The Future Is Neuro-Symbolic: Where Has It Been, and Where Is It Going?

Vaishak Belle[1], **Gary Marcus**[2]

[1]University of Edinburgh
vaishak@ed.ac.uk
[2]New York University (Emeritus)

Abstract

This report explores the evolution and current state of neuro-symbolic artificial intelligence, an approach that integrates neural network capabilities with symbolic reasoning. We trace the historical context from early AI aspirations to modern implementations and successes, highlighting key paradigms, and other logical and semantical considerations. We argue against the "scaling is all you need" hypothesis, and point to persistent challenges in reliable symbolic reasoning with deep and large models. We conclude by suggesting that despite numerous implementation choices and the "broad church" nature of neuro-symbolic AI, these approaches offer the most promising path towards AI systems that combine pattern recognition with robust reasoning, particularly for applications requiring structured knowledge, explainability, and trustworthiness.

Introduction

In its original inception, the field of artificial intelligence (AI) was deeply concerned with how human cognition could be automated on a computer (Turing 1950). John McCarthy, for example, argued for the development of so-called *common-sense programs* that could reason and problem solve, using the mathematical apparatus of logic for formalising the application domain. The idea was arguably radical at that time (although Leibniz had already wondered about thinking as an algebraic process (Levesque 2012)).

Despite continuing extensions of logical formalisms to deal with actions, plans and agents (Levesque 1996; Kelly and Pearce 2008), three representational aspects severely challenge this program: (a) the lack of *complete* knowledge (that is, an exhaustive formalisation) in almost all applications, (b) the difficulty in asserting that logical assertions are *categorically* true, and (c) the need to leverage sensorimotor data and relevant statistical patterns from that data. Sensorimotor data, including visual and auditory information, moreover, is sampled from high-dimensional spaces, which lead to the problem of *scalability* with purely logical frameworks. For example, a 16×16 black and white image would require, say, a data structure with $2^{16\times 16}$ propositions to capture the pixel values, and for a dataset of 100,000 images would require as many instances of such data structures.

To deal with possibly false facts, and uncertainty in general, *probabilistic logics* emerged (Bacchus 1990; Halpern 2003), that allowed for a mixture of probabilistic and logical assertions. However, because they inherited the expressive power of (often first-order) logic but then further extended it for probabilistic knowledge, they too suffered from scalability issues. And they largely glossed over the point of where the probabilities came from. Presumably, these need to be drawn from data, but integrating the learning of statistical information and ensuring its consistency with prior logical knowledge is a non-trivial matter (Valiant 1999).

When limited to graphs, however, Bayesian (and later causal) networks offered a reasonable compromise between expert knowledge together with probabilistic and statistical information (Pearl 1998), and were amenable to certain types of learning from data (Koller and Friedman 2009). Building on this success, the field of statistical relational learning (SRL) aimed to unify logical and probabilistic frameworks by controlling the expressiveness (Raedt et al. 2016), such as by investigating finite-domain relational extensions to Bayesian networks (Koller and Pfeffer 1997; Getoor et al. 2001). Be that as it may, a purely expert-driven paradigm for dealing with high-dimensional data is unlikely to succeed – unless they outsource data-heavy computation to neural networks.

From Neural to Neuro-Symbolic

In the areas of vision and speech, in the last two decades, neural network-based learning began demonstrated outstanding performance in pattern recognition across computer vision (Krizhevsky, Sutskever, and Hinton 2012), natural language processing, and recommendation systems. This is owing to a cascade of improvements. Classically, neural networks were built consisting of input, hidden, and output layers with limited depth, but then "deep" learning architectures introduced multiple hidden layers, likely enabling the learning of deeper hierarchical representations (although these internal constructions are largely opaque to humans). There were also:

1. **Architectural improvements:** The development of convolutional neural networks (CNNs) for computer vision, recurrent neural networks (RNNs) (Goodfellow et al. 2016) and transformers (Vaswani et al. 2017) for sequential data, and graph neural networks for structured data

(Barceló et al. 2020) dramatically expanded the representational capabilities of neural approaches.

2. **Training regimes:** Although not always well understood, techniques such as dropout and batch normalisation (Goodfellow et al. 2016) targeted issues in computing gradients.
3. **Computational advances:** The development of GPU computing provided the necessary computational power to train complex models with thousands of hidden layers on massive datasets.
4. **Data availability:** Partly driven by the ability to now harness such datasets, continued efforts in data collection, including from the internet, provided the training resources needed for such models to learn.

Despite these strides, purely neural models still have important, and perhaps fundamentally irreparable, limitations. On the one hand, these models require large amounts of data to achieve robust predictions, making them unsuitable for many applications where extensive datasets are simply not available. But on the other, does scaling these models to train over larger and larger datasets lead to general-purpose intelligence (Chollet 2019)? That is, would it "hit a wall" in terms of its abilities (Marcus 2018, 2022; Chollet 2017)?

From a technical viewpoint, purely neural models have been noted to exhibit several critical limitations (Marcus 2018; Besold et al. 2021):

1. **Structured reasoning:** Neural networks are powerful at pattern recognition but struggle with hierarchical or composite reasoning (Lake and Baroni 2018) and cannot effectively differentiate between causality and correlation. The community has often addressed that by explicit programmatic bias, which is essentially a neuro-symbolic endeavor (Ellis et al. 2022; Lake, Salakhutdinov, and Tenenbaum 2015), as we discuss below.
2. **Data requirements:** As mentioned above, neural networks demand large amounts of data to achieve robust predictions, making them unsuitable for many applications where extensive datasets are simply not available (Marcus 2018), and where there is existing expert knowledge that is not easily reflected in the collected data.
3. **Knowledge integration:** On the latter point, neural networks do not easily support the integration of expert or common sense knowledge (Maldonado et al. 2018).
4. **Explainability:** Neural networks function as black-box systems, making it difficult to understand how they reach specific predictions for given inputs (Rudin 2019).
5. **Guarantees:** Neural networks compute probability distributions over possible outcomes, potentially predicting outcomes that violate constraints, a serious concern in safety-critical applications (Gajowniczek et al. 2020; Innes and Ramamoorthy 2020).
6. **Distribution drift:** Neural networks, and almost all other statistical models that predict by learning a distribution for the training data, struggle in environments that differ from their training data (Marcus 1998). This is because they do not include any inherent mechanisms to deal with worlds distinct from the controlled settings where the data was collected in the first place. All of this leads to degraded performance or complete failure, some catastrophic. One may resort to (often expensive) continuous retraining or adaptation mechanisms, but a general solution is lacking (Lu et al. 2018; Mallick et al. 2022).

The development of neuro-symbolic AI can be understood as a response to the limitations of purely neural approaches (Sun 2002; Garcez and Lamb 2023; Marcus 2001). Interestingly, to a large extent, it can remain agnostic about how deep learning may mature in the future. One of the motivations of neuro-symbolic AI is to also better understand how symbols emerge (Garcez and Lamb 2023), which, yet again, fits nicely with the deep learning paradigm where the hypothesis is entirely implicit, informed by the hyperparameters. Such symbols would then empower the reasoning processes of the system, and possibly aid in transparency, guarantees, correctness and beyond.

From a practical viewpoint, neuro-symbolic AI seeks to combine the pattern recognition capabilities of neural networks with the explicit reasoning mechanisms of symbolic systems. In this sense, neuro-symbolic AI is not but a single approach, but a family of frameworks that consider the coupling of symbols and reasoning, on the one hand, and connectionism, probabilities and learning, on the other (Hitzler and Sarker 2022; Sarker et al. 2021). From a logical point of view, there is further debate on whether a fuzzy semantics for the representations is better suited (van Krieken, Acar, and van Harmelen 2022) or a probabilistic (De Smet and De Raedt 2025) one, the latter closely borrowing and extending earlier work on statistical relational learning (De Raedt et al. 2020).

As noted in (Garcez and Lamb 2023), precisely because of this broad remit, it relates to debates and discussions on how to integrate algebraic processing with neural networks (Marcus and Davis 2019; Marcus 2001), as well as Kahneman's System 1 vs System 2 thinking (Kahneman 2011).

This sub-field has gained increasing prominence, with dedicated workshops (e.g., hosted by Samsung and IBM), interest groups at major research institutions (e.g., at the Alan Turing Institute), and the recent establishment of a specialized journal. Google DeepMind's Olympiad-level AI solvers are neuro-symbolic (Google DeepMind 2024b). Likewise, the increasing use of symbolic systems such as code interpreters within LLM-based systems represents another neuro-symbolic approach; however, as argued by (Marcus 2025b), this training regime is not always acknowledged, leading the wider community to believe LLM systems have powerful reasoning capabilities despite not having any inbuilt symbolic manipulation engines.[1]

Indeed, as studied by efforts such as (Valmeekam et al. 2022), LLMs' ability to reason about mathematical and symbolic truths is brittle and unreliable. Outside of this, the in-

[1] A nuanced point of debate here is whether integrating, say, a python interpreter counts as neuro-symbolic in just the same way as integrating a SAT solver does. We gloss over the point here, and only note that symbol manipulation, emergence, reasoning and related facets are all major considerations in neuro-symbolic AI.

ability of stochastic learners to understand causal relationships as well as the tendency of large models to hallucinate (or more accurately, confabulate) certainly suggests background expert-led knowledge is crucial for trustworthiness.

Directions

Taking a pragmatic view on neuro-symbolic AI as formalisms and frameworks that combine, enhance or otherwise support neural networks with reasoning mechanisms leads, not surprisingly, to distinct and numerous strategies (Feldstein et al. 2024). For example, an important branch of neuro-symbolic AI builds on SRL by combining probabilistic logical models with neural training (De Raedt et al. 2020), and this leads a well-defined and scoped view of neuro-symbolic AI as the neural extension to weighted model counting (Belle 2017; Chavira and Darwiche 2008; Sang, Beame, and Kautz 2005; Van den Broeck, Meert, and Darwiche 2014). There are, however, other design patterns: integrating symbols into the learning regime is not the only player in town. One could, for example, let the high-level control layer be based on logic that works largely independently of the neural/probabilistic layers (Silver et al. 2023), commonly seen in robotics applications. In fact, if we are to leverage large language models (Vaswani et al. 2017), we would undoubtedly have more instances of such decoupling of neural and symbolic functions (Athalye et al. 2024). In (Wang et al. 2024), for example, it is shown that augmenting LLMs with external working memory via a neuro-symbolic pipeline helps with multi-step deductive reasoning.

As suggested in (Lenat and Marcus 2023), there are several interesting ideas from earlier attempts of building large-scale expert-driven logical knowledge sources, such as CYC (Lenat and Guha 1989), that could play a pivotal role in ensuring that generative models become trustworthy in the future.

Keeping this breadth in mind, we outline some key representative areas of inquiry in the field. (We cannot of course cover or do justice to all, given scope and length limits.)

1. **Knowledge Graphs and Expert Knowledge Integration.** Knowledge graphs represent a fundamental area in neuro-symbolic AI, with applications ranging from protein databases and social networks to common sense knowledge bases in advanced language models (Demir and Ngomo 2023; Hossain, Saghapour, and Chen 2025; Lavin 2021). Research in this area focuses on both learning such graphs from neural techniques and reasoning about the knowledge they contain.
2. **Neuro-Symbolic Programs.** Generally, this line of research can be seen as extending knowledge graphs with non-trivial axioms specified in logic programs. Representative examples include DeepProbLog (Manhaeve et al. 2018), which interfaces neural predicates as external artefacts in probabilistic logic programs, and Logic Tensor Networks (Badreddine et al. 2022), which allows neural outputs as artefacts in first-order theories. Interestingly, the former is based on a probabilistic interpretation of the output labels, and the latter on a fuzzy (or real-valued) truth of the labels. In both these formalisms, the neural training is influenced by the logical structure.
3. **Differential program induction.** Program induction is a major area in computer science (Gulwani 2010). A notable example is work from Google DeepMind (Evans and Grefenstette 2018) on learning explanatory logic rules from noisy data, which upgrades inductive logic programming to model neural predicates learned from noisy data. A related but somewhat orthogonal paradigm uses programs to enforce structure in neural predictions, as seen in the work of concept learning (Lake, Salakhutdinov, and Tenenbaum 2015). For work on how program induction might affect human comprehension, see (Rule et al. 2024).
4. **Training Neural Networks with Logic Formulas.** The idea of constraining the loss function of neural networks with symbolic knowledge is useful in a wide range of applications from physics (Stewart and Ermon 2017) to robotics (Innes and Ramamoorthy 2020). There are a number of developments here, such as the work of semantic loss (Gajowniczek et al. 2020) and MultiplexNet (Hoernle et al. 2022), the former based on a probabilistic interpretation of the output labels, and the latter on a fuzzy (or real-valued) truth of the labels. There is a close relation between loss function constraints versus high-level knowledge representation, and incidentally the semantic loss and DeepProblog perform essentially similar functions when computing gradients.
5. **Semantics.** As hinted above, some neuro-symbolic formalisms use a probabilistic interpretation, and others a fuzzy (or real-valued) interpretation for their logical variables. This can impact the learned hypothesis, and gradient computation (van Krieken, Acar, and van Harmelen 2022). There is also the issue of what sort of semantics best describes neuro-symbolic learning; e.g., see discussions in (Tsamoura, Hospedales, and Michael 2021) that it is a type of abduction.
6. **Static vs Dynamics.** Not surprisingly, adding temporal or dynamic aspects can change the semantics and language, and there are a number of symbolic solutions to dealing with actions. For example, reward machines focuses on training (reinforcement learning) agents with (temporal) logical formulas (Icarte et al. 2022). Differential program induction can be extended to a dynamic setting (e.g., reinforcement learning) too (Belle and Bueff 2023; Bueff and Belle 2024).
7. **Leveraging LLMs.** Exploiting and augmenting LLMs for symbolic reasoning tasks constitutes an entirely new class of approaches. When used as blackbox helper functions, systems like Logic-LM (Pan et al. 2023) employ LLMs to translate natural language into logical formulas, after which symbolic executors compute solutions directly. This serves as an effective countermeasure against LLM confabulations, particularly for mathematical tasks and puzzles. A pre-ChatGPT solution making a similar argument can be found in (Dries et al. 2017), where combinatorial problems formulated in natural language are converted to symbolic constructs solved using logic pro-

gramming. Notably, Wolfram Alpha implemented a comparable pipeline just as ChatGPT was released (Wolfram 2023). These symbolic executor pipelines demonstrate considerable power; they can correctly handle modal reasoning problems such as theory of mind (Tang and Belle 2024). Similarly, (Athalye et al. 2024) build symbolic world models from language models. Importantly, such symbolic guardrails may be essential, as LLMs consistently struggle with reasoning and planning tasks (Valmeekam et al. 2022).

Considerations

The designing, building and deployment of neuro-symbolic approaches undoubtedly brings some added complexity, especially owing to the lack of a unified "framework" or programming regime. But this was perhaps true too during the early days of deep learning. It is worth factoring in a few considerations:

1. **Diversity of Approaches.** The "broad church" nature of neuro-symbolic AI (Belle et al. 2023) raises questions about whether a uniform approach is necessary or whether diversity in paradigms benefits the field. This includes considering if we need a unified mathematical language to bridge various system-building methodologies. Evidence from robotics applications suggests the latter (Athalye et al. 2024; Silver et al. 2023) given the complexity of mixing hardware and software interactions.
2. **Balancing Reasoning and Learning.** A fundamental question concerns the appropriate balance between expert-provided knowledge and data-derived knowledge. This includes considerations about how high-level concepts provided by humans might be mapped onto low-level features, and whether these high-level concepts themselves could be partially or fully learned. There may also be domain-specific considerations about the balance between neural and symbolic components, with some application areas potentially requiring more of one than the other. Coupled with this point is the question of what kind of logic is needed in neuro-symbolic systems, e.g., we already noted differences between a probabilistic vs fuzzy interpretation. Also, propositional logic and finite-domain relational logic have been common in neuro-symbolic AI, but one may need to further consider temporal and dynamic logic, particularly for agents operating in physical environments, and perhaps even epistemic logic (Tang and Belle 2024).
3. **Knowledge vs Correctness.** High-level knowledge specified for loss functions may not be obeyed after training (Gajowniczek et al. 2020). There is even evidence that networks could learn properties that are different from what was intended (Marconato et al. 2023). We may hope that future neuro-symbolic solutions robustly handle such problems, but they are certainly a concern today.

Despite these challenges (of maturity, one might add), unless one accepts the "scaling is all you need" argument, it is difficult to envision approaches that rigorously and categorically address issues of knowledge integration and correctness other than neuro-symbolic ones! Interested readers may also refer to (Marcus 2022) on why one should not be accepting the scaling argument in the first place. As (Marcus 2025a) puts it, it is problematic scientifically, economically, and politically.

Still, neurosymbolic AI should not been seen as a magic bullet for solving general-purpose AI, nor a magic solution for trustworthiness. For example, formal approaches may not capture various non-quantifiable harms in the resposible and trustworthy deployment of technology (Belle 2023). Neuro-symbolic AI is likely necessary for progress – primarily because reliable answers about objects in the world require world models, which neuro-symbolic approaches can provide – but it is not sufficient on its own (Marcus 2020).

Discussion, Developments and Case Studies

AlphaGeometry and AlphaProof We discussed an early example of neuro-symbolic integration in Wolfram Alpha's ChatGPT extension (Wolfram 2023). For example, asking ChatGPT about the distance between Chicago and Tokyo could lead to errors, not least owing to its stochasticity. But the extension parses the natural language question to a symbolic lookup in Wolfram Alpha, plausibly leading to the correct response. In this way, its not different from the symbolic executor approach of (Pan et al. 2023) and (Tang and Belle 2024), discussed earlier.

A significant large-scale development in neuro-symbolic AI is arguably Google DeepMind's work on mathematical problem-solving. They introduced systems AlphaGeometry and its successor, AlphaGeometry 2, as well as a newer system called AlphaProof, all of which achieved impressive results in tackling International Mathematical Olympiad problems (Google DeepMind 2024a). AlphaGeometry (Google DeepMind 2024b), for example, is explicitly described as "a neuro-symbolic system made up of a neural language model and a symbolic deduction engine, which work together to find proofs for complex geometry theorems". This system exemplifies the "thinking, fast and slow" paradigm (as reported by Google DeepMind), where one system provides fast, intuitive ideas, and the other offers more deliberate, rational decision-making. Likewise, the newer AlphaGeometry 2 system maintains this neuro-symbolic structure but incorporates a language model, trained from scratch on significantly more synthetic data than its predecessor. AlphaProof too appears to have an analogous structure, with a language model feeding into a search through formal proofs verified and formulated in Lean, a symbolic proof assistant system (Ying et al. 2024). See (Wang et al. 2025; Xin et al. 2024) for other such efforts. On a related note, Google DeepMind's AlphaFold was cited in a recent Nobel Prize award (Google DeepMind 2024c), which is a differentiable approach but involves numerous geometry and domain constraints along with structural priors in its design, which is very much in the spirit of neuro-symbolic modelling (e.g., loss functions and background knowledge) discussed earlier.

LLMs and Reasoning Recent work has highlighted significant limitations in the reasoning capabilities of LLMs that further motivate the neuro-symbolic approach. For example, studies from both Apple (Shojaee et al. 2025) and

Salesforce (Huang et al. 2025) have demonstrated that LLMs struggle with algorithmic reasoning and multi-turn tasks that require precise execution of procedures. These findings suggest that current LLM technology is likely not reliable for complex reasoning tasks. And as discussed earlier, efforts such as (Valmeekam et al. 2022) demonstrated much earlier about the limitations of LLMs for planning.

It is plausible that LLMs might excel at tasks they have encountered in training but struggle with novel situations or slight variations of familiar problems. But variable and symbol substitutions, symmetry and transitivity, among other things, are hallmarks of logical reasoning, which suggests that a logical oracle or solver that is distinct from the LLM machinery is unavoidable.

Tools Integration Along with Wolfram Alpha's integration with ChatGPT, OpenAI has released plug-ins like "Code Interpreter" to enable the model to perform mathematical calculations or correctness checking using external software.

While these integrations improve performance on straightforward mathematical tasks, they still face challenges in reliably solving word problems that involve a combination of science and mathematics. As argued in (Marcus and Davis 2023), the difficulties arise from two main gaps: LLMs usually lack the commonsense knowledge needed to translate word problems into mathematical calculations, and they do not seem to reliably understand how to use the tools. Testing of these tool-augmented systems seems to show mixed results; see (Marcus and Davis 2023) for discussions. This inconsistency suggests the need for more robust integration between neural and symbolic systems.

Conclusion

Neuro-symbolic AI represents a promising (and perhaps only) approach to addressing the limitations of purely neural or purely symbolic systems. By combining the pattern recognition capabilities of neural networks with the reasoning power of symbolic systems, neuro-symbolic approaches offer potential solutions to challenges in areas such as structured reasoning, knowledge integration, explainability, and reliability.

Recent developments, particularly in mathematical problem-solving systems like AlphaGeometry and AlphaProof, demonstrate the viability and effectiveness of neuro-symbolic approaches. Given the current excitement about LLM technology, we pointed to studies that underscore the necessity of moving beyond purely neural approaches, and assuming scaling will lead to general-purpose AI on its own.

Fundamentally, determining the optimal balance between neural and symbolic components, and understanding the formal linguistic and semantic requirements, might be a valuable place to start towards the development of future AI systems that are more capable, more reliable, more interpretable and more alignable.

References

Athalye, A.; Kumar, N.; Silver, T.; Liang, Y.; Wang, J.; Lozano-Pérez, T.; and Kaelbling, L. P. 2024. From Pixels to Predicates: Learning Symbolic World Models via Pretrained Vision-Language Models. *arXiv preprint arXiv:2501.00296*.

Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press.

Badreddine, S.; Garcez, A. d.; Serafini, L.; and Spranger, M. 2022. Logic tensor networks. *Artificial Intelligence*, 303: 103649.

Barceló, P.; Kostylev, E. V.; Monet, M.; Pérez, J.; Reutter, J.; and Silva, J.-P. 2020. The logical expressiveness of graph neural networks. In *8th International Conference on Learning Representations (ICLR 2020)*.

Belle, V. 2017. Weighted Model Counting With Function Symbols. In *UAI*.

Belle, V. 2023. Knowledge representation and acquisition for ethical AI: challenges and opportunities. *Ethics and Information Technology*, 25(1): 22.

Belle, V.; and Bueff, A. 2023. Deep Inductive Logic Programming meets Reinforcement Learning. In *The 39th International Conference on Logic Programming*. Open Publishing Association.

Belle, V.; Fisher, M.; Russo, A.; Komendantskaya, E.; and Nottle, A. 2023. Neuro-symbolic AI+ agent systems: a first reflection on trends, opportunities and challenges. In *International Conference on Autonomous Agents and Multiagent Systems*, 180–200. Springer.

Besold, T. R.; Bader, S.; Bowman, H.; Domingos, P.; Hitzler, P.; Kühnberger, K.-U.; Lamb, L. C.; Lima, P. M. V.; de Penning, L.; Pinkas, G.; et al. 2021. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-symbolic artificial intelligence: The state of the art*, 1–51. IOS press.

Bueff, A.; and Belle, V. 2024. Learning explanatory logical rules in non-linear domains: a neuro-symbolic approach. *Machine Learning*, 1–36.

Chavira, M.; and Darwiche, A. 2008. On probabilistic inference by weighted model counting. *Artif. Intell.*, 172(6-7): 772–799.

Chollet, F. 2017. The limitations of deep learning. *Deep learning with Python*.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

De Raedt, L.; Dumančić, S.; Manhaeve, R.; and Marra, G. 2020. From statistical relational to neuro-symbolic artificial intelligence. *arXiv preprint arXiv:2003.08316*.

De Smet, L.; and De Raedt, L. 2025. Defining neurosymbolic AI. *arXiv preprint arXiv:2507.11127*.

Demir, C.; and Ngomo, A.-C. N. 2023. Neuro-Symbolic Class Expression Learning. In *IJCAI*, 3624–3632.

Dries, A.; Kimmig, A.; Davis, J.; Belle, V.; and De Raedt, L. 2017. Solving Probability Problems in Natural Language. In *IJCAI*.

Ellis, K.; Albright, A.; Solar-Lezama, A.; Tenenbaum, J. B.; and O'Donnell, T. J. 2022. Synthesizing theories of human language with Bayesian program induction. *Nature communications*, 13(1): 5024.

Evans, R.; and Grefenstette, E. 2018. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61: 1–64.

Feldstein, J.; Dilkas, P.; Belle, V.; and Tsamoura, E. 2024. Mapping the neuro-symbolic AI landscape by architectures: A handbook on augmenting deep learning through symbolic reasoning. *arXiv preprint arXiv:2410.22077*.

Gajowniczek, K.; Liang, Y.; Friedman, T.; Zabkowski, T.; and Van den Broeck, G. 2020. Semantic and generalized entropy loss functions for semi-supervised deep learning. *Entropy*, 22(3): 334.

Garcez, A. d.; and Lamb, L. C. 2023. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11): 12387–12406.

Getoor, L.; Friedman, N.; Koller, D.; and Taskar, B. 2001. Learning Probabilistic Models of Relational Structure. In *ICML*, 170–177.

Goodfellow, I.; Bengio, Y.; Courville, A.; and Bengio, Y. 2016. *Deep learning*, volume 1. MIT press Cambridge.

Google DeepMind. 2024a. AI achieves silver-medal standard solving International Mathematical Olympiad problems. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/. Vaishak Belle: Accessed: 2025-09-01 Vaishak Belle: Accessed: 2025-09-01.

Google DeepMind. 2024b. AlphaGeometry: An Olympiad-level AI system for geometry. https://deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/. Accessed: 2025-09-01.

Google DeepMind. 2024c. Demis Hassabis & John Jumper awarded Nobel Prize in Chemistry. https://deepmind.google/discover/blog/demis-hassabis-john-jumper-awarded-nobel-prize-in-chemistry/. Accessed: 2025-09-01.

Gulwani, S. 2010. Dimensions in program synthesis. In *PPDP*, 13–24. ACM.

Halpern, J. Y. 2003. *Reasoning about Uncertainty*. MIT Press. ISBN 0262083205.

Hitzler, P.; and Sarker, M. K. 2022. Neuro-Symbolic Artificial Intelligence: The State of the Art.

Hoernle, N.; Karampatsis, R. M.; Belle, V.; and Gal, K. 2022. Multiplexnet: Towards fully satisfied logical constraints in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5700–5709.

Hossain, D.; Saghapour, E.; and Chen, J. Y. 2025. NeSyDPP4-QSAR: A Neuro-Symbolic AI Approach for Potent DPP-4-Inhibitor Discovery in Diabetes Treatment. *bioRxiv*, 2025–03.

Huang, K.-H.; Prabhakar, A.; Thorat, O.; Agarwal, D.; Choubey, P. K.; Mao, Y.; Savarese, S.; Xiong, C.; and Wu, C.-S. 2025. Crmarena-pro: Holistic assessment of LLM agents across diverse business scenarios and interactions. *arXiv preprint arXiv:2505.18878*.

Icarte, R. T.; Klassen, T. Q.; Valenzano, R.; and McIlraith, S. A. 2022. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73: 173–208.

Innes, C.; and Ramamoorthy, S. 2020. Elaborating on learned demonstrations with temporal logic specifications. *arXiv preprint arXiv:2002.00784*.

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Kelly, R. F.; and Pearce, A. R. 2008. Complex Epistemic Modalities in the Situation Calculus. In *KR*.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press. ISBN 978-0-262-01319-2.

Koller, D.; and Pfeffer, A. 1997. Object-oriented Bayesian networks. In *Proc. UAI*, 302–313.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lake, B.; and Baroni, M. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, 2873–2882. PMLR.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Lavin, A. 2021. Neuro-symbolic neurodegenerative disease modeling as probabilistic programmed deep kernels. In *International Workshop on Health Intelligence*, 49–64. Springer.

Lenat, D.; and Marcus, G. 2023. Getting from generative AI to trustworthy AI: What LLMs might learn from CYC. *arXiv preprint arXiv:2308.04445*.

Lenat, D. B.; and Guha, R. V. 1989. *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc.

Levesque, H. J. 1996. What Is Planning in the Presence of Sensing? In *Proc. AAAI / IAAI*, 1139–1146.

Levesque, H. J. 2012. *Thinking as computation: A first course*. Mit Press.

Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; and Zhang, G. 2018. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12): 2346–2363.

Maldonado, R.; Goodwin, T. R.; Skinner, M. A.; and Harabagiu, S. M. 2018. Deep learning meets biomedical ontologies: knowledge embeddings for epilepsy. In *AMIA Annual Symposium Proceedings*, volume 2017, 1233.

Mallick, A.; Hsieh, K.; Arzani, B.; and Joshi, G. 2022. Matchmaker: Data drift mitigation in machine learning for large-scale systems. *Proceedings of Machine Learning and Systems*, 4: 77–94.

Manhaeve, R.; Dumancic, S.; Kimmig, A.; Demeester, T.; and De Raedt, L. 2018. Deepproblog: Neural probabilistic logic programming. *Advances in Neural Information Processing Systems*, 31.

Marconato, E.; Teso, S.; Vergari, A.; and Passerini, A. 2023. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. *Advances in Neural Information Processing Systems*, 36: 72507–72539.

Marcus, G. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Marcus, G. 2020. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv:2002.06177.

Marcus, G. 2022. Deep learning is hitting a wall. *Nautilus*, 10: 2022.

Marcus, G. 2025a. The Fever Dream of Imminent Superintelligence Is Finally Breaking. https://www.nytimes.com/2025/09/03/opinion/ai-gpt5-rethinking.html. Accessed: 2025-09-01.

Marcus, G. 2025b. How o3 and Grok 4 Accidentally Vindicated Neurosymbolic AI. https://garymarcus.substack.com/p/how-o3-and-grok-4-accidentally-vindicated?r=8tdk6. Accessed: 2025-09-01.

Marcus, G.; and Davis, E. 2019. *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Marcus, G.; and Davis, E. 2023. Getting GPT to work with external tools is harder than you think. https://garymarcus.substack.com/p/getting-gpt-to-work-with-external. Accessed: 2025-09-01.

Marcus, G. F. 1998. Rethinking eliminative connectionism. *Cognitive psychology*, 37(3): 243–282.

Marcus, G. F. 2001. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Pan, L.; Albalak, A.; Wang, X.; and Wang, W. Y. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Pearl, J. 1998. Graphical models for probabilistic and causal reasoning. In *Quantified Representation of Uncertainty and Imprecision*, 367–389. Springer.

Raedt, L. D.; Kersting, K.; Natarajan, S.; and Poole, D. 2016. Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 10(2): 1–189.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.

Rule, J. S.; Piantadosi, S. T.; Cropper, A.; Ellis, K.; Nye, M.; and Tenenbaum, J. B. 2024. Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications*, 15(1): 6847.

Sang, T.; Beame, P.; and Kautz, H. A. 2005. Performing Bayesian Inference by Weighted Model Counting. In *AAAI*, 475–482.

Sarker, M. K.; Zhou, L.; Eberhart, A.; and Hitzler, P. 2021. Neuro-symbolic artificial intelligence. *AI Communications*, (Preprint): 1–13.

Shojaee, P.; Mirzadeh, I.; Alizadeh, K.; Horton, M.; Bengio, S.; and Farajtabar, M. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Silver, T.; Chitnis, R.; Kumar, N.; McClinton, W.; Lozano-Pérez, T.; Kaelbling, L.; and Tenenbaum, J. B. 2023. Predicate invention for bilevel planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12120–12129.

Stewart, R.; and Ermon, S. 2017. Label-free supervision of neural networks with physics and domain knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Sun, R. 2002. From the Unconscious to the Conscious: A Connectionist-Symbolic Approach. In *From Synapses to Rules: Discovering Symbolic Rules from Neural Processed Data*, 293–313. Springer.

Tang, W.; and Belle, V. 2024. ToM-LM: Delegating Theory Of Mind Reasoning to External Symbolic Executors in Large Language Models. *NeSy*.

Tsamoura, E.; Hospedales, T.; and Michael, L. 2021. Neural-symbolic integration: A compositional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5051–5060.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind*, 59(236): 433–460.

Valiant, L. G. 1999. Robust logics. In *Proceedings of the thirty-first annual ACM symposium on Theory of Computing*, 642–651.

Valmeekam, K.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2022. Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). *arXiv preprint arXiv:2206.10498*.

Van den Broeck, G.; Meert, W.; and Darwiche, A. 2014. Skolemization for Weighted First-Order Model Counting. In *KR*.

van Krieken, E.; Acar, E.; and van Harmelen, F. 2022. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 302: 103602.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Unsal, M.; Lin, X.; Baksys, M.; Liu, J.; Santos, M. D.; Sung, F.; Vinyes, M.; Ying, Z.; Zhu, Z.; et al. 2025. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.11354*.

Wang, S.; Wei, Z.; Choi, Y.; and Ren, X. 2024. Symbolic working memory enhances language models for complex rule application. *arXiv preprint arXiv:2408.13654*.

Wolfram, S. 2023. Wolfram alpha as the way to bring computational knowledge superpowers to ChatGPT. *Stephen Wolfram Writings RSS, Stephen Wolfram, LLC*, 9.

Xin, H.; Guo, D.; Shao, Z.; Ren, Z.; Zhu, Q.; Liu, B.; Ruan, C.; Li, W.; and Liang, X. 2024. Advancing theorem proving in LLMs through large-scale synthetic data. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Ying, H.; Wu, Z.; Geng, Y.; Wang, J.; Lin, D.; and Chen, K. 2024. Lean workbook: A large-scale lean problem set formalized from natural language math problems. *Advances in Neural Information Processing Systems*, 37: 105848–105863.