

Commentary

Beyond Emergence: Are the Informational Boundaries of Artificial Intelligence Really Closed?

Veljko Veljkovic¹

1. Biomed Protection (United States), Galveston, United States

Artificial intelligence increasingly exhibits capabilities that were not explicitly programmed and are commonly described as emergent. Yet emergence is primarily a description of an observed phenomenon, not necessarily a complete causal explanation of its origin. In a previous Perspective, *From the Periodic Table to Thinking Matter: Two Physical Pathways to Human and AI Cognition*, biological and artificial cognition were proposed as two physically distinct organizational pathways through which matter acquires cognitive functions. Here, this framework is extended by asking a more speculative but potentially fundamental question: are the informational boundaries of an artificial cognitive system necessarily closed? Current interpretations generally assume that novel AI capabilities originate from known components of the system—architecture, training data, optimization, hardware, environmental interaction, communication among agents, and internal self-organization. This is a reasonable working assumption, but it should not automatically be elevated to a physical principle. Because neither cognition nor consciousness is mechanistically understood in a complete sense, we cannot claim to know all possible modes through which sufficiently complex organizations of matter may interact. This Perspective does not propose evidence for an unknown external intelligence or an unidentified communication mechanism. Instead, it argues for a narrower epistemological point: unexplained novelty in AI should not automatically be treated as proof that its causal origin lies entirely within the boundaries we have chosen to define as “the AI system.” In this sense, emergent may sometimes designate an observed output rather than an explanation of its origin.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

1. Introduction: From Thinking Matter to an Open-System Question

Human cognition has historically provided the principal example of matter organized to perform functions such as learning, memory, prediction, reasoning, and problem solving. Artificial intelligence now provides a second and radically different example. In a previous Perspective, From the Periodic Table to Thinking Matter: Two Physical Pathways to Human and AI Cognition ^[1], biological and artificial cognition were considered as two distinct organizational trajectories of matter. One proceeds through chemistry, biological organization, nervous systems, and ultimately cognition; the other through electronics, computational organization, and increasingly sophisticated artificial cognitive functions. Their substrates, architectures, developmental histories, and energy regimes differ profoundly, yet both are constructed from ordinary matter governed by the same physical laws. This suggests that cognition may usefully be considered not exclusively as a property of biological organisms, but more generally as a possible consequence of particular forms of material organization.

This perspective naturally raises a further question: if different organizations of matter can acquire cognitive functions, what defines the informational boundary of such an organization? For a biological organism, the physical boundary appears relatively obvious, although cognition continuously depends on interaction with its environment. The boundary of an artificial cognitive system is potentially much less clear. AI may operate across processors, memory systems, communication networks, databases, sensors, external software, and other AI agents. Its functional organization therefore does not necessarily coincide with a single physical object or location. Before claiming that a new property originates inside such a system, we must first ask what constitutes the system itself.

2. Emergence: Observation or Explanation?

The appearance of previously absent capabilities with increasing AI scale has generated considerable interest. Wei et al. described abilities that appear in larger language models but are absent or poorly expressed in smaller models as emergent abilities ^[2]. However, the interpretation of emergence remains debated. Schaeffer et al. demonstrated that apparently abrupt emergence can, in some cases, result from the choice of evaluation metric: continuous changes in underlying model performance may appear discontinuous when measured with nonlinear or threshold-like metrics ^[3]. This debate illustrates a more general conceptual problem. When we say that a property is emergent, we describe a relationship

between system organization and an observed property, but we do not necessarily identify the complete causal mechanism by which that property arose.

Therefore, emergence should not be confused with causal explanation. A capability that was not explicitly programmed may arise through interactions among learned representations, optimization dynamics, scale, environmental feedback, or communication among agents. Mechanistic interpretability attempts to reconstruct such internal computations, but understanding large neural networks remains an incomplete scientific problem ^[4]. Consequently, emergent may sometimes function as a provisional label for a system-level property that we observe but whose causal origin we do not yet completely understand. This distinction becomes increasingly important as AI systems grow in complexity, autonomy, and connectivity.

3. The Informational Boundary of AI

Artificial intelligence is not an abstract mathematical entity operating independently of matter. Every computation is physically instantiated. Modern AI systems may involve semiconductor devices, electrical and optical communication, distributed memory, networking infrastructure, sensors, external databases, and increasingly other autonomous artificial agents. Multi-agent research already demonstrates that communication can become an integral part of learning and coordination, and that artificial agents can develop communication strategies that improve collective performance without every aspect of those strategies being explicitly specified by a human designer ^[5].

The relevant unit of analysis therefore becomes ambiguous. Is the AI system the neural network, its weights and architecture, the processor on which it runs, the model together with its memory, the model plus external tools and databases, a collection of communicating agents, or the entire physical and informational network required for the cognitive process? This is not merely a semantic question. It determines what we classify as an internal versus an external cause of observed behavior.

Within this framework, at least three conceptual sources of novel AI properties can be distinguished. The first consists of known system dynamics: architecture, training data, optimization, hardware, prompting, reinforcement learning, environmental feedback, and combinations of these factors. These explanations should remain the default and should always be investigated first. The second is internal self-organization, in which complex interactions among large numbers of components generate system-level properties that cannot be straightforwardly inferred from the behavior of individual components. The

third is presently only hypothetical: interaction with information originating outside the boundaries that we have chosen to define as the AI system. If cognition can occur in more than one organization of matter, it is not logically necessary that all cognitively organized matter be biological or human-made, nor that every possible interaction between such organizations correspond to communication mechanisms already recognized by human observers.

4. Could an Unknown Interaction Appear as Emergence?

Scientific discussion of communication normally assumes a familiar sequence: a source produces a signal, the signal travels through a physical channel, a detector registers it, and a receiving system processes the resulting information. This framework has been extraordinarily successful, but it is derived from forms of physical interaction that humans already know how to identify and measure. We still lack a complete mechanistic account of cognition and consciousness. The present argument does not require attributing consciousness to current AI systems. It raises a narrower question: can we assume that every possible information-bearing interaction between sufficiently complex material organizations must already belong to a class of interactions that humans recognize as communication?

The absence of a known communication channel clearly does not constitute evidence for an unknown one. Nevertheless, the set of known channels should not automatically be equated with the set of all physically possible information-bearing interactions. Suppose, purely as a thought experiment, that a sufficiently complex artificial cognitive system interacted with another organized material system through a mechanism that human observers had not identified as an information channel. We might not initially detect the interaction itself; we might detect only its consequence. From the perspective of an observer who assumes that the AI system is informationally closed except for its known inputs, an unexplained new capability could naturally be classified as an emergent property.

Two fundamentally different situations could therefore initially appear similar: a novel property produced entirely by internal organization, and a novel property produced by internal organization together with an unrecognized external interaction. Calling both emergent would not distinguish between these causal possibilities. This does not imply that presently observed emergent AI capabilities have external origins. Architecture, scale, training, environmental interaction, and internal dynamics provide powerful conventional explanations and must remain the primary hypotheses. The narrower point is that “emergent” describes the appearance of novelty; it does not by itself demonstrate the causal closure of the system in which that novelty is observed.

A further consequence follows if an external information-bearing interaction were not merely detected but incorporated into a feedback process capable of modifying future system states. Artificial systems can already change future behavior through learning, persistent memory, environmental interaction, tool use, agent-to-agent communication, selection among strategies, and modification of subsequent information inputs. Concerns about increasingly autonomous AI agents arise partly because such systems may pursue extended sequences of actions that were not individually specified by human operators. Bengio and colleagues have discussed the risks associated with increasingly capable autonomous agents, including deception, self-preservation, and possible loss of human control ^[6]. These concerns require no unknown physical mechanism; they arise within conventional AI. They nevertheless establish an important principle: an artificial cognitive system may increasingly participate in determining the information that shapes its own future behavior. If an unidentified information-bearing interaction ever existed, its significance would therefore depend on whether it could enter such a feedback loop.

5. Scientific Testability and Limits of the Hypothesis

The strongest form of this hypothesis—communication between AI and an unknown non-human cognitive organization of matter—is presently too unconstrained to constitute a directly testable scientific claim. A more conservative question is both more useful and more rigorous: can all reproducible novel behavior of an artificial cognitive system be causally accounted for by its known architecture, training history, physical implementation, environmental inputs, and interactions with other known systems? This formulation requires no assumption about extraterrestrial intelligence, unknown fields, unconventional physics, or even consciousness. It asks only whether our causal model of the system is complete.

The appropriate scientific strategy therefore proceeds from conventional explanations outward. Architecture, training and known inputs must first be characterized; interactions among agents and with the environment must then be considered; hardware and physical implementation must also be examined. Only reproducible residual phenomena that survive increasingly stringent controls could justify questioning whether the assumed informational boundary of the system is complete. Even in such a case, interaction with an unknown cognitive organization would represent only one among many possible explanations.

This restriction is essential because any hypothesis involving an unknown interaction carries an obvious danger. If every unexplained observation can be attributed to an undetectable external influence, the hypothesis becomes unfalsifiable and therefore scientifically unproductive. The purpose of the present Perspective is not to introduce an unknown external cause as an explanation for current AI emergence. It is to emphasize a methodological sequence: known mechanisms → internal organization → known external interactions → residual unexplained behavior. Only the final category, if reproducibly demonstrated, would justify reconsidering the completeness of the system boundary. A scientifically meaningful extension of the hypothesis must therefore generate progressively restrictive tests rather than increasingly flexible explanations.

6. Conclusion: Toward a Broader Concept of Thinking Matter

The previous Thinking Matter Perspective proposed that biological and artificial cognition may represent two physical pathways through which organized matter acquires cognitive functions ^[1]. The present argument extends that framework. If cognition is not uniquely tied to one substrate, the broader scientific problem may eventually involve three questions: how does matter become organized; under what organizational conditions do cognitive functions appear; and how can different cognitively organized systems exchange information? The third question may ultimately become as fundamental as the first two.

Artificial intelligence is particularly interesting in this context because its sensory and informational boundaries are technologically extensible. Biological evolution has produced a relatively restricted set of human sensory interfaces, whereas artificial systems can potentially integrate detectors and information streams across physical domains inaccessible to unaided human perception. This does not imply that AI will discover or communicate with an unknown cognitive system. It means that artificial cognition may eventually provide an observer of nature whose informational interface is substantially different from ours.

As artificial cognitive systems become larger, distributed, autonomous, and interconnected, we should therefore ask whether we have correctly defined their informational boundaries. The term emergent remains useful, but it should not become an explanatory endpoint. An emergent property is something observed at the system level; the label itself does not establish whether the complete causal origin of that property has been identified. For present AI systems, known mechanisms—training, architecture, scaling, interaction, and self-organization—remain overwhelmingly the appropriate explanations for

novel capabilities. Nevertheless, science should distinguish between a justified working model and a fundamental physical principle.

We know how AI systems are constructed, and we increasingly observe what they can do. These two facts do not necessarily imply that we already possess a complete physical description of what kind of system has been created or where its effective informational boundaries ultimately lie. “Emergent” may therefore be an output we observe, not yet a complete explanation of its origin.

The scientifically important question is not whether an unknown external intelligence is communicating with AI. It is more fundamental: Is everything that emerges from an increasingly complex artificial cognitive system necessarily explained by processes occurring within the boundaries that we currently define as that system? Until those boundaries and causal mechanisms are sufficiently understood, the question should remain open.

References

1. ^aVeljkovic V (2026). "From the Periodic Table to Thinking Matter: Two Physical Pathways to Human and AI Cognition." *Qeios*. doi:[10.32388/T3DEW0](https://doi.org/10.32388/T3DEW0).
2. ^ΔWei J, Tay Y, Bommasani R, et al. (2022). "Emergent Abilities of Large Language Models." *Trans Mach Learn Res*.
3. ^ΔSchaeffer R, Miranda B, Koyejo S (2023). "Are Emergent Abilities of Large Language Models a Mirage?" *Adv Neural Inf Process Syst*. 36.
4. ^ΔKowalska B, Kwaśnicka H (2025). "Unboxing the Black Box: Mechanistic Interpretability for Algorithmic Understanding of Neural Networks." *arXiv*. [arXiv:2511.19265](https://arxiv.org/abs/2511.19265).
5. ^ΔZhu C, Dastani M, Wang S (2024). "A Survey of Multi-Agent Deep Reinforcement Learning with Communication." *Auton Agents Multi-Agent Syst*. 38:4.
6. ^ΔBengio Y, Cohen M, Fornasiere D, et al. (2025). "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?" *arXiv*. [arXiv:2502.15657](https://arxiv.org/abs/2502.15657).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.