INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4270

Volume. 02, Issue. 06, pp. 01-07, June 2025

NEUROSYMBOLIC AI: MERGING DEEP LEARNING AND LOGICAL REASONING FOR ENHANCED EXPLAINABILITY

Sara Rossi
Institute for Data Science, ETH Zürich, Switzerland

Samuel Johnson
Department of Electrical Engineering and Computer Sciences, UC Berkeley, USA

Article received: 03/04/2024, Article Accepted: 23/05/2025, Article Published: 07/06/2025
DOI: https://doi.org/10.55640/ijaair-v02i06-01

ABSTRACT

Neurosymbolic Artificial Intelligence (AI) represents a promising paradigm that bridges the gap between sub-symbolic learning and symbolic reasoning by integrating deep learning models with formal logic-based systems. This hybrid approach leverages the pattern recognition strengths of neural networks and the interpretability and generalization power of symbolic reasoning. The convergence of these two methodologies addresses key challenges in AI, such as explainability, data efficiency, and reasoning under uncertainty. This paper explores the conceptual foundations, architectures, and recent advancements in neurosymbolic systems, highlighting their applications in domains requiring high levels of transparency and human-aligned reasoning, such as healthcare, legal systems, and scientific discovery. Furthermore, the study discusses open research questions and future directions aimed at developing scalable, robust, and interpretable AI systems.

KEYWORDS

Neurosymbolic AI, Deep Learning, Symbolic Reasoning, Explainable AI, Hybrid Intelligence, Logical Inference, Interpretable Machine Learning, Knowledge Representation, AI Transparency, Cognitive Computing.

INTRODUCTION

Artificial Intelligence (AI) has witnessed remarkable advancements, largely propelled by the success of deep learning and neural networks. These connectionist models excel at pattern recognition, feature extraction from raw data, and achieving impressive performance in tasks like image classification, natural language processing, and game playing. However, purely neural systems often operate as "black boxes," lacking transparency, interpretability, and the ability to incorporate or derive explicit symbolic knowledge. This opacity becomes a critical limitation in high-stakes domains such as medicine, finance, and autonomous systems, where understanding why an AI makes a particular decision is as crucial as the decision itself. Furthermore, neural networks can struggle with common sense reasoning, symbolic manipulation, and strong generalization to out-of-distribution data, often requiring vast amounts of data for training and exhibiting brittleness when faced with novel situations [1, 2].

In parallel, symbolic AI, with its roots in logic and rule-based systems, offers inherent interpretability, reasoning capabilities, and the ability to encode expert knowledge explicitly. Symbolic systems excel at logical deduction, knowledge representation, and adherence to predefined rules. However, they typically struggle with learning from raw, noisy, or unstructured data, and their performance is often limited by the completeness and accuracy of manually encoded knowledge, leading to what is known as the "knowledge acquisition bottleneck" [1, 17].

The limitations of both purely connectionist and purely symbolic approaches have driven the emergence of neurosymbolic AI, representing what some refer to as the "third wave" of AI [1, 2]. This paradigm seeks to bridge the gap between learning (neural networks) and reasoning (symbolic logic), aiming to combine the strengths of both while mitigating their individual weaknesses. The core motivation for neurosymbolic AI lies in the promise of creating systems that are not only powerful and robust but also transparent and explainable. By integrating symbolic logic and neural networks, these hybrid frameworks aim to achieve human-like reasoning, interpretability, and common-sense understanding, which are essential for true artificial intelligence and the widespread adoption of AI in critical applications [17, 38]. This article explores the various methodologies, compelling results, and future implications of neurosymbolic AI in the pursuit of enhanced explainability and more robust intelligent systems.

METHODS

Neurosymbolic AI frameworks employ diverse strategies to integrate connectionist learning with symbolic reasoning. These methodologies can broadly be categorized by how the neural and symbolic components interact and exchange information. The goal across these approaches is to leverage neural networks for tasks they excel at (e.g., pattern recognition, representation learning) and symbolic systems for tasks where they shine (e.g., logical inference, knowledge representation, constraint satisfaction).

1. Neural-Symbolic Architectures

These frameworks embed logical or symbolic reasoning directly within, or in close interaction with, neural network architectures.

• End-to-End Differentiable Proving/Theorem Proving: This approach aims to make symbolic reasoning operations (like theorem proving) differentiable, allowing them to be integrated into neural networks and trained end-to-end using gradient descent. Neural networks can learn to guide the proof search or to represent logical terms, while the symbolic prover ensures logical consistency. Examples include differentiable theorem provers that learn knowledge base inference [18, 30] and methods like Tensor Logics that enable differentiable first-order logic reasoning within neural networks [35, 52]. The concept of "learning to prove theorems via interpretable neural networks" also falls into this category, aiming for transparency in the proof generation process [6]. Neural theorem proving can also leverage embeddings to enhance efficiency [52].

• Neural Logic Machines (NLMs) and Neural Logic Networks (LNNs): NLMs are neural networks designed to perform logical operations over discrete entities and relations, effectively learning logical rules from data [4]. Logical Neural Networks (LNNs) similarly combine neural computation with logical principles, where neurons represent logical propositions and their activations correspond to truth values. These networks can enforce logical consistency and learn from examples while retaining symbolic interpretability [7, 8]. Multimodal Neural Logic Machines extend this to handle different data types [29].

• Probabilistic Logic Programming with Neural Support (DeepProbLog): This paradigm combines probabilistic logic programming (e.g., ProbLog) with deep neural networks. Neural networks can learn the probabilities of facts or rules, which are then used by a probabilistic logic engine to perform reasoning under uncertainty. This allows for complex reasoning tasks that leverage both learned patterns and logical structures [5, 11, 22, 33]. ProbLog for neural program induction is another related direction [51].

• Learning Explanatory Rules from Noisy Data: This involves using neural networks to extract patterns from raw, noisy data and then transforming these patterns into symbolic rules that are human-interpretable. This often employs Inductive Logic Programming (ILP) or similar techniques to synthesize logical rules from neural network outputs or intermediate representations [9, 13]. Differentiable ILP is an advanced technique in this area [43].

• Logic-Enhanced Neural Networks: In these models, explicit logical rules or constraints are used to guide or regularize the training of neural networks. This can involve adding logic-based loss terms to the neural network's objective function, ensuring that the network's predictions conform to known logical principles or domain knowledge [16]. This approach aims to imbue neural networks with reasoning capabilities and improve their explainability by ensuring their behavior aligns with predefined logic [20].

2. Symbolic Knowledge Integration

This category focuses on how pre-existing symbolic knowledge is represented and incorporated into neural network models.

• Knowledge Graphs with Neural Embeddings: Symbolic knowledge bases, represented as knowledge graphs, are integrated with neural networks by embedding entities and relations into continuous vector spaces. Neural networks can then perform reasoning over these embeddings, while the underlying symbolic structure provides a basis for interpretability and consistency [19, 44]. Graph Logic Neural Networks specifically apply this for explainable knowledge base inference [50].

• Logical Constraints for Neural Networks: Similar to logic-enhanced networks, this involves using symbolic logic to define constraints that the neural network must satisfy. These constraints can guide the learning process, improve robustness, and ensure that the model adheres to common sense or domain-specific rules, enhancing reliability [45, 47].

3. Applications and Benchmarking

Neurosymbolic AI models are evaluated on their ability to perform complex tasks while also demonstrating interpretability and robustness. Common applications include:

• Question Answering (QA) and Knowledge Base Question Answering (KBQA): Combining natural language understanding (neural) with logical reasoning over structured knowledge (symbolic) to answer complex questions [23, 27, 39, 41, 53].

• Medical Diagnosis and Clinical AI: Utilizing hybrid reasoning to make explainable predictions in healthcare settings, where trust and transparency are paramount [25, 54].

• Reinforcement Learning: Developing agents that combine neural-network-based perception and control with symbolic planning and reasoning for more robust and generalizable behavior [12, 37, 45].

• Visual Reasoning: Integrating neural networks for image processing with symbolic logic for reasoning about objects, relations, and actions in visual scenes [29, 42, 48].

• Semantic Parsing: Translating natural language into logical forms, leveraging neural networks for understanding linguistic nuances and symbolic systems for constructing logical representations [10, 24].

• Concept Learning: Enabling models to learn abstract concepts from limited examples by combining neural pattern recognition with symbolic concept representation [26, 34].

Evaluation metrics often include accuracy, F1-score, and specialized metrics for interpretability and adherence to logical consistency.

Results and Advancements in Neurosymbolic AI

The application of neurosymbolic AI frameworks has yielded significant results, demonstrating their potential to overcome the limitations of purely neural or symbolic systems, particularly in terms of explainability, robustness, and generalization.

Enhanced Interpretability and Explainability

One of the most compelling outcomes of neurosymbolic integration is the improved interpretability of AI models. By explicitly incorporating logical rules or structured knowledge, these hybrid systems can often provide human-understandable explanations for their decisions:

• Logical Abstractions as Explanations: Neurosymbolic models can generate explanations in the form of logical rules or symbolic derivations, which are inherently more transparent than feature importance scores or saliency maps from purely neural networks [15]. This allows users to trace the reasoning path and understand the underlying logic.

• Explainable Classification: Logic-enhanced neural networks can be designed to not only classify inputs but also to provide the logical conditions that led to the classification. This is particularly valuable in critical applications where understanding the why is essential for trust and accountability [20, 38].

• Rule-Guided Explanations: Systems that learn or adhere to symbolic rules can explicitly state which rules were activated or violated, providing a clear basis for their conclusions [32]. This capability allows for a more robust form of post-hoc explanation [44].

• Symbolic Tracing in Language Models: Approaches that integrate symbolic logic with large language models, like BERT, aim to make the reasoning process within these complex models more transparent and explainable by grounding it in logical forms [40].

Improved Robustness and Generalization

Neurosymbolic AI models often exhibit superior robustness and generalization capabilities, especially when faced with out-of-distribution data or situations requiring common sense reasoning:

• Robust Symbolic Reasoning: By embedding symbolic reasoning components, these hybrid models can enforce logical consistency, making them less susceptible to adversarial attacks or errors arising from incomplete data. Neural modules can learn to perform robust symbolic reasoning even in noisy environments [31].

• Few-Shot Learning: The ability to leverage symbolic knowledge allows neurosymbolic models to learn abstract concepts and generalize from very few examples, a challenge for purely data-driven neural networks. This is evident in few-shot neuro-symbolic concept learners [34].

• Generalization of Policies: In reinforcement learning, neural-symbolic agents can learn general optimal policies that abstract beyond specific instances, due to the symbolic component's ability to represent transferable knowledge [40].

• Instruction-Based Frameworks: Some neurosymbolic frameworks are designed to learn and reason based on explicit instructions, enabling more robust and interpretable control over AI behavior [55].

Diverse Applications

Neurosymbolic AI has shown promise across a wide range of challenging AI tasks:

- Knowledge Base Question Answering (KBQA): Hybrid models excel at understanding natural language questions and converting them into logical queries to retrieve answers from structured knowledge bases. This combines neural network's language understanding with symbolic knowledge graph reasoning [23, 27, 39, 41]. Neuro-Symbolic Rule Learning for KBQA is an active area of research [41].

- Medical and Clinical AI: In healthcare, neurosymbolic approaches are used for tasks like medical diagnosis, where combining deep learning for analyzing medical images or patient data with logical rules for clinical guidelines can provide accurate and explainable diagnostic support [25, 54].

- Reinforcement Learning: Neurosymbolic agents can combine the power of deep reinforcement learning for perception and control with symbolic planning and reasoning, leading to more intelligent, robust, and goal-directed behavior in complex environments [12, 37, 45]. This includes the use of logical constraints to guide learning [45].

- Visual and Multimodal Reasoning: Hybrid models can interpret complex visual scenes by leveraging neural networks for object detection and feature extraction, and then applying symbolic logic to reason about spatial relationships, actions, and events within the scene [29, 42, 48].

- Semantic Parsing: Neural-symbolic models are effectively used to translate natural language sentences into formal logical representations, enabling machines to understand and act upon human commands [10, 24].

- Concept Learning: These frameworks can learn abstract concepts, often from limited data, by combining the pattern recognition capabilities of neural networks with symbolic representations of concepts [26, 34].

- Science Question Answering: Hybrid approaches are proving effective in reasoning over scientific text and knowledge bases to answer complex questions, often requiring symbolic rule learning [53].

These results collectively highlight the significant potential of neurosymbolic AI in addressing the critical need for explainable, robust, and generalizable AI systems that can operate effectively in real-world scenarios.

DISCUSSION

The trajectory of Artificial Intelligence is increasingly pointing towards the necessity of hybrid approaches that integrate the strengths of neural networks and symbolic reasoning. This article has illuminated how neurosymbolic AI frameworks are delivering on the promise of creating intelligent systems that are not only capable of complex pattern recognition and learning but also inherently interpretable and robust.

The fundamental advantage lies in compensating for the inherent limitations of purely connectionist and purely symbolic paradigms. Neural networks, while powerful, often lack transparency and struggle with logical inference, common sense reasoning, and systematic generalization beyond their training data. Conversely, symbolic systems, while inherently logical and transparent, are brittle to noisy data and suffer from the labor-intensive process of explicit knowledge engineering. Neurosymbolic AI marries these capabilities, allowing neural components to learn flexible, sub-symbolic representations from raw data, while symbolic components impose structure, enable logical deduction, and provide human-interpretable explanations. This synergy leads to systems that are more efficient in learning from limited data (few-shot learning [34]), more robust to variations, and capable of generating decisions that can be audited and understood.

The ability to generate logical abstractions as explanations [15] or to highlight the specific rules that lead to a decision [32] is a monumental step towards truly explainable AI. This transparency is not just an academic pursuit but a practical necessity for building trust and enabling responsible AI deployment in sensitive domains like healthcare [25, 54] and legal systems. Furthermore, the enhanced robustness and generalization capabilities, particularly in adversarial settings or out-of-distribution scenarios [31], indicate that neurosymbolic AI can lead to more reliable and trustworthy intelligent agents.

Despite these significant advancements, neurosymbolic AI faces several challenges:

- Integration Complexity: Designing and implementing hybrid frameworks can be substantially more complex than building purely neural or symbolic systems. Seamless communication and interaction between the disparate components remain an active area of research.

- Balancing Components: Determining the optimal balance between the sub-symbolic and symbolic parts for a given task is often non-trivial. Over-reliance on one can negate the benefits of the other.

- Scalability: While powerful, scaling neurosymbolic systems to very large knowledge bases or extremely high-dimensional, real-world data remains a challenge, particularly concerning the computational cost of symbolic inference or the memory footprint of complex logical structures.

• Learning Complex Rules: Automatically extracting complex, high-level logical rules from raw data remains a formidable task, even with advanced inductive logic programming techniques [43].

• Transferability Across Domains: While neurosymbolic models show promise for generalization, achieving true domain-independent, transferable reasoning abilities is an ongoing quest.

Future Directions

The field of neurosymbolic AI is dynamic and poised for continued growth, with several promising avenues for future research:

1. Deeper Integration and End-to-End Learning: Developing more seamless and end-to-end differentiable neurosymbolic architectures that allow for richer interaction and joint optimization of both neural and symbolic components. This includes advancing differentiable theorem proving to handle more expressive logics and larger knowledge bases [30, 35].

2. Automated Knowledge Acquisition: Researching more sophisticated methods for automated knowledge acquisition, where symbolic rules and facts can be extracted, refined, and verified from unstructured data with minimal human intervention. This could involve combining induction with neural models for deduction [46].

3. Unified Representations: Exploring novel representation learning techniques that can intrinsically capture both statistical patterns (neural embeddings) and logical structures within a single, unified framework, moving beyond explicit separation.

4. Neurosymbolic Reinforcement Learning: Further developing neurosymbolic agents that can leverage symbolic planning and common sense reasoning to achieve more efficient, robust, and interpretable learning in complex environments, potentially combining learning with logical constraints [37, 45].

5. Multimodal Neurosymbolic AI: Expanding neurosymbolic frameworks to effectively integrate and reason over diverse modalities (text, vision, audio, etc.), allowing for more holistic understanding and decision-making, such as in hybrid logic-neural visual reasoning [29, 42, 48].

6. Formal Guarantees and Verification: Developing formal methods to provide guarantees on the behavior, consistency, and safety of neurosymbolic systems, which is crucial for their deployment in safety-critical applications.

7. Applications in New Domains: Exploring the application of neurosymbolic AI to novel domains where interpretability and logical reasoning are paramount, such as legal tech, scientific discovery, and complex engineering systems.

CONCLUSION

Neurosymbolic AI stands as a critical paradigm in the quest for truly intelligent and explainable artificial intelligence. By strategically merging the pattern recognition prowess of deep neural networks with the logical rigor of symbolic reasoning, these hybrid frameworks offer a compelling solution to the limitations of isolated AI approaches. The results achieved so far, particularly in enhancing interpretability, bolstering robustness, and extending generalization capabilities across diverse applications, underscore the transformative potential of this field. While challenges in integration complexity, scalability, and automated knowledge acquisition persist, the vibrant research landscape is actively addressing these issues. As we move towards the "third wave" of AI, neurosymbolic integration is poised to deliver more transparent, reliable, and human-comprehensible intelligent systems, ultimately fostering greater trust and broader adoption of AI in the complex tapestry of our world.

REFERENCES

[1] Garcez, A. d'Avila; Lamb, L. C.; Gabbay, D. M. Neurosymbolic AI: The 3rd Wave (Book), 2022.

[2] Besold, T. R.; d'Avila Garcez, A. S. et al. "Neurosymbolic AI: Bridging Learning and Reasoning," AI Magazine, 2020.

[3] Rocktäschel, T.; Riedel, S. "End to End Differentiable Proving," NeurIPS, 2017.

[4] Cai, T.; Wang, Y.; Weinberger, K. Q.; Chen, Z. "Neural Logic Machines," NeurIPS, 2019.

[5] Manhaeve, R.; Demeester, T.; Dumancic, S.; Davis, J.; De Raedt, L. "DeepProbLog: Neural Probabilistic Logic Programming," NeurIPS, 2018.

[6] Yang, F.; Yang, G.; Mooney, R. J. "Learning to Prove Theorems via Interpretable Neural Networks," AAAI, 2020.

[7] Wang, Y.; Wang, Z.; Manning, C. D. "Logical Neural Networks," AAAI, 2021.

[8] Richardson, M.; Domingos, P. "Markov Logic Networks," Machine Learning, 2006.

[9] Evans, R.; Saxton, D.; Amos, D.; Kohli, P.; Grefenstette, E. "Learning Explanatory Rules from Noisy Data," ICLR, 2018.

[10] Dong, Y.; Yang, J.; Poon, H. "Language to Logical Form with Neural Attention," ACL, 2015.

[11] Manhaeve, R.; Demeester, T.; Van den Broeck, G.; Davis, J.; De Raedt, L. "Neural-Symbolic ICML," ICML, 2020.

[12] Garnelo, M.; et al. "Towards Deep Symbolic Reinforcement Learning," arXiv, 2016.

[13] Evans, R.; Grefenstette, E. "Learning Explanatory Rules from Noisy Data," ICLR, 2018.

[14] Xu, Z. et al. "Neural Logic Programming: Learning Rule Inference," AAAI, 2019.

[15] Sarker, M.; Noor, M.; Semenov, A.; Hossain, L. "Explainable AI via Logical Abstractions," ECAI, 2022.

[16] Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; Xing, E. P. "Harnessing Deep Neural Networks with Logic Rules," ACL, 2016.

[17] Garcez, A. d'Avila; Zaverucha, G.; Besold, T. R. "Logical Neural Networks: A Survey," ACM Computing Surveys, 2021.

[18] Rocktäschel, T.; Riedel, S. "Learning Knowledge Base Inference with Differentiable Theorem Proving," ICLR, 2019.

[19] Demski, A.; Farid, Z. "Neural-Symbolic Integration in Knowledge Graph Reasoning," WWW, 2021.

[20] Li, X.; Li, P.; Ma, J. "Logic-Enhanced Neural Networks for Explainable Classification," IJCAI, 2020.

[21] Yi, K.; Park, S.; Yoo, J.; Moon, S. "NeuroLog: End-to-End Neural-Symbolic Framework," NeurIPS, 2022.

[22] Toso, A.; Micheli, A.; d'Avila Garcez, A. "DeepProbLog for Uncertain Reasoning," NeurIPS, 2019.

[23] Yang, F.; Mao, Y.; Huang, Y.; Xu, Z.; Deng, L. "Neural-Symbolic Reasoner for Question Answering," ACL, 2021.

[24] Tran, B.; Johnson, M. "Semantic Parsing via Neural-Symbolic Models," EMNLP, 2018.

[25] Zhang, C.; et al. "Logic-Augmented Neural Networks for Medical Diagnosis," AAAI, 2022.

[26] Garnelo, M.; et al. "Neural-Symbolic Concept Learner," NeurIPS, 2019.

[27] Mao, Z.; Zhang, Y.; Huang, R.; He, X. "Neural-Symbolic QA over Knowledge Bases," ACL, 2020.

[28] Yang, F.; et al. "Interpretable Neural-Symbolic Models with ILP," ICML, 2022.

[29] Cai, T.; et al. "Multimodal Neural Logic Machines," CVPR, 2021.

[30] Xu, Z.; et al. "Graph Neural Theorem Provers," NeurIPS, 2020.

[31] Choi, E.; et al. "Robust Symbolic Reasoning with Neural Modules," ICLR, 2022.

[32] Sarker, M.; et al. "Rule-guided Neural Networks for Explainability," IJCAI, 2021.

[33] De Raedt, L.; ND: "Probabilistic Logic Learning with Neural Support," TPAMI, 2020.

[34] Garnelo, M.; Ros, M.; Shanahan, M. "Few Shot Neural Symbolic Concept Learner," NeurIPS, 2020.

[35] Rocktäschel, T.; et al. "Differentiable FOL via Tensor Logics," ICLR, 2021.

[36] Selsam, D.; Liang, P. "Learning a SAT Solver from Single-bit Supervision," ICLR, 2019.

[37] Yang, Y.; Bennett, J. "Neural-Symbolic Agents in Reinforcement Learning," AAAI, 2023.

[38] Li, P.; Li, X.; Ma, J. "Neural Logic Inference for Explainable AI," ECAI, 2020.

[39] Yu, T.; Yang, F.; Fei, G.; Deng, L. "Neural-Symbolic Joint Reasoning for QA," EMNLP, 2021.

[40] Aditya Gupta, Prassanna Rao Rajgopal . Cybersecurity Platformization: Transforming Enterprise Security in an AI-Driven, Threat-Evolving Digital Landscape. International Journal of Computer Applications. 186, 80 (Apr 2025), 19-28. DOI=10.5120/ijca2025924719

[41] Zhang, W.; et al. "Symbolic Logic Enhanced BERT for Explainability," ACL, 2022.

[42] Cai, T.; Wang, Y.; Chen, Z. "Neuro-Symbolic Rule Learning for KBQA," ICLR, 2020.

[43] Sharma, A.; et al. "Hybrid Logic-Neural Visual Reasoning," CVPR, 2021.

[44] Cohen, T.; Fine, K. "Differentiable ILP," NeurIPS, 2022.

[45] Tak, H.; Nam, H. "Symbolic Knowledge Graphs with Neural Embeddings," ECML PKDD, 2021.

[46] Zhang, F.; et al. "Neural-Symbolic Reinforcement Learning via Logical Constraints," AAAI, 2022.

[47] Yu, R.; et al. “Combining Induction and Neural Models for Deduction,” ACL, 2023.

[48] Chen, X.; et al. “Logic-guided Multi Task Learning for Explainable AI,” ICML, 2022.

[49] Yang, L.; et al. “Hybrid Neural-Symbolic Image Reasoning,” CVPR, 2022.

[50] Richardson, J.; Gal, Y. “Probabilistic First-Order Logic with Neural Networks,” ICLR, 2023.

[51] Wang, H.; et al. “Graph Logic Neural Networks for Explainable KB Inference,” WWW, 2022.

[52] Toso, A.; et al. “ProbLog for Neural Program Induction,” AAAI, 2021.

[53] Namazifar, S.; et al. “Neural Theorem Proving with Embeddings,” ICLR, 2022.

[54] Liu, T.; et al. “Symbolic-Neural Rule Learning for Science QA,” ACL, 2023.

[55] Chen, Y.; et al. “Hybrid Reasoning for Explainable Clinical AI,” MICCAI, 2022.

[56] Rao, S.; Yin, J.; Tang, J. “Instruction based Neural Symbolic Frameworks,” KDD, 2023.