Report from Dagstuhl Seminar 25291

(Actual) Neurosymbolic AI: Combining Deep Learning and Knowledge Graphs

Pascal Hitzler*[1], Cogan Shimizu*[2], Daria Stepanova*[3], and Frank van Harmelen*[4]

1 **Kansas State University – Manhattan, US.** phitzler@googlemail.com
2 **Wright State University – Dayton, US.** cogan.shimizu@wright.edu
3 **Bosch Center for AI – Renningen, DE.** daria.stepanova@de.bosch.com
4 **VU Amsterdam, NL.** frank.van.harmelen@vu.nl

Abstract

In the past decade, both deep learning (DL) and knowledge graphs (KGs) have seen astonishing growth and groundbreaking milestones – DL due to newly available resources (e.g., accessibility of (modern) web scale data), previously un-scalable techniques (e.g., transformers), and modern hardware; KGs due to successful standardization, web-scale integration, and previously un-scalable techniques for querying and inference. This has brought new and increased interest to both fields, and especially in how they can complement each other. This report documents the program and the outcomes of Dagstuhl Seminar 25291 "(Actual) Neurosymbolic AI: Combining Deep Learning and Knowledge Graphs". This Dagstuhl Seminar brought 34 internationally recognized experts together to examine the gap between deep learning and knowledge graphs, and architect their integration: neurosymbolic AI.

Seminar July 13–18, 2025 – http://www.dagstuhl.de/25291
2012 ACM Subject Classification Information systems → Semantic web description languages; Theory of computation → Automated reasoning; Theory of computation → Description logics; Theory of computation → Semantics and reasoning; Human-centered computing → Human computer interaction (HCI); Computing methodologies → Machine learning
Keywords and phrases deep learning, knowledge graphs, neurosymbolic ai
Digital Object Identifier 10.4230/DagRep.15.7.53

1 Executive Summary

Cogan Shimizu (Wright State University, US, cogan.shimizu@wright.edu)
Pascal Hitzler (Kansas State University, US, phtizler@googlemail.com)
Daria Stepanova (Bosch Center for AI, DE, daria.stepanova@de.bosch.com)
Frank van Harmelen (VU Amsterdam, NL, frank.van.harmelen@vu.nl)

Run un-conference style, with merely three set presentations for topic introductions, the participants decided on themes of discussion groups within the theme of the seminar, and on the goal of providing a written account, found in this report, of the emerging themes, structured into definitions, ambitions, challenges, and the state of the art. The themes that emerged are the themes of the Breakout Group Reports found herein: Defining Neurosymbolic Systems; Symbol Emergence; Small Data and Neurosymbolic AI; Explainable AI; Neurosymbolic AI in the Age of Generative AI; Knowledge Graphs and Ontologies in Neurosymbolic

* Editor / Organizer

(Actual) Neurosymbolic AI: Combining Deep Learning and Knowledge Graphs, *Dagstuhl Reports*, Vol. 15, Issue 7, pp. 53–123
Editors: Pascal Hitzler, Cogan Shimizu, Daria Stepanova, and Frank van Harmelen
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Figure 1 A picture of the NeSy Theory cluster of topics. The specific colors do not have meaning. The dots signify a "vote" rather than repeating a new version of the sticky.

Systems; Cognition and Neurosymbolic AI; Benchmarks in the Neurosymbolic Ecosystem; and Real-World Applications in Neurosymbolic Artificial Intelligence. Additional discussions evolved around the general question of how the two major outlets for Neurosymbolic AI – the Neurosymbolic Learning and Reasoning Conference,[1] and the Neurosymbolic AI journal[2] – can best support the nascent community. Key drivers of both outlets were in attendance at the seminar and have already begun to set some of the discussion results into motion.

The Seminar

Thirty-four participants were finally able to attend the seminar. We believe that the diversity of our attendees was both fair, broad, and representative of the breadth of the fields: bridging seniority, gender diversity, geographic location, industry vs. academia, and expertise in neural or symbolic (or both) AI systems. But even through the variety, there were interesting through-lines and other connections.

We began introductions through a novel experience: announcing an interesting or otherwise memorable failure [of our own]. In particular, we were interested in "What have you tried, that just didn't seem to work?" This process set us on even ground: we are equal in our setbacks, and we were there to help each other overcome them.

This seminar followed an *unconference*-style. This means that there was only a very loose structure. There were only three imposed talks, given at the start of the first three days. These talks were given by well-regarded figures (see Section 3, to give broad, deep,

[1] https://nesy-ai.org/
[2] https://neurosymbolic-ai-journal.com/

and historical perspectives of neurosymbolic AI. The remainder of each day was organized into breakouts. These breakout groups were self-selected and even self-generated. After the initial roll-call, we performed a group exercise:

- We wrote down any number of topics that we wanted to tackle this week onto a sticky.
- We placed the topic-sticky onto the chalkboard at the front of the seminar room.
- Small dot stickers were provided for attendees to *upvote* specific topic-stickies.
- We collectively clustered the stickies and identified a theme.

While this reliance on attendee-participation was high-risk, it was also certainly high-reward. However, it was our intuition that the selected participants would be both amenable to this style, but also collegiate in their collective regard for each other, allowing for open, fluid, and vibrant discussion. Indeed, we believe that our risk paid off, and has culminated in this report, as follows.

Overview of the Report

The report is organized into two parts. Section 3 provides the abstracts for the opening talks of the first three days of this seminar. Section 4 contains reports that are jointly written by each breakout group, as described above. Each report provides an overview of the topic, addresses ambitions and challenges, and describes the state of the art.

2 Table of Contents

3 Overview of Opening Talks

For the first three days the seminar, we had an invited talk that would set the stage of Neurosymbolic AI from different perspectives. We provide the abstracts for the latter two. The first talk, given by Frank van Harmelen provided a conversational platform for discussing many of the open problems that our field faces. The second talk, given by Artur d'Avila Gartez, provided both an in-depth discussion on the *technical* aspects of neurosymbolic AI and a historical perspective. Our final invited talk was given by Ute Schmid, who provided context from the cognitive science, as well as a critical look at how knowledge engineering (broadly defined) has underwritten the last four decades of AI research.

3.1 Neurosymbolic AI

Artur d'Avila Garcez (City – University of London, GB)

Artificial Intelligence (AI) has become the focus of large-scale research endeavors in industry and has changed business practice. This led to important debates around the impact of AI on education and society. Concerns around the reliability, fairness, energy efficiency and accountability in AI were raised by influential thinkers [1]. Many identified the need for well-founded knowledge representation and reasoning to be added to the neural-network approach to AI called deep learning. Neurosymbolic AI has been an active area of research for many years seeking to do just that, bringing together robust learning in neural networks with reasoning and symbolic computation.

It has been argued that building AI systems capable of reliable reasoning and safe and trustworthy AI will require neurosymbolic AI systems capable of integrating sound reasoning and deep learning (DL). Parallels have been drawn between Daniel Kahneman's research on human reasoning and decision making [2] and the so-called AI system 1 and system 2. In this keynote, I review the research in neurosymbolic AI and seek to identify promising directions and challenges for the next decade of machine learning (ML) research from the perspective of neurosymbolic computation.

Specifically, I seek to place more than 20 years of research in the area of neurosymbolic AI known as neural-symbolic integration [3] in the context of the recent explosion of interest and excitement around the combination of deep learning and reasoning. I revisit early theoretical results of fundamental relevance to shaping the latest research, such as the proof that recurrently connected, neural networks compute the semantics of various logic formalisms [4]. I also identify bottlenecks and the most promising technical directions in my view towards the sound representation of learning and reasoning in neural networks.

As well as pointing to the various related and promising techniques in neurosymbolic AI, we aim to help organize some of the terminology commonly used around AI, ML and DL. DL is now recognized as being the efficient computational mechanism upon which data-driven AI is to be realized. ML includes DL but also other forms of machine learning such as decision trees, and AI includes machine learning but also reasoning, planning and other abstract cognitive processes. This distinction is important at this exciting time when AI becomes popularized among researchers and practitioners coming from multiple areas of computer science and from other fields altogether, psychology, cognitive science, economics, medicine, engineering and neuroscience.

Recent months have seen a proliferation of releases of Large Language Models by AI companies culminating with the release of GPT5 by ChatGPT's owner OpenAI. Following various announcements of large-scale government and private investments in AI infrastructure, model training and alignment, and heated debates around the safety and risks of AI such as at the World Economic Forum, it has become clear that AI's lack of reliability persists as Large Language Models continue to "hallucinate". The adoption of Agentic AI also failed to solve the problem despite the claims of reduced hallucinations. It turns out that one bad hallucination is sufficient to destroy trust for a long time.

A lot of the evaluation of current AI is based on benchmarking on reasoning tasks with a rather vague definition of reasoning. Neurosymbolic AI has been studying and formalizing reasoning in neural networks for many years [3]. Neurosymbolic AI has as a requirement treating DL as a computational model capable of learning efficiently from data, but also of reasoning logically from what has been learned, incorporating data and Knowledge, formally specified, and satisfying certain verifiable system properties such as correctness.

I survey some of the prominent forms of neural-symbolic integration from the perspective of distributed and localist representations based on the assumption that representation precedes learning. This is possible without having to create a separation between neural and symbolic approaches, as was the motivation of the founding fathers of AI even before the term AI was coined (as in the case of the 1942 paper by McCulloch and Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity, and Von Neumann's 1952 Lectures on Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components, indicating that the gap between distributed vector representations (embeddings) and localist symbolic representations (logic) was not as large as some might imagine. Even Alan Turing's 1948 Intelligent Machinery introduced a type of neural network called a B-type machine). The term Artificial Intelligence was coined ahead of the now famous Dartmouth Workshop, New Hampshire, in 1956. Since then the field has separated into two: symbolic AI and connectionist AI (or neural networks). I argue that this separation into two has delayed progress.

I show how (parts of a) neural network can be coupled with symbolic descriptions with well-defined semantics. I give examples of how propositional, first-order, modal and temporal logic map to and from generative, encoder-decoder and recurrent networks. I illustrate the application of the neurosymbolic cycle (instil and distil knowledge, reason formally, iterate) in a medical diagnosis scenario. The cycle allows domain experts to explain, ask what-if questions and if necessary intervene in the neural network. I argue that the current Chain-of-Thought (CoT) approach to reasoning is misguided in that it tries to improve reasoning by stepping through the input prompts, ignoring the infinite uses of finite means afforded by the combinations of the inputs and its associated accumulation of errors. In neurosymbolic AI, differently from CoT, knowledge is added to neural networks' architectures or loss functions along with a proof of network convergence to stable states. Whether the interpretation is probabilistic or based on many-valued fuzzy logic [5], the neural network learns a probability distribution and its symbolic counterpart provides explainability, knowledge consolidation across multiple tasks, and even extrapolation beyond data distribution. The intended goal is to avoid diverging results and eliminate hallucinations.

References

1 AI Debate 3: The AGI debate, 24 Dec 2022. `https://www.youtube.com/watch?v=JGiLz_Jx9uI`

2 D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

3 A. d'Avila Garcez, K. Broda, D. Gabbay, *Neural-Symbolic Learning Systems: Foundations and Applications*. Springer, New York, 2002.
4 A. d'Avila Garcez, Luis C. Lamb. Neurosymbolic AI: The 3rd Wave. *Artif Intell Rev* **56**, Springer Nature, New York, 2023.
5 S. Odense and A. d'Avila Garcez. A semantic framework for neurosymbolic computation. *Artificial Intelligence* **340**, Elsevier, 2025.

3.2 A Cognitive Perspective on Neurosymbolic AI

Ute Schmid (Universität Bamberg, DE)

Cognitive and AI research mutually inspire each other (very much in early AI, not so much later on, but currently with rising interest again) – giving rise to novel AI methods (e.g., in XAI) as well as influencing cognitive theories. In the talk I first give a reminder on early work on neural learning and symbolic machine learning focusing on relational and rule learning. I will point out crucial differences between neural and symbolic learning with respect to compositionality and productivity, giving illustrations with learning the recursive rule for solving Tower of Hanoi problems and number series induction by pattern abstraction from few examples (rather than generate-and-test). Then I will focus on interleaving implicit and explicit learning with examples from rule-based explanations for blackbox models. I will point out relations to explainable AI for blackbox models. I will discuss abstract visual reasoning as a challenge problem for neurosymbolic approaches and end the talk by pointing out what I think are core challenges for interdisciplinary NeSy research: (1) learning productive rules by abstraction/generalization from few examples, (2) interleaving perception learning and learning complex rules, (3) better understanding of the human inductive bias (generalize the relevant aspects), (4) introducing meta-cognition to control and evaluate generated output (error monitoring, faithfulness), (5) aligning human and machine learning and reasoning for efficient joint problem solving and decision making.

4 Breakout Group Reports

In the following sections, we describe the outcomes of our Breakout Groups. For each outcome report, we specifically ask four questions (in some variation and in some order):

1. What is it?
2. What is the ambition?
3. What are the challenges?
4. Where are we now?

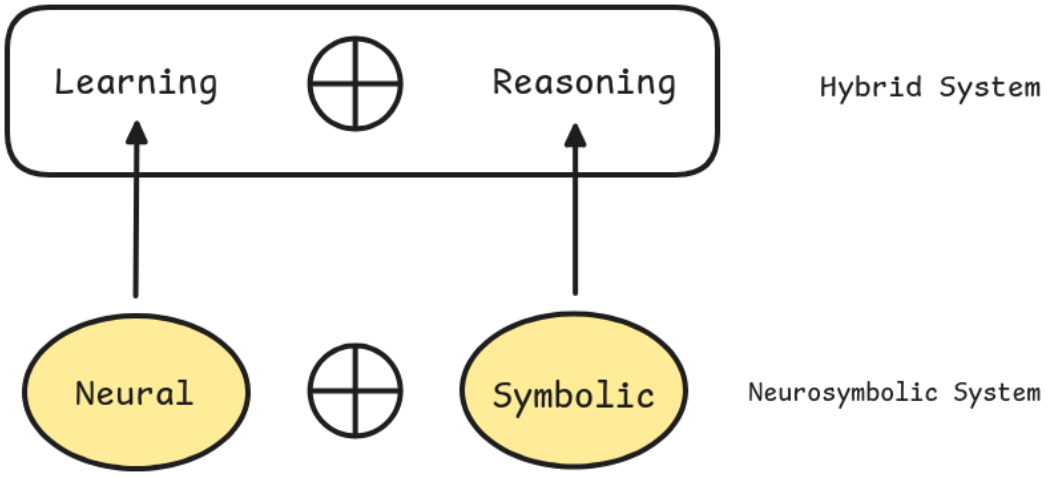

Figure 2 Caption.

4.1 Defining Neurosymbolic Systems

Cogan Shimizu (Wright State University – Dayton, US), Annette ten Teije (VU Amsterdam, NL), Frank van Harmelen (VU Amsterdam, NL)

4.1.1 What are Neurosymbolic Systems?

Neurosymbolic (NeSy) systems aim to unify the strengths of connectionist approaches, such as neural networks, with symbolic reasoning frameworks, such as knowledge graphs and logical inference. This particular variety of hybrid system is motivated by the longstanding observation that neither purely neural nor purely symbolic systems are sufficient to meet the broad demands of general intelligence. While neural networks excel in perception tasks and statistical pattern recognition, they struggle with explainability, data efficiency, and reasoning under constraints. Conversely, symbolic systems offer strong formal guarantees, transparency, and the capacity to integrate explicit knowledge, but are often brittle and lack flexibility in uncertain or noisy environments. This is just one demarcation, which varies in formality, specificity, and perspective.

4.1.1.1 Informal Demarcation

At a high level, neurosymbolic systems operate at the interface of two distinct paradigms, often informally and conveniently characterized along dual axes:

- **System 1 vs. System 2:** fast, unconscious, non-verbal inference (neural) versus slow, deliberate, verbal reasoning (symbolic).
- **Implicit vs. Explicit:** learned representations from data versus human-interpretable rules and knowledge.
- **Black-box vs. Explainable/Interpretable:** opaque computation versus structured reasoning paths.

Yet, for that convenience, we do miss considerable nuance: it is not always the case that either paradigm falls easily or directly onto either axis. Nonetheless, this demarcation

highlights the complementary nature of neural and symbolic components and motivates their integration. A representative description, as articulated by Gartner, defines neurosymbolic AI as "*...a form of composite AI that combines machine learning methods and symbolic systems (for example, knowledge graphs) to create more robust and trustworthy AI models. This combination allows statistical patterns to be combined with explicitly defined rules and knowledge to give AI systems the ability to better represent, reason and generalize concepts. This approach provides a reasoning infrastructure for solving a wider range of business problems more effectively.*"

The key element in this informal demarcation is the combination of a learning component (inferring general patterns from specific instances) and a reasoning component (inferring specific instances from general patterns), as depicted in Figure 2. It does not necessarily hold that the neural (symbolic) component must be the learning (reasoning) component, as there are many ways to compose them or otherwise integrate their properties and functionality.

4.1.1.2 Demarcation & Categorization

While the informal intuition behind neurosymbolic systems is widely shared, precise demarcations remain contested and have changed over time [2, 12]. Key questions include:

- Must the learning component be neural? (e.g., is inductive logic programming [4] included?). This concerns the choice of the learning component, which in the figure has the form of a neural component.
- Must reasoning be symbolic in a strict logical sense? (e.g., is differentiable reasoning [21, 13] in scope? Should LLMs be considered inherently neurosymbolic? Are deep learning systems that solve symbolic tasks neurosymbolic?). This concerns the choice of the reasoning component in the figure.
- How tight should the coupling between the learning and the reasoning component be? This concerns the choice of the $\oplus$ operator in the figure that combines particular learning and reasoning components at the bottom of the figure (e.g., is graph-RAG with an LLM [15] in scope?).
- What constitutes a *symbol*? That is, what is the boundary between symbolic and neural knowledge?

Note that Figure 2 simplifies (for graphical reasons) a neurosymbolic architecture as the combination of a *single* learning component with a *single* reasoning component, while in the literature, more complex combinations have been shown to be useful [10].

Answering these questions is essential for defining the scope of neurosymbolic AI and constructing a taxonomy of system architectures. These include variations in how knowledge is represented in (the architecture [9]), the loss function [5]), the formalism used (fuzzy logic [3]), or probabilistic logic [19]), and the coupling between learning and reasoning components (tight vs. loose integration [14]).

4.1.2 What is the ambition of a theory of NeSy?

The overarching ambition of the neurosymbolic community is to understand and formalize the design space of neurosymbolic systems. While best practices are currently grounded in empirical insights, the field is maturing toward a more principled foundation.

This includes the development of a body of knowledge that:

- Explains why certain architectural choices of components ("neural" n and "symbolic" s in Figure 2) and operators ($\oplus$) work for particular tasks or domains,

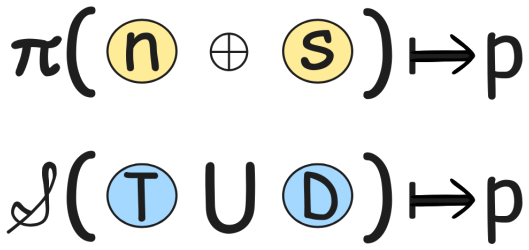

Figure 3 Caption.

- Identifies desirable properties (e.g., safety, robustness, efficiency, scalability, and expressivity) depicted by p in Figure 3.
- Enables formal reasoning about system architecture and behavior, as depicted in Figure 2 as a procedure π that establishes properties p given a specific neurosymbolic architecture.
- Identify tasks and domains (T and D, respectively) for which neurosymbolic approaches would be particularly suitable via $\mathcal{S}$.

In other words, a central goal is to establish a theory of neurosymbolic systems that allows for formal specification, design, and verification. This involves defining input-output behavior, architectural configurations, and desirable properties, and developing proof techniques that ensure certain system properties (safety, efficiency, etc).

4.1.3 What are the challenges for a Theory of NeSy?

The fundamental challenge to a theory of NeSy is the unification of the different views on what NeSy is into a coherent definition, which can be subsequently poured into an appropriate formal form. Such a theory should form the basis for formalisms, methods, and tools for:

- **Specification:** formal descriptions of functional I/O behaviour
- **Design:** formal description of architectures that implement the specification (i.e., the bottom layer in Figure 2).
- **Requirements:** formal definition of desired properties (p in Figure 3).
- **Verification:** proof methods π that derive whether the requirements p hold.

Illustratively, consider a self-driving system deciding whether to stop (specification). A symbolic constraint might enforce that safety conditions must never be violated (requirement). Depending on how these constraints are implemented – embedded in the loss function or enforced as a final reasoning layer (design) – guarantees may differ in strength: a loss function may increase the likelihood of the requirement being met, but will not guarantee it, whereas a final reasoning layer would (verification).

Such a theory of neurosymbolic systems should satisfy the following desiderata, and by their nature, each a challenge unto themselves:

- Descriptive: the theory must describe a broad family of NeSy systems.
- Predictive: the theory must allow us to analytically predict (derive) properties, using operations such as refinement, composition and abstraction, going beyond the purely empirical observations in the current literature.
- Prescriptive: such properties should form the basis for design guidelines that will for the first time advise practitioners which architecture to use for a given set of quality attributes.

- Compositional: the theory would have to be compositional, in the sense that any properties p that the theory derives (π) about a particular neurosymbolic configuration should be a function of the individual components (neural, symbolic) and the way they have been composed ($\oplus$).

4.1.4 Where are we now?

Although individual systems are typically based on solid theoretical grounds (e.g., probabilistic logic for DeepProbLog, Description Logic for Ontolearn, fuzzy logic for LTNs), theories about the entire class of neurosymbolic systems have until now been limited to categorizations of such systems, and have not yet had the shape of a theory as we outlined in two preceding sections. These existing categorizations remain largely informal and only fulfill part of the desired properties from the preceding section:

Table 1 Caption.

Desiderata	**[8]**	**[20]**	**[10]**	**[6]**	**[1]**	**[17, 16]**
Descriptive	+	-	+	-	-	+
Predictive	-	+/-	+/-	+	-	-
Prescriptive	-	+/-	-	+	+	-
Compositional	-	-	-	-	-	+
Formal	-	-	-	+	-	-

Recent developments propose more formal theories of neurosymbolic systems [14, 11], with [14] capturing one of the six patterns proposed in [8]. Although very different in nature, a final noteworthy recent attempt is Uller [18], a Python library for neurosymbolic systems, where (although not intended as such), the abstractions proposed in the library can be seen as "theory" about neurosymbolic architectures.

We see all of these as important precursors for an all-encompassing theory of neurosymbolic systems that is at the same time powerful, has broad coverage, and will be the basis for a well-founded design practice for neurosymbolic systems.

References

1 Elvira Amador-Domínguez, Emilio Serrano, and Daniel Manrique. Neurosymbolic system profiling: A template-based approach. *Knowl. Based Syst.*, 287:111441, 2024.

2 Sebastian Bader and Pascal Hitzler. Dimensions of neural-symbolic integration - A structured survey. In Sergei N. Artëmov, Howard Barringer, Artur S. d'Avila Garcez, Luís C. Lamb, and John Woods, editors, *We Will Show Them! Essays in Honour of Dov Gabbay, Volume One*, pages 167–194. College Publications, 2005.

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

4 Andrew Cropper and Sebastijan Dumancic. Inductive logic programming at 30: a new introduction. *CoRR*, abs/2008.07912, 2020.

5 Claudia d'Amato, Nicola Flavio Quatraro, and Nicola Fanizzi. Injecting background knowledge into embedding models for predictive tasks on knowledge graphs. In Ruben Verborgh, Katja Hose, Heiko Paulheim, Pierre-Antoine Champin, Maria Maleshkova, Óscar Corcho, Petar Ristoski, and Mehwish Alam, editors, *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings*, volume 12731 of *Lecture Notes in Computer Science*, pages 441–457. Springer, 2021.

6 Charles Dickens, Connor Pryor, Changyu Gao, Alon Albalak, Eriq Augustine, William Yang Wang, Stephen J. Wright, and Lise Getoor. A mathematical framework, a taxonomy of

modeling paradigms, and a suite of learning techniques for neural-symbolic systems. *CoRR*, abs/2407.09693, 2024.

7 Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic ai: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, Nov 2023.

8 Henry A. Kautz. The third AI summer: AAAI robert s. engelmore memorial lecture. *AI Mag.*, 43(1):93–104, 2022.

9 Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: neural probabilistic logic programming. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3753–3763, Red Hook, NY, USA, 2018. Curran Associates Inc.

10 Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, 328:104062, 2024.

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

12 Md. Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-symbolic artificial intelligence. *AI Commun.*, 34(3):197–209, 2021.

13 Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

14 Lennert De Smet and Luc De Raedt. Defining neurosymbolic AI. *CoRR*, abs/2507.11127, 2025.

15 Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560, 2024.

16 Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. Modular design patterns for hybrid learning and reasoning systems. *Appl. Intell.*, 51(9):6528–6546, 2021.

17 Frank van Harmelen and Annette ten Teije. A boxology of design patterns forhybrid learningand reasoning systems. *J. Web Eng.*, 18(1-3):97–124, 2019.

18 Emile van Krieken, Samy Badreddine, Robin Manhaeve, and Eleonora Giunchiglia. ULLER: A unified language for learning and reasoning. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jiménez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, *Neural-Symbolic Learning and Reasoning - 18th International Conference, NeSy 2024, Barcelona, Spain, September 9-12, 2024, Proceedings, Part I*, volume 14979 of *Lecture Notes in Computer Science*, pages 219–239. Springer, 2024.

19 Emile van Krieken, Thiviyan Thanapalasingam, Jakub M. Tomczak, Frank van Harmelen, and Annette ten Teije. A-nesi: A scalable approximate method for probabilistic neurosymbolic inference. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

20 Laura von Rüden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning - A taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.*, 35(1):614–633, 2023.

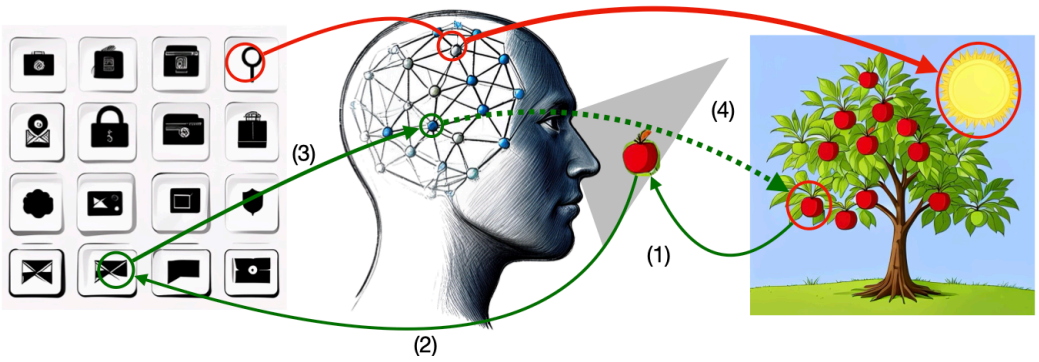

Figure 4 The Symbol Emergence Loop: An agent (1) observes a new concept in the environment, tries to map to an existing symbol, (2) realises that there is a semantic gap with the current symbol systems, and (3) decides to create a new symbol.

21 Po-Wei Wang, Priya L. Donti, Bryan Wilder, and J. Zico Kolter. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. *CoRR*, abs/1905.12149, 2019.

4.2 Symbol Emergence

Riccardo Tommasini (INSA Lyons, FR), Luciano Serafini (Fondazione Bruno Kessler (FBK) – Trento, Italy), Giuseppe Marra (KU Leuven, BE), Jay Pujara (University of Southern California, US), Gustav Šír (Czech Technical University, CZ), Natalia Díaz-Rodríguez (University of Grenada, ES)

4.2.1 What is Symbol Emergence?

This section investigates the conditions under which new symbols should be introduced in the learning process, and how this impacts the semantics and performance of NeSy. Symbol emergence (SE) refers to a symbol being identified from data, gaining recognition as a discrete concept, and becoming useful for reasoning or communicating to one or more agents, according to some pre-defined utility measure. In NeSy, symbols are objects that sit at the boundary between neural and symbolic components, enabling the coupling of these two systems. Symbol emergence provides the seeds for symbol grounding, enabling agents to establish a shared vocabulary for communication.

The symbol emergence problem fills a gap in the "symbol grounding problem" [6], which is defined as endowing agents with the means to autonomously create internal representations that link their manipulated symbols to corresponding referents in the external world. However, the origin of these symbols remains underspecified in the AI community.

4.2.2 What is the NeSy Ambition?

Unlike approaches where symbols are predefined, symbol emergence aims to capture how agents can autonomously form, adapt, and communicate symbolic structures in a manner

that is both grounded in data and useful for reasoning. To sharpen this vision, we suggest some questions:

- What are the desiderata for "good" symbols, and what processes benefit from them?
- How are symbols formed, and what drives their emergence?
- What are the principles according to which symbols emerge?
- What are the processes that underlie symbol creation, adaptation, and forgetting?
- Is a symbol sufficient or necessary for reaching the agent's goal?

Our analytical discussion around these questions led us to formulate the **GRACE^{2}ful** NeSy ambition, a vision for neurosymbolic systems that are generative, robust, adaptable, communicative, efficient, and explainable.

- **Generative (formerly Creative):** Neurosymbolic systems should be capable of inventing new symbols whose semantics are not pre-specified by humans, thereby avoiding human biases. Such creativity goes beyond recombination: it entails analogical reasoning that reuses existing symbolic structures to generate novel conceptualisations. Example: Extending the relation between "apple" and "peel" to reason about "earth" and "crust".
- **Robust:** Emergent symbols must remain stable across variation, noise, and context. A robust NeSys generalises reliably, identifying symbols consistently even under changes in environment or modality. Example: Recognising "apple" whether it appears in sunlight, shadow, or partial occlusion, as well as if it appears in different colours, like in the image above (red and green).
- **Autonomous (formerly Adaptable):** Symbols should not be fixed a priori. Instead, NeSy will continuously refine and expand its symbolic boundaries by autonomously adapting its symbolic repertoire in response to task demands and new input. Example: Introducing "apple" when predicting whether fruits have peels, without prior specification.
- **Communicative:** Symbols must be transferable and composable across agents, humans, or systems, enabling alignment and effective knowledge exchange. Therefore, NeSy will enable knowledge alignment with other systems, agents, or humans by supporting symbols that are appropriately abstracted. For example, a system can provide a description of an apple in either English or French, depending on the user's nationality with whom it is interacting.
- **Efficient:** Emergent symbols should decrease data requirements and computational costs while improving reasoning. For example, a system will be able to use the (emergent) symbol of "apple" to more quickly classify healthy meals. NeSy will benefit from operating on Small Data, reducing reliance on labels, or increasing the efficiency of reasoning.
- **Explainable:** Symbols should make reasoning transparent. Example: Tracing a misclassification of pears as apples back to the boundaries of the symbol "apple". NeSy will be debugged by breaking down processes by their relevant features or sub-processes. For example, mistakes in a prediction can be better understood by inspecting the symbol "apple" and whether it also refers to pears.

4.2.3 What are the challenges, and how to address them?

Realizing the **GRACE^{2}ful** ambition requires overcoming several challenges. Below, we highlight them and relate them to the ambition via Table 2.

- **Criteria for Symbol Emergence.** A central challenge is determining when a new symbol should be introduced. Without clear criteria, systems risk ambiguity, redundancy, or overly narrow symbols. The solution must balance abstraction (generalizing across

Challenge/Property	G	R	A	C	E	E
Criteria for Symbol Emergence	✓		✓			✓
Continuous and Efficient Induction	✓		✓		✓	
Managing Complexity		✓			✓	✓
Semantic Alignment		✓		✓		✓
Shared World Models and ToM		✓	✓	✓		

Table 2 GRACE^2-ful symbols: **G** – generative; **R** – robust; **A** – autonomous; **C** – communicative; **E** – efficient ; **E** – explainable.

trivial variations) with sensitivity (capturing task-relevant distinctions). This challenge is closely tied to the **Generative**, **Autonomous**, and **Explainable** dimensions.
  - *Example:* a fruit recognition system may only need "apple", while a grocery system must distinguish between "Fuji" and "Honeycrisp".
- **Continuous and Efficient Induction.** Beyond criteria, NeSy systems need mechanisms for automatic, continuous, and efficient induction of symbols. Such processes must model symbol semantics (relations between new and existing concepts) and remain scalable. This connects to **Generative**, **Autonomous**, and **Efficient**.
  - *Example:* an agent monitoring sensor streams introduces an "anomaly cluster" when distributions shift.
- **Managing Complexity.** Symbolic reasoning risks combinatorial explosion, where possible symbol combinations grow unmanageably. Systems must introduce symbols that aid efficiency through pruning or prioritisation. This challenge relates to **Robustness**, **Efficiency**, and **Explainability**.
  - *Example:* avoiding fruit–colour–size combinations unless task-relevant.
- **Semantic Alignment.** Emergent symbols must be grounded in perceptual regularities and aligned across two or more agents (who may be humans or machines). Misalignment risks semantic drift. This ties to **Robust**, **Communicative**, and **Explainable**.
  - *Example:* one system uses "apple" only for red apples, another for all apples.
- **Shared World Models and Theory of Mind.** Symbol emergence requires overlapping knowledge or experiences among agents to disambiguate a novel symbol. Indeed, some overlap in experiences is needed to disambiguate potential polysemanticity in symbols. Yet some emergent concepts (textures, scales, properties) resist symbolic capture. This links to **Robust**, **Autonomous**, and **Communicative**.
  - *Example:* two autonomous vehicles must share a compatible notion of 'obstacle'.

4.2.4 Where are we now?

Neurosymbolic systems perform symbol grounding in various ways, including fine-tuning neural networks, clustering, human-guided mapping, or assuming a priori neural-symbolic grounding exists. As NeSy systems incorporate SE, these grounding techniques must be extended to support newly introduced symbols. Several explorations of symbol emergence provide guideposts for designing SE modules in NeSy systems. In this aim, numerous research topics in related fields can benefit NeSy. Taniguchi et al. provide a wide-ranging study of symbol emergence across many fields and consider symbol emergence in robotics [8]. Silver and Mitchell examine the interaction of concepts and symbols in human cognition, utilising evidence from functional magnetic resonance imaging (fMRI) experiments. When examining the literature in Machine Learning, Cognitive Psychology, Social Science, and related fields,

we find a range of techniques closely connected to the problem of symbol emergence (SE). Below, we offer a necessarily incomplete list of approaches from various domains that relate to this issue.

- **Machine learning** has investigated different techniques along which symbolic, sparse or discrete representations arise, independent of possible symbolic processes (reasoning). According to information-theoretic principles, discrete symbols arise from the balance between minimal description length (for compression) and maximal mutual information with the task (for expressiveness).
- When learning **sparse and compressed representations**, guided by principles like information bottleneck and regularisation, naturally promote the emergence of discrete, symbol-like structures by enforcing abstraction and compactness [2, 1].
- **Object-centric learning** aligns with symbol emergence by promoting the disentanglement of scenes into discrete, compositional entities that can be referenced symbolically (e.g., [5]).
- **Causal and disentangled representations** aim to isolate underlying generative factors, which are prime candidates for interpretable and symbolic abstractions [9].
- **Conceptual clustering** partitions continuous data into discrete groups, effectively inducing symbolic categories that can serve as building blocks for higher-level reasoning in neurosymbolic systems. **Incremental clustering** extends clustering methods to cope with data streams that require cluster update (extension/contraction) depending on the data coming in.
- **Structure learning** in the form of program induction or predicate invention (as in Inductive Logic programming) directly engages with symbolic representations, formalizing the emergence of new symbols and relations from data. Symbols can emerge through interaction with humans or other agents, as shared meanings are gradually negotiated and grounded in communicative behaviour and mutual experience [4, 7].
- **Architectural biases** in machine learning models, such as discrete relaxations (Gumbel-softmax [3]) or sparsity constraints, encourage neural models to adopt symbolic-like behavior, facilitating differentiable approximations of symbolic reasoning.

References

1 Javier Blanco-Romero, Vicente Lorenzo, Florina Almenares Mendoza, and Daniel Díaz-Sánchez. Machine learning predictors for min-entropy estimation, 2024.

2 Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

3 Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.

4 Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

5 Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

6 Dairon Rodríguez, Jorge Hermosillo, and Bruno Lara. Meaning in artificial agents: The symbol grounding problem revisited. *Minds Mach.*, 22(1):25–34, 2012.
7 Luc Steels. Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312, 2003.
8 Tadahiro Taniguchi, Justus H. Piater, Florentin Wörgötter, Emre Ugur, Matej Hoffmann, Lorenzo Jamone, Takayuki Nagai, Benjamin Rosman, Toshihiko Matsuka, Naoto Iwahashi, and Erhan Öztop. Symbol emergence in cognitive developmental systems: A survey. *IEEE Trans. Cogn. Dev. Syst.*, 11(4):494–516, 2019.
9 Kevin Xia and Elias Bareinboim. Neural causal abstractions. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 20585–20595. AAAI Press, 2024.

4.3 Small Data and Neurosymbolic AI

Filip Ilievski (VU Amsterdam, NL), Axel-Cyrille Ngonga Ngomo (University of Paderborn, DE), Hande McGinty (Kansas State University – Manhattan, US), Valentina Tamma (University of Liverpool, UK)

4.3.1 What is it?

This section assumes NeSy systems applied to supervised machine learning. In this setting, the systems are provided with a set of training data items $D \subseteq X \times Y$, where X is drawn from the set of all possible inputs U, and Y is the corresponding output. The systems then aim to approximate an optimal target function $f : U \to Y$ by learning $\hat{f} : U \to Y$. The performance of $\hat{f}$ is finally measured in terms of bounded performance measures such as accuracy, F-measure, mean reciprocal rank, and hits@n. Within supervised machine learning, small data is commonly understood in terms of limited training data in terms of **size** [7, 26, 60] or **statistical moments** prior to training [57]. While other definitions related to data models are also found in the literature [58, 48], these go beyond the scope of this section.

The first interpretation of small data is often seen as a counterpart to the large amount of data needed to train deep learning systems. For example, transformer-based large language models are routinely trained with $> 10^{12}$ tokens before being aligned using thousands of annotated data samples [8]. Given that the provision of such samples is often costly, the ability of systems to achieve high performance after having been trained on limited magnitude data sets D is aimed at reducing annotation time and costs, training time, and even improving energy efficiency. The second definition of small data is correlated with – but not equivalent to – the first one and refers to data with a small data statistic T (e.g., entropy) in comparison to the dimensionality of the problem [57]. This definition equates to regarding data as small if the distribution of samples within or across classes is skewed with respect to some statistical moment (e.g., frequency, standard deviation). Such limitations are predominant when data annotation costs are high (for example, in medicine, where image annotation costs can reach $50 USD/image) or when data availability is limited before training (e.g., when building industrial plants).

4.3.2 What is the NeSy Ambition?

NeSy systems seem particularly well suited for being trained with small data (according to both meanings of the term) as they promise to address a key limitation of purely data-driven systems: to limit their dependence on training data to discriminate across instance-class pairs by an explicit use of symbolic (and subsymbolic) background knowledge. Specifically, NeSy AI aims to support small-data learning, reasoning, and evaluation. First, NeSy AI aims to **learn generalizable features** from the data, including implicit features captured in the subsymbolic space (e.g., prototype neurons [22]), concepts (e.g., learning with less than 10 samples for classification tasks [10, 46, 37, 13]), and explicit symbolic features that are meaningful to humans (e.g., using object parts to recognize sketches in computer vision [9]). Meaningful decision-making requires that the features in the symbolic and subsymbolic space (i.e., neurons and symbols) should be at least alignable [50, 43]. Second, NeSy AI promises to enable **knowledge-enriched explicit symbolic reasoning** over the derived symbolic features. Since small data is, in general, unlikely to be sufficient for models to learn how to solve the task at hand, abstracting to symbols provides an opportunity for symbolic reasoning and the incorporation of contextual commonsense and causal knowledge [28]. Third, NeSy AI aims to **develop principled procedures for evaluating small data** use cases. NeSy frameworks can provide meaningful auditing of systems in line with quality attributes [56]. The focus on symbolic representations enables the adoption and adaptation of cognitive mechanisms, such as prototypes, analogy, and rule learning, for benchmarking purposes [29]. Carefully designed human intelligence tests have already inspired a long list of AI benchmarks, such as those in abstract visual reasoning [24, 40]. Conversely, the strong link of NeSy to machine learning enables the introduction of statistical methods that can support meaningful evaluation, such as sampling methods that derive subsets from existing datasets that fit specific statistical properties and possibly simulate long-tail phenomena [47].

Drawing on SotA methods for learning, reasoning, and evaluation makes the NeSy methods a promising fit for a variety of **applications where data collection may be laborious, expensive, or dangerous**. This is apparent for sensitive applications where data collection can be challenging and potentially dangerous to one's health, like medical trials [1, 12]; applications with privacy, security, and ethics concerns like molecular science [15]; and applications where the high cost of error hinders large-scale data acquisition, such as autonomous driving [21, 27]. More broadly, NeSy methods aim to support any task in domains with high irregularities that heavily rely on causal and what-if reasoning, such as traffic [59] and ecology [54] domains. Advancements in learning, reasoning, and evaluation in NeSy methods can improve data efficiency, interpretability, and safety in robotic domains, especially for tasks that require embodied decision-making and structured planning from limited demonstrations [55].

4.3.3 What are the challenges and their mitigations?

When working with small datasets, a fundamental challenge arises from their **limited depth and breadth**. Small data often fails to capture the full range of possible variations and alternative perspectives in a domain. For example, clinical trial data may have some variability within a cohort, it is unlikely that the data set is representative of the demographic diversity, potential side effects, or rare complications. As these datasets cannot be easily extended to include what is missing, injecting background knowledge can help bridge gaps and enrich the dataset with relevant contextual information that is otherwise unavailable. However, this process underscores the necessity of adequate abstractions that preserve essential patterns

while generalizing beyond the sparse data available. Without careful abstraction, the utility of both the limited dataset and the supplementary knowledge may be diminished. Key challenges are applying the most **appropriate abstraction**, the need to **incorporate discrete and continuous representations** (e.g., knowledge graphs into neural networks), and ensuring **trustworthy background knowledge** (e.g., using provenance to compensate for absent or underrepresented aspects of the data).

Various NeSy methods offer a trade-off between the richness of prior knowledge and the burden of integrating it (§4.3.4). *Prototypes* are an approach to abstraction, with a question of where they come from: they can be specified by experts, which is relatively cheap, or learned from data, which requires sufficient coverage and well-structured input. Class *hierarchies* can guide the learning of prototypes [22], but it remains uncertain how well these approaches perform under truly small data constraints. *Ontology building* through middle-out approaches [44, 53, 5, 42, 49, 25], where the class definition is refined to cover some sample data that domain experts provide, may enable prototype learning and support flexible abstraction. While *case-based reasoning* can generalize to training data samples, a key challenge is representing the cases at the right level of granularity/abstraction and devising a corresponding similarity metric [23]. *Causal relationships* can inform generalization in new situations, such as using drug side effects data to infer multiple drug-drug interactions. While causal knowledge injection has strong generalization power, especially for counterfactual or rare-event reasoning, causal knowledge is challenging to obtain and validate, requiring strong assumptions. Across the approaches, a clear regularity emerges: to compensate for limited labeled data, we inject additional knowledge into the learning process, whether in the form of abstractions, background knowledge, prototypes, cases, or constraints, which invariably carry costs: acquiring knowledge, representing it effectively, and ensuring its relevance are non-trivial challenges.

Another increasingly recognized challenge concerns evaluation. Conventional performance metrics, such as accuracy, precision, recall, and F_1 score, enable well-understood ground for comparison, but do not account for data size and can lead to misleading assessments of model reliability in low-resource settings. In the context of NeSy systems, this issue is particularly significant given the interplay between symbolic reasoning and statistical learning. To address this, there is a growing need for evaluation **metrics that explicitly consider the size and distributional characteristics of the training data**. These might include data-aware performance bounds, sample-efficiency metrics, or metrics that incorporate statistical moments. However, while performance-related metrics are relatively well-established, other vital **properties of NeSy systems, such as safety, robustness, and explainability, remain challenging to quantify**. Existing proxy measures for explainability (e.g., fidelity, sparsity, or concept alignment) are still far from standardized and often fail to capture user-centered or domain-specific needs [2]. Similarly, there is **a lack of principled and widely adopted metrics** for evaluating the safety of model outputs, particularly in distribution shift or in high-stakes decision-making contexts. Developing rigorous, interpretable, and context-sensitive metrics for these properties remains an open and pressing challenge in the field.

4.3.4 Where are we now?

NeSy AI approaches have shown promise in learning generalizable features, reasoning over them using knowledge-enriched symbolic methods, and enabling principled evaluation tailored to these capabilities. The integration of symbolic structures (e.g., logic rules, background knowledge, compositional constraints) into neural learning processes can enhance the inherent

semantics of limited or redundant data. Structured priors help constrain the hypothesis space, guide generalization, and support abstraction, all of which are critical in data-scarce and low-diversity settings. We review SotA benchmarks and methods next.

4.3.4.1 Benchmark Datasets and Tasks for Small Data

Benchmarks are essential to evaluate and drive progress in NeSy methods, especially under small-data or low-diversity conditions. Notable examples include:

1. **Abstract visual reasoning benchmarks** [24, 40]
   a. *Abstraction and Reasoning Corpus (ARC)* [11]: ARC focuses on abstract patterns and limited training sets. As such, ARC poses challenges that naturally align with NeSy approaches due to its emphasis on compositionality and generalization from minimal data.
   b. *Raven's Progressive Matrices (RPMs)* [45, 3] test abstract visual reasoning through matrix pattern completion and has been widely adopted to study systematic generalization. NeSy methods excel here by incorporating explicit relational and logical reasoning.
   c. *Bongard Problems* [6]: Classic tests of concept learning and analogical reasoning, Bongard problems require understanding subtle, often symbolic, distinctions from very few examples, making them a canonical challenge for neurosymbolic AI.
   d. *MARVEL* [31]: designed to generalize abstract visual reasoning tasks and to separate perception and inference explicitly. MARVEL tests the ability to learn from limited visual data while leveraging symbolic reasoning modules.
2. **Sketch recognition tasks** [39, 18]: While sometimes debated as a small-data problem, sketch recognition benchmarks involving human-like sparse sketches can require efficient abstraction and symbolic interpretation.
3. **Concept learning benchmarks** [35]: Tasks that test the ability to learn symbolic concepts from few examples, often requiring compositional generalization and abstraction capabilities.
4. **Lateral thinking puzzles** [30, 34]: Tasks that test the ability of models to overwrite commonsense associations, by the virtue of adequate abstractions in textual brain teasers and multimodal rebus puzzles.
5. **Structural analogies** [52, 4] are important drivers of knowledge transfer in humans. Recent work has introduced benchmarks for analogies in structures, specifically in images and stories, inspired by cognitive science theories of analogy like structural mapping [20].

4.3.4.2 Methods for Small Data Scenarios

In both small and low-diversity data regimes, symbolic structures serve as a form of *semantic injection*: providing constraints, abstractions, or prior knowledge that compensates for the limitations of the raw data. Several promising categories of NeSy methods that inject symbolic structures have emerged (summarized in Table 3):

1. **Inductive logic programming (ILP) and differentiable reasoning**: Differentiable ILP frameworks and neural theorem-provers can learn logical rules from limited datasets, particularly when symbolic background knowledge is encoded [17]. These systems benefit from structured search spaces and gradient-based learning to optimize over logical rules.
2. **Transfer learning in symbolic contexts**: In transfer learning, neural networks pre-trained on large corpora are fine-tuned with limited task-specific data. Systems such as DeepProbLog [41] integrate neural predicates into logic programs, allowing gradients to

flow through symbolic reasoning pipelines. Domain-adversarial training [19] can further enhance generalization when no labeled data are available in the target domain.

3. **Prototype-based learning and concept bottlenecks**: Few-shot models such as Prototypical Networks [51] summarize each class with a prototype in latent space. When guided by symbolic descriptors (e.g., logic-derived attributes), they enforce structure and boost generalization. Similarly, Concept Bottleneck Models [33] use human-defined concepts to align internal representations, reducing sample complexity and improving interpretability, especially with low-diversity data.
4. **Neurosymbolic case-based reasoning (CBR)** systems combine neural encoders with symbolic memory or prototype mechanisms [38]. By comparing new inputs with stored prototypes or cases, they improve generalization from few or homogeneous examples, often using auto-encoders or prototype networks to learn structured latent representations.
5. **Knowledge injection and logic-based regularization**: Models like Logic Tensor Networks (LTNs) [14] and DeepProbLog [41] integrate logical constraints and knowledge bases into learning objectives, while others inject background knowledge to guide embedding representations [16] or adapt LLMs [28]. These priors regularize training and improve robustness under data sparsity or low variability.
6. **Causal and compositional generalization**: NeSy models increasingly target abstract reasoning, compositionality, and causality, which are key capacities for robust generalization from limited or redundant data [36]. Structural inductive biases (e.g., causal graphs, symbolic decompositions) enhance learning efficiency in such regimes. Entailment trees, that may be modeled as belief graphs [32], further support learning under small-data conditions by enforcing consistency and structure in reasoning. By explicitly modeling inferential dependencies between symbolic statements, they enable the system to generalize from sparse evidence through accurate, compositional inference.

References

1 Scott Askin, Denis Burkhalter, Gilda Calado, and Samar El Dakrouni. Artificial intelligence applied to clinical trials: opportunities and challenges. *Health and technology*, 13(2):203–213, 2023.

2 Roberto Barile, Claudia d'Amato, and Nicola Fanizzi. Lp-dixit: Evaluating explanations for link predictions on knowledge graphs using large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 4034–4042, 2025.

3 David G T Barrett, Felix Hill, Adam Santoro, Ari S Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. *International Conference on Machine Learning (ICML)*, 2018.

4 Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. VASR: visual analogies of situation recognition. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

5 Eva Blomqvist, Karl Hammar, and Valentina Presutti. Engineering ontologies with patterns–the extreme design methodology. In *Ontology Engineering with Ontology Design Patterns*, pages 23–50. IOS Press, 2016.

6 M. M. Bongard. *Pattern Recognition*. Herzen State Pedagogical Institute, 1970.

7 Lorenzo Brigato and Luca Iocchi. A close look at deep learning with small data. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2490–2497. IEEE, 2021.

8 Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered:

Table 3 Expanded summary of NeSy approaches and their mechanisms for learning from small data.

Approach	Symbolic Component	Neural Component	Small Data Advantage
1. Inductive Logic Programming (ILP) and Differentiable Reasoning			
Inductive Logic Programming (ILP)	Logical rules, background knowledge, Prolog-style programs	Differentiable rule scoring, logic unfolding via neural models	Reduces hypothesis space and enables rule induction from few examples
Differentiable ILP / Neural Theorem Proving	Soft logic operators, logic programs	End-to-end differentiable optimization over rule space	Enables gradient-based learning of logical structures under weak supervision
Logic Tensor Networks (LTN) / Logic-Based Regularization	First-order logic rules translated into differentiable constraints	Neural networks trained under logic-derived loss terms	Encodes symbolic priors as soft constraints, improving sample efficiency
2. Transfer and Probabilistic Logic Models			
DeepProbLog / NeSy Probabilistic Logic	Probabilistic logic programs with symbolic predicates and rules	Neural modules embedded as learnable predicates	Allows few-shot or zero-shot learning by leveraging symbolic logic during training and inference
Domain Adaptation / Transfer Learning with Symbolic Regularization	Source-target mappings, task-level symbolic structure	Domain-invariant neural representations, often adversarially trained	Transfers pretrained knowledge with minimal labeled target data by aligning feature spaces
3. Prototype and Concept-Based Models			
Prototype Networks	Optionally logic-guided feature selection, prototype semantics	Latent space embedding and prototype-based metric learning	Learns class-specific prototypes that support generalization from few labeled examples
Concept Bottleneck Models (CBM)	Concept taxonomy or human-annotated concept space	Neural concept predictors with interpretable bottlenecks	Supervizes intermediate representations, improving generalization and interpretability
Neurosymbolic Case-Based Reasoning (CBR)	Symbolic case library or prototype memory, similarity metrics	Neural encoder-decoder or autoencoder with prototype layer	Enables retrieval-based generalization and reconstruction from small training sets
4. Program Synthesis and Neuro-Guided Search			
Program Synthesis (e.g., DeepCoder, DreamCoder)	Domain-specific language (DSL), symbolic grammars, type constraints	Neural models that guide search or program induction	Solves tasks with few I/O examples by learning efficient symbolic programs
5. Causal and Compositional Generalization			
Causal and Compositional Reasoning Models	Causal graphs, compositional rule structures, abstract hierarchies	Modular neural architectures or neural-symbolic hybrids	Supports abstraction and robust generalization under distributional shifts or novel combinations

A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 2024.

9 Kezhen Chen, Kenneth D Forbus, Balaji Vasan Srinivasan, Niyati Chhaya, and Madeline Usher. Sketch recognition via part-based hierarchical analogical learning. In *IJCAI*, pages 2967–2974, 2023.

10 Kezhen Chen, Irina Rabkina, Matthew D McLure, and Kenneth D Forbus. Human-like sketch object recognition via analogical learning. In *Proceedings of the AAAI Conference*

on Artificial Intelligence, volume 33, pages 1336–1343, 2019.

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

12 Hitesh Chopra, Dong K Shin, Kavita Munjal, Kuldeep Dhama, Talha B Emran, et al. Revolutionizing clinical trials: the role of ai in accelerating medical breakthroughs. *International Journal of Surgery*, 109(12):4211–4220, 2023.

13 Caglar Demir and Axel-Cyrille Ngonga Ngomo. Neuro-symbolic class expression learning. In *IJCAI*, pages 3624–3632, 2023.

14 Ivan Donadello, Luciano Serafini, and Artur d'Avila Garcez. Logic tensor networks for semantic image interpretation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1596–1602, 2017.

15 Bozheng Dou, Zailiang Zhu, Ekaterina Merkurjev, Lu Ke, Long Chen, Jian Jiang, Yueying Zhu, Jie Liu, Bengong Zhang, and Guo-Wei Wei. Machine learning methods for small data challenges in molecular science. *Chemical Reviews*, 123(13):8736–8780, 2023.

16 Claudia d'Amato, Nicola Flavio Quatraro, and Nicola Fanizzi. Injecting background knowledge into embedding models for predictive tasks on knowledge graphs. In *European Semantic Web Conference*, pages 441–457. Springer, 2021.

17 Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

18 Yan Fu, Yibing Xu, Yanning Chen, and Alan L. Yuille. Learning to recognize sketches by spatio-temporal graph convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):3007–3019, 2018.

19 Yaroslav Ganin and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

20 Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

21 Leilani H Gilpin and Filip Ilievski. Neuro-symbolic reasoning in the traffic domain. *J AI Res*, 15(3):123–145, 2021.

22 Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 32–40, 2019.

23 Jerry Zhi-Yang He, Zackory Erickson, Daniel S Brown, Aditi Raghunathan, and Anca Dragan. Learning representations that enable generalization in assistive tasks. In *Conference on Robot Learning*, pages 2105–2114. PMLR, 2023.

24 José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L. Dowe. Computer models solving intelligence test problems: Progress and implications (extended abstract). In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 5005–5009. ijcai.org, 2017.

25 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

26 Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

27 Atalay Mert Ileri, Nalen Rangarajan, Jack Cannell, and Hande McGinty. Vel: A formally verified reasoner for owl2 el profile, 2024.

28 Filip Ilievski. Human-centric ai with common sense, 2024.

29 Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, et al.

Aligning generalisation between humans and machines. *arXiv preprint arXiv:2411.15626*, 2024.

30 Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. Brainteaser: Lateral thinking puzzles for large language models. *EMNLP*, 2023.

31 Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 46567–46592. Curran Associates, Inc., 2024.

32 Nora Kassner, Øyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schütze, and Peter Clark. Language models with rationality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14190–14201. Association for Computational Linguistics, 2023.

33 Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

34 Koen Kraaijveld, Yifan Jiang, Kaixin Ma, and Filip Ilievski. COLUMBUS: Evaluating cognitive lateral understanding through multiple-choice rebuses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4410–4418, 2025.

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

36 Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

37 Jens Lehmann, Sören Auer, Lorenz Bühmann, and Sebastian Tramp. Class expression learning for ontology engineering. *Journal of Web Semantics*, 9(1):71–81, 2011.

38 Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

39 Marcus Liwicki and Horst Bunke. A novel approach to on-line handwriting recognition based on combined symbol and structural information. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 211–216. IEEE, 2005.

40 Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736, 2023.

41 Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31, pages 3749–3759, 2018.

42 Hande Küçük McGinty. *KNowledge Acquisition and Representation Methodology (KNARM) and its applications*. PhD thesis, University of Miami, 2018.

43 Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024.

44 Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

45 John C. Raven. *Raven manual: Section 3: Standard progressive matrices*. Pearson, 1998.

46 Giuseppe Rizzo, Nicola Fanizzi, and Claudia d'Amato. Class expression induction as concept space exploration: From dl-foil to dl-focl. *Future Generation Computer Systems*, 108:256–272, 2020.

47 Michael Röder, Pham Thuy Sy Nguyen, Felix Conrads, Ana Alexandra Morim da Silva, and Axel-Cyrille Ngonga Ngomo. Lemming - example-based mimicking of knowledge graphs. In *15th IEEE International Conference on Semantic Computing, ICSC 2021, Laguna Hills, CA, USA, January 27-29, 2021*, pages 62–69. IEEE, 2021.

48 Hadas Shalit Peleg and Anat Milo. Small data can play a big role in chemical discovery. *Angewandte Chemie*, 135(26):e202219070, 2023.

49 Cogan Shimizu, Karl Hammar, and Pascal Hitzler. Modular ontology modeling. *Semantic Web*, 14(3):459–489, 2023.

50 Daniel L. Silver and Tom M. Mitchell. The roles of symbols in neural-based ai: They are not what you think!, 2023.

51 Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4077–4087, 2017.

52 Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. Arn: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086, 2024.

53 Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The neon methodology for ontology engineering. In *Ontology engineering in a networked world*, pages 9–34. Springer, 2011.

54 William J. Sutherland, Jake M. Robinson, David C. Aldridge, tim Alamenciak, Matthew Armes, Nina Baranduin, Andrew J. Bladon, Martin F. Breed, Nicki Dyas, Chris S. Elphick, Richard A. Griffiths, Jonny Hughes, Beccy Middleton, Nick A. Littlewood, Roger Mitchell, William H. Morgan, Roy Mosley, Silviu O. Petrovan, Kit Prendergast, Euan G. Ritchie, Hugh Raven, Rebecca K. Smith, Sarah H. Watts, and Ann Thornton. Editorial: Creating testable questions in practical conservation: a process and 100 questions. *Conservation Evidence Journal*, 19, Jan 2022.

55 Emre Ugur, Alper Ahmetoglu, Yukie Nagai, Tadahiro Taniguchi, Matteo Saveriano, and Erhan Oztop. Neuro-symbolic robotics. *IEEE Transactions on Robotics (review article)*, 2025. Comprehensive review of neural–symbolic integration in robotic architectures.

56 Frank Van Harmelen and Annette Ten Teije. A boxology of design patterns for hybrid learning and reasoning systems. *Journal of Web Engineering*, 18(1-3):97–123, 2019.

57 Edoardo Vecchi, Lukás Pospísil, Steffen Albrecht, Terence J. O'Kane, and Illia Horenko. espa+: Scalable entropy-optimal machine learning classification for small data problems. *Neural Comput.*, 34(5):1220–1255, 2022.

58 Alvaro Velasquez, Neel Bhatt, Ufuk Topcu, Zhangyang Wang, Katia Sycara, Simon Stepputtis, Sandeep Neema, and Gautam Vallabha. Neurosymbolic ai as an antithesis to scaling laws. *PNAS Nexus*, 4(5):pgaf117, 05 2025.

59 Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9878–9888, 2021.

60 Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9(1):42, 2023.

4.4 Explainable AI

Pascal Hitzler (Kansas State University – Manhattan, US), Catia Pesquita (Universidade de Lisboa, PT), Michael L. Raymer (Wright State University – Dayton, US), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Daria Stepanova (Bosch Center for AI, DE)

4.4.1 What is Explainable AI?

Explainable AI (XAI) [9, 27, 1, 5, 20, 25, 26, 2] aims to communicate information about an AI system that can help to understand or assess validity of system output. The need for XAI arrives out of the fact that AI systems such as those based purely on deep learning are often "black boxes" with obscure internal mechanisms that do not readily allow for an understanding how certain inputs lead to certain outputs.

Explanations in XAI can be characterized along several interrelated dimensions, each shaping the nature of the explanation, its integration with the underlying model, and its utility for various stakeholders. While these dimensions are not entirely independent, they offer a valuable conceptual framework for the design of explanation techniques. One critical consideration is the purpose of the explanation (see e.g., [6]) – whether it is intended to expose the internal structure and operation of the model, to support debugging and refinement, to justify or validate model outputs in a way that fosters trust and confidence, or to offer insights into the underlying data-generating process or real-world phenomena.

The form that an explanation takes may also vary, ranging from machine-readable outputs that enable model introspection, comparison, or reasoning, to human-interpretable formats such as visualizations (e.g., heatmaps or causal graphs) or natural language descriptions. The intended audience further shapes the design of an explanation, which might be tailored for domain experts, lay users, model developers, automated agents, or regulatory bodies. Additionally, explanations may differ in their relationship to the model itself: some systems are explicitly designed to produce inherently interpretable outputs [11, 22], while others rely on post-hoc techniques to extract explanations from otherwise opaque reasoning processes [8, 28, 7, 12, 18, 17, 16]. Finally, explanations may convey meaning at varying levels of abstraction. While low-level, granular accounts (e.g., visualizations of activation patterns or weight dynamics) can offer insight into the internal mechanics of a model, more meaningful and actionable explanations often emerge at higher levels of abstraction (see also Section 4.2). This is analogous to thermodynamics, where macroscopic properties like temperature and pressure provide more interpretable and practically relevant information than the exhaustive specification of each particle's motion and energy.

4.4.2 What is the NeSy Ambition?

Logic-based AI systems are inherently explainable to some extent, e.g., they allow for the examination of reasoning chains (like proof trees), identification of key facts influencing outputs (like abductive reasoning), and the tracing of conclusions back to foundational assumptions (such as whether the system operates under an open or closed world view). Neurosymbolic AI systems that are hybrid (i.e., consist of coupled neural and symbolic systems) are therefore inherently partially explainable because their symbolic components are.

In contrast, post-hoc explanation methods are used after an AI system (say, one based on deep learning) has been trained. Neurosymbolic post-hoc methods typically produce a set of logical axioms that may (partially) capture the network's input-output behavior, and/or internal activation propagation (say, in the form of logical rules [28, 7], and/or labels for hidden node activation patterns (say, as description logic concepts [8]). The resulting logical axioms can then be used for additional reasoning or analysis, such as detecting contradictions, exploring inference paths, model improvements, bias detection, or verifying assumptions, ultimately improving quality of output and trust in the system, for the end user or within collaborative human-AI teaming efforts [14]. Explanations that take hidden node activations and their propagation into account can also help ground the explanation in the actual run-time mechanics of the neural model, increasing trust and interpretability. Post-hoc explanations can also bring in implicit knowledge that is not part of the task input [13].

Neurosymbolic explainability can enable domain experts – like a biochemist – to understand underlying mechanisms, such as identifying biomarkers for a disease. Through its use of logic-based knowledge representation, it also opens up XAI to to full power and versatility of formal semantics. As such, neurosymbolic approaches to XAI can also draw from external knowledge in order to provide independently verifiable evidence for AI claims in responses; a possible pathway to establish external validity of post-hoc rationalizations. Ultimately, hybrid NeSy supports not just explainable models, but systems that invite deeper human understanding and interaction.

4.4.3 What are the challenges?

There are several challenges to fulfilling the NeSy ambition that are rooted in the overarching challenges of explainability [23] but are tied to the particulars of NeSy.

One major issue lies in ensuring the actual usefulness of explanations [4]: while NeSy methods are able to generate explanations that are both faithful to the model's output and coherent with the knowledge model, they may not serve the user's needs or context. In fact, current research in explainable AI has faced criticism for being driven by researchers' assumptions rather than the actual needs of end users, often overlooking who the explanations are truly for [23]. As a result, there is growing momentum toward adopting a human-centered approach in XAI to ensure explanations are meaningful and tailored to specific stakeholders [20, 21].

This misalignment often stems from mismatched expectations around explanatory depth, abstraction level, and complexity – users may receive explanations that are either too general to be informative, too detailed, or too large to be understandable. Balancing depth and simplicity becomes crucial, particularly as different users (e.g., experts vs. laypersons) require different levels of abstraction. A clear example would be an explanation that lacks relevance, e.g., when explaining a specific drug recommendation (e.g., sunitinib[3]) for a cancer patient, a faithful, coherent, and useless explanation would be (patient -has→ cancer -is treated by→ anti-cancer drugs ← is a - sunitinib). This explanation does not afford sufficient explanatory depth to fulfill the purpose of validating the recommendation and supporting a medical decision. On the other hand, explanations at a similar abstraction level may not serve the specific user's purpose. For example, while the following could be an appropriate explanation targeting a medical doctor: patient -has-diagnosis→ clear cell renal carcinoma -has-guideline-treatment→ sunitinib; it would not suit the purposes of a drug development researcher, who likely requires an explanation focusing on the molecular mechanisms.

[3] https://en.wikipedia.org/wiki/Sunitinib

In turn, this also introduces the challenge of the intelligibility of an explanation, i.e., a symbolic explanation may be faithful, coherent, and relevant, but not be comprehensible to the user. Symbolic reasoning chains are indeed not automatically intelligible; translating them into a communicable format (be it natural language or a diagram, etc) that reflects human explanatory norms is challenging, particularly if the symbols do not map intuitively to a user's understanding (re. symbol grounding problem).

Additional challenges are introduced by the incompleteness or outdatedness of the knowledge model [4]. When domain knowledge evolves quickly, explanations might lag behind, giving rise to temporal drift/timeliness by relying on obsolete or potentially incorrect information. Moreover, NeSy explanations should be able to gracefully manage the tension between adhering to established knowledge and incorporating novel patterns that contradict it, to reveal new insights, supporting an unconfined exploration/dynamic-scoped exploration.

NeSy explainability also faces the transversal issue of scalability. As the knowledge base grows or the domain evolves rapidly – as in healthcare or finance – maintaining timely, relevant, and accurate explanations becomes increasingly difficult. Real-time symbolic reasoning is computationally intensive, especially when personalized or context-sensitive explanations are required. In dynamic domains, outdated knowledge can lead to stale or misleading explanations unless the system can adapt to temporal changes. Furthermore, as NeSy systems are deployed more broadly, they must support a diverse range of users with different informational needs and levels of expertise. This necessitates scalable user modeling and context-aware explanation strategies. Ultimately, the promise of NeSy explainability hinges not just on fidelity to reasoning processes, but on the system's ability to communicate those processes effectively and adaptively to humans.

In order to make progress on XAI, concerted benchmarking and evaluation efforts are also required, that should take the above mentioned aspects into account. At this point in time high-quality benchmarks remain to be established for this purpose (see also Section 4.8).

4.4.4 Where are we now?

Research into neurosymbolic explainable AI within hybrid neurosymbolic systems is underway, with promising early work across several key areas [11, 22, 8, 28, 7, 12, 14, 4, 15, 10, 3, 24, 5, 19]), some of which we already touched upon above. However the state of the field is clearly still exploratory, in that a multitude of methods is currently being proposed and tried out. And while individual approaches show promise, they need to be rigorously validated, better integrated into coherent systems, and evaluated through strong proof-of-concept implementations. Crucially, future work must demonstrate how these approaches enhance decision quality and foster user trust.

References

1 Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

2 Sajid Ali, Tamer Abuhmed, Shaker H. Ali El-Sappagh, Khan Muhammad, Jose Maria Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz Rodríguez, and Francisco Herrera. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion*, 99:101805, 2023.

3 Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frédéric Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In Andreas Krause,

Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 1801–1825. PMLR, 2023.

4 Roberto Barile, Claudia d'Amato, and Nicola Fanizzi. LP-DIXIT: evaluating explanations for link predictions on knowledge graphs using large language models. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov, editors, *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 4034–4042. ACM, 2025.

5 Adrien Bennetot, Gianni Franchi, Javier Del Ser, Raja Chatila, and Natalia Díaz Rodríguez. Greybox XAI: A neural-symbolic learning framework to produce interpretable predictions for image classification. *Knowl. Based Syst.*, 258:109947, 2022.

6 David J. Chalmers. Propositional interpretability in artificial intelligence. *CoRR*, abs/2501.15740, 2025.

7 Gabriele Ciravegna, Francesco Giannini, Pietro Barbiero, Marco Gori, Pietro Lio, Marco Maggini, and Stefano Melacci. Learning logic explanations by neural networks. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 547–558. IOS Press, 2023.

8 Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Samatha Ereshi Akkamahadevi, Avishek Das, Cara Widmer, Eugene Y. Vasserman, Md Kamruzzaman Sarker, and Pascal Hitzler. Towards a neurosymbolic understanding of hidden neuron activations. *Neurosymbolic Artificial Intelligence*, 2025. To Appear.

9 Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *CoRR*, abs/2006.11371, 2020.

10 David Debot, Pietro Barbiero, Francesco Giannini, Gabriele Ciravegna, Michelangelo Diligenti, and Giuseppe Marra. Interpretable concept-based memory reasoning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

11 Anna Himmelhuber, Stephan Grimm, Mitchell Joblin, Sonja Zillner, and Thomas A. Runkler. Combining sub-symbolic and symbolic methods for explainability. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 559–576. IOS Press, 2023.

12 Vitor A. C. Horta and Alessandra Mileo. Explaining cnns using knowledge extraction and graph analysis. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 577–608. IOS Press, 2023.

13 Filip Ilievski. *Human-Centric AI with Common Sense*. Synthesis Lectures on Computer Science (SLCS). Springer, 2024.

14 Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, Pascal Hitzler, Giuseppe Marra, Pasquale Minervini, Martin Mundt, Axel-Cyrille Ngonga Ngomo, Alessandro Oltramari, Gabriella Pasi, Zeynep G. Saribatur, Luciano Serafini, John Shawe-Taylor, Vered Shwartz, Gabriella Skitalinskaya, Clemens Stachl, Gido M. van de Ven, and Thomas Villmann. Aligning generalization between humans and machines. *Nature Machine Intelligence*, 7(9):1378–1389, Sep 2025.

15 Filip Ilievski, Kaixin Ma, Alessandro Oltramari, Peifeng Wang, and Jay Pujara. Building robust and explainable AI with commonsense knowledge graphs and neural models. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 178–209. IOS Press, 2023.

16 Youmna Ismaeil, Jan-Hendrik Metzen, Trung-Kien Tran, Hendrik Blockeel, and Daria Stepanova. Beyond manual labels: Unsupervised graph-based explanations for error analysis in image classifiers. *ISWC*, 2025. To Appear.

17 Youmna Ismaeil, Daria Stepanova, Trung-Kien Tran, and Hendrik Blockeel. Feabi: A feature selection-based framework for interpreting KG embeddings. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I*, volume 14265 of *Lecture Notes in Computer Science*, pages 599–617. Springer, 2023.

18 Youmna Ismaeil, Daria Stepanova, Trung-Kien Tran, Piyapat Saranrittichai, Csaba Domokos, and Hendrik Blockeel. Towards neural network interpretability using commonsense knowledge graphs. In Ulrike Sattler, Aidan Hogan, C. Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato, editors, *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, volume 13489 of *Lecture Notes in Computer Science*, pages 74–90. Springer, 2022.

19 Katarzyna Kaczmarek-Majer, Gabriella Casalino, Giovanna Castellano, Monika Dominiak, Olgierd Hryniewicz, Olga Kaminska, Gennaro Vessio, and Natalia Díaz Rodríguez. PLENARY: explaining black-box models in natural language through fuzzy linguistic summaries. *Inf. Sci.*, 614:374–399, 2022.

20 Xiangwei Kong, Shujie Liu, and Luhao Zhu. Toward human-centered XAI in practice: A survey. *Mach. Intell. Res.*, 21(4):740–770, 2024.

21 Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. Questioning the AI: informing design practices for explainable AI user experiences. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–15. ACM, 2020.

22 Mohammad Saeid Mahdavinejad, Peyman Adibi, Amirhassan Monajemi, and Pascal Hitzler. Towards explainable depression detection: A neurosymbolic approach to uncover social media signals with generative AI. In *19th International Conference on Neurosymbolic Learning and Reasoning*, 2025.

23 Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

24 Natalia Díaz Rodríguez, Alberto Lamas, Jules Sanchez, Gianni Franchi, Ivan Donadello, Siham Tabik, David Filliat, Policarpo Cruz, Rosana Montes, and Francisco Herrera. Explainable neural-symbolic learning (*X-NeSyL*) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Inf. Fusion*, 79:58–83, 2022.

25 Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.*, 38(5):3043–3101, 2024.

26 Timo Speith. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2239–2250. ACM, 2022.

27 Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans. Neural Networks Learn. Syst.*, 32(11):4793–4813, 2021.

28 Joe Townsend, Esma Mansouri-Benssassi, Kwun Ho Ngan, and Artur d'Avila Garcez. Discovering visual concepts and rules in convolutional neural networks. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 337–372. IOS Press, 2023.

4.5 Neurosymbolic AI in the Age of Generative AI

Daria Stepanova (Bosch Center for AI, DE), Mehwish Alam (Télécom Paris, Institut Polytechnique de Paris, FR), Stefan Ollinger (Schloss Dagstuhl – Leibniz Center for Informatics, DE)

4.5.1 What is GeNeSy AI?

Generative AI (GenAI) refers to AI systems, typically based on large-scale neural networks, that can **generate new content** such as natural language text, images, or code based on patterns learned from existing data. While the generation of **natural language text** remains the most well-known application, the scope of GenAI is much broader [47]. Typical symbolic outputs that can be produced by GenAI methods include code, formal problem specifications, structured data like graphs, tabular data, machine-readable semantic representations, logical rules and constraints (e.g., SHACL, OWL, FOL), etc. Thus, GenAI has great potential for addressing the modeling and knowledge acquisition bottleneck of symbolic AI (see Section 4.2).

4.5.1.1 GenAI Development Pipeline

A structured approach to GenAI development can be broken down into several key stages that build upon one another. It begins with data preparation, which involves sourcing, cleaning, and transforming data to create high-quality inputs for training or fine-tuning. This data may come from internal databases, third-party sources, or user-generated content. Feature engineering plays a crucial role here, helping to extract meaningful patterns while eliminating irrelevant or noisy information to ensure the model learns from the most accurate data available.

Following this, the model training phase involves developing a base generative model, either from scratch or by initializing an existing model architecture. Using large-scale datasets, the model learns to capture general patterns in multimodal data. This step typically requires extensive computing resources and represents the core of the model development process.

Once the base model is established, it is fine-tuned and customized using domain-specific or task-specific data. This tailoring process may involve supervised fine-tuning, domain adaptation, instruction tuning, or parameter-efficient techniques such as Low-Rank Adaptation, all aimed at improving the model's relevance, accuracy, and safety in real-world applications.

With a fine-tuned model, attention shifts to prompting and output control. Prompts are carefully engineered to guide the model's responses, using techniques such as zero-shot, few-shot, or chain-of-thought prompting. Sampling parameters like temperature, top-k, and

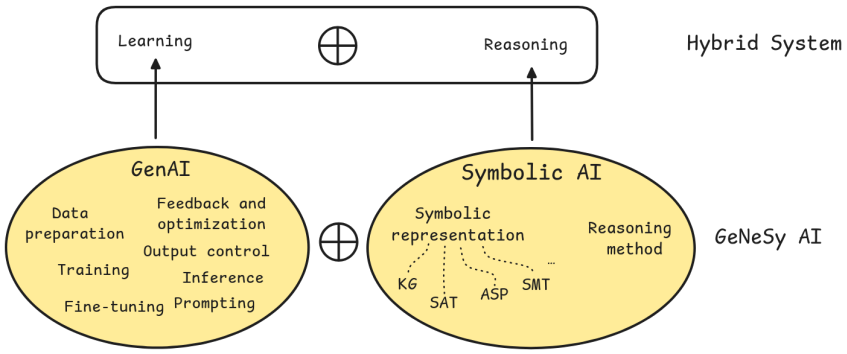

Figure 5 Illustration of a NeSy AI system, where a neural component is instantiated with the GenAI method; we refer to the resulting system as GeNeSy AI. Integration of symbolic representations, e.g., KGs, propositional formulas (SAT), answer set programs (ASP), satisfiability modular theories (SMT), etc. and the respective reasoning methods can be exploited at different stages of the GenAI development pipeline.

top-p are adjusted to manage the diversity and coherence of outputs. In the post-generation phase, outputs may undergo filtering to remove hallucinations or irrelevant content using rules, classifiers, or moderation APIs. Additional steps like re-ranking help to ensure that the most appropriate response is selected.

Finally, the feedback and optimization stage focuses on improving the model over time. Feedback is gathered through user interactions, surveys, or manual reviews to identify and address weaknesses such as bias or factual inaccuracies. This input can lead to prompt revisions, sampling tweaks, or further fine-tuning. Reinforcement learning techniques, such as Reinforcement Learning from Human Feedback (RLHF), may also be applied to better align model outputs with human preferences or task requirements.

Once development is complete, the model is deployed and integrated into applications and workflows. Post-deployment monitoring and adjustments are essential to ensure ongoing safety, reliability, and alignment with intended use.

4.5.1.2 GeNeSy AI as a Variation of NeSy AI

When it comes to NeSy AI systems in the context of GenAI, naturally the neural (learned) component of a NeSy AI system corresponds to the GenAI method itself, while the symbolic component can take various forms, such as ontologies, Knowledge Graphs (KGs), logical constraints or rules, structured schemas, or grammars (see Section 4.1). At multiple stages of the GenAI development pipeline described in Section 4.5.1.1, symbolic structures and reasoning systems can be integrated, enhancing the capabilities and reliability of generative models. We illustrate this point in Figure 1, and discuss possible integration strategies in what follows.

- **Data Preparation.** Symbolic knowledge can play a key role in generating data that conforms to predefined constraints, which can then be used–among other purposes–for training generative models. This includes, for example, the generation of graphs with specific topologies or properties, such as [30], or the synthesis of images guided by constrained scene graphs [42]. Incorporating knowledge graphs (KGs) during the data synthesis process can enhance the representativeness and structure of the generated data, e.g., see [29, 3] for question answering applications or [20] for story generation.

- **Training.** During training, symbolic structures can constrain and guide the learning process. This may involve incorporating KG embeddings, where the most prominent techniques include joint learning of Large Language Models (LLMs) and knowledge graph embeddings [9], incorporation of joint loss functions that account for both neural and symbolic objectives, embedding concatenation (see, e.g., [36] for overview), or knowledge distillation using teacher-student setups, where a symbolic model acts as a teacher [25].
- **Fine-tuning.** Symbolic knowledge, such as formal rules, specifications, or domain-specific logic, can be used to fine-tune generative models, e.g., to ensure consistency with a set of predefined rules and constraints [7]. Integration of symbolic knowledge at the fine-tuning stage is also often realized for improving performance of LLMs on structured or specialized tasks, e.g., generation of logic programs from natural language text [10]. Moreover, structural graph-based knowledge has been integrated into LLMs during fine-tuning stage to improve its performance on graph-specific tasks [38].
- **Prompting.** Symbolic knowledge can also be effectively utilized at the prompting stage, e.g., the variety of GraphRAG-based approaches are prominent examples here [14]. In the most simple scenarios factual knowledge in KGs is represented in a textual form and injected into the context during LLM prompting [5]. Prompts can also be enriched symbolically by incorporating ontological knowledge, e.g., for motion planning [11] or news summarization [43]. Other more sophisticated strategies use dedicated reasoning modules to traverse KGs and then guide the LLMs through chain-of-thought prompting [50].
- **Inference.** Finally, there are several approaches proposed in the literature that exploit symbolic reasoning to constrain or guide the output of the model during inference. Most prominent techniques include logic-based sampling [2] or decoding [40, 15, 28] to maintain consistency with the available background knowledge. Additionally, outputs produced by LLMs can be post-filtered to enforce symbolic constraints, such as exclusion of certain words or ensuring syntactic and semantic validity of generated code [27]. Several works leverage KG structures or embeddings to influence and refine the output generated by LLMs, e.g., [8]. Another promising direction is concerned with utilizing LLMs for translating textual problem statements to symbolic artifacts (see also Section 4.2), passing them to dedicated solvers for reasoning, and subsequently translating the output generated by the reasoner back to natural language using LLMs [35, 49, 12, 32].

4.5.2 What is the GeNeSy AI Ambition?

As GenAI becomes increasingly used across a variety of applications such as web search, content creation, and code generation, the goal of Neurosymbolic AI is not to replace GenAI, but to enhance it by incorporating symbolic methods, particularly in downstream tasks where GenAI tends to underperform. Recently, significant effort has been devoted to identifying and classifying such challenging tasks, as well as developing methods to address them. Much of this analysis is empirical in nature, relying on diverse benchmarks designed to highlight and characterize areas where GenAI models consistently struggle, e.g., ARC benchmarks[4]. At the same time, theoretical research has focused on approaching the problem of identifying weaknesses of GenAI methods from computational complexity and expressivity perspective.

The high-level ambition of neurosymbolic AI is that by integrating symbolic knowledge at various stages in the GenAI development pipeline as described above, GenAI systems can achieve greater accuracy, interpretability, and control, particularly in domains requiring

[4] https://arcprize.org/

structure, reasoning, or domain-specific compliance, e.g., medical applications, production systems or legal cases. Importantly, while there is a great potential for NeSy AI approaches to address many limitations of GenAI, clearly it is not a silver bullet, and precisely detecting scenarios and tasks where NeSy AI methods would have the largest impact is crucially important. Below we discuss limitations of GenAI methods, and outline possible ways how symbolic methods can be potentially utilized to address them.

Combinatorial Problems. Recently, a number of benchmarks have been introduced to evaluate the performance of GenAI methods–particularly reasoning-capable LLMs–on combinatorial tasks such as SAT solving [17], logical puzzles [26], and classical planning [21]. These studies consistently reveal a core limitation of current LLMs: their inability to handle increasing problem complexity, a phenomenon often referred to as the curse of complexity. Even with larger models and more compute, reasoning capabilities of LLMs often plateau or degrade.

While LLMs can solve certain well-known problems, such as the "25 horses" puzzle [19], their performance drops sharply on semantically equivalent but syntactically unfamiliar variants, e.g., changing "horses" to "bunnies" significantly reduces accuracy [19]. Evaluations on out-of-distribution instances [39] suggest that, in the context of combinatorial problems, many correct outputs can be attributed to memorization rather than genuine reasoning.

In contrast, symbolic AI has made considerable progress in solving these combinatorial problems effectively and accurately, as evidenced, e.g., by outcomes from SAT competitions[5]. While symbolic methods still struggle with scalability (see also Section 4.2), they offer a superior trade-off between accuracy and runtime compared to purely generative AI approaches.

This points toward the promise of neurosymbolic AI systems, which integrate LLMs with classical solvers by using LLMs to translate natural language problem descriptions into formal representations and then pass them to symbolic solvers for generating solutions [13, 32, 12, 35, 49].

Spatial and Temporal Reasoning. Recent studies have highlighted the poor performance of GenAI models on spatial and temporal reasoning tasks [4, 48, 34]. Symbolic AI, in contrast, offers mature frameworks precisely for these types of reasoning–such as RCC8 [24, 22] and Allen's interval algebra [18]. Once again, a promising direction lies in using LLMs to convert natural language inputs into formal representations that these symbolic frameworks can process [4].

Hallucinations. Hallucinations remain a critical and largely unresolved limitation of GenAI systems [6]. One potential mitigation strategy involves incorporating symbolic methods to verify or filter generated outputs. For example, [33] uses ontological reasoners to identify and eliminate hallucinations that violate logical consistency within a given ontology. Similarly, knowledge graphs have been explored as tools for grounding and validating LLM outputs [1].

Analogy and Abstraction. Human analogical reasoning involves transferring relational structures from known to novel contexts, often by applying abstract rules. This ability emerges early in development and operates across domains–from linguistic analogies (e.g., "body : feet :: table : ?") to visual ones (e.g., "(:) :: < : ?") [45]. GenAI models, however, struggle with such tasks, particularly when they involve non-standard symbols or those from low-resource languages such as Greek [45]. Analogy and abstraction has been actively studied in the area of symbolic AI, so there is a high potential for utilizing these results also in combination with GenAI.

[5] `https://satcompetition.github.io/2025/index.html`

Creative Problem Solving. The MACGYVER benchmark [46] evaluates the creative problem-solving abilities of LLMs across 1,600+ real-world scenarios that require innovative object use and out-of-the-box thinking. The results reveal that LLMs frequently suggest implausible or physically impossible actions. To address this, one promising approach involves enforcing commonsense constraints–formulated in symbolic AI languages such as first-order logic–on top of the LLM outputs.

Theoretical Limitations of Transformers. A growing body of theoretical work has identified fundamental limitations of transformer architectures, which underpin current GenAI models. For instance, it has been proven that transformers cannot natively model compositional functions [37]. As a result, they struggle with questions requiring the composition of multiple relations–such as: "What is the birthday of Frédéric Chopin's father?", which requires chaining the facts that Chopin's father is Nicolas Chopin and that Nicolas was born on April 15, 1771 [16].

These theoretical limitations have also been empirically validated in multiple studies, e.g., [23]. A promising neurosymbolic solution is to augment transformers with external knowledge representations, such as knowledge graphs [16], which naturally support compositional reasoning.

Beyond semantically rich tasks, transformers also falter on abstract algorithmic problems. For example, [41] demonstrates that even multi-head transformers cannot solve the 3-Matching problem (i.e., checking whether any three integers in a sequence sum to zero modulo a large number). While this task is synthetic, it theoretically explains the inherent limitations of transformers for combinatorial search and reasoning. In [31], the expressive power of transformers with chain-of-thought prompting is analyzed, showing that their capabilities map to the complexity class P, thus providing a theoretical upper bound on what such models can compute.

4.5.3 What are the challenges?

A central challenge in developing effective GeNeSy AI systems lies in identifying the kinds of problems where pure generative AI tends to fail, but can be meaningfully enhanced by integrating symbolic methods. Understanding these limitations is key to designing hybrid architectures that are both powerful and practical (see Section 4.8 for more details). Certain types of tasks clearly illustrate the added value of symbolic reasoning. As discussed, combinatorial problems, for instance, often go beyond the capabilities of GenAI alone and require symbolic solvers to find valid solutions or verify generated ones. Despite the combinatorial explosion of symbolic reasoning (see Section 4.2), the existing methods are still more effective than any alternatives when it comes to computing provably correct solutions.

Symbolic methods also shine in other domains, such as resolving inconsistencies across multiple sources, explaining and repairing data errors or answering complex queries with formal constraints. However, it is important to recognize that symbolic tools are not a cure-all. For example, there are failure cases, like gaps in physical world knowledge (e.g., analysis of novel diseases or structural properties of unknown materials) for which neither GenAI nor symbolic reasoning methods might be suited. The ultimate goal in this area is to develop intuitive, task-specific guidelines: for any given problem, define the GeNeSy AI configuration that is most likely to yield success.

A second major challenge lies in the emergence of suitable symbolic knowledge, both with and for GenAI. While symbolic representations like rules, constraints, or ontologies are powerful tools, they remain difficult to scale, primarily due to the bottleneck of manual

curation (see Section 4.2). A compelling direction is to explore whether GenAI itself can help construct symbolic knowledge from raw text or examples. Can GenAI be used to induce formal structures automatically? And if such structures are generated, does feeding them back into the GenAI pipeline offer measurable benefits? These are open questions that demand rigorous benchmarks to assess where and when symbolic intermediates truly improve performance.

The issue of scalability presents another critical concern. While symbolic solvers can offer precise reasoning capabilities, they often struggle to scale, especially when embedded into larger NeSy AI systems. This becomes particularly problematic when GenAI is used to generate symbolic artifacts–such as logical formulas or knowledge bases–that must then be validated or processed at scale. Tasks like solving constraints over large outputs, repairing expressive ontologies, or reasoning over extensive knowledge graphs exemplify these challenges.

In addition, aligning symbolic reasoning with multimodal GenAI systems presents unique challenges. A model processing both text and images may develop inconsistent internal representations across modalities—for instance, interpreting the concept of a "bat" as an animal in text but as a sports object in images. Such divergence can lead to errors in tasks that demand consistent cross-modal understanding. Integrating symbolic reasoning could help reconcile these discrepancies and promote coherence across modalities, but designing effective approaches remains a complex and open problem requiring careful attention.

Finally, usability of GeNeSy AI systems remains a key concern. One of GenAI's biggest strengths is its accessibility–users can interact with it naturally, without needing specialized knowledge. Symbolic systems, in contrast, often require fluency in formal languages, limiting their reach. For GeNeSy AI approaches to gain broader adoption, symbolic logic must be abstracted away from users whenever possible. The system's internal complexity should be hidden, unless deeper, expert-level access is specifically required. Ideally, interfaces with GeNeSy AI systems should be developed that retain the rigor and correctness of symbolic reasoning, while offering the ease-of-use that defines the GenAI experience. Together, these challenges frame a roadmap for building more powerful, scalable, and usable GeNeSy AI systems. They especially underscore the importance of understanding when and how to blend symbolic reasoning with GenAI to achieve the most benefit.

4.5.4 Where are we now?

A key development in addressing the limitations of generative AI lies in its growing ability to translate natural language into formal logic. This translation allows symbolic solvers, traditionally reliant on structured, logic-based input, to be effectively leveraged in tasks that originate from unstructured human language. As a result, a new wave of NeSy systems, also referred to as prompt-symbolic [44] is emerging. These systems often follow a common architecture: GenAI is used to convert natural language into precise problem specifications, which are then processed by symbolic solvers to derive solutions. In more advanced configurations, LLMs function as agents within search-based reasoning frameworks, coordinating steps in complex problem-solving tasks. This hybrid approach not only enables more rigorous reasoning but also leverages the flexibility and accessibility of GenAI to front-load formalization.

Symbolic representations are increasingly being incorporated into GenAI to enhance performance and factual accuracy. Advances such as GraphRAG illustrate this trend by integrating KGs into retrieval-augmented generation pipelines, allowing models to ground their outputs in structured knowledge. This leads to improved consistency, relevance, and logical coherence–especially useful in domains like law, where resolving contradictions across

multiple documents is critical. Moreover, symbolic methods, including the use of ontologies and KGs, support more effective integration of structured and unstructured data within GenAI systems.

Additionally, researchers are exploring closed-loop integrations of GenAI and symbolic AI, where the two components iteratively inform and refine each other. In these setups, GenAI may propose candidate solutions or knowledge structures, which are then evaluated and corrected using symbolic methods–feeding the results back into the generative process. While still an area of active exploration, such feedback loops hold promise for creating systems that not only generate and reason, but also self-correct in structured and explainable ways (see Section 4.4 for detailed discussions on explainability). Together, these developments point toward a powerful synthesis: leveraging GenAI's language capabilities to interface with symbolic tools, while embedding symbolic structure within GenAI pipelines to enhance reasoning depth and factual robustness.

References

1 Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. Can knowledge graphs reduce hallucinations in llms? : A survey. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3947–3960. Association for Computational Linguistics, 2024.

2 Kareem Ahmed, Kai-Wei Chang, and Guy Van den Broeck. Controllable generation via locally constrained resampling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

3 Fernando Amodeo, Fernando Caballero, Natalia Díaz-Rodríguez, and Luis Merino. Og-sgg: ontology-guided scene graph generation—a case study in transfer learning for telepresence robotics. *IEEE Access*, 10:132564–132583, 2022.

4 Konstantine Arkoudas. GPT-4 can't reason. *CoRR*, abs/2308.03762, 2023.

5 Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *CoRR*, abs/2306.04136, 2023.

6 Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this. *CoRR*, abs/2409.05746, 2024.

7 Diego Calanzone, Stefano Teso, and Antonio Vergari. Logically consistent language models via neuro-symbolic integration. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

8 Lang Cao. GraphReason: Enhancing reasoning capabilities of large language models through a graph-based verification approach. In Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Ben Lipkin, Danilo Neves Ribeiro, Lionel Wong, Xi Ye, and Wenting Zhao, editors, *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 1–12, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

9 Erica Coppolillo. Injecting knowledge graphs into large language models. *CoRR*, abs/2505.07554, 2025.

10 Erica Coppolillo, Francesco Calimeri, Giuseppe Manco, Simona Perri, and Francesco Ricca. LLASP: fine-tuning large language models for answer set programming. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, 2024.

11 Muhayy Ud Din, Jan Rosell, Waseem Akram, Isiah Zaplana, Máximo A. Roa, Lakmal D. Seneviratne, and Irfan Hussain. Ontology-driven prompt tuning for llm-based task and motion planning. *CoRR*, abs/2412.07493, 2024.

12 Marius-Constantin Dinu, Claudiu Leoveanu-Condrei, Markus Holzleitner, Werner Zellinger, and Sepp Hochreiter. Symbolicai: A framework for logic-based approaches combining generative models and solvers. In Vincenzo Lomonaco, Stefano Melacci, Tinne Tuytelaars, Sarath Chandar, and Razvan Pascanu, editors, *Conference on Lifelong Learning Agents, 29-1 August 2024, University of Pisa, Pisa, Italy*, volume 274 of *Proceedings of Machine Learning Research*, pages 869–914. PMLR, 2024.

13 Subhabrata Dutta, Ishan Pandey, Joykirat Singh, Sunny Manchanda, Soumen Chakrabarti, and Tanmoy Chakraborty. Frugal lms trained to invoke symbolic solvers achieve parameter-efficient arithmetic reasoning. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17951–17959. AAAI Press, 2024.

14 Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130, 2024.

15 Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10932–10952. Association for Computational Linguistics, 2023.

16 Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18126–18134. AAAI Press, 2024.

17 Rishi Hazra, Gabriele Venturato, Pedro Zuidberg Dos Martires, and Luc De Raedt. Have large language models learned to reason? A characterization via 3-sat phase transition. *CoRR*, abs/2504.03930, 2025.

18 Duygu Sezen Islakoglu and Jan-Christoph Kalo. Chronosense: Exploring temporal understanding in large language models with time intervals of events. *CoRR*, abs/2501.03040, 2025.

19 Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie Su, Camillo J. Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4722–4756. Association for Computational Linguistics, 2024.

20 Yifan Jiang, Filip Ilievski, and Kaixin Ma. Transferring procedural knowledge across commonsense tasks. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1156–1163. IOS Press, 2023.

21 Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

22 Irtaza Khalid, Amir Masoud Nourollah, and Steven Schockaert. Large language and reasoning models are shallow disjunctive reasoners. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 8843–8869. Association for Computational Linguistics, 2025.

23 Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7675–7688. Association for Computational Linguistics, 2024.

24 Fangjun Li, David C. Hogg, and Anthony G. Cohn. Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6342–6349. ijcai.org, 2024.

25 Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, and Jun Zhao. Neural-symbolic collaborative distillation: Advancing small language models for complex reasoning tasks. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 24567–24575. AAAI Press, 2025.

26 Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *CoRR*, abs/2502.01100, 2025.

27 Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. "we need structured output": Towards user-centered constraints on large language model output. In Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, and Corina Sas, editors, *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 10:1–10:9. ACM, 2024.

28 Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 780–799. Association for Computational Linguistics, 2022.

29 Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press, 2021.

30 Manuel Madeira, Clément Vignac, Dorina Thanou, and Pascal Frossard. Generative modelling of structurally constrained graphs. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

31 William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

32 Kostis Michailidis, Dimos Tsouros, and Tias Guns. Cp-bench: Evaluating large language models for constraint modelling. *CoRR*, abs/2506.06052, 2025.

33 Marco Monti, Oliver Kutz, Nicolas Troquard, and Guendalina Righetti. Improving the accuracy of black-box language models with ontologies: A preliminary roadmap. In

Proceedings of the Joint Ontology Workshops (JOWO) - Episode X: The Tukker Zomer of Ontology, and satellite events co-located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), Enschede, The Netherlands, July 15-19, 2024, volume 3882 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

34 Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. Mapwise: Evaluating vision-language models for advanced map queries. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 9348–9378. Association for Computational Linguistics, 2025.

35 Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3806–3824. Association for Computational Linguistics, 2023.

36 Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599, 2024.

37 Binghui Peng, Srini Narayanan, and Christos H. Papadimitriou. On limitations of the transformer architecture. *CoRR*, abs/2402.08164, 2024.

38 Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Seyed Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for llms. *CoRR*, abs/2402.05862, 2024.

39 Zhenting Qi, Hongyin Luo, Xuliang Huang, Zhuokai Zhao, Yibo Jiang, Xiangjun Fan, Himabindu Lakkaraju, and James R. Glass. Quantifying generalization complexity for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

40 Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 794–805. Association for Computational Linguistics, 2020.

41 Clayton Sanford, Daniel J. Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

42 Giacomo Savazzi, Eugenio Lomurno, Cristian Sbrolli, Agnese Chiatti, and Matteo Matteucci. Neuro-symbolic scene graph conditioning for synthetic image dataset generation. *CoRR*, abs/2503.17224, 2025.

43 A. R. S. Silva and Y. H. P. P. Priyadarshana. Ontology-based prompt tuning for news article summarization. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2025.

44 Adam Stein, Aaditya Naik, Neelay Velingker, Mayur Naik, and Eric Wong. The road to generalizable neuro-symbolic learning should be paved with foundation models. *CoRR*, abs/2505.24874, 2025.

45 Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. Can large language models generalize analogy solving like people can? *CoRR*, abs/2411.02348, 2024.

46 Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. Macgyver: Are large language

models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5303–5324. Association for Computational Linguistics, 2024.

47 Son Tran, Edjard Mota, and Artur d'Avila Garcez. Reasoning in neurosymbolic AI. *CoRR*, abs/2505.20313, 2025.

48 Yuqing Wang and Yun Zhao. TRAM: benchmarking temporal reasoning for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6389–6415. Association for Computational Linguistics, 2024.

49 Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. Satlm: Satisfiability-aided language models using declarative prompting. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

50 Ruilin Zhao, Feng Zhao, Long Wang, Xianzhi Wang, and Guandong Xu. Kg-cot: Chain-of-thought prompting of large language models over knowledge graphs for knowledge-aware question answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6642–6650. ijcai.org, 2024.

4.6 Knowledge Graphs and Ontologies in Neurosymbolic Systems

Roberto Confalonieri (University of Padova, IT), Raghava Mutharaju (IIIT-Delhi, IN), Ernesto Jimenez-Ruiz (City St George's, University of London, UK), Catia Pesquita (Universidade de Lisboa, PT), Cogan Shimizu (Wright State University – Dayton, US)

4.6.1 What is it?

OWL-based knowledge graphs (KGs) [15] may play a pivotal role in NeSy systems by combining symbolic knowledge representation with formal logical reasoning. Grounded in Description Logics [2], OWL enables automated and sound reasoning for both inference (e.g., deducing new relationships) and consistency checking, offering reliable and interpretable symbolic capabilities that complement learning models. OWL-based KGs can guide learning, validate predictions, and enrich sparse or weak annotations – particularly useful in few-shot and zero-shot learning settings [8].

A major strength of OWL lies in its ability to express complex logical constraints (e.g., domain and range restrictions, class disjointness, property inverses) in a concise and reusable form. OWL-based reasoning can be embedded into neural training pipelines or used to filter semantically invalid outputs at inference time. These capabilities are supported by a rich ecosystem of open standards (e.g., [17, 7, 13]), tools (e.g., [1]), and public ontologies (e.g., [9, 3, 23]), enabling researchers to rapidly prototype and build on shared symbolic infrastructure without starting from scratch.

Reusing existing OWL ontologies and tools not only accelerates development but also promotes interoperability and reproducibility in NeSy research. Publicly available OWL-based resources provide high-quality domain knowledge and reasoning frameworks that can

be directly integrated or extended [14]. The (re)use of large ontologies and knowledge graphs to enhance neurosymbolic systems, as well as general learning and reasoning systems, is slowly gaining attention. Traditional neurosymbolic systems typically encode a limited number of rules and do not scale effectively with large modern ontologies [20].

In contrast, [5, 12] emphasize the role of OWL ontologies in generating intelligible, human-centered explanations in neurosymbolic AI. They propose a new conceptual framework highlighting three key roles of ontologies: as formal reference models, as enablers of commonsense reasoning, and as tools for abstraction and complexity management in explanations. Additionally, they introduces the idea of ontological unpacking [12] to enhance the semantic transparency of symbolic artifacts and discusses emerging challenges, such as integrating ontologies with large language models to improve trust and mitigate hallucinations.

Our NeSy position advocates for knowledge formalized through ontologies. We treat ontologies and knowledge graphs as equivalent concepts, possibly expressed in OWL or one of its fragments.[6] Our vision for NeSy falls into a more general hybrid learning and reasoning framework that (i) integrates ontologies as core components, (ii) clearly distinguishes and integrates between deductive (certain) logical reasoning and other (uncertain) learning and reasoning methods, and (iii) accommodates various forms of learning beyond neural networks. Yet, as in Section 4.1, it is also otherwise important to note that this NeSy vision does not neatly encapsulate all possible couplings between neural and symbolic components. For example, LLMs (or other neural systems capable of manipulating or interpreting text) which produce knowledge graphs or learn ontologies can be considered NeSy, as well. The nature of the composition of the components is critical, as well as their purpose.

4.6.2 Where are we now?

The neurosymbolic approach is often described in terms of a dichotomy between symbolic methods, which are human-readable and writable, and neural-based methods, which leverage connectionist training techniques [7] (see also Section 4.1). This approach employs an iterative integration cycle between symbolic and neural methods, where symbolic (expert) knowledge is embedded into neural models, and refined knowledge learned by neural networks is extracted back into symbolic form [7]. The knowledge extracted in this cycle supports the continuous refinement and modification of predefined symbolic rules through reasoning [22]. This iterative cycle is structured around three core dimensions: translation, extraction (generation), and consolidation, each emphasizing a distinct role of knowledge.

In the realm of knowledge translation, symbolic knowledge informs neural network training by embedding logical constraints directly into neural network loss functions [11] or structuring neural architectures based on established background knowledge [18]. Additionally, knowledge translation includes semantic data augmentation, which enriches training data by applying symbolic reasoning to existing datasets. This process not only improves neural network generalization [16] but also enhances the interpretability and human-understandability of subsequent explanations [6].

The **knowledge extraction** dimension focuses explicitly on deriving symbolic knowledge from trained neural networks (see also Section 4.2). Knowledge extraction is a transformation from learned representations into comprehensible rules that support explanation and reasoning tasks [21]. This extracted symbolic knowledge optimizes criteria such as accuracy, fidelity, consistency, and comprehensibility.

[6] While we choose this particular definition, other ways to define them might include that a KG is a populated ontology, or that an ontology might act as a schema for KG.

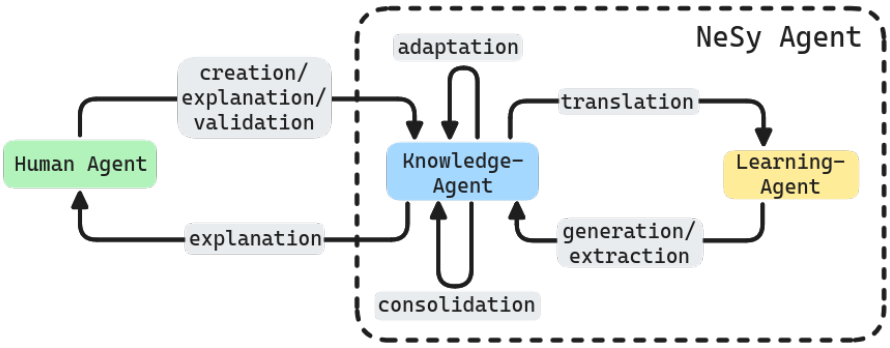

Figure 6 A proposed architecture for an ontology-mediated neurosymbolic AI.

Knowledge consolidation integrates the newly extracted symbolic knowledge back into existing knowledge structures. The revised and consolidated symbolic knowledge provides meaningful semantics to explanations, significantly facilitating human-machine interaction. Consolidated knowledge can also be reintegrated into neural models, enhancing their overall predictive performance and interpretability.

However, a more integrated approach is required, where knowledge graphs (KGs) and ontologies are treated as first-class components in neurosymbolic systems. Such integration actively involves users in the neurosymbolic cycle, empowering them to both consume and produce knowledge and explanations, thereby enabling a more robust and dynamic knowledge ecosystem.

Within this enriched integration, specific roles of knowledge emerge distinctly:

- **Knowledge creation** employs ontologies to support the formalisation of human expertise by providing a common language. This formalisation aims to capture implicit knowledge, insights, and expertise from human experts and domain specialists, structuring them explicitly using ontological languages and knowledge graphs.
- **Knowledge adaptation** employs ontology-based complexity management techniques such as abstraction, clustering, and refinement to adjust the granularity of knowledge graphs according to the demands of learning tasks and explanatory requirements.
- **Knowledge translation** leverages ontological reasoning to enrich and expand training data, guiding neural learning by explicitly embedding semantic constraints into model architectures and loss functions.
- **Knowledge generation and extraction** ensures that neural model predictions and symbolic knowledge remain consistent and interoperable, thus enhancing the reliability and transparency of derived explanations.
- **Knowledge consolidation** ensures the coherent integration of new symbolic knowledge with existing knowledge structures, maintaining consistency and semantic integrity across the entire knowledge base.
- **Knowledge explanation** grounds human-centered explanations in symbolic knowledge provided by ontologies and knowledge graphs, ensuring explanations are contextually meaningful and adapted to user expertise.

4.6.3 What is the NeSy ambition?

We envision NeSy as a hybrid agent-based architecture for continuous learning, composed of three key agents: the human agent, the knowledge-based agent, and the learning agent (Figure 6). This architecture aligns with contemporary approaches that leverage knowledge

graphs (KGs) [4] and is inherently ontology-mediated. We note that this is relatively exclusive whereby a learning agent (e.g., an LLM) is capable of *generating* or otherwise performing knowledge engineering (as in [19]). On the other hand, as these alternate methods grow in complexity (i.e., utilizing intermediate representations during the knowledge engineering tasks), it perhaps asymptotically approaches the first stated vision.

The **human agent** creates and contributes domain expertise, leveraging existing knowledge to solve specific tasks. It interacts directly with the knowledge-based agent by creating new knowledge or by **learning** new knowledge (in the form of predictions and explanations) about the system's reasoning process. Additionally, it can request further clarification if initial explanations are insufficient, actively participating in a human-in-the-loop manner to refine, consolidate and validate new knowledge.

The **knowledge-based agent** stores verified and consolidated knowledge, represented formally through ontologies and knowledge graphs. It serves the human agent by providing direct answers through symbolic inference or, when necessary, requesting new knowledge from the learning agent. Such interactions may involve **adapting** knowledge specifically for a given learning task or **translating** symbolic knowledge into semantically enriched data, or constraints into neural architectures and loss functions. The knowledge-based agent receives newly generated knowledge from the learning agent. When neural models are used, they **extract symbolic representations** in the form of structured symbolic elements, such as facts, rules, and logical statements. It **consolidates** this knowledge into existing ontologies, enriching its knowledge base while ensuring semantic consistency. This agent also delivers semantically enriched **explanations** tailored to human users.

The **learning agent** operates agnostically, capable of **generating new knowledge** either symbolically (e.g., via Inductive Logic Programming) or neurally (through deep learning methods). When neural models are used, **knowledge generation** is accompanied by the **extraction of symbolic representations** in structured symbolic forms.

In this ontology-mediated neurosymbolic framework, ontologies play pivotal roles across multiple knowledge-centric dimensions, specifically creation, adaptation, translation, generation and extraction, consolidation, and explanation.

In terms of **knowledge creation**, ontologies support the **formalization of human expertise** by providing a common language. This process captures implicit knowledge, insights, and expertise from human experts and domain specialists, structuring them explicitly using ontological languages and knowledge graphs. Ontologies facilitate knowledge creation by offering **standardized vocabularies** and **formal semantic frameworks**, enabling domain experts to express complex concepts, relationships, constraints, and rules in a precise, interoperable, and reusable manner. Such formalization ensures that previously tacit knowledge becomes explicitly available for reasoning, learning, validation, and explanation within neurosymbolic systems.

In terms of **knowledge adaptation**, ontologies enable **complexity management** through abstraction, clustering, and refinement techniques. These methods allow **fine-grained adjustments** to the granularity of the knowledge graph, aligning it closely with the specific requirements of learning tasks and the explanatory needs of end-users.

Regarding **knowledge translation**, ontologies facilitate the enrichment of training data through **semantic data augmentation**, leveraging deductive reasoning guided by ontological knowledge to enhance model inputs. Additionally, ontologies **guide feature selection** based on domain-specific semantics, thus improving the generalizability and interpretability of learning models. Furthermore, ontologies enable the **injection of logical constraints** directly into neural network training by encoding symbolic knowledge into loss functions, and they can **influence neural network architectures** by structuring them based on ontological categories and relationships, introducing effective inductive biases.

Within knowledge consolidation, ontologies facilitate the verification and validation of new knowledge generated by learning models, ensuring consistency and resolving conflicts with existing knowledge graphs. They facilitate incremental expansion of ontological frameworks, integrating newly learned symbolic knowledge in a coherent manner. Semantic interoperability is maintained by aligning new knowledge with established and standardised ontologies, promoting reusability across systems.

In terms of **knowledge generation and extraction**, ontologies help **align neural and symbolic representations**, thereby maintaining interoperability between predictions generated by neural models and the symbolic knowledge. This alignment supports the **extraction of symbolic rules and knowledge** from neural-based models, enhancing the clarity and global coherence of explanations.

Finally, for **knowledge explanation**, ontologies provide explicit **semantic grounding for explanations**, resulting in justifications that are human-understandable and contextually meaningful. Through ontology-driven abstraction, clustering, and refinement, the granularity and form of **explanations can be tailored** to the user's expertise. **Explanation-driven consolidation** further ensures transparency in knowledge integration processes by highlighting inconsistencies or unforeseen logical entailments, strengthening trust and interpretability.

4.6.4 What are the challenges, and how can we address them?

Using KGs in NeSy systems presents a range of fundamental challenges.

A crucial challenge is transforming **implicit information into explicit, machine-interpretable form**. Knowledge that is relevant to real-world tasks can potentially be inferred from data through learning or deductive reasoning, but establishing a procedure to extract and formalize this implicit knowledge, and then integrate it with established ontologies and KGs, is an open challenge (see also Section 4.2). Furthermore, properly mapping structured knowledge to raw data is crucial to ensure that the system effectively leverages the KG during both training and inference.

Another critical challenge is the **mismatch between the level of abstraction or granularity in the KG and the specific demands of the task**. KGs may contain very broad domain representations, while learning tasks often require fine-grained, context-specific features. This misalignment can degrade performance and introduce irrelevant knowledge, which adds complexity or noise. For learning agents, this may lead to suboptimal generalization or misinterpretation. For human users, it can result in confusing or overly complex explanations, undermining trust in the system's outputs. The difficulty of isolating modular, task-relevant knowledge further complicates system design and interpretability (see also Section 4.4).

Timeliness and **consistency** between the KG and real-world data also impact system performance. When new knowledge arises in the data that is not yet reflected in the KG, or when the KG evolves, but the model has not been updated. These discrepancies can lead to contradictions or inconsistencies in predictions and explanations. Effective knowledge integration mechanisms are needed to verify new information, resolve contradictions, and preserve logical consistency. **Explanations** may play a dual role here – both in surfacing inconsistencies and in making updates more interpretable to human users.

Finally, KGs pose **scalability** challenges in NeSy systems. There is often a trade-off between preserving semantic richness and interpretability, on the one hand, and achieving computational efficiency and seamless integration with neural components, on the other. High-fidelity semantic models may enable more precise reasoning and explanation, but they can be computationally intensive and harder to align with vector-based representations used

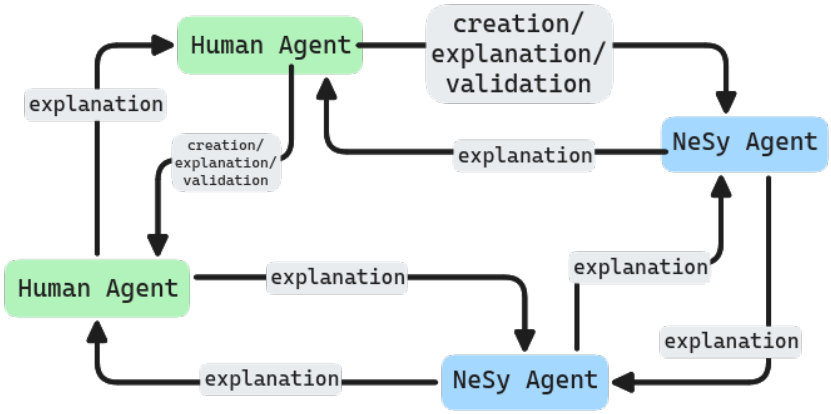

Figure 7 Multi-agent view of ontology-mediated neurosymbolic system.

in learning models. Balancing these tensions remains a central design challenge in building robust, scalable neurosymbolic AI. Figure 7 emphasizes a multi-agent view of the NeSy architecture, illustrating interactions between different types of agents (human, knowledge-based, learning), highlighting bidirectional knowledge flow and iterative refinement processes. NeSy agents can broadly include various forms, such as symbolic reasoning agents, neural agents, or even large language models (LLMs), given their symbolic output. Interaction between agents continuously updates each agent's beliefs and knowledge representations, emphasizing dynamic, evolving knowledge ecosystems (A self-loop arrow for knowledge creation might indicate self-driven learning or iterative internal refinement within human and NeSy agents).

References

1 Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

2 Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

3 BioPortal. https://bioportal.bioontology.org/.

4 Piero A. Bonatti, John Domingue, Anna Lisa Gentile, Andreas Harth, Olaf Hartig, Aidan Hogan, Katja Hose, Ernesto Jimenez-Ruiz, Deborah L. McGuinness, Chang Sun, Ruben Verborgh, and Jesse Wright. Towards computer-using personal agents, 2025.

5 Roberto Confalonieri and Giancarlo Guizzardi. On the multiple roles of ontologies in explanations for neuro-symbolic ai. *Neurosymbolic Artificial Intelligence*, 1:NAI–240754, 2025.

6 Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471, 2021.

7 Richard Cyganiak, David Wood, and Markus Lanthaler, editors. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation 25 February 2014, 2014. Available from http://www.w3.org/TR/rdf11-concepts/.

8 Claudia d'Amato, Louis Mahon, Pierre Monnin, and Giorgos Stamou. Machine learning and knowledge graphs: Existing gaps and future research challenges. *TGDK*, 1(1):8:1–8:35, 2023.

9 Dbpedia. https://wiki.dbpedia.org/.

10 Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic ai: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, Nov 2023.

11 Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. In Luc De Raedt, editor, *Proceedings of the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, IJCAI-ECAI 2022, Survey Track, Vienna, Austria, July 23-29, 2022*, pages 5478–5485. IJCAI/AAAI Press, July 2022.

12 Giancarlo Guizzardi and Nicola Guarino. Explanation, semantics, and ontology. *Data Knowl. Eng.*, 153:102325, 2024.

13 Steven Harris and Andy Seaborne. SPARQL 1.1 query language. W3C recommendation, W3C, March 2013. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

14 David Herron, Ernesto Jiménez-Ruiz, and Tillman Weyde. On the potential of logic and reasoning in neurosymbolic systems using owl-based knowledge graphs. *Neurosymbolic Artificial Intelligence*, 1:29498732251320043, 2025.

15 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021.

16 Majlinda Llugiqi, Fajar J Ekaputra, and Marta Sabou. Semantic-based data augmentation for machine learning prediction enhancement. *Neurosymbolic Artificial Intelligence*, 1:29498732251340160, 2025.

17 Boris Motik, Peter Patel-Schneider, and Bijan Parsia. OWL 2 web ontology language structural specification and functional-style syntax (second edition). W3C recommendation, W3C, December 2012. https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/.

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

19 Cogan Shimizu and Pascal Hitzler. Accelerating knowledge graph and ontology engineering with large language models. *J. Web Semant.*, 85:100862, 2025.

20 Gunjan Singh, Sumit Bhatia, and Raghava Mutharaju. Neuro-symbolic RDF and description logic reasoners: The state-of-the-art and challenges. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 29–63. IOS Press, 2023.

21 Geoffrey G. Towell and Jude W. Shavlik. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 13(1):71–101, Oct 1993.

22 Son Tran, Edjard Mota, and Artur d'Avila Garcez. Reasoning in neurosymbolic ai, 2025.

23 Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

4.7 Cognition and Neurosymbolic AI

Mena Leemhuis (University of Bozen-Bolzano, IT), Mehwish Alam (Télécom Paris, Institut Polytechnique de Paris, FR), Dagmar Gromann (University of Vienna, AT), Alessandro Oltramari (Bosch Center for AI, US & Carnegie Bosch Institute, US), Ute Schmid (University of Bamberg, DE), Eugene Vasserman (Kansas State University – Manhattan, US)

4.7.1 What is it?

Despite super-human performance of AI systems in very specific areas such as board games or object recognition and above average performance for tasks such as translating or summarizing texts, AI systems are lacking the flexibility, robustness, and generality of human cognition. Human cognition is based on causal models of the world allowing for understanding and explaining which is crucially different from solving pattern-recognition problems. Learning is guided by previous knowledge and experience and allows for flexible adaptation and revision based on novel information [16]. In consequence, humans can often generalize complex and productive rules from few examples [11], teach other humans about what they have learned, for instance by explanations [18], interleave learning and meta-cognition to monitor and evaluate generalizations and conclusions [32], and can apply/transfer existing knowledge to new problems and situations [17].

When human intelligence is compared with GenAI models, cognitive mechanisms such as memory, attention, and generalization are often simplified, if not completely misunderstood. For instance, the famous paper "attention is all you need", is not at all about attention, but rather focused on computational processes of working memory [31]. We tend to ascribe cognitive properties to Large Language Models (LLMs) and foundation models as a result of their human-like behavior/performance across well-defined tasks and benchmarks, although the types of errors that affect those systems are generally not observed in humans and hard to reconcile with experimental evidence from cognitive psychology. Common sense tasks, e.g., question answering, show this phenomenon quite clearly.

While LLMs show unprecedented linguistic performance, their language use is not grounded in the physical world, thus words are not connected to their real-world referents. The degree to which pattern-recognition in LLMs constitutes understanding language is debatable [7]. Human linguistic competences are formal, using the form of a language correctly, and functional, which refers to goal-directed language use [19]. To emulate both competences, models require world knowledge, ability to track changes over time, reasoning and problem-solving skills, and consideration of situative, pragmatic context [19].

Humans are often able to generalize complex and productive rules from a very small set of examples [20]. This ability is covered in many intelligence test problems such as Raven Progressive Matrices or induction of number series [10], in solving puzzles like the Tower of Hanoi [15, 28], and in generalizing relations such as "greater than" to different domains (e.g. numbers and sizes of objects). It is also apparent in language learning, for instance when learning the regular form of the past tense of verbs [20]. This cognitive ability is related to the human ability of analogy making in visual as well as in semantic domains [23]. In contrast, the data need is a severe issue of many modern AI approaches (as discussed in section "small data"). To reduce the data need, neurosymbolic architectures need to integrate perception and knowledge into generalizing from a few examples and should be based on single, generalized core mechanisms.

As discussed in section "XAI", in the context of the requirement of transparency and human control for AI systems, explainable AI (XAI) has become an active area of research [22]. Current approaches to explainability are typically one-shot and "one-size-fits-all". In contrast, explanations are central to human understanding and for the communication of causal knowledge and beliefs [18]. Although human explanations might sometimes be ex-post constructed justifications, they are considered as an important constituent of human conceptual representations [18]. Explanations in human communication are often a sequence of elucidations, varying in level of detail and modality [22, 8]. Neurosymbolic architectures should provide approaches to explainability which allow this type of context-specific adaptation to the specific information needs of a person in a given situation.

Neurosymbolic approaches have the promise to realize AI systems with more human-like abilities of learning and reasoning. AI approaches which are more closely aligned with the characteristics of human information processing may solve tasks which are better solved by humans, and also contribute to cognitive science research by providing computational models of human cognition. Finally, more human-like AI systems support better human-AI alignment for joint problem solving and decision making tasks [30].

Therefore, tackling the major goals of neurosymbolic AI, such as the data need, explainability or symbol emergence could profit from cognitively inspired approaches. This especially hints towards neurosymbolic AI, as human cognition has a tight connection to neurosymbolic ideas. In this regard, e.g., Kahneman [13] proposes with system 1 and system 2, that human cognition combines implicit and explicit reasoning. There are several viewpoints on how humans are able to model this tight integration between this implicit and explicit information. For example, Gärdenfors proposed with his Conceptual Spaces [9] one way of interpreting this human ability by modeling explicit information as geometric structures directly in the feature space.

Cognitively inspired AI approaches have a long standing tradition, e.g., prototype theory by Rosch [27] dates back to the 70's and inspired many AI approaches since then.

Other cognitively inspired AI approaches are, e.g., by [34], enhancing LLMs by incorporating attention, memory, reasoning, learning, and decision-making mechanisms or using conceptual spaces as basis for a learning approach [4]. However, they do not reach the level of tight integration needed, neither for a human-like performance nor for making up a human-machine interface.

Cognition is, however, not only of importance for reaching human-like performance due to cognitively-inspired AI. Next to that, NeSy opens up the opportunity to bridge the gap between humans and machines also in another way: by enabling human-machine collaboration by defining a neurosymbolic world model shared by human and agent.

4.7.2 What is the NeSy ambition?

Neurosymbolic AI could be used to bridge humans and machines by providing a sort of "cognitive API", thus an application programming interface for human-machine collaboration. This would allow humans and machines to negotiate a common understanding of the world ("world model") that can be expressed in a shared language. Such a world model needs to be based on a deep integration of subsymbolic and symbolic representation and reasoning: It is necessary to tightly connect explicit (maybe also shared) knowledge, e.g., in form of ontological information and implicit, feature or similarity-based information. Thus, there needs to be a grounding of the symbolic information. This especially requires an interpretation of meaning in natural language: Symbolic approaches that capture and explicitly represent world knowledge need to be integrated with powerful language representations that cater to the formal competences. This holds the promise of bringing AI systems one step closer to natural language understanding and interpretation of meaning encoded in language rather

than emulating or mimicking language behavior. This allows for both a deeper machine understanding of the human's world model and the ability of symbol emergence in the machine's world model to facilitate negotiation with a human. Construction of world models occurs by learning and generalizing from experience / data, constantly acquiring and updating knowledge.

One vital human ability needed in this context is negotiation. It is a process of disambiguation that can enable explanation, knowledge transfer (teaching, education), and collaborative decision-making. Negotiation occurs at the symbolic level, and is the iterative process used to communicate world models, achieve consensus, and make decisions.

A cognitive API should not only be able to react to symbolic negotiations but should also be able to react to non-verbal negotiations, e.g., via physical cues (especially when humans "teach" an embodied AI model how to perform physical tasks). These models must be able to "generalize", e.g., not repeat every little movement but differentiate the necessary movements from incidental ones (for instance, inferring the human intention or plan; filtering out session-specific actions as noise in multi-session training scenarios). When such inference is symbolic, we can achieve generalizable "understanding" rather than situation-specific mimicry. Physical interaction enables (semi-)autonomous control of the physical world by machine models, e.g., autonomous driving and flying. Thus, the API should be able to adapt its shared world-model in line with the human's needs.

Note that for this ambition, a cognitively inspired AI system could be beneficial, but is, however, not necessary.

4.7.3 Where are we now?

Agentic AI frameworks (see [1] for a recent survey) are used today to improve human-machine collaboration. These frameworks allow humans to execute complex tasks using LLMs connected with various software systems, such as web services, and data processing pipelines. For instance, you can ask Google Gemini to book a hotel close to where your conference is, which requires the model to use location-based services, third party websites for price comparison, reservation, etc. De facto, these systems are unidirectional: although dialog-based interaction is possible, and oftentimes necessary, there's neither an assumption nor a requirement for mutual explanation between machines and human, for bidirectional knowledge acquisition (a concept learned by a human being transferred – "taught" to a large model, and vice versa).

LLMs can simulate aspects of understanding and creativity by processing large linguistic or multimodal inputs and generating context-aware outputs. However, they often fall short in tasks involving reasoning and causal inference, which demand generalization beyond their training data [29]. To address these limitations, next to the points discussed in section "GenAI", integrating LLMs into cognitive architectures, thus systems modeled on human cognition, can enhance their robustness, adaptability, and reasoning. In this context, knowledge-grounded LLMs that use structured external knowledge are especially effective. Cognitive agents that manage reasoning, memory, or symbolic operations can further boost LLMs' capabilities and bring them closer to human-like intelligence. Still, linking LLMs to human cognition requires both theoretical and empirical validation.

Expanding cognitive architectures with multimodal inputs like eye-tracking and fMRI data can enhance alignment between artificial systems and human cognition. A benchmark dataset allowing for such examinations has recently been published [35]. Eye-tracking provides insights into attention, reading behavior, and decision-making, while fMRI captures brain activity related to language, memory, and reasoning. These allow on the one hand

to examine the alignment of AI approaches and human thought processes (as discussed, e.g., in [6]). On the other hand, integrating these signals allows cognitive models to ground language understanding in both perceptual and neural data, leading to richer, more human-like representations and responses. This approach supports the development of agents with more sophisticated, cognitively informed world models and thus communication options. However, this is a research area that is still in its infancy.

4.7.4 What are the challenges, and how to address them?

Although neurosymbolic approaches to learning and reasoning have the representational advantage of modeling both the sub-symbolic/neural and symbolic/reasoning facets of human cognition more faithfully, many of the aspects of human cognition and especially of a cognitive API discussed above are currently not or only partially addressed.

Language and Cognition. LLMs excel at mimicking human writing [7] and language. However, even in terms of formal competences, complex grammatical tasks or those that require correct semantic interpretation remain challenging, such as semantically illegal cardinality comparisons as in Fewer athletes have been to Beijing than I have [24]. Tasks that target the functional competences, including Natural Language Inference (NLI), fact checking/verification, and multi-hop question answering, are generally self-contained and represent only a small slice of the real-world. Such an approximation of meaning in AI systems might not necessarily represent understanding real-world referents [19]. A more accurate evaluation of the language-cognition alignment in AI requires more realistic benchmarks that go beyond linguistic surface forms or only small proportions of functional competences. IBM Watson's participation in Jeopardy! is a well-known example of such a challenging, knowledge-rich testbed for AI systems. A cognitive API would require such a deep real-world understanding.

Flexible re-representations for learning productive rules and abstractions from few examples. Learning complex rules and abstractions have been addressed with different, not human-like strategies [11]. For instance, generate-and-test approaches have been used to tackle the abstract reasoning challenge[7]. In contrast, humans often seem to know immediately what the relevant aspects are for generalization [14]. Purely symbolic approaches often cannot deal with noisy or imperfect input and rely on carefully tailored representations. Some autonomous driving systems, especially those that integrate world modeling / recognition and control are frequently fragile in the face of "minor" unexpected real-world artifacts, e.g., a few white dots pasted onto a stop sign or lane marker thoroughly confuses models, which may no longer see it as a traffic sign or lane marker at all; it is also possible to induce recognition as a different sign altogether [2, 3, 26, 25]. Humans are more robust in recognizing road signs, lane markers, etc., as well as inferring the plans of other drivers and pedestrians, but humans are worse at paying continuous attention to driving tasks. Furthermore, humans flexibly re-represent information in such a way that examples can be suitably aligned, that is in a goal directed manner. This has been addressed by Douglas Hofstadter with the Copycat system [12] and also by making use of background theories [33]. Current approaches to solve visual abstract reasoning problems are often neurosymbolic by combining representation learning with rule learning [21, 36]. While this seems a promising way to go, these approaches need large sets of training data and, again, re-representation is

[7] https://arcprize.org/

not addressed, restricting rule learning to the class of entities which has been present in the training examples. This especially challenges a shared human-agent world model: a shared world model is only possible when the agent is able to figure out the relevance of and can abstract from the given information.

Integrated symbolic and subsymbolic systems. Cognitive models such as Gärdenfors conceptual spaces allow for modeling a tight connection between implicit and explicit information and thus could be considered as a good strategy for modeling a tightly integrated neurosymbolic system or as a starting point for a representation of a shared world model. However, this tight connection comes to the cost of a high modeling complexity. One solution strategy is to find a trade-off between tightness of the connection between symbolic and subsymbolic and usability. Research in the area of knowledge base embeddings (KBE) [5] can be seen as a special case of conceptual spaces, where the geometric space is not aimed to model feature information but is solely focused on similarity information. This enables link prediction and query answering by modeling concept conjunction, however, it comes at the cost of losing semantic information in the form of features. This tradeoff between learnability and the representation of semantic information is a fundamental design consideration that needs to be carefully addressed when developing these cognitive inspired approaches. This also directly points towards the problem of symbol rising as discussed in the section of "symbol emergence": how can the human ability of not only reasoning with given symbols but also introducing new symbols could be handled?

Human Machine Collaboration. To conclude, all these aspects are relevant for modeling a world model shared between human and agent. However, for such a "cognitive API", there are currently not even specified requirements available. For Negotiation, what kind of language do we need? Or should we base this process on a library of languages, depending on the tasks or domains under consideration? For Construction of world models: learning is clearly a capability that both humans and machine models need, in order to distill and consolidate their knowledge. But the differences between the two modalities of learning are stark: for instance, while humans learn from few examples, neural networks require massive amounts of data. How do these different ways of learning affect the representation of the world models? Does mapping human symbols to machine-generated symbols in a common world model require mapping their different construction processes? Evaluation: how can we effectively evaluate that human-machine collaboration was successful, given a task x? Most of the current metrics focus on very specific quantitative properties of tasks (hits@k, Bleu score), however new qualitative metrics may be needed to measure collaboration. This topic is discussed in more detail in Section 4.8.

References

1 Deepak Bhaskar Acharya, Karthigeyan Kuppan, and Divya Bhaskaracharya. Agentic AI: autonomous intelligence for complex goals - A comprehensive survey. *IEEE Access*, 13:18912–18936, 2025.
2 Evan Ackerman. Three Small Stickers on Road Can Steer Tesla Autopilot Into Oncoming Lane. *IEEE Spectrum*, 2019.
3 Ross Anderson and Ilia Shumailov. Situational awareness and adversarial machine learning – robots, manners, and stress. *Apollo*, 2021.
4 Lucas Bechberger. *Using Conceptual Spaces for Artificial Intelligence*. PhD thesis, Universität Osnabrück, 2023.
5 Camille Bourgaux, Ricardo Guimarães, Raoul Koudijs, Victor Lacerda, and Ana Ozaki. Knowledge base embeddings: Semantics and theoretical properties. In Pierre Marquis,

Magdalena Ortiz, and Maurice Pagnucco, editors, *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, 2024.

6 Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), February 2022.

7 Christine Cuskley, Rebecca Woods, and Molly Flaherty. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083, 08 2024.

8 Bettina Finzel, David E. Tafler, Stephan Scheele, and Ute Schmid. Explanation as a process: User-centric construction of multi-level and multi-modal explanations. In Stefan Edelkamp, Ralf Möller, and Elmar Rueckert, editors, *KI 2021: Advances in Artificial Intelligence - 44th German Conference on AI, Virtual Event, September 27 - October 1, 2021, Proceedings*, volume 12873 of *Lecture Notes in Computer Science*, pages 80–94. Springer, 2021.

9 Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, 03 2000.

10 José V. Hernández-Conde. A case against convexity in conceptual spaces. *Synthese*, 194(10):4011–4037, Oct 2017.

11 José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L. Dowe. Computer models solving intelligence test problems: Progress and implications (extended abstract). In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 5005–5009. ijcai.org, 2017.

12 Douglas R. Hofstadter and Melanie Mitchell. The copycat project: A model of mental fluidity and analogy-making. In *Analogical connections*, Advances in connectionist and neural computation theory, Vol. 2, pages 31–112. Ablex Publishing, Westport, CT, US, 1994.

13 Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

14 Craig A Kaplan and Herbert A Simon. In search of insight. *Cognitive Psychology*, 22(3):374–419, 1990.

15 K Kotovsky, J.R Hayes, and H.A Simon. Why are some problems hard? evidence from tower of hanoi. *Cognitive Psychology*, 17(2):248–294, 1985.

16 Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *CoRR*, abs/1604.00289, 2016.

17 J. Loewenstein, L. Thompson, and D. Gentner. Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, 6(4):586–597, dec 1999.

18 Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470, 2006.

19 Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024.

20 Gary F. Marcus. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Learning, Development, and Conceptual Change. MIT Press, Cambridge, MA, 2003.

21 Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736, 2023.

22 Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

23 Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *CoRR*, abs/2102.10717, 2021.

24 Elliot Murphy, Evelina Leivada, Vittoria Dentella, Fritz Guenther, and Gary Marcus. Fundamental principles of linguistic structure are not represented by o3. *CoRR*, abs/2502.10934, 2025.

25 Ben Nassi and Yuval Elovici. Protecting autonomous cars from phantom attacks. *Communications of the ACM*, 66:56–69, March 2023.

26 Ben Nassi, Yisroel Mirsky, Dudi Nassi, Raz Ben-Netanel, Oleg Drokin, and Yuval Elovici. Phantom of the ADAS: securing advanced driver-assistance systems from split-second phantom attacks. In Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna, editors, *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 293–308. ACM, 2020.

27 Eleanor Rosch. *Principles of Categorization*, pages 312–322. Elsevier, 1988.

28 Ute Schmid and Emanuel Kitzelmann. Inductive rule learning on the knowledge level. *Cogn. Syst. Res.*, 12(3-4):237–248, 2011.

29 Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025.

30 Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, Dec 2024.

31 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

32 Marcel V. J. Veenman, Bernadette H. A. M. Van Hout-Wolters, and Peter Afflerbach. Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, 1(1):3–14, Apr 2006.

33 Stephan Weller and Ute Schmid. Solving proportional analogies by E -generalization. In Christian Freksa, Michael Kohlhase, and Kerstin Schill, editors, *KI 2006: Advances in Artificial Intelligence, 29th Annual German Conference on AI, KI 2006, Bremen, Germany, June 14-17, 2006, Proceedings*, volume 4314 of *Lecture Notes in Computer Science*, pages 64–75. Springer, 2006.

34 Yuanzhen Xie, Tao Xie, Mingxiong Lin, WenTao Wei, Chenglin Li, Beibei Kong, Lei Chen, Chengxiang Zhuo, Bo Hu, and Zang Li. Olagpt: Empowering llms with human-like problem-solving abilities, 2023.

35 Yunhao Zhang, Xiaohan Zhang, Chong Li, Shaonan Wang, and Chengqing Zong. Mulcog-bench: a multi-modal cognitive benchmark dataset for evaluating chinese and english computational language models. *Language Resources and Evaluation*, 59(3):3005–3028, May 2025.

36 Shukuo Zhao, Hongzhi You, Ru-Yuan Zhang, Bailu Si, Zonglei Zhen, Xiaohong Wan, and Da-Hui Wang. An interpretable neuro-symbolic model for raven's progressive matrices reasoning. *Cognitive Computation*, 15(5):1703–1724, Sep 2023.

4.8 Benchmarks in the Neurosymbolic Ecosystem

Claudia d'Amato (University of Bari, IT), Jennifer D'Souza (TIB Leibniz Information Centre for Science and Technology, DE), Anna Lisa Gentile (IBM Research Almaden, US), Hande McGinty (Kansas State University – Manhattan, US)

4.8.1 What is Benchmarking?

In computing, a benchmark is the act of running a computer program, a set of programs, or other operations, in order to assess the relative performance, normally by running a number of standard tests against it [35]. The goal of benchmarks is to provide a quantitative consensus on what constitutes good performance and to represent a shared framework for the comparison of methods [31]. Depending on the problem, there are several types of benchmarking techniques, including behavioral frameworks [34], task completion assessment [41, 3, 45], human-in-the-loop evaluations [12, 1], game-based [30, 17], etc. A dataset-based benchmark is a standardized performance test, usually consisting of a dataset or a set of datasets, a collection of questions or tasks, and a scoring mechanism including one or more metrics [31]. The task is a particular specification of the problem (as represented in the dataset). A metric is a way to summarize system performance over datasets of task(s) as a single number or score. The metric provides a means of counting success and failure at the level of individual system outputs and summarizing those counts over the full dataset [31].

4.8.1.1 Benchmark Properties, Quality and Best Practices

Various studies have attempted to provide guidelines and best practices to create effective and meaningful benchmarks, and study their desired properties. A dataset based benchmark can be formalized as a triple <dataset, task, metric>. The quality of a benchmark depends on these three components, their combination, and their usage. In [32], four main properties for minimum quality assurance have been defined, encompassing: (i) downstream utility, as grounded in real-life scenarios; (ii) validity, including size for statistical significance of results; (iii) regular updating the benchmark over time to prevent overfitting; (iv) interpretability of the score; and (iv) accessibility. A few studies propose essential guidelines to ensure good benchmarking, some addressing high level qualities such as scope or representativeness [42]; some with very detailed and granular "checklists" [7]. A well known and accepted methodology has been proposed by BetterBench, that used 46 criteria to assess the quality of a benchmark [33].

4.8.1.2 Metrics and Evaluation Methods

A metric can be defined as a computable way for measuring something quantitatively e.g, reuse metrics in software engineering, search engine quantification metrics, classification metrics, and many others. The choice of the metrics depends on the task and the properties we are interested in (for a given dataset). An example is the task of assessing the quality of software, which is typically grounded on the computation of metrics such as lines of code, cyclomatic complexity, average nesting depth, software length, effort, and time metrics [13].

In the specific perspective of neurosymbolic (NeSy) systems, multiple tasks may be of interest, including classification, semantic parsing, knowledge graph completion, visual reasoning, logical entailment, program synthesis, and symbolic planning, among others –

spanning fields such as language, vision, robotics, and ontology engineering (see Section 4.8.4 for more details). Each task can be evaluated along one or, more often, multiple metrics. For example, Classification-based metrics, such as precision, recall, F_1-score, and accuracy, are prevalent in tasks such as ontology learning, ontology alignment, semantic parsing, and visual reasoning, where discrete predictions (e.g., relation labels, answers) are evaluated against ground truth. Ranking-based metrics, such as mean reciprocal rank (MRR) and Hits@N, are standard for knowledge graph completion tasks, where models rank candidate entities. Execution-based metrics dominate in program synthesis and instruction-following benchmarks, measuring whether predicted programs (e.g., SQL queries, action plans) produce the correct output or state transition. Success rate and goal-conditioned success are often used in embodied reasoning and planning tasks, indicating whether an agent achieves the desired final state. Some benchmarks (e.g., EntailmentBank [10], ProofWriter [40]) additionally assess the quality of explanations besides the correctness of the answers, recognizing the importance of interpretable and logically coherent reasoning chains.

Beyond task-specific accuracy and ranking measures, several language generation metrics are used in translation, summarization, and structured prediction tasks. Exact match is a strict criterion, particularly valuable in question answering and program synthesis. Perplexity, originally introduced in the context of speech recognition [19], is a measure of uncertainty: the larger the perplexity, the less likely it is that an observer can guess the value which will be drawn from the distribution – in the context of LLMs it assesses how well a model predicts the next token, with lower values indicating better fluency. BLEU (Bilingual Evaluation Understudy [28]) is a measurement of the difference between an automatic translation and human-created reference translations of the same source sentence, specifically, it measures n-gram overlap in translation, while ROUGE [26], including ROUGE-N and ROUGE-L, evaluates summarization by capturing recall and sequence similarity. Together, these metrics capture complementary aspects of performance – from structural correctness and symbolic consistency to fluency and semantic alignment – highlighting the multifaceted nature of system evaluation. This diversity of metrics underscores the need for task-specific evaluation while also highlighting opportunities for unified evaluation protocols. Nevertheless, as discussed in the next section, in the specific perspective of evaluating neurosymbolic systems, there are challenges that need to be addressed, particularly aiming at assessing both neural performance and symbolic faithfulness.

4.8.2 What are the challenges, and how to address them?

Current benchmarks adopted for evaluating NeSy solutions are mostly grounded on performance with respect to defined metrics (e.g., F_1, MRR, hits@k) (see Section 4.8.4 for details). These certainly contribute to providing an objective, measurable, comparable quantification of system performances on a given task; however, they only provide a partial picture/answer to the problem. Even more so, particularly for the case of Machine Learning related tasks, evaluation revolves around inappropriate averaging/aggregation of the results [35].

For example, given two NeSy systems having rather similar F_1 values, the kind of failures may be due to different causes (e.g. errors that are due to one system being sensitive to the data distribution, errors due to the inability of the other system to capture similarities in the data space). A NeSy benchmark needs to be able to dissect the architectural components of NeSy systems in order to the assess the influence they have in the final assessment of the performances.

Similarly, given two NeSy systems having rather similar F_1 values, the current benchmarks and metrics fail to capture the influence of the neural/symbolic component respectively, on the final performance and even more so, how sensitive a system is with respect to the

(semantic) quality of each component. Indeed, by the use of symbolic (semantic) components of the NeSy systems, NeSy aims to unbox the black-box systems and inject explainability to the systems.

Another shortcoming of the current metrics and benchmarks revolve around the solutions NeSy systems can provide with small datasets. By design, NeSy systems can have the ability to provide value when datasets are small in amount or there is great internal variation within the dataset. However, current performance-based metrics may be failing to identify how a NeSy system may be able to provide better predictions based on the abstraction and symbolic components working together with numeric machine learning approaches. This kind of deeper understanding requires more fine-grained properties and metrics to be evaluated.

The analysis reported above suggests that, when talking about properties, not only benchmark properties need to be considered (see Section 4.8.1.1) but also properties referring to the (NeSy) system to be evaluated need to be taken into account, and both of them (jointly with their respective measurable metrics) should be part of the benchmark description, with a clear distinction.

Examples of benchmark properties are: dynamic benchmarking (evolving over time with respect to the dataset, keeping different dataset versions, metrics for measuring the variation within the dataset) vs. static benchmarking (stable over the time or updated from time to time) vs. real time benchmarking (benchmarks that might not exist, but need to be built in real time); data consistency vs. data inconsistency (assessed logically); benchmarks evaluating different steps along solving the targeted task (e.g. in abstract visual puzzle it is possible to distinguish the steps: perception, abstraction (what matters?), strategy; or in hypothetical reasoning different steps could be: counterfactuality, anticipatory thinking, causality) vs. benchmarks evaluating holistic solutions for the targeted task (e.g. classification).

Examples of system properties to be evaluated are: the ability to cope with/learn from small data collections vs. large data collections; the ability of explaining the learning process to the solution vs. the ability to explain/justify the solution itself disregarding the learning process vs. lack of ability of providing explanations/justifications; (lack of) robustness against data drift; verifiability, that is to be coherent with respect to existing domain knowledge; scalability and possibly also actionability, that is what can be done to change the system decision.

Last but not the least, a deeper understanding and identification of concrete problems/tasks for which the adoption of NeSy solutions in the first place is particularly beneficial is also needed.

4.8.3 What is the NeSy ambition?

Moving from the main need of having a systematic way for evaluating system/agent performances on (established) datasets by computing metrics that are relevant for assessing the system/agent ability to solve a targeted task, the main desiderata from the NeSy perspective are the ability to dissect benchmarks and evaluation under the multiple dimensions analyzed in Sect. 2, and most of all, assessing the impact and sensitivity of the characterizing (semantic) symbolic component of NeSy solutions.

Particularly, the ambition is on the definition of a generalized methodology for building benchmarks that can be operationalized and automatically customized providing the task on interest and the desired benchmark/system properties as input. While we recognize the non trivial challenge of our ambition, we evaluate it very promising both in terms of impact and feasibility. Indeed, preliminary results considering specific tasks already exist [5]. Additionally, we intend to build on existing alternative and complementary assessment

methods [31] used in different systems. These may include (but are not limited to) systematic development of test suites, audits, and adversarial testing, analyzing failure modes: system output analysis, behavioral testing, error analysis, disaggregated analysis, and counterfactual analysis, ablation testing, and analyzing model properties that are orthogonal to system outputs, such as profiling energy consumption, memory requirements, and stability in the face of perturbations to training data.

A complementary goal is the definition of a framework for checking the match/compliance of existing benchmarks with respect to (selected) properties which may enable studying/assessing the impact of these (missing) properties in solving the selected task. Additionally, assessing the fitting properties for an existing benchmark may be exploited for determining the complexity of the benchmark.

4.8.4 Where are we now?

In this section we analyze state of the art benchmarks across multiple task categories. Specifically, an overview of representative benchmarks across these task categories, including their symbolic components, domains, and evaluation metrics, is provided in Table 1. These benchmarks span a diverse set of tasks central to neurosymbolic AI, including ontology matching, knowledge graph completion, program synthesis, visual and relational reasoning, and embodied planning. Many integrate symbolic structures – such as ontologies, logical forms, scene graphs, or formal programs – with tasks that test reasoning, generalization, or action planning. Metrics range from precision/recall in classification tasks to exact match, logical form accuracy, Hits@N, and task success rate, reflecting the multifaceted nature of evaluation in this space (see also Section 4.3).

Despite this breadth, notable gaps remain. First, many benchmarks rely on static symbolic formalisms (e.g., predefined ontologies or logic rules) without testing models' ability to construct, revise, or explain symbolic representations. Second, existing datasets tend to isolate symbolic reasoning from learning under uncertainty or naturalistic conditions. Few benchmarks systematically evaluate alignment between neural and symbolic outputs, or explicitly measure interpretability, trustworthiness, or symbolic faithfulness. Moreover, coverage across scientific domains and real-world applications remains uneven – domains like biology, law, or social science are underrepresented. Addressing these limitations calls for the design of next-generation benchmarks that jointly test symbolic competence, robustness, and alignment with real-world reasoning, including multi-modal grounding.

Table 4 Benchmarks Overview.

Dataset	Task	Symbolic component	Domain	Scale	Metrics
Ontology Alignment Evaluation Initiative (OAEI) [29]	Ontology Alignment		Anatomy, Conference, Multifarm, Food, Bio-ML, Biodiversity and Ecology, Digital Humanities, Archaeology, Circular Economy, Knowledge Graphs, Pharmacogenomics		P, R, F1, Semantic P., Semantic R. [14], Runtime, consistency and conservativity [20, 37]
Large Language Models for Ontology Learning (LLMs4OL) [4, 15]	Ontology Learning		Biomedicine, Material Science, Earth and Environmental Science, Medicine, Food, Plant, Chemistry, Web		P, R, F1
FB15k-237 (from Freebase) [6]	Knowledge Graph Entity Completion	The data is essentially an ontology/graph; models often use embedding techniques but can incorporate ontological constraints or logical rules (e.g. transitivity).			Link prediction hits@N and mean reciprocal rank (MRR)
YAGO3-10	Knowledge Graph Entity Completion	As a curated ontology-derived KG, it includes type hierarchies and relational schema that methods can exploit (e.g. hasCitizenship implies a type constraint on object = Country).		123,182 entities, 37 relations, and ~1,179,040 triples, focusing heavily on persons (with relations like bornIn, hasProfession, etc.)	link prediction metrics (MRR, Hits@1/3/10)
VQA v2.0 dataset [16, 2]	Visual QA and Visual Reasoning	Questions often imply structural reasoning (counting objects, identifying attributes, etc.), though no explicit knowledge base is given.		265K images, ~1.1M questions	Acc.
Compositional Language and Elementary Visual Reasoning (CLEVR) [21]	Visual QA and Visual Reasoning	Each question comes with a functional program that specifies the reasoning steps, and ground-truth scene graphs are provided. The task is to answer complex compositional questions about the scene (counting, comparing attributes, logical operations) designed to require multi-step reasoning with minimal dataset bias.		100K images, ~865K questions	Acc.
Collision Events for Video Representation and Reasoning (CLEVRER) [43]	Video QA and Video Reasoning	Queries are designed to test understanding of physical causality; a symbolic program (event logic) can be used to derive answers		20,000 videos, annotations, and questions	Acc., per option acc., per ques acc.
Cornell Natural Language Visual Reasoning (NLVR) [38]	Visual Reasoning			92,244 pairs of examples of natural statements grounded in synthetic images with 3,962 unique sentences	Acc.
NLVR 2 [39]	Visual Reasoning			107,292 examples of English sentences paired with web photographs	Acc.
Question Answering on Image Scene Graphs (GQA Dataset) [18]	Graph QA	Images are labeled with objects, attributes, relations (scene graph); questions are represented as functional programs, enabling symbolic reasoning over the scene		113K images, 22M compositional questions	

Continued on next page

Name	Task	Symbolic component	Domain	Scale	Metrics
Outside-Knowledge VQA [27]	Visual QA and Visual Reasoning	Often leverages a knowledge base or ontology (e.g. WordNet, Wikipedia) for reasoning		~14K questions	Acc.
Spider (Text-to-SQL) [44]	Program synthesis and semantic parsing	The target output is a SQL program, and the task tests the model's ability to incorporate schema knowledge (table/column names) and logic.		10,181 natural questions and 5,693 unique SQL queries across 200 databases	Acc.
WikiSQL	Program synthesis and semantic parsing	The task is essentially mapping natural language to a SQL logical form, requiring understanding of conditions and mappings to table schema.		80,654 questions and SQL queries over 24,241 Wikipedia tables	Execution acc. (whether the predicted SQL yields the correct answer on the table) and logical form acc.
Compositional Freebase Questions (CFQ) [23]	semantic parsing	Uses a fixed ontology (Freebase schema) and logical queries (SPARQL)		240k natural questions generated from Freebase (with answers), each paired with a SPARQL query against a knowledge graph	Acc. and compound divergence metric.
SCAN (Simplified versions of the CommAI Navigation tasks) [24]	semantic parsing	The action sequence can be seen as a program the agent must generate, and generalization requires following combinatorial rules (e.g. learning the meaning of "twice").			Exact match of the output action sequence; various splits test zero-shot compositional generalization
ARC (Abstraction and Reasoning Corpus) [9]	semantic parsing	The underlying solution for each puzzle is effectively a small program or rule (e.g. "reflect the pattern", "count and place objects") that the AI must infer. No training examples are provided per task, emphasizing few-shot abstract reasoning		contains 1,000 grid-based puzzles where a small set of input-output examples is given and the model must output the result for a new input	Number of tasks solved perfectly (traditionally, execution on hidden test cases)
The "Compositional Language Understanding and Text-based Relational Reasoning" benchmark (CLUTRR) [36]	Logical and Relational Reasoning	Under the hood, each story corresponds to a small knowledge graph of family relations and a logical proof chain; the dataset systematically varies the number of hops and adds distracting facts to test inductive reasoning robustness			Relation prediction acc.
Higher-order Logic Theorem Proving (HolStep) [22]	Logical and Relational Reasoning	Each example is a formal logic formula or theorem; solving tasks involves logical inference in a strict symbolic sense (neural models must interface with formal rules).		2 million logical statements and 10,000 theorems from higher-order logic proofs in the HOL Light system	Usually reported as precision/recall for premise selection or accuracy of proof step prediction.
TPTP library ("Thousands of Problems for Theorem Provers")	Logical and Relational Reasoning	All problems are given in formal logic syntax (CNF or TPTP format); automated theorem provers (ATPs) or neuro-symbolic reasoners attempt to prove a conjecture from given axioms.			number of problems solved and proof quality

Continued on next page

Name	Task	Symbolic component	Domain	Scale	Metrics
EntailmentBank [11] & ProofWriter [40]	Logical and Relational Reasoning	The provided explanations are structured as logical entailment proofs (often in natural language form, but representing a symbolic proof structure) which models are encouraged to reproduce. These benchmarks push models to perform symbolic reasoning (like chaining facts) in addition to answering correctly.			twofold – accuracy of the final answer and some measure of explanation quality
BabyAI (instruction-following in a gridworld) [8]	Robotics and Planning	The environment is modeled as a grid with objects; an instruction corresponds to a structured plan (sequence of actions) that could be represented in a formal language. The BabyAI platform even includes a built-in verifier that symbolically checks if an agent's actions satisfy the command		There are 19 BabyAI levels with increasing complexity, testing sequence of skills and compositional learning.	Success rate on completing the instruction correctly, as well as sample efficiency (learning from few demonstrations).
ALFRED (Action Learning From Realistic Environments and Directives) [34]	Robotics and Planning	Tasks have an underlying plan structure (open fridge -> grab apple -> heat apple -> etc.); they can be represented as sequences of high-level actions. ALFRED annotations include step-by-step action sequences as the ground-truth plan.		1,000+ household tasks described by natural language instructions and visual observations	Success rate and goal-condition success (did the final world state meet the objective). This tests integration of vision (for perception) with symbolic planning/execution.
Neuro-Symbolic VQA/NS-CL [43]	Robotics and Planning	The model uses an explicit program (in a logical DSL) parsed from language, which is then executed on a neural representation of the scene			Measures include question-answering accuracy and generalization to new combinations of attributes or novel tasks.
Neuro-Symbolic Action Planning (NS-AP)	Robotics and Planning	Tasks are described as sequences of subgoals that can be represented by a symbolic program, analogous to a subroutine in a classical planner			Success rate on 10 benchmarking scenarios, which require correctly executing all subgoals.
LogiCity – A simulated urban environment benchmark for neuro-symbolic reasoning in dynamic scenes [25]	Robotics and Planning	The environment's dynamics are defined by FOL rules (e.g. logic clauses for when an entity must stop or yield)	It defines an urban world with cars, pedestrians, etc., and customizable first-order logic rules governing their behavior (e.g. traffic rules, right-of-way)		In navigation, success rate of reaching the goal safely; in action prediction, accuracy of predicting correct agent actions per the logic.

References

1 Maryam Amirizaniani, Jianli Yao, Andrew Lavergne, Edward S. Okada, Aman Chadha, Tina Roosta, and Chirag Shah. Llmauditor: A framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*, 2024.

2 Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

3 Negar Arabzadeh, Siqing Huo, Nikhil Mehta, Qingyun Wu, Chi Wang, Ahmed Awadallah, Charles L.A. Clarke, and Julia Kiseleva. Assessing and verifying task utility in llm-powered applications. *arXiv preprint arXiv:2405.02178*, 2024.

4 Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol: Large language models for ontology learning. In *International Semantic Web Conference*, pages 408–427. Springer, 2023.

5 Roberta Barile, Claudia d'Amato, and Nicola Fanizzi. Lp-dixit: evaluating explanations for link predictions on knowledge graphs using large language models. In *Proceedings of the ACM on Web Conference 2025 (WWW)*, WWW '25, pages 4034–4042, Sydney, NSW, Australia, 2025. ACM.

6 Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

7 Junjie Cao, Yiu-Kwan Chan, Zhihua Ling, Wei Wang, Shaowu Li, Ming Liu, Rong Qiao, Yunlong Han, Chenglei Wang, Bowei Yu, Peng He, Shilin Wang, Zhenming Zheng, Michael R. Lyu, and S.C. Cheung. How should we build a benchmark? revisiting 274 code-related benchmarks for llms. *arXiv preprint arXiv:2501.10711*, 2025.

8 Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Sacha Lahlou, Lucas Willems, Chimera Saharia, Thang H. Nguyen, and Yoshua Bengio. BabyAI: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272*, 2019.

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

10 Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. *EMNLP*, 2021.

11 Bhushan Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnie Xie, Hannah Smith, Lalit Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*, 2022.

12 I. Drori and D. Te'eni. Human-in-the-loop ai reviewing: feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109, 2024.

13 H.E. Dunsmore. Software metrics: An overview of an evolving methodology. Technical report, Technical Report, 1984.

14 Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 348–353, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

15 Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol 2024 overview: The 1st large language models for ontology learning challenge. In *Open Conference Proceedings*, volume 4, pages 3–16, 2024.

16 Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

17 Lichuan Hu, Qian Li, An Xie, Ning Jiang, Ion Stoica, Hai Jin, and Haichao Zhang. Gamearena: Evaluating llm reasoning through live computer games. *arXiv preprint arXiv:2412.06394*, 2024.

18 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

19 F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity – a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005.

20 Ernesto Jiménez-Ruiz, Christian Meilicke, Bernardo Cuenca Grau, and Ian Horrocks. Evaluating mapping repair systems with large biomedical ontologies. In Thomas Eiter, Birte Glimm, Yevgeny Kazakov, and Markus Krötzsch, editors, *Informal Proceedings of the 26th International Workshop on Description Logics, Ulm, Germany, July 23 – 26, 2013*, volume 1014 of *CEUR Workshop Proceedings*, pages 246–257. CEUR-WS.org, 2013.

21 Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

22 Cezary Kaliszyk, François Chollet, and Christian Szegedy. HOLStep: A machine learning dataset for higher-order logic theorem proving. *arXiv preprint arXiv:1703.00426*, 2017.

23 Daniel Keysers, Nathanael Schärli, Nathan Scales, H. Buisman, D. Furrer, S. Kashubin, N. Momchev, D. Sinopalnikov, L. Stafiniak, and T. Tihon et al. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations (ICLR)*, 2020.

24 Brendan M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*, 2018.

25 Bowen Li, Zhaoyu Li, Qiwei Du, Jinqi Luo, Wenshan Wang, Yaqi Xie, Simon Stepputtis, Chen Wang, Katia P. Sycara, Pradeep Kumar Ravikumar, Alexander Gray, Xujie Si, and Sebastian Scherer. LogiCity: Advancing neuro-symbolic ai with abstract urban simulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

26 Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, July 2004.

27 Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, 2022.

28 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

29 M. A. N. Pour, A. Algergawy, E. Blomqvist, P. Buche, J. Chen, P. G. Cotovio, A. Coulet, J. Cufi, H. Dong, D. Faria, L. Ferraz, S. Hertling, Y. He, I. Horrocks, L. Ibanescu, S. Jain, E. Jiménez-Ruiz, N. Karam, F. Kraus, P. Lambrix, H. Li, Y. Li, P. Monnin, H. Paulheim, C. Pesquita, A. Sharma, P. Shvaiko, M. Silva, G. Sousa, C. Trojahn, J. Vataščinová, B. Yaman, O. Zamazal, and L. Zhou. Results of the ontology alignment evaluation initiative 2024. In *19th International Workshop on Ontology Matching*, volume 3897, pages 64–97, January 2025. © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

30 Deyi Qiao, Chengming Wu, Yilun Liang, Junyi Li, and Nan Duan. Gameeval: Evaluating llms on conversational games. *arXiv preprint arXiv:2308.10032*, 2023.

31 Inioluwa Deborah Raji, Emily M. Denton, Emilly M. Bender, Amandalynne Paullada, and A. T. Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of*

the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS), volume 1, 2021.

32 Andrew Reuel, Anthony Hardy, Carson Smith, Matthew Lamparth, Michelle Hardy, and Mykel J. Kochenderfer. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. *arXiv preprint arXiv:2411.12990*, 2024.

33 Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912. Association for Computational Linguistics, 2020.

34 Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yejin Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *arXiv preprint arXiv:1912.01734*, 2020.

35 Edgar H. Sibley, Philip J. Fleming, and John J. Wallace. Computing practices how not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM*, 29:218–221, March 1986.

36 Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*, 2019.

37 Alessandro Solimando, Ernesto Jiménez-Ruiz, and Giovanna Guerrini. Minimizing conservativity violations in ontology alignments: algorithms and evaluation. *Knowl. Inf. Syst.*, 51(3):775–819, 2017.

38 Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July 2017. Association for Computational Linguistics.

39 Alane Suhr, Siyuan Zhou, Angli Zhang, Iris Zhang, Huajie Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019.

40 Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online, August 2021. Association for Computational Linguistics.

41 Ziyu Wang, Siyuan Zhao, Yutong Wang, Hao Huang, Sicheng Xie, Yi Zhang, Jiashi Shi, Zehua Wang, Hai Li, and Jianjun Yan. Re-task: Revisiting llm tasks from capability, skill, and knowledge perspectives. *arXiv preprint arXiv:2408.06904*, 2024.

42 Lukas M. Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P. Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D. Robinson. Essential guidelines for computational method benchmarking. *Genome Biology*, 20(125), 6 2019.

43 Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural- symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2019.

44 Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, 2018.

45 Rui Zhou, Liang Chen, and Ke Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024.

4.9 Real-World Applications in Neurosymbolic Artificial Intelligence

Ernesto Jimenez-Ruiz (City St George's, University of London, UK), Roberto Confalonieri (University of Padova, IT), Mena Leemhuis (University of Bozen-Bolzano, IT), Catia Pesquita (Universidade de Lisboa, PT), Daria Stepanova (Bosch Center for AI, DE)

4.9.1 What is it?

Over the past decade, AI development has largely focused on data-driven, output-oriented models. More recently, however, there has been a notable shift toward approaches that prioritize trust, transparency, and control. In response to this shift, neurosymbolic AI has gained increasing attention, as it integrates symbolic reasoning with data-driven methods to support more interpretable and reliable systems. As a result, a growing number of neurosymbolic approaches are now being explored and adopted in practice.

4.9.1.1 Key Domains of Application

NeSy AI is gaining traction across a wide range of real-world application domains, particularly in areas where interpretability, explainability, and trust are critical.

In **scientific discovery**, for instance, drug development benefits from systems like IBM's RXN[8], which pair neural networks for reaction prediction with symbolic logic to ensure chemical validity. Similarly, physics research leverages NeSy AI models to derive symbolic equations from experimental data, aiding in interpretable and data-driven discovery. Recently, also startups, e.g., extensity[9] started developing neurosymbolic platforms to accelerate scientific research work by integrating LLMs and knowledge graph technologies.

In **healthcare and medical diagnostics**, radiology applications combine image-based neural diagnosis with symbolic clinical guidelines to support interpretable decision-making, e.g., in [9] this approach has been applied to the problem of Alzheimer's disease diagnostics. Another application is concerned with infection tracking and patient flow, where data about patients and staff movement (collected via tracking devices) is combined with hospital constraints and domain knowledge represented as clinical rules (capturing infection policies) and knowledge graphs (hospital roles, wards, rooms, points of interest, etc.) [14]. In [13] Logical Neural Networks (LNNs) are employed to embed domain-specific rules as weighted logical formulas for explainable diagnosis. Additionally, [7] demonstrates how symbolic reasoning enhances automated radiology report generation by grounding neural outputs in structured clinical logic.

In **industrial automation**, quality control benefits from neural image inspection systems supported by symbolic rule enforcement to ensure product standards. For example, in [1] a neuro symbolic AI approach for automating the compliance verification of the electrical control panels has been proposed. This method combines Deep Learning techniques for recognizing the electrical components from the images of the electrical control panels with an Answer Set Programming-based system for comparing the scheme reconstructed from the picture with its original version to discover possible errors. The work [5] focuses on

[8] https://rxn.app.accelerate.science/
[9] https://www.extensity.ai/

the problem of predictive maintenance, where a neurosymbolic architecture is designed to both detect anomalies and explain their causes. It combines a state-of-the-art unsupervised autoencoder for anomaly detection with an online rule-learning algorithm that provides symbolic explanations for anomalies.

Neurosymbolic AI methods are also applied for **system configuration problems**, e.g., for configuration of E-Drives, where answer set programming (ASP) based approaches are used for finding conceptual designs of E-Drives that satisfy user requirements, and LLMs are invoked to facilitate the interaction with the ASP solver by translating formal explanations to natural language [6]. In a similar way, there is a tendency of combining more classical approaches for scheduling and optimization, e.g., [3, 4] with LLMs to facilitate the interaction with the system. Other NeSy AI applications include, e.g., welding quality monitoring [19] or root cause analysis [17].

In **supply chain optimization**, traditional predictive models often fail to consider complex legal, contractual, and sustainability constraints, leading to infeasible or non-compliant recommendations. Neurosymbolic AI offers a solution by integrating neural forecasting models with symbolic rule-based systems that enforce supplier agreements, trade compliance laws, and environmental policies. These systems can generate optimized yet compliant strategies across global supply chains. For example, knowledge graph embeddings have been applied for retrieving supplier candidates that are similar to the currently available suppliers, and ASP has been used for finding the most optimal selection of suppliers which minimize a certain objective, e.g., CO2 emission [18]. Moreover, product intelligence and trusted traceability tools have been developed that leverage symbolic logic to ensure regulatory compliance while optimizing logistics operations [16].

Other growing domains of neurosymbolic AI applications include **finance**, **e-commerce**, and **cybersecurity**, where hybrid models help to enforce regulations, detect anomalies, and explain decisions, which are critical capabilities in high-risk and fast-evolving environments. Traditional fraud detection systems often rely on historical transaction patterns, which can quickly become outdated as fraud tactics evolve. Neurosymbolic AI addresses this limitation by combining neural anomaly detection, which can scan millions of transactions for suspicious behavior, with symbolic reasoning that enforces logical constraints, such as regulatory anti-money laundering (AML) rules. This hybrid approach flags potentially fraudulent activity and provides interpretable explanations of which compliance rules were violated, increasing transparency for investigators and auditors. The work [2] integrates a transformer-based neural model with a symbolic Belief-Desire-Intention (BDI) reasoning layer, thus significantly improving the interpretability and decision-making capabilities of fraud detection pipelines. Recently, several startups have emerged that provide hybrid solutions for banking and financial applications, e.g., chatbots that comply with audit-based regulations[10].

Robotics and autonomous systems benefit from neurosymbolic planning, where neural perception handles sensory input, and symbolic reasoning governs task execution and long-term strategy. In human-robot interaction, language understanding (via neural networks) is paired with symbolic logic for command execution.

In the **legal domain**, contract analysis systems use neural NLP models to process text and symbolic reasoning to validate legal clauses. Legal compliance platforms also combine machine learning with rule-based systems to ensure business processes align with regulations. Emerging research even explores how large language models (LLMs) can be evaluated for legal soundness by linking their outputs with specific legal texts through symbolic frameworks

[10] `unlikely.ai`

and prompting techniques. Some works analyze the legal implication of answers provided by LLMs and possibly alert the user. Prompting techniques and RAG are mostly adopted while actual references to the interested law and corresponding articles can be provided by suitable KGs [8]

In **education**, intelligent tutoring systems use neural models to understand inputs like handwriting or speech and symbolic solvers to guide students through structured reasoning in subjects like mathematics. Adaptive learning platforms track student behavior and apply logic-based strategies to personalize learning paths [11].

In **games and strategy applications**, Neurosymbolic AI combines neural network learning with symbolic reasoning to enable dynamic planning and strategy formulation, as seen in complex gaming environments (for example, SwarmBrain in StarCraft II, which uses LLMs for macro-strategy and symbolic or rule-like control modules for tactical execution) [15].

4.9.2 What is the NeSy ambition?

Neurosymbolic AI aims to address several core limitations of purely neural systems. First and foremost, it enhances **interpretability and explainability** (see Section 4.4), particularly crucial in domains like healthcare, finance, or law, where opaque models are often unsuitable due to safety, ethical, or regulatory concerns. By incorporating symbolic reasoning, systems become more transparent and human-compatible, enabling more robust human-in-the-loop applications.

Another key advantage is **better generalization from small datasets** (see Section 4.3). Traditional deep learning often requires large volumes of data, which is not always available in practice due to rarity, privacy concerns, or high acquisition costs. NeSy AI models can bridge data gaps by incorporating domain knowledge through symbolic representations, allowing for meaningful inferences even with limited data.

Regulatory compliance and ethical alignment are also central to the promise of NeSy AI. In domains governed by strict rules and societal expectations, the ability to directly encode laws, guidelines, or ethical norms into the symbolic component of a model ensures more predictable and auditable behavior.

From a development perspective, **debugging and maintenance** are easier in neurosymbolic systems compared to opaque neural models. The symbolic component offers points of inspection and control, simplifying the identification and resolution of issues.

NeSy AI systems also show promise in **handling out-of-distribution inputs**. Real-world environments are unpredictable, and systems must handle novel scenarios. Symbolic rules–such as safety constraints or physical laws–act as safeguards, enabling the system to maintain functionality even in unfamiliar situations. Interestingly, symbolic knowledge doesn't just constrain outputs but can also expand them, guiding the model toward new, valid solutions not directly observed in training data.

Additionally, in many real-world applications there is a growing need to enable effective collaboration between humans and neural models–particularly large language models (LLMs)–in fields such as systems engineering and production optimization [12]. Addressing this challenge requires hybrid, agent-based architectures capable of supporting continuous learning, reasoning, and knowledge exchange. The architectures proposed in Section 4.6, which incorporate not only neural and human agents but also a knowledge-based agent–aim to fulfill this need by facilitating meaningful interaction between humans and machines. This enables effective and transparent knowledge sharing and decision support. Unlike purely neural approaches (e.g., LLM-based systems) or purely symbolic systems (e.g., classical expert systems), these hybrid architectures combine learning and reasoning capabilities, thereby enabling a continuous cycle of adaptation and improvement.

4.9.3 What are the challenges?

Despite their promise, NeSy AI approaches face several notable challenges. One of the main ones among them is the **lack of a clear methodology** for designing and building such systems (see also Section 4.1). Transforming raw data or implicit expert knowledge into symbolic representations is often labor-intensive and requires deep domain expertise (see also Section 4.2).

Many current NeSy AI implementations are still limited to small datasets or artificial benchmarks, lacking demonstration of **scalability and real-world applicability**. Real-world environments change rapidly, raising questions about how to maintain **synchronization between evolving knowledge models and learning systems**.

Moreover, while symbolic reasoning enhances explainability, it can sometimes reduce the raw performance or efficiency of systems, especially when fast inference is critical–as in robotics, finance, or emergency medicine. Symbolic components can also add computational overhead, particularly when combined with large-scale neural models (see again Section 4.2).

There are also practical limitations regarding **tooling and infrastructure**. Unlike the mature ecosystems for deep learning (e.g., PyTorch, TensorFlow), few production-ready libraries support neurosymbolic development. This has recently started to change, however, as new startups focusing on neurosymbolic integrations start developing their own advanced platforms, e.g., Imandra[11]. Still evaluating these systems at large scale in real-world settings remains difficult, with no standard benchmarks and challenges around integrating human factors like usability and interface design (see Section 4.8).

4.9.4 Where are we now?

A foundational requirement for neurosymbolic AI is access to structured knowledge–ontologies, knowledge graphs, argumentation structures–that serve as the backbone of the symbolic component. Historically, the scarcity of such data has limited NeSy AI systems in real-world contexts (see Section 4.2).

However, recent advances in large language models have transformed this landscape. Research efforts, such as those by Hu et al. [10], demonstrate how LLMs can be used to automatically generate structured data, including knowledge graphs and ontologies. This unlocks neurosymbolic applications even in niche domains where no prior structured knowledge existed.

While this marks significant progress, it introduces new concerns around **data quality, provenance, and correctness**. Human-in-the-loop mechanisms become essential to validate and refine LLM-generated knowledge, ensuring the symbolic backbone of NeSy AI systems remains reliable. Nonetheless, these developments substantially broaden the scope of neurosymbolic AI, making its core benefits–interpretability, trustworthiness, and data efficiency–more accessible than ever before.

Neurosymbolic AI represents a promising paradigm shift in the development of intelligent systems. By integrating neural and symbolic approaches, it offers a path toward AI that is not only powerful but also transparent, adaptable, and aligned with human values and regulatory frameworks. While the field still faces methodological and practical hurdles, the rapid progress in LLM-driven knowledge generation and growing interest across domains signal a strong trajectory forward. As the ecosystem matures, NeSy AI approaches may become foundational in building AI we can truly understand and trust.

[11] `https://imandra.ai/`

References

1. Vito Barbara, Massimo Guarascio, Nicola Leone, Giuseppe Manco, Alessandro Quarta, Francesco Ricca, and Ettore Ritacco. Neuro-symbolic AI for compliance checking of electrical control panels. *Theory Pract. Log. Program.*, 23(4):748–764, 2023.
2. Parul Dubey, Pushkar Dubey, and Pitshou N. Bokoro. A unified transformer–bdi architecture for financial fraud detection: Distributed knowledge transfer across diverse datasets. *Forecasting*, 7(2), 2025.
3. Thomas Eiter, Tobias Geibinger, Nysret Musliu, Johannes Oetsch, Peter Skocovský, and Daria Stepanova. Answer-set programming for lexicographical makespan optimisation in parallel machine scheduling. *Theory Pract. Log. Program.*, 23(6):1281–1306, 2023.
4. Thomas Eiter, Tobias Geibinger, Nelson Higuera Ruiz, Nysret Musliu, Johannes Oetsch, Dave Pfliegler, and Daria Stepanova. Adaptive large-neighbourhood search for optimisation in answer-set programming. *Artif. Intell.*, 337:104230, 2024.
5. João Gama, Rita P. Ribeiro, Saulo Martiello Mastelini, Narjes Davari, and Bruno Veloso. A neuro-symbolic explainer for rare events: A case study on predictive maintenance. *CoRR*, abs/2404.14455, 2024.
6. Tobias Geibinger, Tobias Kaminski, and Johannes Oetsch. Explanations for guess-and-check asp encodings using an llm (extended abstract). In *Program TAASP. Workshop on Trends and Applications of Answer Set Programming (TAASP 2024)*, page 6, Klagenfurt, Austria, 2024.
7. Zhongyi Han, Benzheng Wei, Xiaoming Xi, Bo Chen, Yilong Yin, and Shuo Li. Unifying neural learning and symbolic reasoning for spinal medical report generation. *Medical Image Anal.*, 67:101872, 2021.
8. George Hannah, Rita T. Sousa, Ioannis Dasoulas, and Claudia d'Amato. On the legal implications of large language model answers: A prompt engineering approach and a view beyond by exploiting knowledge graphs. *Journal of Web Semantics*, 84:100843, 2025.
9. Yexiao He, Ziyao Wang, Yuning Zhang, Tingting Dan, Tianlong Chen, Guorong Wu, and Ang Li. Neurosymad: A neuro-symbolic framework for interpretable alzheimer's disease diagnosis. *CoRR*, abs/2503.00510, 2025.
10. Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. Enabling llm knowledge analysis via extensive materialization, 2025.
11. Chris Davis Jaldi, Eleni Ilkou, Noah L. Schroeder, and Cogan Shimizu. Education in the era of neurosymbolic AI. *J. Web Semant.*, 85:100857, 2025.
12. Sebastian Krakowski. Human-ai agency in the age of generative ai. *Information and Organization*, 35(1):100560, 2025.
13. Qiuhao Lu, Rui Li, Elham Sagheb, Andrew Wen, Jinlian Wang, Liwei Wang, Jungwei W. Fan, and Hongfang Liu. Explainable diagnosis prediction through neuro-symbolic integration. *CoRR*, abs/2410.01855, 2024.
14. Chi Him Ng, Annette ten Teije, and Frank van Harmelen. A boxology-based analysis of design patterns for neuro-symbolic medical decision making systems. In Riccardo Bellazzi, José Manuel Juarez Herrero, Lucia Sacchi, and Blaz Zupan, editors, *Artificial Intelligence in Medicine – 23rd International Conference, AIME 2025, Pavia, Italy, June 23-26, 2025, Proceedings, Part I*, volume 15734 of *Lecture Notes in Computer Science*, pages 333–343. Springer, 2025.
15. Xiao Shao, Weifu Jiang, Fei Zuo, and Mengqing Liu. Swarmbrain: Embodied agent for real-time strategy game starcraft ii via large language models, 2024.
16. Siemens Digital Industries Software. Product intelligence & trusted traceability, 2023.
17. Nicholas Tagliapietra, Juergen Luettin, Lavdim Halilaj, Moritz Willig, Tim Pychynski, and Kristian Kersting. Causalman: A physics-based simulator for large-scale causality. *CoRR*, abs/2502.12707, 2025.

18 Cuong Chu Xuan, Mohamed H. Gad-Elrab, Trung-Kien Tran, Marvin Schiller, Evgeny Kharlamov, and Daria Stepanova. Supplier optimization at bosch with knowledge graphs and answer set programming. In *The Semantic Web: ESWC 2023 Satellite Events – Hersonissos, Crete, Greece, May 28 – June 1, 2023, Proceedings*, volume 13998 of *Lecture Notes in Computer Science*, pages 200–204. Springer, 2023.

19 Baifan Zhou, Zhipeng Tan, Zhuoxun Zheng, Dongzhuoran Zhou, Yunjie He, Yuqicheng Zhu, Muhammad Yahya, Trung-Kien Tran, Daria Stepanova, Mohamed H. Gad-Elrab, and Evgeny Kharlamov. Neuro-symbolic AI at bosch: Data foundation, insights, and deployment. In Anastasia Dimou, Armin Haller, Anna Lisa Gentile, and Petar Ristoski, editors, *Proceedings of the ISWC 2022 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 21st International Semantic Web Conference (ISWC 2022), Virtual Conference, Hangzhou, China, October 23-27, 2022*, volume 3254 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

Participants

- Mehwish Alam
Institut Polytechnique de Paris, FR
- Vaishak Belle
University of Edinburgh, GB
- Roberto Confalonieri
University of Padova, IT
- Claudia d'Amato
University of Bari, IT
- Artur d'Avila Garcez
City – University of London, GB
- Jennifer D'Souza
Leibniz Universität Hannover, DE
- Luc De Raedt
KU Leuven, BE
- Natalia Díaz-Rodríguez
University of Granada, DE
- Anna Lisa Gentile
IBM Almaden Center – San Jose, US
- Dagmar Gromann
Universität Wien, AT
- Pascal Hitzler
Kansas State University – Manhattan, US
- Filip Ilievski
VU Amsterdam, NL
- Ernesto Jiménez-Ruiz
City – University of London, GB
- Mena Leemhuis
Free University of Bozen-Bolzano, IT
- Bertram Ludäscher
University of Illinois at Urbana-Champaign, US
- Giuseppe Marra
KU Leuven, BE
- Hande McGinty
Kansas State University – Manhattan, US
- Raghava Mutharaju
IIITD – New Dehli, IN
- Axel-Cyrille Ngonga Ngomo
Universität Paderborn, DE
- Stefan Ollinger
Universität Trier, DE
- Alessandro Oltramari
Carnegie Bosch Institute – Pittsburgh, US
- Catia Pesquita
University of Lisbon, PT
- Jay Pujara
USC – Marina del Rey, US
- Michael L. Raymer
Wright State University – Dayton, US
- Ute Schmid
Universität Bamberg, DE
- Luciano Serafini
Bruno Kessler Foundation – Trento, IT
- Cogan Matthew Shimizu
Wright State University – Dayton, US
- Daria Stepanova
Bosch Center for AI – Renningen, DE
- Valentina Tamma
University of Liverpool, GB
- Annette ten Teije
VU Amsterdam, NL
- Riccardo Tommasini
INSA – Lyon, FR
- Frank van Harmelen
VU Amsterdam, NL
- Eugene Vasserman
Kansas State University – Manhattan, US
- Gustav Šír
Czech Technical University in Prague, CZ