Neurosymbolic AI: towards sound reasoning and causal learning and the road to AGI

Artur d'Avila Garcez

December 19, 2025

Abstract

As society contemplates AI's impact on everyday life and work, the opacity surrounding AI development since the release of ChatGPT contributes to fears of existential risk and fuels claims of an upcoming AI bubble burst. In this article, I argue that the emerging field of neurosymbolic AI can address the lack of reliability of current AI. Instead of ever increasing compute power, use of chain-of-thought prompting and performing alignment via reinforcement learning, neurosymbolic AI promotes model compression, symbolic knowledge reuse and alignment via knowledge sharing. I discuss how the persisting problem of reliability can be addressed by neurosymbolic AI with the use of formal reasoning, causal inference and extrapolation towards artificial general intelligence.

Keywords: Neurosymbolic AI, Machine Learning, Logical Reasoning, Generative AI.

1 Introduction

Neurosymbolic AI can be defined pragmatically as the application of the neurosymbolic cycle, that is, the change of representation between what is commonly known as a neural network and what is commonly called a symbolic system in AI. In neurosymbolic AI, given a symbolic system as input, a *translation algorithm* will produce a corresponding neural network as output; an *extraction algorithm* will produce a symbolic description as output given a trained neural network as input. Studying the various forms of symbolic and neural representation and how they map to each other helps organize the field with the derivation of expressiveness results and evaluations of the various existing AI models within a domain of application.

Generally speaking, symbolic systems are defined around a formalization of knowledge representation as required for formal reasoning. Neural networks, by contrast, excel at learning from data as an efficient computational model. As a result, neurosymbolic AI is well-placed to address the current question of reasoning in neural networks, particularly Large Language Models (LLMs), to produce AI systems that combine in a principled way statistical learning and formal reasoning with data and knowledge. Whether to improve system performance, check for safety properties, increase fairness and trust or reduce data and energy requirements, neurosymbolic AI approaches have been proposed in recent years, obtaining promising results in each of these areas [14, 15]. Some of these approaches seek to integrate formal reasoning into neural networks [24]. Others make use of a hybrid system having a symbolic component and a neural network component, such as a knowledge graph that communicates with a transformer or graph neural network or an LLM whose outputs are checked by a symbolic proof assistant [16]. In what follows, I will discuss and explore the reach of both possibilities: neurosymbolic integration and neurosymbolic hybrid systems.

2 Scaling Neurosymbolic AI

The value of neurosymbolic AI has been illustrated with the use of discrete search at test time (e.g. Google's AlphaGeometry system [30]) and the use of model compression when distillation is combined with LLMs, e.g. DeepSeek [11]. Despite its promise, neurosymbolic AI has not been adopted as mainstream by the AI industry leaders. I will argue that the time is right to scale up neurosymbolic AI as a decentralized alternative approach to current AI. Scaling in neurosymbolic AI is very different from scaling up neural networks. The so-called *scale is all you need* approach of neural networks requires unimaginable amounts of data to satisfy an ever-increasing number of network parameters. It brings with it very high energy costs and pressing questions around copyright violation. Scaling in neurosymbolic AI is about increasing the number of times that the neurosymbolic cycle is repeated. In this setting, a successful combined application of data and knowledge should enable network compression by knowledge reuse, instead of the ever increasing number of parameters required by *scale- is all you need*. When the neurosymbolic cycle is applied successfully to multiple tasks, it should require fewer data to produce the same results as neural networks [9].

Starting from a trained network, the application of the neurosymbolic cycle requires effective extraction of symbolic knowledge from the network, typically using the network as an oracle to produce a simpler network (sometimes called a *student* network [13]) or a decision tree (known to be equivalent to propositional logic statements in disjunctive normal form) [22], a graph [19] or computer program [5], or by querying (probing) the network to derive symbolic knowledge or logic programs [32]. Any symbolic representation to be extracted should be measured using a standard metric known as the *fidelity* of the symbolic description with respect to the trained network [24], as discussed next.

2.1 The Knowledge Extraction Problem

Extraction of knowledge with precise semantics from neural networks is a major challenge and the main bottleneck in the application of the neurosymbolic cycle at scale. Knowledge extraction from large networks such as LLMs is a daunting task, if not impossible, due to the sheer scale of current models. Approaches seeking to extract knowledge from multimodal networks, combining for example learning from text and images, seem to be more viable. That is because one modality can help ground the explanation of the other modality, although this too continues to be a major challenge [3].

An alternative to the task of explaining LLMs is offered by the application of the neurosymbolic cycle. Knowledge can be extracted from small parts of the network as they are being trained followed by reasoning about what has been learned based on partial knowledge, followed by further training as more data become available including in other related tasks (multitask learning). This process permits measuring how well a symbolic description may approximate parts of the neural network (fidelity), it is amenable to the presence of potentially incorrect background knowledge that can be revised through learning, and it allows domain experts to ask *what-if* questions and data scientists to intervene during the process based on sound reasoning results [23].

2.2 Scale is not all we need

The neurosymbolic cycle uses network probing, knowledge extraction, manipulation and distillation to obtain more compact networks with associated knowledge, the opposite of scaling by increasing network size. Instead of scaling in the traditional sense, scaling the neurosymbolic cycle is about applying this recipe: *learn a little*, *reason a little*, *repeat*. Extraction of knowledge opens the possibility of knowledge reuse and sharing across application domains. Knowledge reuse reduces the training data requirements over time.

As a concrete simple example of knowledge sharing, consider the learning of the transitive relation *greater-than* in computer vision, as used in [28]. Two toy application domains are considered: a blocks world scenario where a tower of blocks is taller than another, and a relational version of the MNIST data set where two handwritten digits are provided as input and the goal is to learn whether a digit is larger than the other. In both scenarios, the relation learned is transitive: if one tower (or digit) is greater than another tower (or digit) and this other tower (or digit) is greater than yet another tower (or digit) then the first tower (or digit) must be greater than the last tower (or digit). Whether the learned relation is called *taller-than* in the case of towers, *larger-than* in the case of digits, or simply *greater-than*, irrespective of whether it applies to images of blocks or digits, the general rule structure to be learned is the same: given any three objects (X, Y, Z) of the same object type, if X is greater than Y and Y is greater than Z then X is greater than Z.[1]

Differently from the towers of blocks where it can be inferred from the images whether a tower is taller than another, nothing in the MNIST data set indicates when a digit is greater than another. The *greater-than* relation is said to be at a higher-level of abstraction. As with children who may benefit from learning

[1] In logic, $\forall X, Y, Z((GreaterThan(X, Y) \wedge GreaterThan(Y, Z) \rightarrow GreaterThan(X, Z))$.

arithmetic using blocks before moving to more abstract descriptions of digits and digit manipulation, here too a neurosymbolic system capable of learning the general rule should benefit from its application across domains. Interestingly, and of great relevance to exemplifying the distinction between deep learning and neurosymbolic AI, once a description can be obtained in symbolic logic given a trained neural network, reasoning that is provably correct (as in the case of the *greater-than* relation) is guaranteed for any number of objects, regardless of the sizes of the towers or the magnitudes of the numbers being compared. Knowledge reuse here should enable extrapolation beyond the observed data, both within application domain and across domains having different abstraction levels (as in the case of blocks and digits). Reasoning that makes use of the symbolic descriptions extracted will be correct by design, learned from data but richer than reasoning by similarity, as discussed next.

3 Causality and Extrapolation

In [25], Pearl argues that neural networks are only capable of handling the first layer of his proposed three-layers of causal hierarchy: association (*what is*), intervention (*what if*), counterfactuals (*why*). This interpretation aligns with John McCarthy's earlier account of the capabilities of neural networks, which McCarthy referred to as the *propositional fixation* of neural networks [20] in response to Paul Smolensky's neurosymbolic treatment of connectionism [27].

Even if neural networks were only capable of learning correlations, as argued by Pearl and McCarthy, knowledge extraction offers an additional possibility of model intervention by user interaction with the AI system. This includes even counterfactual reasoning as illustrated in the next paragraph. Knowledge extraction here only seeks to make sense of the network, not to represent the real world, which is a much harder task. By iterating the neurosymbolic cycle, measuring the results and providing feedback, one hopes to obtain better approximations of the real world over time, but there are no guarantees. This is why scaling up and validating the application of the neurosymbolic cycle is so important. This way, the data-driven network learning is evaluated continuously, with formal reasoning informing the network alignment process ahead of further training with data. An example of how this is done was provided in [32], combining Logic Tensor Networks [2] with Testing with Concept Activation Vectors [18] used as a network probing technique for knowledge extraction.

Assume, for the sake of argument, that neural networks are only capable of association, that is, neural networks are learners of correlations. Given a trained network as input, suppose that a knowledge extraction algorithm produces the following symbolic propositional knowledge as output: $A \rightarrow B$ (the activation of a set of neurons denoted by A implies the activation of a set of neurons denoted by B). With this knowledge extraction, it is clear that interventions become possible. Given $A \rightarrow B$, a data scientist or domain expert may ask the questions *What if* B *was false* ($\neg B$)? What is the minimal change I need to make to A that would make the probability of B go below 0.5? Is B true only if A is true ($B \to A$)? (see [22] for a concrete application in medical diagnosis). Such manipulations enable the extraction of counterfactual explanations from neural networks, even if the networks themselves were just association learners (see [34]). The symbolic descriptions $A \to B, \neg B, B \to A, ...$ denote the behavior of the network up to fidelity error. The main argument here is that a rich and compact extracted symbolic description can augment the capabilities of the trained neural network from which it is extracted, as well as our understanding of the network's computations. In the first-order logic[2] case, the provision of a symbolic description allows for extrapolation of the data-driven model to infinite domains. A first-order rule extracted from a trained network, e.g. $\forall x P(x)$, despite being obtained from the finite set of observed values of x, applies to any value of variable x.[3]

There is something particularly relevant about the above change of representation from distributed (combinations of partial functions) to localist (executable symbolic descriptions) [32]. In this cycle of representation change, the building over time of relevant abstract descriptions derives from learning with data followed by reasoning about what has been learned. Done successfully, this should produce compact representations with an ability to extrapolate to new cases for which there is a shortage of data, as in the case where a recursive definition is learned that is applicable in general (e.g. the well-known Tower of Hanoi puzzle in AI whose learned description should be applicable to towers of any height). In neurosymbolic AI, the symbolic descriptions are expected to be obtained following the application of an efficient neural learning algorithm such as gradient descent. Furthermore, user-validated or consolidated knowledge from multiple experiments must be instilled back into an efficient and ideally compact distributed network representation.

[2] John McCarthy coined the term *propositional fixation* of neural networks [20] to refer to the challenge of first-order logic computation by neural networks. The computational efficiency that we observe in neural networks may well be due to the possibility that this propositional fixation is a theoretical limit of networks that are always *grounded* on data and would be, therefore, incapable of representing first or higher-order logic.

[3] The neural networks that we are referring to are implemented e.g. in Python on a computer simulating a Turing machine. Implementation in a given programming language does not make the system neurosymbolic. Similarly, the fact that LLMs have symbols as input and output does not make LLMs neurosymbolic. For this reason, our definition and analysis of neurosymbolic AI takes place at the level of the common definitions of artificial neural networks and logical systems. First-order logic descriptions with logical variables, relations and quantification represent infinite domains (e.g. $\forall x P(x), x \in \mathbb{N}$). Such descriptions are typically grounded onto propositional instances (e.g. enumerating the values of x) for the sake of efficient learning. Generation of a computer program by an LLM that learned a partial recursive function from data is a case-in-point. The network is grounded on the data, but the symbolic output allows symbol manipulation to take place (and code to be executed, as part of what became known as *agentic AI*). Put together with symbolic manipulation, agentic AI becomes neurosymbolic. The justified safety concerns around agentic AI stem from the unknown, unintended consequences of code execution without a formal semantics. Neurosymbolic AI seeks to address this problem by offering a semantics to deep learning [24].

4 Formal Reasoning with Neural Networks

Theoretical results about the reasoning capabilities of neural networks have shown that various forms of reasoning can be carried out within neural networks [8]. This is typically achieved by proving that the stable states of a network correspond to a fixed-point semantics of a given logic formalism. It is very different from the approach that became known as Chain-of-Thought (CoT) reasoning [33]. Reasoning in LLMs with CoT prompting generates synthetic data at run-time using the underlying transformer network itself in a chain that breaks down a prompt into sub-prompts. It is reasonable to expect this increase in *test-time compute* to improve reasoning performance because reasoning tasks are typically solved by breaking down a problem into sub-problems. These LLM models were claimed to "think before they answer", when in fact very little is known about how such systems improved on reasoning and code generation benchmarks. Let's assume that such LLMs are a kind of "GPT-Go" system, a generative pre-trained transformer to which a tree search is incorporated in the style of Google DeepMind's earlier Alpha-Go system. The tree search uses CoT prompting to break down the prompts into sub-prompts. The system's "thinking" time is presumably needed to build the tree for the CoT. Leaving aside the practical question of how long users will be happy to wait for an answer, by itself CoT offers a *trial-and-error* approach to reasoning. Without some form of control of the learned function, small changes made to the input can produce diverging results, therefore inconsistencies. When LLMs produce such inconsistencies and *make stuff up* such as referring to non-existing articles on the internet, they are said to hallucinate. The problem is two-fold: a lack of reliability of the synthetic data generation, known as the curse of recursion [26], and the combinatorial nature of the CoT input, best described as *infinite uses of finite means* (a finite dictionary giving rise to infinite possible texts). Small changes in input may produce diverging results due to the inevitable accumulation of errors in the calculations of the neural network. As a result, CoT will solve one reasoning task today, only to fail at a very similar reasoning task tomorrow. This is best illustrated by the examples provided in [21] showing that a mere change in naming convention can affect reasoning performance dramatically.

By contrast with CoT, neurosymbolic integration seeks to control the architecture or the loss function of the network, instead of adjusting the input, in order to learn to reason. Based on symbolic descriptions that are either learned or already known, the combination of data and knowledge is expected to increase reliability. Sound and approximate reasoning can be achieved in a neural network either by engineering the network architecture in modular fashion or by regularizing the loss function with respect to a formal language and intended semantics [8]. Following a long tradition of combining neural networks with symbolic knowledge, the task at hand is to ascribe meaning to neural computation by offering a formal semantics to neural networks.

Neurosymbolic hybrid systems have also shown improvement in reasoning performance when learning is combined with search-based reasoning. In [22], knowledge distillation is used to build a decision tree from a trained neural network, with the corresponding propositional logic descriptions of the tree used to carry out formal reasoning at test time. The goal is to combine learning and reasoning to make model development parsimonious by: (1) extracting symbolic descriptions as learning progresses, (2) reasoning formally about what has been learned, and (3) compressing the network as knowledge is instilled back in the network. Measuring the capabilities of the neural network w.r.t. formally-defined provably-sound reasoning offers a much needed measure of the accumulation of errors within the network, as discussed below.

Consider again the above example of learning a transitive relation and querying a trained network to extract first-order knowledge [2, 32]. This creates a symbolic description derived from the data-driven trained network, but this description enables extrapolation to infinite domains. Once a first-order logic rule such as $\forall X, Y\ P(X, Y)$ is extracted from a network, this rule is applicable to any of the values that variables X and Y can take. Even though the rule might have been learned *in-distribution* from a finite number of examples, it also applies *out-of-distribution* as in the case of the transitivity of towers of blocks and MNIST digits seen earlier. Of course, this extrapolation requires a measure of how well the rules approximate the neural network, the above fidelity measure, which has been applied in practice e.g. in [34, 35].

4.1 Controlling the accumulation of errors

We have seen that knowledge extraction is a challenge and the bottleneck of neurosymbolic integration. The above problem of the accumulation of errors in a learning system was identified by Leslie Valiant in [31]. Auto-regressive models compound approximation errors as the calculations iterate. Left unchecked, this causes the overall learning system to diverge rather than converge to the intended stable states with well-defined semantics. The same problem is seen in graph neural networks where addressing the so-called *multi-hop reasoning* problem - information retrieval requiring multiple reasoning steps, possibly across different data sources - proved to be very difficult to solve. The prototypical example of the multi-hop problem is the retrieval of the name of the mother of the singer of *Superstition*, requiring a reasoning step (one *hop*) to conclude that Stevie Wonder is the singer of *Superstition*, before another *hop* may retrieve the name of Stevie Wonder's mother. Reasoning over qualitative statements described symbolically in a chain of rules such as $A \rightarrow B, B \rightarrow C, C \rightarrow D, ...$ applies consistently across any number of steps (hops) in the chain. The problem is solved by avoiding the accumulation of errors seen in continuous space that is required for efficient learning with gradient descent, with a mapping of the neural network onto a discrete space of qualitative rules (the above chain) where sound reasoning applies easily. This is what Leslie Valiant ultimately referred to as *reconciling the statistical* (continuous) *nature of learning and the logical* (discrete) *nature of reasoning* in [31]. In other words, the availability of an approximate symbolic description for a trained network eliminates the accumulation of errors typically seen in the numerical or probabilistic network.

The symbolic rules are a qualitative approximation of the learned probabilistic model. Symbolic computation controls the accumulation of errors by making the application of the rules precise for the purpose of reasoning. Probability distributions can be learned efficiently inside the networks, but don't have to be expressed symbolically in every case. Uncertainty can be expressed qualitatively through rules with confidence values, e.g. in [29], or preference relations among the rules that represent the underlying probabilistic learning model.

4.2 Integrating Reasoning and Learning towards AGI

We have seen increasing attention devoted in recent years to the question of reasoning in neural networks. The principled integration of reasoning and learning in neural networks is a main objective of the field of neurosymbolic AI. When in neurosymbolic AI an algorithm is used to translate a form of symbolic knowledge into the architecture and initial set of weights of a neural network, the intention is to prove that the network is a massively-parallel model of computation for exactly that knowledge. Trained with data on top of that knowledge, the network is expected to produce better learning and generalization performance (faster training and higher accuracy) than if it were trained from scratch with data. Symbolic knowledge is provided to the network in the form of general rules which are believed to be true in a domain, or rules which are expected to be true across application domains. When rules are not available to start with, they can be extracted from a trained network. When rules are contradicted by data, they can be revised by learning and can't be assumed to be always true. This has been shown to offer a flexible framework with knowledge and data combined, expected to lead in the long run to a better understanding of the capabilities of complex networks used for learning and reasoning [10, 4, 12].

Although LLMs are a great engineering achievement, impressive at text summarization, code generation and language translation, three years on from the release of ChatGPT, it is clear that network hallucinations are not going away. Fixing LLM's reliability issues case-by-case with Reinforcement Learning with Human Feedback (RLHF) has proved to be too costly, both financially and in terms of human costs when bad mistakes are made. Far too many exceptions exist in the data and a single bad hallucination is sufficient to destroy trust. Consider LLMs' ability to produce code in agentic AI mode, where the LLM isn't deployed as a stand-alone computer program, but in a loop where code is executed and data is collected for further training, therefore blurring the distinction that currently exists between model training and run-time decision making (a useful separation of concerns in the engineering of current large AI models). One can see immediately how such *self-improving LLM with agency* without guardrails may pose a serious risk to computational systems worldwide. Recent experiments indicated that system *self-impairing*, rather than self-improving, may actually take place, producing a degradation in performance over time. There must be a better way, other than very costly, post-hoc RLHF of instilling reliability into LLMs, a better way of achieving AI that can offer certain logical guarantees to network training models.

AI has been associated with the idea of an autonomous self-improving system. This idea, in turn, has been associate with the term Artificial General Intelligence (AGI) when applied across multiple complex tasks. LLMs may be seen to be such a general purpose system because they will provide an answer to any question. They do this by predicting the probability of the next word (token) in a sentence. Having made a choice of the next word, LLMs will apply the same calculations again, recursively, to predict the probability of the next token, and so on. For this reason, LLMs are called auto-regressive models: the learned function f is applied recursively to the input sequence to choose the token at the next time point (x_{t+1}) such that $x_{t+1} = f(x_t)$. Artificial General Intelligence, however, is best measured by the ability to adapt to novelty from only a few examples. It requires creativity, abstract reasoning and intuition to identify the best choices among various alternatives, as when spotting the *most beautiful* mathematical derivations among a number of all functionally correct options. As such, AGI will require effective learning from fewer data than LLMs, the ability to reason reliably about the knowledge that has been learned, the extraction of compact (beautiful) descriptions from trained networks and the consolidation of knowledge learned from multiple tasks, using analogy to enable extrapolation to new situations at an adequate level of abstraction. Despite the vast financial investment, scaling up of LLMs didn't produce AGI as was hoped. It is fair to say that the "scale is all you need" approach has failed.

An important distinction needs to be made between AGI and domain-specific AI systems that already exist and can exhibit intelligence at the level of humans or higher. These domain-specific systems exhibit intelligence in specialized tasks: targeted medical diagnoses, protein folding, various closed-world two-player strategy games, and offer huge potential value, discussed next in the context of a brief historical perspective on the development of AI.

5 Neurosymbolic AI: The third wave of AI

The first wave of AI goes back to the 1980s and was said to be knowledge-based and well-founded, but the systems of the time - expert systems - were inefficient by comparison with deep learning. The second wave of AI from the 2010s used data-driven and efficient deep neural networks. Based on distributed learning, these networks were unsound if compared with knowledge-bases. Today, it is clear that neural networks are a main component of AI, but the problems with deep learning have been stubbornly difficult to fix. Next, I discuss how solving these problems require the use of symbolic AI alongside neural networks, known as the third wave of AI: neurosymbolic AI [7].

In order to understand the achievements and limitations of AI, it is helpful to consider the *AGI debate*[4] with its focus on what is missing from current AI systems, i.e. the technological innovation that may bring about reliable general purpose AI or even AGI. Simply put, such innovation may be described as the ability to apply knowledge learned by a neural network from one task to a novel

[4] https://www.youtube.com/watch?v=JGiLz_Jx9uI.

task without requiring too much data. With AI experts John Hopfield and Geoff Hinton awarded the 2024 Nobel Prize for Physics, and AI expert Demis Hassabis awarded the 2024 Nobel Prize for Chemistry (with David Baker and John Jumper), it can be said that the era of computation as the language of science has begun. Hassabis led the team at Google DeepMind that created AlphaFold, an AI model capable of predicting with high accuracy the 3D structure of proteins given their amino acid sequence. AlphaFold is arguably the greatest achievement of AI to date. It is squarely an application specific (also called *narrow*) AI system, not general purpose AI. From particle physics to drug discovery, energy efficiency and novel materials, AI is being adopted as the process by which scientific research is carried out, with vast potential for very relevant targeted breakthroughs to take place in the near future. However, the lack of a description or an explanation capable of conveying a deeper understanding of the solutions being offered by AI is very unsatisfactory. Computer scientists in a great feat of engineering will solve to a high degree of accuracy very challenging problems in science without necessarily improving our understanding of the solutions being discovered. That will be the case if those solutions are provided by very large neural networks, trained on vast amounts of data, that are not humanly possible to inspect. This unsatisfactory lack of explainability of generative AI coupled with the risks of agentic AI confirm the need for neurosymbolic AI. As discussed, neurosymbolic AI uses the technology of knowledge extraction to interpret, ask *what-if* questions and if necessary intervene in the AI system, controlling learning in ways that can offer correctness or fairness guarantees and, with this process, producing a more compact and data efficient system. In 2025, we start to see a shift towards such explainable neurosymbolic AI systems being deployed in domain-specific AI solutions [30, 6].

The history of neurosymbolic AI goes back more than 20 years[5]. Already around the turn of the century, the importance of artificial neural networks as an efficient computational model for learning and reasoning was obvious to a small group of researchers [1]. The value of symbol manipulation and abstract reasoning afforded by symbolic logic was also clear to that group of people. Many others before the turn of the century contributed to the development of neurosymbolic AI, even if not specifically within the field. In fact, it could be argued that neurosymbolic AI starts together with connectionism itself, judging from the title of the 1943 paper by McCulloch and Pitts, *A Logical Calculus of the Ideas Immanent in Nervous Activity*, and with the work of John von Neumann that led to his 1952 *Lectures on Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components*. Their work indicate that the gap between neural network's continuous representation and logic's discreteness was not seen as a large gulf that later separated AI into symbolic and sub-symbolic AI. Even Alan Turing's 1948 *Intelligent Machinery* introduced a type of neural network called a B-type machine. All of this, of course, before the term Artificial Intelligence was coined ahead of the now famous Dartmouth Workshop in 1956. After Dartmouth the field separated in two: symbolic AI (with

[5] See `www.nesy-ai.org`

its expert systems becoming very successful in the 1980's) and connectionism (which led to the exceptional success of deep networks since 2015). This very unfortunate separation of the field in two has slowed down progress as the research community went their separate ways with different conferences, journals and associations.

I argue that the time is now right for revisiting the approaches of the founding fathers of computer science and developing neurosymbolic AI that is fit for the 21st century. As a concrete step in this direction, I refer the reader to [29] which takes a variation of the neural networks for which the Nobel Prize in Physics was awarded in 2024 and provides a constructive proof of the correspondence between such networks and propositional logic.

6 Conclusion

Agentic AI will require guardrails once it is allowed to take action on anyone's behalf, even if that action is something as simple as organizing a holiday, paying for the air tickets and deciding on the sightseeing tour schedule. Three years from the release of ChatGPT, LLMs continue to hallucinate even with skyrocketing costs of post-hoc model alignment. AI will require systems that never hallucinate, that reason reliably, that can handle novelty and treat exceptions requiring less data.

Neurosymbolic AI integrates learning and reasoning as part of model development by following a development cycle known as the neurosymbolic cycle: (i) extract symbolic knowledge descriptions from partially trained networks, (ii) reason formally about what has been learned, (iii) compress the network by instilling consolidated knowledge back into the network, closing the cycle ahead of further training with data. Reasoning in neurosymbolic AI follows the tradition of knowledge representation founded on logic and formal definitions of a semantics for deep learning [24], rather than based on informal evaluations of reasoning capabilities using benchmark data. Evaluating neural networks with respect to formally-defined, sound or approximate reasoning allows for a much needed controlling of the accumulation of errors within the network.

The use of distillation is a step in the direction of neurosymbolic AI in that distillation is a form of knowledge extraction to obtain network compression (see steps (i) and (iii) of the neurosymbolic cycle). The use of test-time compute is also a step in the direction of neurosymbolic AI because reasoning is, in essence, implemented in a computer by means of a search process (step (ii) of the neurosymbolic cycle). However, there are many forms of knowledge representation and reasoning to be mapped onto networks (analogy, modal, epistemic and higher-order logic, temporal, normative, abductive, abstract reasoning). Distillation without explicit knowledge, that is, without semantic description, is limited in what it can offer to reliability, explanation, knowledge reuse, transfer and curriculum learning.

Broadly speaking, neurosymbolic AI is tasked with the theory, algorithms and tools capable of establishing a principled understanding of the correspondence that exists between neural and symbolic representations. Ultimately, the outcome of neurosymbolic AI has to include achieving safety via verifiable descriptions of network modules, energy efficiency via parsimonious learning from data and knowledge, fairness by imposing requirements specified in logic, and trust by empowering users of AI with the help of explainability and improved system interaction.

AI will soon require systems that adapt to novelty from only a few examples, that check their understanding, that can multi-task and reuse knowledge to improve data efficiency and that can reason in sophisticated ways using first-order and higher-order logic. Adapting to novelty requires an ability to create abstract, simple representations (whether in the brain or the mind) but also to change representations from time to time as the need arises [17]. Change of representation allows looking at a problem from a different angle to obtain new insight and handle analogous situations. An explicit description is one that can be manipulated by asking the question *what might happen if I were to make this change?*, without making the change. Hence, a description is required to be amenable to symbolic manipulation as much as a neural network is required to handle data efficiently.

Finally, AI is not only changing the world of employment, but also education. Learning at schools and universities will need to change in order to instill a culture of critical and creative thinking from the start, of learning from the history of science to cultivate the values of scientific inquiry, reasoning under uncertainty and skeptical interrogation. AI will soon be taught at schools with the goal of creating a more discerning and informed society. Leaders, decision makers and domain experts should probably also learn the basics of AI. Learning whether or not to trust the output of LLMs is hard and will require the help of technological advancements such as explainable and neurosymbolic AI, but also a new economic and social contract, empowering local democracy, accountability, requiring fast policy decisions in a very fast-changing world. The next decade of research in neurosymbolic AI should make key strides towards addressing these era-defining challenges.

References

[1] Recurrent neural networks: Models, capacities and applications. Dagstuhl Seminar 08041, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2008.

[2] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

[3] Adrien Bennetot, Ivan Donadello, Ayoub El Qadi El Haouari, Mauro Dragoni, Thomas Frossard, Benedikt Wagner, Anna Sarranti, Silvia Tulli, Maria Trocan, Raja Chatila, Andreas Holzinger, Artur S. d'Avila Garcez, and Natalia Díaz-Rodríguez. A practical tutorial on explainable ai techniques. *ACM Computing Surveys*, 57(2):1–44, January 2025.

[4] Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jiménez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors. *Neural-Symbolic Learning and Reasoning*, volume 14979 of *Lecture Notes in Computer Science*, Barcelona, Spain, 2024. Springer.

[5] Swarat Chaudhuri. Neurosymbolic program synthesis. In *Handbook on Neurosymbolic AI and Knowledge Graphs*, volume 400 of *Frontiers in Artificial Intelligence and Applications*, pages 532–549. IOS Press, 2025.

[6] Byron Cook. Your AI strategy needs mathematical logic. *Fortune, https://fortune.com/2025/09/20/aws-scientist-why-hallucinations-are-ai-llm-greatest-asset/*, September 2025. Opinion piece.

[7] Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *Artif. Intell. Rev.*, 56(11):12387–12406, 2023.

[8] Artur d'Avila Garcez, Luis C. Lamb, and Dov M. Gabbay. *Neural-Symbolic Cognitive Reasoning.* Springer, 2008.

[9] Artur d'Avila Garcez and Simon Odense. Neurosymbolic deep learning semantics. *arXiv preprint 2511.02825*, 2025.

[10] Artur S. d'Avila Garcez, Tarek R. Besold, Marco Gori, and Ernesto Jiménez-Ruiz, editors. *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning*, volume 3432 of *CEUR Workshop Proceedings*, La Certosa di Pontignano, Siena, Italy, July 3-5 2023. CEUR-WS.org.

[11] DeepSeek-AI, Xiao Bi, Deli Chen, and Guanting Chen et al. Deepseek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, January 2024. Submitted January 5, 2024.

[12] Leilani H. Gilpin, Eleonora Giunchiglia, Pascal Hitzler, and Emile van Krieken, editors. *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, volume 284 of *Proceedings of Machine Learning Research*, UC Santa Cruz, Santa Cruz, CA, USA, 2025. PMLR.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Pascal Hitzler and Md Kamruzzaman Sarker, editors. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications.* IOS Press, Amsterdam, 2021.

[15] Pascal Hitzler, Md Kamruzzaman Sarker, and Aaron Eberhart, editors. *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications.* IOS Press, Amsterdam, 2023.

[16] Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z. Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, Hussain Masoom, Ottavia Bertolli, Tom Zahavy, Amol Mandhane, Jessica Yung, Iuliya Beloshapka, Borja Ibarz, Vivek Veeriah, Lei Yu, Oliver Nash, Paul Lezeau, Salvatore Mercuri, Calle Sönne, Bhavik Mehta, Alex Davies, Daniel Zheng, Fabian Pedregosa, Yin Li, Ingrid von Glehn, Mark Rowland, Samuel Albanie, Ameya Velingker, Simon Schmitt, Edward Lockhart, Edward Hughes, Henryk Michalewski, Nicolas Sonnerat, Demis Hassabis, Pushmeet Kohli, and David Silver. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 2025. Published online: 12 November 2025.

[17] Daniel Kahneman. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, 2011.

[18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 2018.

[19] Luís C. Lamb, Artur S. d'Avila Garcez, Marco Gori, Marcelo O. R. Prates, Pedro H. C. Avelar, and Moshe Y. Vardi. Graph neural networks meet neural-symbolic computing: A survey and perspective. *CoRR*, abs/2003.00330, 2020.

[20] John McCarthy. Epistemological problems for connectionism. *Behavioral and Brain Sciences*, 11(1):44, 1988. Commentary on Smolensky's "On the proper treatment of connectionism".

[21] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. 2024.

[22] Kwun Ho Ngan, James Phelan, Esma Mansouri-Benssassi, Joe Townsend, and Artur S. d'Avila Garcez. Closing the neural-symbolic cycle: Knowledge extraction, user intervention and distillation from convolutional neural networks. In *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, Siena, Italy, July 3-5, 2023*, volume 3432 of *CEUR Workshop Proceedings*, pages 19–43. CEUR-WS.org, 2023.

[23] Kwun Ho Ngan, James Phelan, Joe Townsend, and Artur d'Avila Garcez. Symbolic knowledge extraction and distillation into convolutional neural networks to improve medical image classification. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Yokohama, Japan, June 30–July 5 2024.

[24] Simon Odense and Artur d'Avila Garcez. A semantic framework for neurosymbolic computation. *Artif. Intell.*, 340:104273, 2025.

[25] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[26] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross J. Anderson. The curse of recursion: Training on generated data makes models forget. *CoRR*, abs/2305.17493, 2023.

[27] Paul Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23, 1988.

[28] Harald Strömfelt, Luke Dickens, Artur d'Avila Garcez, and Alessandra Russo. Formalizing consistency and coherence of representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6873–6885. Curran Associates, Inc., 2022.

[29] Son Tran, Edjard Mota, and Artur d'Avila Garcez. Reasoning in neurosymbolic AI. *arXiv preprint 2505.20313*, 2025.

[30] Trieu Trinh, Yuhuai Wu, Quoc Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 2024.

[31] Leslie G. Valiant. Robust logics. *Artificial Intelligence*, 117(2):231–253, 2000.

[32] Benedikt J. Wagner and Artur d'Avlia Garcez. A neurosymbolic approach to AI alignment. *Neurosymbolic AI journal*, 1:1–12, August 2024. IOS Press.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[34] Adam White and Artur S. d'Avila Garcez. Measurable counterfactual local explanations for any classifier. In Giuseppe De Giacomo et al., editor, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 Aug - 8 Sep 2020, Santiago de Compostela, Spain*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2529–2535. IOS Press, 2020.

[35] Adam White, Kwun Ho Ngan, James Phelan, Kevin Ryan, Saman Sadeghi Afgeh, Constantino Carlos Reyes-Aldasoro, and Artur S. d'Avila Garcez. Contrastive counterfactual visual explanations with overdetermination. *Machine Learning*, 112(9):3497–3525, May 2023.