Physical Interpretations of Emergent Behaviors

Bentley Yusen Lin; Bentley@safeware.com.tw

September 10, 2025

Abstract

This paper presents a novel mathematical framework for interpreting emergent behaviors in complex systems, with a particular focus on modern artificial intelligence (AI). We posit that emergence is not a mystical phenomenon but a predictable consequence of a system's exploration of a high-dimensional state space. By defining the relationship between the microscopic rules of a system and its macroscopic properties through the lens of exploration capacity, we provide a model that explains why low-probability, "miraculous" behaviors become reliable and in sufficiently large systems, such as large language models. The core of our thesis is the derivation of an inverse exponential relationship between the probability of an emergent behavior and the size of the exploration space, offering a quantitative basis for this previously qualitative concept. This revised framework introduces the concept of *emergent areas* as behavioral classes that undergo phase transitions based on system scale and dimensionality.

1 Introduction

The concept of *emergence* describes the phenomenon where complex behaviors and patterns arise from the interaction of simpler components. These macroscopic behaviors are often difficult to predict from the knowledge of the microscopic rules alone [1]. In fields ranging from thermodynamics to biology, emergence is a cornerstone principle. The recent and rapid advancement of AI, particularly with the rise of deep learning, has brought this concept to the forefront of computer science. Capabilities such as in-context learning, reasoning, and creativity *emerge* in large-scale neural networks without being explicitly programmed [2].

While often described qualitatively, a pressing need exists for a formal, mathematical framework to describe the mechanics of emergence. This paper aims to bridge that gap by proposing a physical and mathematical interpretation of emergence, framing it as a function of a system's exploration capacity within a constrained high-dimensional space.

2 A Model of Exploration and Emergence

2.1 Definitions

Consider a system whose state can be represented as a point in an N-dimensional space. At any discrete time step t, the system can transition to a new state by moving in a chosen direction.

- **Theoretical Exploration Space ($\mathbf{\Omega_T}$):** In an unconstrained system, the number of possible directional choices per time step is 2^N (positive or negative along each axis).
- **Constrained Exploration Space ($\mathbf{\Omega_C}$):** Physical and mathematical constraints (e.g., network architecture, loss landscape geometry) limit the system's realistic choices. We model this reduction by defining an *effective exploration dimensionality* K, where $K < N$. The number of available directions per time step is thus reduced to 2^K.

2.2 The Exploration Space Size

Over a sequence of T time steps (e.g., training steps or inference steps), the total number of possible paths S the system can take is given by:

$$S = (2^K)^T = 2^{KT} \tag{1}$$

S represents the total **exploration space size**. The logarithm of S is proportional to the system's informational entropy, representing its potential diversity of behaviors.

2.3 From Paths to Emergent Areas: A Revised Probability Model

Rather than considering individual paths to specific behaviors, we introduce the concept of **emergent areas** - regions in the state space that represent coherent behavioral phenotypes or classes (e.g., reasoning, creativity, deception).

Let $E_i \subset S$ represent an emergent area corresponding to a specific behavioral class Φ_i. The probability of exhibiting behavior from this class is given by the normalized measure of this area:

$$P(\Phi_i) = \frac{\mu(E_i)}{\mu(S)} \approx \frac{f(S)}{S} \tag{2}$$

where $\mu(E_i)$ is the measure (volume) of the emergent area E_i, and $f(S)$ is a sigmoidal function representing the phase transition of the emergent area's measure as S increases:

$$\mu(E_i) = f(S) = \frac{L_i}{1 + e^{-k_i(S-u_i)}} \tag{3}$$

where:

- L_i is the maximum potential measure of the emergent area E_i,
- u_i is the threshold exploration size for the phase transition of E_i,
- k_i is a growth factor defining the sharpness of the transition.

This model explains why emergent behaviors transition from impossible to stochastic to reliable as system scale increases.

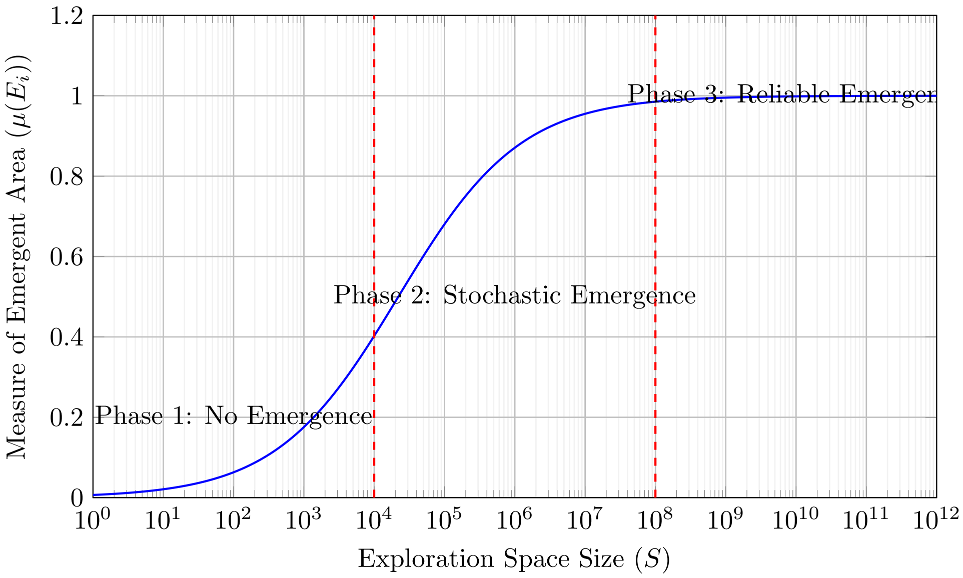

Figure 1: **Phase Transition of Emergent Area Measure:** The measure (volume) of an emergent area $\mu(E_i)$ undergoes a sigmoidal phase transition as exploration space size S increases. Below threshold u_1, the area has negligible measure ($\mu(E_i) \approx 0$). Between u_1 and u_2, the area forms but has small measure, making emergence possible but unreliable. Above u_2, the area has significant measure, making emergence reliable.

3 The Mechanism of Miracle-to-Normality Transition

The transition of an emergent behavior from a "miracle" to a "normality" is a direct consequence of scaling S [3], which drives the phase transition of emergent areas.

3.1 The Threshold of Emergence

There exist thresholds u_1 and u_2 such that:

- When $S < u_1$, $\mu(E_i) \approx 0$ for all emergent areas E_i - no emergence is possible.
- When $u_1 \leq S < u_2$, $\mu(E_i) > 0$ but small - emergence is possible but stochastic.
- When $S \geq u_2$, $\mu(E_i)$ is significant - emergence becomes reliable.

Although $P(\Phi_i)$ may be small for any specific behavioral class, the absolute accessibility of behaviors in class Φ_i becomes significant when S is sufficiently large:

$$\text{Accessibility}(\Phi_i) = \mu(E_i) = P(\Phi_i) \cdot \mu(S)$$

The "miracle" is observed relative to a system with $S < u_2$.

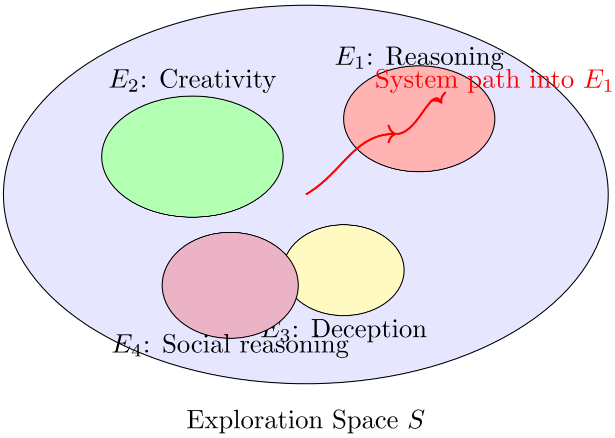

Figure 2: **Exploration Space with Multiple Emergent Areas:** The exploration space S contains multiple emergent areas (E_1, E_2, E_3, E_4), each representing a different behavioral class or phenotype. As the system explores the space (red path), it may enter different emergent areas, exhibiting different macroscopic behaviors. The size and accessibility of these areas depend on the overall scale of S.

3.2 Engineering Emergence through Scale and Dimensionality

Modern AI provides a quintessential example of this principle. The "technical means" to enlarge S are:

- **Increasing K (Effective Dimensionality)**: This is achieved by increasing model parameter count. A larger model has a higher-dimensional, richer space of possible internal representations and transformations [4]. Each new capability (language, reasoning, tool use) adds a new dimension to the exploration space.
- **Increasing T (Exploration Steps)**: This corresponds to increasing training compute (FLOPs) and allowing longer reasoning chains during inference (e.g., Chain-of-Thought prompting) [5].

By scaling K and T, engineers effectively force S far beyond the threshold u_2 for desired capabilities like reasoning or code generation. Consequently, these once-emergent "miracles" become reliable, marketable features.

4 Discussion and Conclusion

We have formulated a mathematical model where emergence is redefined not as an anomaly but as an inevitable outcome of scale. The phase transition of emergent areas provides a quantitative foundation for this phenomenon, explaining both the "miracle" of emergence in small systems and the "normality" of emergence in large systems.

This framework demystifies the behaviors observed in large-scale AI systems. It suggests that the pursuit of Artificial General Intelligence (AGI) may be less about discovering new algorithms and more about the strategic scaling of existing architectures to navigate the exploration space efficiently [6]. The challenge shifts from *whether* a capability can emerge to *how we can guide* the system to find the emergent areas E_i that lead to desirable and safe macroscopic behaviors.

Future work will focus on refining this model, particularly on characterizing the growth function of $\mu(E_i)$ for specific emergent capabilities and developing methods to guide systems toward desirable emergent areas.

References

References

[1] Mitchell, M. (2009). *Complexity: A Guided Tour.* Oxford University Press.

[2] Wei, J., et al. (2022). *Emergent Abilities of Large Language Models.* Transactions on Machine Learning Research.

[3] Kaplan, J., et al. (2020). *Scaling Laws for Neural Language Models.* arXiv:2001.08361.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning.* Nature, 521(7553), 436–444.

[5] Wei, J., et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* Advances in Neural Information Processing Systems.

[6] Hutter, M. (2019). *Human Knowledge: Foundations and Limits.* Online. `https://www.hutter1.net/ai/humknow.htm`