Symbolic AI and Big Data

Andrew Li

October 2024

The landscape of artificial intelligence (AI) research has shifted dramatically in recent years with the rise of large language models (LLMs) like GPT-4 and Gemini. In this research statement, I argue why symbolic approaches to AI are more relevant than ever in the age of big data and share my vision for future AI systems that augment LLMs with symbol manipulation. I describe how my past research on neurosymbolic AI supports this vision before concluding with a set of proposals for the future.

Why Should We Care About Symbols?

Before the deep learning revolution of the 2000s, AI systems that operated over symbolic inputs and outputs were ubiquitous in AI research. In his seminal work *Programs with Common Sense* [1], John McCarthy proposed the idea of an "advice-taking" program that accepts and stores knowledge as declarative facts and rules, then manipulates sentences in a formal language to reason and adapt to new situations. These ideas culminated in the development of *expert systems* that pushed AI capabilities along several frontiers—from automatically providing medical diagnoses [2] and configuring computer systems [3] to playing chess at a superhuman level [4].

Today, with the advent of advanced chatbots and generative models, one is less likely to think of systems operating over symbols when they think of AI[1]. Does this mean that modern, data-driven machine learning systems have superseded symbolic methods in every way imaginable? Not at all. We continue to delegate a host of important problems to systems that are symbolic or rule-based, rather than data-driven. Compilers are complex programs written by humans, not learned from data. Shortest path problems continue to be delegated to explicit graph search algorithms like Dijkstra's, not neural nets trained on millions of graphs. Even in domains like chess, where the introduction of neural nets proved to be revolutionary, explicit search techniques remain indispensable for achieving strong performance [5, 6].

Systems that procedurally manipulate symbols have strengths that are difficult to emulate with neural nets alone. Compilers and search algorithms like Dijkstra's behave reliably for *any* valid program and *any* graph, while neural nets, which often struggle to extrapolate beyond their training data, would require an inordinate amount of data and training

[1] The "AI Effect" is a phenomenon where behaviours previously achieved by AI are no longer perceived as requiring intelligence. (https://en.wikipedia.org/wiki/AI_effect)

compute to reach a similar degree of reliability. Furthermore, symbolic approaches afford a level of interpretability and transparency that neural nets do not.

Abstraction: Bridging Data and Symbols

Many real-world problems of interest operate over far more granular representations than the symbols that classical AI systems manipulate. For example, a robot system might take camera and haptic sensor inputs comprising hundreds of thousands of data points per second, while outputting torques for each individual actuator. I believe this disconnect poses the most significant barrier to the broader application of symbolic methods in AI systems.

This barrier is overcome through a process called a*bstraction*, in which low-level inputs (e.g. images) are mapped to high-level manipulable symbols [7]. For instance, modern autonomous vehicle systems achieve this by explicitly determining their position and orientation in the environment (*localization*), detecting nearby objects (*perception*), and inferring the intentions of other road users (*prediction*), all from sensory inputs, before traditional search techniques are applied to plan a route [8].

Past Work. Broadly speaking, my research investigates how we can automatically abstract high-dimensional data into symbols in order to bring to bear classical AI techniques towards important real-world problems over complex inputs and outputs.

My early graduate research focuses on temporal abstraction from long time-series data. In our AAAI 2021 paper, *Interpretable Sequence Classification via Discrete Optimization* [9], we train compact sequence classifiers in the form of finite-state automata that support explanation, counterfactual reasoning, and human-in-the-loop modification. A follow-up work [10] shows that learning automata representations of a reward function in the training loop of a reinforcement learning (RL) agent expedites learning.

Inspired by the disparity between modern deep RL systems and traditional belief state techniques for POMDPs, our ICML 2023 work *Learning Belief Representations for Partially Observable Deep RL* [11] introduces a state abstraction technique for deep RL agents to capture the notion of a belief state. Operating over these abstract state representations drastically improves agents' ability to seek and remember salient information.

The core of my PhD research explores how symbolic and compositional representations of RL tasks can be leveraged to improve the reasoning ability of RL agents while exposing the rationale behind their decision making. In our ICML 2021 work *LTL2Action: Generalizing LTL Instructions for Multi-Task RL* [12], we train an RL agent to understand temporal and logical relationships expressed in the formal language Linear Temporal Logic (LTL). Using a prespecified symbolic abstraction, we enable the agent to model its own task progress and solve a wide array of never-before-seen robotic tasks in a single try. Our NeurIPS 2022 paper *Learning to Follow Instructions in Text-Based Games* [13] extends this approach to a challenging text-based game. We equip an RL agent with an internal LTL representation of tasks to improve its instruction-following capabilities and propose an LLM-based translation tool from natural language to LTL.

Symbols and Abstraction in the Era of Large Language Models

The rise of LLMs has unlocked a wealth of new opportunities for neurosymbolic AI systems. If abstraction is indeed the most critical prerequisite to adopting techniques from symbolic AI, then LLMs are a massive leap forward—language is a nearly universal abstraction encompassing everyday concepts and underpinning much of human commonsense reasoning [14]. Modern multimodal LLMs make it possible to seamlessly map common data-rich modalities like vision or audio to text. Thus, I envision that many future AI systems will employ LLMs for the purpose of abstraction into everyday, human-interpretable concepts, supported by an under-the-hood symbolic manipulation system for reasoning.

Future Work. The challenges and opportunities related to this vision are as follows.

Firstly, neural nets (including LLMs) are notoriously error-prone when exposed to inputs not encountered during training, and language is notoriously ambiguous [15]. Thus, symbolic AI systems will need to be robust to a noisy abstraction process. Our recent NeurIPS 2024 paper [16] provides an initial study into how neurosymbolic RL systems are impacted by an imperfectly specified symbolic abstraction, and how this impact can be mitigated.

Secondly, LLMs open the door to the development of AI that can reason in many domains. Until recently, AI capabilities have largely been limited to singular domains or problem settings—AlphaGo could play Go but nothing more. However, the true beauty of human intelligence lies in our ability to rapidly master new behaviours. Humans can learn to drive, to play a new sport, or to code in a new programming language remarkably quickly, and Nigel Richards (a former Scrabble World Champion) famously also conquered the *French* Scrabble World Championship after studying French for only *nine* *weeks*. Our ability to rapidly adapt to new situations is in part thanks to our ability to *analogize*—to reuse abstract concepts across different domains [17]. Thus, to develop AI that excels at many problems and not just a few, we must adopt symbolic representations that are transferable across many problems, accordingly.

Thirdly, while language offers a good starting point for general-purpose reasoning, we must also consider abstractions beyond language, as it is not always the ideal substrate for reasoning. For instance, language models for chess often struggle to even find legal moves without access to a spatial representation [18]. Ideally, AI that rapidly adapts must develop and refine its own abstract representations that are conducive to reasoning.

Bibliography

(*) denotes equal contribution.

[1] McCarthy, J. (1959). Programs with Common Sense.

[2] Shortliffe, E. H., & Buchanan, B. G. (1975). A Model of Inexact Reasoning in Medicine. Mathematical biosciences, 23(3-4), 351-379.

[3] McDermott, J. P. (1980, August). RI: An Expert in the Computer Systems Domain. In AAAI (Vol. 1, pp. 269-271).

[4] Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep Blue. Artificial Intelligence, 134(1-2), 57-83.

[5] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature, 529(7587), 484-489.

[6] Jones, A. L. (2021). Scaling Scaling Laws with Board Games. arXiv preprint arXiv:2104.03113.

[7] Giunchiglia, F., & Walsh, T. (1992). A Theory of Abstraction. Artificial intelligence, 57(2-3), 323-389.

[8] Faisal, A., Kamruzzaman, M., Yigitcanlar, T., & Currie, G. (2019). Understanding Autonomous Vehicles. Journal of Transport and Land Use, 12(1), 45-72.

[9] Shvo, M., **Li, A. C.**, Icarte, R. T., & McIlraith, S. A. (2021, May). Interpretable Sequence Classification via Discrete Optimization. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 11, pp. 9647-9656).

[10] Christoffersen, P. J., **Li, A. C.**, Icarte, R. T., & McIlraith, S. A. (2023). Learning Symbolic Representations for Reinforcement Learning of non-Markovian Behavior. arXiv preprint arXiv:2301.02952.

[11] Wang, A.*, **Li, A. C.***, Klassen, T. Q., Icarte, R. T., & McIlraith, S. A. (2023). Learning Belief Representations for Partially Observable Deep RL. In International Conference on Machine Learning (pp. 35970-35988). PMLR.

[12] Vaezipoor, P.*, **Li, A. C.***, Icarte, R. A. T., & Mcilraith, S. A. (2021). LTL2Action: Generalizing LTL Instructions for Multi-Task RL. In International Conference on Machine Learning (pp. 10497-10508). PMLR.

[13] Tuli, M., **Li, A.**, Vaezipoor, P., Klassen, T., Sanner, S., & McIlraith, S. (2022). Learning to Follow Instructions in Text-Based Games. Advances in Neural Information Processing Systems, 35, 19441-19455.

[14] Vygotsky, L. S. (2012). Thought and Language. MIT press.

[15] Vogt, P. (2007). Language Evolution and Robotics: Issues on Symbol Grounding and Language Acquisition. In Artificial Cognition Systems (pp. 176-209). IGI Global.

[16] **Li, A. C.**, Chen, Z., Klassen, T. Q., Vaezipoor, P., Icarte, R. T., & McIlraith, S. A. (2024). Reward Machines for Deep RL in Noisy and Uncertain Environments. arXiv preprint arXiv:2406.00120.

[17] Shanahan, M., & Mitchell, M. (2022). Abstraction for Deep Reinforcement Learning. arXiv preprint arXiv:2202.05839.

[18] Toshniwal, S., Wiseman, S., Livescu, K., & Gimpel, K. (2022, June). Chess as a Testbed for Language Model State Tracking. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 10, pp. 11385-11393).