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

Kieran Greer

Potential competing interests: No potential competing interests to declare.

I think that this paper is not ready for publication. There is some very good work, but ultimately, you have maybe summarised too much other work, maybe laboring some points and then tried to put it into a theory that does not quite work for me. I am not convinced that you fully understand the underlying problems, so I would do more research first and try to simplify what you currently have, before adding your own theory.

I think that complexity theory and emergence theory are in fact major topics in AI that have been looked at over the years. Studying insects or swarm behaviour, agent-based systems or autonomous systems, for example. But these systems are still difficult to write and control, so any contribution there would be welcome. You might want to consider some of the earlier work on the topic, such as Complex Adaptive Systems (John Holland), fractals, or cellular automata.

Your summary of AI is very good, up to maybe section 4.1.1. With Goodfellow, you might want to mention Generative networks explicitly. No figure of an ANN?

I am not sure that you should call neural networks emergent. While they are black boxes in that individual nodes cannot be understood by themselves, their behaviour is for the most part predictable and they are essentially statistical. The neurons carry out the same function each time and so the behaviour does not change. You would want to consider something like cellular automata, where the function does change, or distributed systems where the nodes can behave independent of each other.

Section 4

For this reason, I think that section 4.1.1 is incorrect and also the following sections. Once you get onto the notion of emergence, I think that you are not correct. You have to allow the components to behave independently of each other and be able to adapt. A neural network is distribtued but behaves in a holistic way. The new ChatGPT networks may become unpredictable simply because of the numbers of nodes involved (billions), but I think that is a different topic.

Section 5

This contains the sort of information that you would need to consider, but the equation is not practical at the moment. For emergent behaviour, you would also have to consider how the different agents interact with each other - maybe that is the

environment and history parts. I will leave the review there, because I think that you still have some work to do before this can be published.