Review of: "Towards a Comprehensive Theory of Aligned Emergence in AI Systems: Navigating Complexity towards Coherence"

Yasin Kaygusuz

Potential competing interests: No potential competing interests to declare.

1. Introduction

"*Complexity theory provides a multidisciplinary perspective for analysing complex systems characterized by intricate hierarchies and network patterns that arise from simple rules. It emphasizes that understanding the individual components of a complex system is insufficient for predicting its overall behaviour, as the system exhibits distinct properties that emerge from the interactions among its components. This holistic view underscores the interconnectedness and dynamics of the system as a whole. (Larsen-Freeman, 2013)*"

The above paragraph is the approach of Larsen and Freeman. Normally complexity tehory has nothing to do with "understanding the individual components of a complex system is insufficient for predicting its overall behavior". Understanding the individual components is mandatory in understanding the overall behaviour, yet not solely sufficient. Therefore, I propose to add some additional introduction on coplexity theory. This description seems vague to me personally. also, it seems the descriptions given for complexity theory and emergence theory are strongly similar. The difference int between needs to be described.

"*Additionally, the study emphasizes the dynamic nature of alignment within AI systems. Rather than perceiving alignment as a static, one-time achievement, it advocates for understanding alignment as an ongoing and evolving construct.*"

Above paragraph need to be updated with a proper, neat and clean definition of alignment. What is alignment? what do you understand from it? Is it the direction, heading, behaviour? please clarify.

2. AI

""*Artificial Intelligence (AI), at its core, is an expansive field dedicated to creating machines that can mimic human intelligence.*" My personal opinion is that current AI is used to various tasks beyond human intelligence. No need to mimic human intelligence. The goal is to provide *some* intelligence. Not human intelligence.

1. **2.4. Artificial Neural Networks** *(There is a typo)*

I dont get why you needed such a surface mathematical representation of ANNs. The description gives nothing but the output computations. I propose to give a visual, a picture to show that an ANN is and not to provide equations. If you think eqs are required, then you need to give backpropagation or cost functions etc whatever training method is used. Than it

becomes a big task.

2.5.1

"*To achieve AGI, a system would need to combine a variety of AI techniques, including machine learning, deep learning, reinforcement learning, and more*." This sentence needs either reference or proof. One can easily claim single method may work and not joint method needed. ex. I propose the most promising method is SPAs and SNNs provided by NEF by Eliasmith from Waterloo.

3.

"*In the context of AI and especially neural networks, emergence can refer to behaviours or capabilities of the system that arise from the interaction and organization of individual artificial neurons. Such behaviours or capabilities can be complex tasks like recognizing objects in an image or understanding natural language, which are not predictable based on the individual neurons alone. (Bubeck et.al, 2023)*". this paragraph is quite the hearth of the proposal in the manuscript. I liked it. Chalmers asks the same question. "how does a V4 neuron knows that the answer is 4?" I propose you to widen your paragraph.

4.1

"*One of the defining hallmarks of emergent phenomena in AI is the ability to generate behaviors that are more than the sum of their parts. AI models, consisting of interconnected components such as neural networks, exhibit emergent behavior that extends beyond the capabilities of their individual neurons or algorithms. This emergent behavior encompasses the model's ability to learn, adapt, generalize, and exhibit intelligent decision-making in complex environments. (Gershenson & Fernández, 2012)*" I need to remind you that the neurons in ANNs are not real neurons and are just equations. Therefore, one can claim that the behaviour is not emerging from neurons, but it is simply the behaviour of the ANN. It is quite discussable that emergence is not needed when the neuron is not there as s subcomponent.

"*AI systems also demonstrate self-adaptation, the capacity to modify their structure and behavior in response to changing conditions or experiences. Through mechanisms like machine learning and reinforcement learning, AI models continuously refine their internal parameters, decision-making policies, or predictive capabilities. This self-adaptation empowers AI systems to improve their performance, learn from data, and autonomously adapt to new challenges or tasks.*" What you mention is something different than training or weight adjustment? this claim needs strong evidence. Please show an AI that changes its own structure, not the weights or some basic parameters solely. **Weak claim.** Also; "*This capacity for pattern recognition, abstraction, and generalization enables AI systems to discover novel insights, solve complex problems, and provide innovative solutions. (Bubeck et.al, 2023)*" for some writers scientific curiosity is solely belonging to human in our universe. You claim of innovating AI requires evidence. Please show an AI application that can innovate. Most of such claims is misunderstandings. There is no such thing like emerging creativity when the outputs are just a summary of the web. chatGPT does not innovate, just replicate or mix what exists in web. Additionally, one cannot

say deduction is properly innovation.

4.1.2

"*Moreover, AI systems may find novel solutions to problems that were not explicitly programmed into them. (Bubeck et.al, 2023)*" see above. needs evidence.

"*Consider the case of reinforcement learning, where AI agents learn to perform tasks by maximizing a reward signal. There have been numerous instances where these AI agents have discovered solutions that were unanticipated by their human creators. For example, in playing a game, an AI might discover a completely new strategy that humans had not thought of, or in a physics simulation, an AI might find an innovative way to achieve its goal.*" In the general structure of the manuscript, there are multiple examples written that way. Each needs an example from a real life manuscript. e.g. jason's (200x) gambling bot etc….

5.1

"*For example, an AI system designed to understand and respond to human language—such as OpenAI's GPT-4—does not simply process input and generate output based on pre-defined rules.*" GPT4 does not understand human language. This is what a weak AI claimant would tell. this sentence is quite under discussion world wide in cognitive science. If you believe so, it makes you a strong AI claimist. For q weak AI claimist, GPT4 just scans and replicate precoded.

6. and 6.1. 6.2

for all 6, I propose to use a better representation for the equations. B_t is not a good way. B(t) or $B_t$may work better.

For 6.1.

There is no clue before the graph for the first emergence example? what is the function outputing that result? please clarify. If F is just hypothetical as you said, how can it provide a graph output? Therefore, I understand you adjust some basic functions to compute such an output.

6.5.

Before that chapter, the system that you constitute is already a dynamical system because the states of the system already depends on previous states. So attractors or repellers will exist and so you have a dynamic complex system. in all your previous paragraphs B(t) is a function of B(t-1). So I do not get why you needed 6.5. I propose you to open the reasons.

7. I think this is a well written chapter and describes the goal. But there are things that I did not get.

G'i: how do we get it for a complex system, when its previous state is not known; your estimation assumes it is on a known state.

Notice such integral systems for dynamical complex systems are very difficult to solve. I in past tried to solve some while

studying DFTs and only steady state solutions could be solved. I propose you to assume a steady state, such as a 0 attractor or a divergence point and then solve the system for it. It will be very useful since your system will be integrable.