Emergent AI: A Position Paper on Extracting Novel Strategies from Superhuman AI Systems

Adam Rida
adamrida.ra@gmail.com

March 21, 2025

Abstract

Beyond AI's widely acknowledged predictive capabilities lies a profound and underexplored potential: its role as a tool for epistemological discovery. This paper argues for the use of emergent multi-agent AI systems as instruments to uncover novel strategies and deepen human comprehension of complex, self-organizing systems. Inspired by Edgar Morin's concepts of emergence and holistic complexity, and illustrated through recent AI achievements such as AlphaFold [3] and Altera's [10] spontaneous social simulations, we advocate for a transdisciplinary methodology leveraging explainability techniques. By systematically extracting insights from superhuman AI models, this approach seeks to bring to light previously unkown strategies and concepts to humans, enhancing both scientific discovery and our broader understanding of complex systems.

1 Introduction

Since its inception in the 1950s, AI has transformed how we approach problem-solving by giving us solutions to increasingly complex problems. Being able to model and predict complex phenomena has seen immense application fields across the years, from the United States Postal Service using convolutional neural networks [1] to parse zip codes all the way to getting close to solving natural language [2]. Yet, beyond AI's impressive capabilities to predict lies another interesting potential. What if AI could be used as an epistemological tool to increase human knowledge and discover new strategies in the most complex systems of our universe. Echoing Edgar Morin's [5] view of complexity and emergence as a holistic phenomenon, this paper presents and positions emergent AI as a discovery engine capable of expanding human knowledge. Inspired by recent breakthroughs in AI —such as DeepMind's AlphaFold [3], which revolutionized biology by solving the protein-folding problem, and Altera's multi-agent simulations capable of spontaneously exhibiting sophisticated social behaviors [10]—this perspective proposes to leverage knowledge extraction techniques (e.g. Explainability) to better understand how AI models with superhuman ability understand about the phenomenon they model that we don't as humans.

We propose that emergent multi-agent AI systems can serve as a playground for discovering novel strategies, concepts, and principles governing complex and self-organizing systems (following Morin's definition [5]). Ultimately, if we manage to capture AI's emergent capabilities through a transdisciplinary lens, it could not only open the door to novel scientific and practical discoveries but also help steer AI development towards humanity's everlasting quest to understand ourselves and our universe. This position paper aims to spark new ideas and hint at potentially promising directions, rather than offer definitive claims or solutions.

2 Background

Emergence has been a subject of study for decades, with pioneers like Edgar Morin emphasizing the limitations of reductionism and the importance of understanding complex, self-organizing systems. He argues that the way science was done in the 20th century had dangerous limitations due to the fact that it was focused on reductionism and discrimination, making us blind to the true nature of complex systems. He took different examples from biology to sociology and showed that in most cases, complex systems can only be apprehended as a whole, especially considering their self-organizing nature. Laws that emerge from one layer of abstraction to another are very hard to predict at our scale (modulo the Laplassian devil) - we cannot predict politics or wars by describing atom-to-atom interactions.

The next very hard problems to model for humanity are complex and self-organizing ones. Whether it's protein folding, weather forecast or economical crisis, modeling has become challenging but not impossible. Although, as the mathematician Box said, "All models are wrong, some are useful".

We have seen multiple attempts at modeling chaotic or pseudo chaotic phenomenon both in academia and in the industry. Most managed to reach performance on-par with top human performers in their respective fields and even sometimes outperform them. One of the most significant examples is DeepMind's breakthrough with AlphaFold, which accurately predicts protein structures from an amino-acid sequence, a long-standing challenge in biology. This is typically a successful attempt at modelling an emergent phenomenon from nature.

On another note, video games have been a good playground to self-train models to perform and significant capabilities were achieved. For instance, OpenAI's Five model demonstrated superhuman performance in Dota 2, a game with significant strategic complexity due to its numerous degrees of freedom. Similarly, MetaAI released around the same time agents that mastered Starcraft II. The interesting aspect about video games is that it's relatively easy to interpret as humans what are the strategies used by the agent in front of us. For dota for example, it could be translated as simply making the decision of buying a specific item and then later down the line realizing that because of this item, there is now a damage gap.

Additionally, Schut et al. [14] have shown that AlphaZero had learned chess strategies that even grandmasters weren't aware of, and demonstrated these concepts could be successfully taught to top human chess players. McGrath et al. [9] similarly analyzed how the AI system acquired chess concepts during training, with insights from former world champion Vladimir Kramnik. More recently, Altera, with their multiagent minecraft model [10], has been able to recreate emerging sociological phenomena without enforcing it at the beginning. This could open the door to exploring phase transitions in these emerging behaviors.

The field that studies human-AI interactions and understanding, Explainable AI (XAI), has been flourishing for the past years with an increasing awareness of dangers behind black-box models. Whether it's to detect spurious correlations, mitigate biases or provide actionability (counterfactual reasoning), XAI has become very proficient at extracting knowledge from complex ML models.

This lays the path to what could be an interesting opportunity behind complex systems and AI. Can we use XAI to extract knowledge from those superhuman models to make us more knowledgeable? What if we can, as humans, understand more deeply our universe through them? More formally, we hypothesize that concept-based interpretability methods applied to emergent multi-agent AI can systematically reveal novel, human-usable strategies in domains such as biology, complex systems, and beyond.

3 Research Directions and Applications

While most of research and industry have been focusing on automating things, we are missing out on understanding underlying mechanisms of emergence. Ultimately, this can help bring us closer to better understanding our universe and its laws. Instead of building superintelligence only to solve hard problems, we propose to work on extracting knowledge from this superintelligence.

Evidence from Schut et al. work on AlphaZero [14] shows that it's possible to extract novel chess concepts from superhuman AI systems and even transfer this knowledge to top human players. McGrath et al. [9] similarly demonstrated how linear probes can extract concepts learned by these models. Kim et al.'s T-CAV [13] method provides ways to abstract out concepts from deep learning models. More generally, we propose exploring how we can devise new approaches designed specifically to help extract insights from AI models that teach humans new concepts. This new epistemological approach would allow us to leverage AI as a pure discovery engine at the service of humanity.

More practically, servicing the power of AI to discovery opens up different wide application fields. With such approaches, we could imagine having human-AI strategy transfer. Similarly to what has been done to MOBA video games ([6], [7]), we could imagine extending it to games that are closer to concepts from other sciences. An interesting area to start on this could be to explore any synergies between the SIMA team [16] and Dr Been Kim team, both teams at DeepMind. Extending T-CAV to agents that learn to play in different types of games could help explain the results found by SIMA, namely the fact that an agent trained on different games was still able to perform decently on an unseen game. Which abstract concepts were learned?

Extending it as well to Altera's minecraft would for instance help understand better how emerging patterns appeared. Further down the line, we could learn more about markets (e.g. how financial crises emerge), logistics or even policymaking ([10] - was able to reproduce emerging policy making by allowing agents to vote and even influence votes).

Adding real-time analysis of the current state of a complex system could help as well improve decision making by understanding better the impact of both macro and micro decision. In scientific discovery, we could study chaotic systems in biology or physics under a new lens and understand new properties. In oncology for instance, the immense complexity of possible factors that could trigger a cancer within a cell makes it a hard problem to solve. What if we can bridge the gap between mollecular's interaction complexity and human understanding through abstract concepts? By its nature, life is a self-organizing system that has too many components that interact with each other, making it very challenging to understand and uncover its mechanisms.

As mentioned in the previous section, we are getting closer to having models that mimic complex systems (at least to an extent [3]) and having methods that could help see through this complexity by extracting the learned concepts from such models would be invaluable to medical practitioners.

Looking ahead, these potential directions do not claim to cover all aspects of emergent AI research. Rather, they aim to encourage further exploration and collaboration across disciplines, ultimately laying the groundwork for bridging the gap between superhuman AI intelligence and genuine human comprehension. By systematically harnessing emergent properties and interpretability methods, we may be able to push beyond the current boundaries of what we understand about complex systems and ourselves.

4 Conclusion

This position paper presents a different perspective where AI would not only be used for automation but rather as a collaborator to help us answer the most important questions about our universe by understanding complex, self-organizing systems. Through abstracting out emergent properties with explainable AI, we can tap into the ”black box” insights of superhuman AI models and transform them into human-usable knowledge.

Crucially, this paper does not claim to have all the answers; instead, its aim is to spark further questioning, research and encourage ongoing, transdisciplinary collaboration. Whether in biology, the social sciences, or the design of advanced multi-agent systems, AI's emergent capabilities offer a chance to shed light on previously unknown new strategies and concepts. If we can bridge the current gap between AI's superhuman performance and genuine human comprehension, we may find ourselves on the threshold of discoveries far exceeding the scope of today's approaches.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[2] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[3] J. Jumper, R. Evans, A. Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[4] M. Baker, J. Z. Leibo, L. Barreira, et al. Emergent Social Learning and Coordination via Multi-Agent Reinforcement Learning. *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2020:583–591, 2020.

[5] E. Morin. *La Méthode*. Seuil, 1977.

[6] OpenAI Team. OpenAI Five. `https://openai.com/blog/openai-five/`, 2018.

[7] O. Vinyals, I. Babuschkin, W. M. Czarnecki, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[8] D. Silver, J. Schrittwieser, K. Simonyan, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[9] T. McGrath, A. Kapishnikov, N. Tomašev, et al. Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022.

[10] Altera.AL, A. Ahn, N. Becker, et al. Project Sid: Many-agent simulations toward AI civilization. *arXiv preprint arXiv:2411.00114*, 2024.

[11] Z. C. Lipton. The Mythos of Model Interpretability. *Commun. ACM*, 61(10):36–43, 2018.

[12] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv preprint arXiv:1702.08608*, 2017.

[13] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 80:2668–2677, 2018.

[14] L. Schut, N. Tomasev, T. McGrath, D. Hassabis, U. Paquet, B. Kim. Bridging the Human-AI Knowledge Gap: Concept Discovery and Transfer in AlphaZero. *arXiv preprint arXiv:2310.16410*, 2023.

[15] S. Badal, T. Gordon, et al. Machine Learning in Oncology: Emerging Concepts and Methodologies. *Comput. Struct. Biotechnol. J.*, 18:2416–2429, 2020.

[16] A. S. Vezhnevets, J. Z. Leibo, P. Kohli, et al. SIMA: General-Purpose Interfaces for Intelligent Agent Development. *ArXiv preprint arXiv:2304.07937*, 2023.