabora, L.; and Steel, M. 2017. Autocatalytic networks in cognition and the origin of culture. *Journal of Theoretical Biology*, 431: 87–95.

Gabora, L.; and Steel, M. 2022. From uncertainty to insight: An autocatalytic framework. In Beghetto, R. A.; and Jaeger, G. J., eds., *Uncertainty: A Catalyst for Creativity, Learning and Development*, 125–158. Berlin: Springer.

Harkness, D.; and Keshavan, A. 2019. Is the Bayesian brain computationally tractable? *Computational Brain & Behavior*, 2(3-4): 253–256.

Hohwy, J. 2013. *The Predictive Mind*. Oxford: OUP.

Hordijk, W.; Hein, J.; and Steel, M. 2010. Autocatalytic sets and the origin of life. *Entropy*, 12(7): 1733–1742.

Hordijk, W.; Smith, J. I.; and Steel, M. 2015. Algorithms for Detecting and Analysing Autocatalytic Sets. *Algorithms for Molecular Biology*, 10: 15.

Hordijk, W.; and Steel, M. 2004. Detecting Autocatalytic, Self-Sustaining Sets in Chemical Reaction Systems. *Journal of Theoretical Biology*, 227(4): 451–461.

Hordijk, W.; Steel, M.; and Kauffman, S. 2012. The structure of autocatalytic sets: Evolvability, enablement, and emergence. *Acta Biotheoretica*, 60(4): 379–392.

Kauffman, S. A. 1993. *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford University Press.

Russell, S. J.; and Norvig, P. 2021. *Artificial Intelligence: A Modern Approach*. Hoboken, NJ: Pearson, 4 edition.

Seth, A. K.; and Bayne, T. 2022. Theories of consciousness. *Nature Reviews Neuroscience*, 23(7): 439–452.

Steel, M.; Hordijk, W.; and Smith, J. 2013. Minimal autocatalytic networks. *Journal of Theoretical Biology*, 332: 96–107.

Steel, M.; Xavier, J. C.; and Huson, D. H. 2020. The Structure of Autocatalytic Networks, with Application to Early Biochemistry. *Journal of the Royal Society Interface*, 17: 20200488.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2 edition.

Vaswani, A.; Shazeer, N.; Parmar, N.; et al. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.