Bridging the Gap – A Grand Challenge for Explainable AI

Gregory Gelfond

Noeon Research
Midtown Tower 34F, 9-7-1 Akasaka
Minato-ku, Tokyo, Japan
gregory@noeon.ai

Abstract

The recent surge of activity spurred by the development of advanced GPU technologies and subsequent enabling of LLMs has brought AI to the fore in the popular and industry consciousness. In addition, it has brought to light some important shortfalls of the current subsymbolic approach to AI. We at Noeon Research believe that the next evolution of AI is dependent on solving the grand challenge of bridging together the symbolic and subsymbolic approaches to AI, especially to meet the needs of domains which involve a high degree of explainability and verifiability, such as the energy sector.

Shortly after the field's inception, two major paradigms emerged in artificial intelligence research: *Symbolic AI*, which focuses on rule-based reasoning and logic; and approaches rooted in numerical and statistical methods, often emphasizing pattern recognition and data-driven learning that we refer to as *subsymbolic* (Ilkou and Koutraki 2020; Platzer 2024). These paradigms diverged on their foundational assumptions regarding the nature of human reason. While there has been considerable ping-ponging in the scientific community's consciousness between them, the recent surge of activity spurred by the inreasing applicability of large language models LLMs to a broad range of tasks (Minaee et al. 2024) has deservedly brought considerable attention to the fruits of the latter's labor.

It can be argued, that the Symbolic AI school had it's origins with the work of McCarthy (McCarthy 1959), where he posits that "a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows." This idea, coupled with a strong intuition that humans reason by the application of internal rules, led to a hypothesis that these rules could *be discovered*, and *formalized in a mathematical logic*, and *realized by computable functions*. This started a line of inquiry that led to numerous discoveries, among which are: *the discovery of non-monotonic logics* and *equivalence classes* between them; *default reasoning* as a distinct natural kind; precise formulations of and approaches to modeling and solving various reasoning tasks (e.g., *temporal projection*, *planning*, *diagnostic reasoning*, and *counterfactual reasoning*); and even the development of *new programming languages and paradigms* (Lifschitz 2008) and their attendant algorithms and methodologies (Gelfond and Kahl 2014). These methodoligies have been further tested and refined in non-trivial settings such as diagnosing and remedying flaws in complex systems (such as the reaction control system in the NASA's space shuttle (Balduccini et al. 2001)), and finding finding explanations for program/theory behavior (Cabalar, Fandinno, and Muñiz 2020; Cabalar et al. 2021).

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The school of Subsymbolic AI embodies a different set of core intuitions, also based on an observation of that certain forms of human reason operate on the recognition of patterns and their application towards *discrimination* (i.e., classification), and *continuation* (i.e., generation). This led to a host of innovations in the form of curve-fitting algorithms and the discovery and implementation of probabilistic reasoning methods. Subesequent advancements in data-driven algorithms and the growth of both data and computation resources led to the application of LLMs in problems ranging from from code generation (Jiang et al. 2024) to certain kinds of mathematical reasoning (Ahn et al. 2024). One such area that is of increasing importance given not only the global state of affairs, but the rapid development of new technologies, is the *energy sector*.

One of the things that makes the energy sector a challenging application space for the current crop of LLMs and their technological siblings and cousins, is the need for a combination of *correctness*, *precision*, *replicability*, and *explainibility*. Added into this mix is the need for a high degree of trust as biases or inaccuracies in the training data have the potential to lead to catastrophic results given the nature of the domain itself. Not only is it necessary to *reliably diagnose* (i.e., replicably diagnose) potential root causes for aberrant symptoms, but also to find actionable and effective *plans* for their resolution/remedy. Furthermore, the causes of the symptoms must be *explicable to human operators*. While LLMs and their cousins have had prominent success with regards to finding diagnoses based on even potentially unobservable patterns of certain phenomena, their probabilistic nature draws into question the replicability of these diagnoses (particularly in light of the propensity of LLMs to hallucinate (Zhang et al. 2023)). The opaque nature of the LLMs has also made the interpretation and explication of their behavior troublesome despite recent efforts (Bereska and Gavves 2024; Singh

et al. 2024). LLMs have also recently undergone considerable scrutiny with regards to their failure to solve planning problems (Kambhampati et al. 2024; Valmeekam, Stechly, and Kambhampati 2024) which have long been considered toy domains in the symbolic world. This again, presents a stumbling block when entering this particular arena. Lastly, it should also be mentioned that the training of LLMs is also a costly endeavor, not only when it comes to monetary cost, but also in terms of energy output (Samsi et al. 2023). This can pose a problem from the perspective of *sustainability*, but also can make the training set costly given the nature of the domain (e.g., meaningful training to find and diagnose failure within the energy grid will by necessity take vast amounts of data, which may not be trivial to obtain a priori).

At Noeon Research, we believe that the task of deploying AI into the energy sector (as well as other problem domains) presents a vivid example of the need for the next evolution of AI to *merge* the strengths of both the Symbolic and Subsymbolic AI communities. While some work has been done in this regard, partly in the form of bringing together answer-set programming and causal bayesian networks (Baral, Gelfond, and Rushton 2009; De Raedt, Kimmig, and Toivonen 2007) or LLM-modulo frameworks (Kambhampati et al. 2024), we at Noeon believe in developing a *common computational substrate* which enables to us to approach the task in a novel way, in which we treat the task of KR/domain modeling in the Symbolic AI sense, and parameterization and training in the Numeric AI sense as special cases of a broader *knowledge representation* problem which can be adquately described in our framework. While believe that the new approach we are purusing at Noeon is potentially revolutionary, we believe that this attempt is the next *grand challenge* facing the AI community writ large, and that the energy sector provides a rich class of problem domains for us to being exploring together.

References

Ahn, J.; Verma, R.; Lou, R.; Liu, D.; Zhang, R.; and Yin, W. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Balduccini, M.; Gelfond, M.; Watson, R.; and Nogueira, M. 2001. The USA-Advisor: A Case Study in Answer Set Planning. In Eiter, T.; Faber, W.; and Truszczyński, M. l., eds., *Logic Programming and Nonmotonic Reasoning*, 439–442. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-45402-1.

Baral, C.; Gelfond, M.; and Rushton, N. 2009. Probabilistic reasoning with answer sets. *Theory and Practice of Logic Programming*, 9(1): 57–144.

Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety–A Review. *arXiv preprint arXiv:2404.14082*.

Cabalar, P.; Fandinno, J.; and Muñiz, B. 2020. A System for Explainable Answer Set Programming. *Electronic Proceedings in Theoretical Computer Science*, 325: 124—-136.

Cabalar, P.; Muñiz, B.; Pérez, G.; and Suárez, F. 2021. Explainable Machine Larning for liver transplantation. *CoRR*, abs/2109.13893.

De Raedt, L.; Kimmig, A.; and Toivonen, H. 2007. ProbLog: a probabilistic prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, 2468–2473. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Gelfond, M.; and Kahl, Y. 2014. *Knowledge Representation, Reasoning, and the Design of Intelligent Agents: The Answer-Set Programming Approach*. Cambridge University Press.

Ilkou, E.; and Koutraki, M. 2020. Symbolic vs sub-symbolic ai methods: Friends or enemies? In *CIKM (Workshops)*, volume 2699.

Jiang, J.; Wang, F.; Shen, J.; Kim, S.; and Kim, S. 2024. A Survey on Large Language Models for Code Generation. *arXiv preprint arXiv:2406.00515*.

Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L.; and Murthy, A. 2024. LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. arXiv:2402.01817.

Lifschitz, V. 2008. What is answer set programming? In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, 1594–1597. AAAI Press. ISBN 9781577353683.

McCarthy, J. 1959. Programs with Common Sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 75–91. London: Her Majesty's Stationary Office.

Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; and Gao, J. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Platzer, A. 2024. Intersymbolic AI: Interlinking symbolic AI and subsymbolic AI. In *International Symposium on Leveraging Applications of Formal Methods*, 162–180. Springer.

Samsi, S.; Zhao, D.; McDonald, J.; Li, B.; Michaleas, A.; Jones, M.; Bergeron, W.; Kepner, J.; Tiwari, D.; and Gadepally, V. 2023. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–9. IEEE.

Singh, C.; Inala, J. P.; Galley, M.; Caruana, R.; and Gao, J. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Valmeekam, K.; Stechly, K.; and Kambhampati, S. 2024. LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench. arXiv:2409.13373.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.