Perspectives

MAX PLANCK LAW

ISSN 2748-3924

What Is Emerging in Artificial Intelligence Systems?

Daria Kim
17 July 2024
DOI: 10.17176/20240717-115050-0

ABSTRACT

This note examines in what sense qualities often attributed to AI can be considered 'emergent' by drawing on perspectives on complexity. It goes on to identify the predictive capacity of neural networks as weakly emergent, aligning with explanatory reductionism, and highlights the need to appreciate human causality in normative assessment.

1 Today's discussions around artificial intelligence (AI) often invoke the notion of 'emergence' to capture an intuitive perception of AI functionalities loosely associated with intelligence, behaviour, creativity, agency, and even personality. Given that emergence is intertwined with autonomy, these accounts further reinforce the framing of AI as 'agents'.

2 Much like the agentic framing of AI, emergence in AI systems tends to be anthropomorphised and mystified, being conflated with the idea that AI acquires 'a mind of its own' and exhibits emergent 'behaviour' which can be 'bewildering', 'exciting and a bit scary', 'fascinating', and even akin to 'magic' and 'alchemy'. Others have viewed the attribution of emerging 'behaviour' to AI as 'a huge cause of concern (and hype)' and argued that the perception of emerging AI capabilities is a matter of the evaluative benchmarks applied, 'rather than an inherent property' of AI models.

3 From a normative perspective, the perception of emerging personality, intelligence, creativity, agency, and autonomy in AI systems is both captivating and puzzling, and has been driving debates about its legal, ethical, and societal ramifications. A robust analysis of the normative implications should build on a credible explanatory account of the underlying technological phenomena. Acknowledging concerns

that conceptual and terminological imprecision can perpetuate misconceptions about AI, this note provides a mapping of the concept of emergence onto machine learning (ML) models based on artificial neural networks (ANNs) to examine what might be 'emergent' about these models, and in which sense. Understanding ANN-based systems through the lens of emergence can further inform the analysis of their legal and regulatory treatment.

Emergence in Complex Systems

4 While the concept of emergence traces its philosophical origins back to Aristotle, it experienced a relatively recent revival with the rise of research on nonlinear complex systems, both biological and artificial. In general terms, emergents are phenomena observed at the system level and arising from the interaction of the system's constitutive elements. These phenomena—such as structures and properties—may not be readily understood or predicted solely from knowledge of the properties of the system's individual elements. The limited predictability, and hence control, of emergents is attributed to incomplete knowledge of the rules governing elements' interactions, nonlinearity, feedback loops within a system, and the system's capacity to adapt to its environment over time. The perceived distinctiveness of emergent properties is captured by expressions such as 'the whole is more than a sum', '[p]arts behave differently in wholes' or 'more is different'. Everyday examples include patterns in ice, flocking birds, stock market trends, and social movements.

5 Complex systems appear to *depend* on their components while also developing *distinct* properties and functionalities that are autonomous relative to the laws governing those components. Depending on the understanding of this relationship between the system and its components—particularly in terms of their dependence vs autonomy and distinctness vs identity—perspectives on emergence differ, leaning towards either weak or strong accounts of emergence. The strong version posits that emergent properties are imposed in a top-down manner by new causal forces not present in the system's elements, while the weak version suggests that emergent properties arise from the bottom-up organization of the system's lower-level components. Weak emergence is compatible with ontological reductionism—the view that system-level emergent phenomena are just lower-level components arranged in specific ways without the interference of new higher-level forces or causal forces, as well as explanatory reductionism—the view that higher-level emergent phenomena can be 'exhaustively' explained by referring to their lower-level components.

6 The literature on emergence offers diverse and sophisticated perspectives; however, providing a systematic assessment of analytical accounts of emergence would exceed the scope of this article. Instead, my goal is to highlight that emergence need not be mystified. While complex systems can be perceived as exhibiting novel emergent qualities, emergence is not 'a mysterious non-causal enigma'; it does not involve 'mysteries popping out of the undergrowth'. Even with strong emergence, there are 'no strange new fundamental forces or kooky élan vital'. Rather, emergence means that a comprehensive understanding of

interactions among the system components can explain 'why a complex system has properties its parts lack on their own' (which, for a strong emergentist, would be a contentious point).

Predictive Capacity as an Emergent Capability of ANN Models

7 The characteristics of complex systems summarized above may ring a bell when considering ANN models. These correspondences between the definitional aspects of complex systems and the characteristics of ANN models can be mapped as follows:

- **The multiplicity of system elements**: At the elementary level, an ANN model consists of a myriad of *numeric values* that are initially derived from numeric representations of training data and subsequently transformed (ie optimized or 'learned') during the model training.

- **The interaction of system components**: An ANN model is built through mathematical operations performed on numeric values. These values interact within an ANN during the model training, influencing each other's adjustment, just as numbers interact in calculus.

- **Rules governing interaction**: The interaction of numeric values within an ANN model is determined by an algorithm, encompassing mathematical operators, such as activation and cost functions. Currently, the training of ANNs is based on mathematical optimization—more precisely, error minimization.

- **The phenomenon perceived as *emergent* on the system level**: A trained ANN model is capable of generating output when presented with new data. In its raw form, this output consists of numeric values. More precisely, the generation of the output refers to the model's prediction of numeric values enabled through model training.

- **Unpredictability of emergent phenomena**: The output of ANN models is often perceived and described as unpredictable or surprising, even to their developers.

- **Nonlinearity**: An ANN model is a nonlinear function where model inputs and outputs are related in a nonlinear way. Nonlinearity results from the application of nonlinear activation functions and the iteration of artificial neurons outputs throughout model training. Nonlinearity enables ANN models to capture complex relationships and patterns in real-world data that are beyond the grasp of linear models.

- **Adaptability of a system to the environment**: A trained ANN model can continue 'learning'—further optimizing model parameters—by being exposed to new data, which improves its predictive power or accuracy.

8 Thus, from a brute-fact point of view, the emergent capability in ANN models boils down to their capacity to produce numeric output—generate predictions—when exposed to new data at the inference stage. This output can take the form of a single numeric value (typical for binary classification tasks), a vector of numeric values (typical for classification tasks such as image recognition), or an ordered sequence of vectors of numeric values (common for large language models that predict the next token 'over and over again'). The common denominator among these types of model output is that they constitute numeric values predicted by the model. It is worth emphasizing that the capacity of a trained ANN to predict numeric values is based on the process of mathematical optimization—the error minimization—performed on the numeric representation of the input data during the model training phase.

Reductively Explainable and (Weakly) Emergent?

9 Though lacking a universal definition, emergence is generally associated with the characteristics of being novel, irreducible, and unpredictable. Thus, the analysis of emergence involves examining the relationship between a system as a whole and its components in terms of these features. The novelty aspect denotes the identity versus distinctness of the system's properties relative to those of its parts, while irreducibility reflects the autonomy versus dependence of the system as a whole relative to the system's components. Both novelty and irreducibility bear on the predictability of emergent properties.

10 As for the identity vs distinctness dichotomy, one could view the predictive capability of ANN models as a *distinct* property compared to a single artificial neuron of an ANN, as a single neuron simply lacks this capacity. This functional distinctiveness of a single artificial neuron relative to a trained model is akin to how a single brick lacks the property of shelter provided by a building composed of thousands of bricks, or the trajectory taken by flocking birds may differ compared to that taken by an individual bird.

11 In terms of autonomy vs dependence of system properties on system components, the predictive capacity of ANN models is entirely dependent on the model elements—numeric representation of the training data—and how the training data is fitted into a model. With their layered structure, ANN models serve as a good example to illustrate the above-cited proposition that one can explain the genesis of the phenomena arising at the system level with a sufficient understanding of interactive processes within a system and the laws governing these interactions. In each processing layer, the inputs to the neurons are transformed through mathematical operations—by applying bias to weighted sums followed by an activation function to produce the output of the layer. This output serves, in turn, as the input value for the next processing layer. In other words, lower-level computational processes *predetermine* higher-level computational states. The transformations of numeric values, as the input data is passed through the layers of an ANN architecture resulting in a trained model, determine the model's capability to generate predictions. Thus, both an ANN model as the result of training and a prediction generated by a model are the unfolding of a cause-and-effect chain,

whereby models do not have the power to act counterfactually, violating the antecedent causal factors.

12 The limited explainability and predictability of an ANN's output are functions of the model complexity (the sheer number of numeric transformations within a model), nonlinearity, deployment factors (particularly new data to which a model reacts), and the model's capacity to adjust to the environment over time. While ANN models are often described as 'black boxes', a characterization which unfortunately contributes to their mystification, computational operations are executed according to the pre-defined instructions, including mathematical operators, that cannot be violated by the computer itself. To emphasize, the 'black-box' characteristic refers to limited insight into how numeric values and operations translate to semantic meaning, rather than implying agentic or mystical powers at work. In other words, despite the limited predictability of the numeric output, the underlying computational processes are deterministic in the sense that the same inputs under the same ontologically objective causal conditions would produce the same output. (Admittedly, replicating all the conditions necessary for reproducing the processes within an ANN may be challenging and potentially unfeasible in practice if randomization techniques, particularly those based on natural entropy, are employed.)

13 Irrespective of challenges to predictability and reproducibility of the model output, the predictive capability of a trained ANN is *reducible* to—ie exhaustively determined by—the model's constitutive elements (numeric values) and mathematical and other rules governing their interactions comprised within the training algorithm. From this perspective, the predictive capability of ML models aligns with the weak account of emergence, particularly the proposition that all relevant causation shaping system-level phenomena occurs through interactions of the system's elements at the lower levels. This is also consistent with explanatory reductionism, which posits that higher-level emergent phenomena can be exhaustively explained in terms of lower-level component interactions.

A Normative Orientation

14 The pitfall is that 'emergence' easily becomes mystified in its use as a 'go-to' term to articulate an intuitive perception of intelligence, decision-making capacity, creativity, mind, and will 'emerging' within AI systems. The initial analysis presented here suggests that the predictive capability of an ANN model is *reductively explainable* and can be deemed as 'emerging'—distinctively efficacious compared to its elements—under the weak account of emergence. This view rejects the idea postulated by strong emergentists that system-level phenomena arise due to an intervention of new (relative to the system constituent components) 'top-down' powers, including agency and subjectively experienced qualities such as mental states. From this perspective, attributing qualities such as human-like agency and personality to ANN-based systems appears as bizarre as personifying other emergent phenomena, such as the movement of water molecules, protein folding, transition from gas to liquid, or crystal growth.

The reductionist account of ANNs' capabilities does not deny or belittle the challenges of mitigating technological risks and allocating legal consequences for the outcomes of ANN-based systems. Their limited predictability and explainability stem from multiple factors, particularly the hidden patterns within data and the randomization applied in model training and deployment. However, these factors neither negate nor conflict with the deterministic relationship between the training of an ANN model and its predictive capacity. The reductionist perspective suggests that instead of grappling with the elusive notion of emergent agency *within* AI systems, a normative analysis should shift the focus to examining legal concepts and rules to address the increasingly dispersed and distanced human causation in AI artefacts and their applications.

Acknowledgement

I want to thank Man Wai Kwok for his meticulous, multi-round review of the paper's technical aspects and our ongoing discussions about AI. The paper has also benefited from Allison Felmy's sharp editorial eye. Any errors are my own.

Sources Cited

PW Anderson, 'More Is Different' (1972) 177 Science 393

Devansh, 'The Biggest Lie about Language Models – AI Emergence. Breaking down the Machine Learning Paper of the Year' (*Medium*, 28 December 2023) <https://medium.com/the-modern-scientist/the-biggest-lie-about-language-models-ai-emergence-b2fe807de5ac> accessed 3 July 2024

Carl Gillett, *Reduction and Emergence in Science and Philosophy* (CUP 2016)

Marshall Gunnell, 'Emergent Behavior in AI' (*Technopedia*, 26 December 2023) <https://www.techopedia.com/definition/emergent-behavior> accessed 3 July 2024

Henrik Jeldtoft Jensen, 'Complex Systems and Emergent Phenomena' in Robert A Meyers (ed), *Encyclopedia of Complexity and Systems Science* (Springer 2009) 1268

AH Louie AH and Roberto Poli, 'Complex Systems' in R Poli (ed), *Handbook of Anticipation* (Springer 2019) 17

Klaus Mainzer, 'Complex Systems' in ALC Runehov, L Oviedo (eds), *Encyclopedia of Sciences and Religions* (Springer 2013) 437

Nestor Maslej and others, 'Artificial Intelligence Index Report 2024' (2024) <https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf> accessed 3 July 2024

Timothy O'Connor, 'Emergent Properties' in Zala EN (ed), *The Stanford Encyclopedia of Philosophy* (2021) <https://plato.stanford.edu/archives/win2021/entries/properties-emergent/> accessed 3 July 2024

Claude Sammut and Geoffrey I Webb (eds), 'Agent', *Encyclopedia of Machine Learning and Data Mining* (Springer 2017) 40

Achim Stephan, 'Emergence, Theories Of' in Anne LC Runehov and Lluis Oviedo (eds), *Encyclopedia of Sciences and Religions* (Springer 2013) 714

The Ezra Klein Show, 'How Should I Be Using A.I. Right Now? (*The New York Times*, 2 April 2024) <https://www.nytimes.com/2024/04/02/opinion/ezra-klein-podcast-ethan-mollick.html?showTranscript=1> accessed 3 July 2024

WIPO Conversation on Intellectual Property and Frontier Technologies—Ninth Session (13 March 2024) <https://webcast.wipo.int/video/WIPO_IP_CONV_GE_24_2024-03-13_AM_122112> accessed 3 July 2024

Wesley J Wildman and F LeRon Shults, 'Emergence: What Does It Mean and How Is It Relevant to Computer Engineering?' in Saurabh Mittal, Saikou Diallo, and Andreas Tolk (eds), *Emergent Behavior in Complex Systems Engineering: A modeling and simulation approach* (Wiley 2018) 21

Kai Zenner, 'A Law for Foundation Models: The EU AI Act Can Improve Regulation for Fairer Competition' (*OECD.AI Policy Observatory*, 20 July 2023) <https://oecd.ai/en/wonk/foundation-models-eu-ai-act-fairer-competition> accessed 3 July 2024

SUGGESTED CITATION: Daria Kim, 'What Is Emerging in Artificial Intelligence Systems?', *Max Planck Law Perspectives* (17 July 2024), https://law.mpg.de/perspectives/what-is-emerging-in-artificial-intelligence-systems/, DOI: 10.17176/20240717-115050-0

CC BY NC ND