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

Adam Rida
adamrida.ra@gmail.com

March 28, 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 unknown 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 and Edgar Morin's complexity 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. Morin argues in his book "La Méthode"[5] 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. Fragmenting complex systems into smaller, seemingly manageable parts would constrain scientific discovery and innovation that stems from trans-disciplinarity. Instead, emergence must be understood as a holistic phenomenon, where the whole exhibits properties that none of the parts possess individually. This perspective is further elaborated in the work of Baas and Emmeche from the Santa Fe Institute, who define emergence mathematically and discuss its implications for explaining complex systems [20].

Morin emphasizes that emergence is not a mere combination of components but a process where new organizational principles arise spontaneously from interactions between simpler elements. These emergent properties have characteristics that are irreducible and unpredictable from the individual parts alone. For Morin, understanding emergence requires acknowledging the dual nature of organization and disorganization, where disorder and randomness are not merely noise but essential drivers of creativity and complexity. This perspective confirms the need for a transdisciplinary approach to studying complex systems, where emergence is a dynamic process that continuously evolves and reshapes itself, challenging any attempt to capture it fully through reductionist methods [5].

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.

Morin's perspective is especially relevant today as we venture into multi-agent AI systems and other models that inherently express emergent properties. The challenge lies not only in creating models that exhibit emergence but also in decoding the underlying principles that give rise to these unexpected behaviors. By adopting Morin's holistic view, we can better appreciate the epistemological value of leveraging AI to explore and understand emergent phenomena in complex systems.

Hard-to-model systems The next set of 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. As the statistician George Box said, "All models are wrong, some are useful".

Significant attempts have been made aiming 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. Historically, the research community has tried to leverage Conway's Game of Life [17] to do so. Despite being a discrete system defined by simple rules, the Game of Life shows how complexity and unexpected behavior can emerge from simplicity. More recently, this concept has been extended to the continuous space (Lenia [18] and Flow-Lenia [19]) with the goal of trying to capture and generate creatures that model biological systems more accurately (factoring in dynamics, robustness to perturbation, healing/reconstruction capabilities, and 'reasoning').

On another note, DeepMind's breakthrough with AlphaFold [3] is another example of successful modeling of an emerging phenomenon from nature. Unlike agent-based modeling or cellular automata like Lenia [18] and Flow-Lenia [19], AlphaFold relies on deep learning architectures, particularly attention mechanisms, to predict protein structures. This approach is purely data-driven, learning from vast amounts of structural data, rather than simulating interactions between agents or cells and was able to yield impressive results. AlphaFold revolutionized structural biology by being able to accurately predict protein structures

from an amino-acid sequence, a long-standing challenge in biology. They showed that AI can achieve superhuman performance in understanding highly complex chaotic systems.

Video games as the perfect playground 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 [6], 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 [7]. 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 multi-agent Minecraft environment [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.

Explainable (XAI) 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.

A notable method in this domain is Testing with Concept Activation Vectors (TCAV), introduced by Kim et al. [13]. TCAV provides a framework to interpret the internal state of neural networks through human-understandable concepts. It quantifies the influence of predefined abstract concepts on a model's predictions, offering insights beyond traditional feature attribution methods. By applying TCAV, practitioners can assess the extent to which a concept affects a model's output without requiring access to the training data or retraining the model. This is particularly valuable in domains where understanding the rationale behind AI decisions is crucial, such as healthcare and finance.

For example, in medical imaging, TCAV has been employed to interpret deep learning models analyzing physiological data, like electroencephalograms (EEGs). By defining concepts corresponding to specific waveform patterns, researchers can determine how these patterns influence the model's predictions, thereby enhancing trust and facilitating collaboration between AI systems and medical professionals [21].

While TCAV traditionally relies on predefined concepts, Ghorbani et al. [22] introduced an unsupervised method that automatically discovers and explains concepts without needing predefined labels.

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 could, as humans, understand our universe more deeply 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 Emergent AI as a discovery engine

We propose a different research direction that considers AI not just as a predictive instrument, but as an epistemological tool. Rather than solely seeking superintelligent solutions to hard problems, we argue for systematic *knowledge extraction* from these models. Building on work such as Schut et al. [14], who showed

that AlphaFold had discovered strategies even grandmasters were unaware of, we see a wealth of untapped potential in using T-CAV and related methods to uncover emergent strategies and concepts.

Specifically, we hypothesize that **concept-based interpretability methods** (e.g., T-CAV) applied to complex, multi-agent AI systems can systematically reveal previously unknown but human-usable insights. We envision a framework wherein:

1. **Develop Emergent Models.** Train multi-agent systems or complex models in domains with high-dimensional interactions, such as social simulations, biological processes, or strategic games. Emergence is not forced but allowed to arise organically through the interactions between agents or components.

2. **Monitor and Capture Emergent Phenomena.** Continuously analyze models during training to identify novel patterns or behaviors that cannot be predicted from initial conditions. This includes both qualitative observations (e.g., unexpected coordination strategies) and quantitative measures (e.g., new attractor states in dynamical systems).

3. **Extract and Formalize Concepts.** Apply techniques such as T-CAV, concept bottleneck models, or attention-based probing to extract human-understandable concepts from emergent behaviors. This involves detecting concepts but also creating abstractions that can be integrated into human knowledge frameworks.

4. **Iterative Concept Refinement.** Present extracted concepts to domain experts for validation, refinement, and formalization. This feedback loop ensures that discovered knowledge is both accurate and usable, enhancing human comprehension of complex systems.

5. **Application and Cross-Domain Transfer.** Test the validity of discovered concepts in related domains. For instance, strategies derived from AI-driven biological models could potentially inspire new techniques in socio-economic modeling or vice versa.

This iterative framework considers knowledge extraction from emergent AI systems is a continuous process of discovery, validation, and refinement. Rather than a static model, it also echoes Morin's call for a holistic approach to complexity, where emergent properties are integrated into broader conceptual frameworks.

When AI is approached as a tool for discovery rather than just prediction, emergent insights can reveal valuable knowledge that would otherwise remain hidden.

4 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.

We base the following on evidence from Schut et al. work on AlphaZero [14], which shows that it's possible to extract novel chess concepts from superhuman AI systems and even transfer this knowledge to top human players. Similarly, McGrath et al. [9] similarly demonstrated how linear probes can extract concepts learned by these models. Finally 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.

On concept-based interpretability A key challenge in leveraging emergent multi-agent systems as tools for discovery is systematically extracting the underlying knowledge that drives their superhuman capabilities. While most interpretability efforts in AI focus on static models (e.g., image classification, language understanding), the dynamic and self-organizing nature of multi-agent systems requires novel explainability techniques that can capture and elucidate emergent phenomena as they unfold over time.

To bridge this gap, we propose extending concept-based interpretability methods like T-CAV and linear probes to track the evolution of high-level concepts during the learning process of multi-agent systems. By applying these techniques to emergent phenomena, it becomes possible to identify abstract principles and strategies that arise spontaneously during training.

For instance, monitoring how coordination, communication, or even conflict resolution mechanisms emerge from simpler interactions could yield insights applicable far beyond the original training environment.

Furthermore, the interpretability challenge in emergent AI systems is not just about detecting what concepts exist, but understanding how they interact and influence each other within the system. This requires developing tools that can visualize and decompose complex behaviors into simpler, understandable parts without losing sight of the holistic nature of emergence. Techniques such as causal inference within neural networks, combined with traditional XAI approaches, could provide a powerful framework for mapping emergent strategies into actionable human knowledge.

Crucially, this approach is not limited to theoretical understanding; it can serve as a practical tool for human-AI collaboration. By continuously extracting and validating emergent concepts from AI systems, we can effectively establish a feedback loop where human knowledge is enhanced by AI discoveries and vice versa. This iterative process could be particularly valuable in domains where human expertise is limited, such as high-dimensional biological systems or socio-economic simulations where the underlying rules are not fully understood.

Application fields 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 Deepmind's SIMA team [16] or Altera [10] and Dr Been Kim's team also 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 and transferred?

Extending it as well to Altera's minecraft agents [10] would, for instance, help to understand better how emerging patterns appeared - how did agents agree on specific things and take collective decisions?

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 for any humans to interpret holistically, 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 sys-

tems (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.

5 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.

More importantly, 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 an uncharted territory of discoveries.

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.

[17] M. Gardner. The Fantastic Combinations of John Conway's New Solitaire Game 'Life'. *Scientific American*, 223(4):120–123, 1970.

[18] B. W.-C. Chan. Lenia: Biology of Artificial Life. *Complex Systems*, 28(3):251–286, 2019.

[19] E. Plantec, G. Hamon, M. Etcheverry, P.-Y. Oudeyer, C. Moulin-Frier, B. W.-C. Chan. Flow-Lenia: Towards Open-Ended Evolution in Cellular Automata Through Mass Conservation and Parameter Localization. *arXiv preprint arXiv:2212.07906*, 2022. https://arxiv.org/abs/2212.07906.

[20] N. Baas and C. Emmeche. On Emergence and Explanation. *Santa Fe Institute Working Paper*, 97-02-008, 1997. https://www.santafe.edu/research/results/working-papers/on-emergence-and-explanation.

[21] A. Janik, J. Dodd, G. Ifrim, K. Sankaran, K. Curran. Interpretability of a Deep Learning Model in the Application of Cardiac MRI Segmentation with an ACDC Challenge Dataset.

[22] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.