Benchmarking in Neuro-Symbolic AI

Robin Manhaeve[1(✉)], Francesco Giannini[2], Mehdi Ali[3,4], Damiano Azzolini[5], Alice Bizzarri[5], Andrea Borghesi[6], Samuele Bortolotti[7], Luc De Raedt[1], Devendra Dhami[8], Michelangelo Diligenti[9], Sebastijan Dumančić[10], Boi Faltings[11], Elisabetta Gentili[5], Alfonso Gerevini[12], Marco Gori[9], Tias Guns[1], Martin Homola[13], Kristian Kersting[14], Jens Lehmann[15], Michele Lombardi[6], Luca Lorello[16,17], Emanuele Marconato[7], Stefano Melacci[9], Andrea Passerini[7], Debjit Paul[11], Fabrizio Riguzzi[5], Stefano Teso[7], Neil Yorke-Smith[10], and Marco Lippi[16]

[1] Department of Computer Science, KU Leuven, Leuven, Belgium
robin.manhaeve@kuleuven.be
[2] Scuola Normale Superiore, Pisa, Italy
[3] Fraunhofer IAIS, Sankt Augustin, Germany
[4] Lamarr Institute, Dortmund, Germany
[5] University of Ferrara, Ferrara, Italy
[6] University of Bologna, Bologna, Italy
[7] University of Trento, Trento, Italy
[8] Eindhoven University of Technology, Eindhoven, The Netherlands
[9] Università di Siena, Siena, Italy
[10] TU Delft, Delft, The Netherlands
[11] EPFL, Lausanne, Switzerland
[12] Università degli Studi di Bresci, Brescia, Italy
[13] Comenius University in Bratislava, Bratislava, Slovakia
[14] Technical University of Darmstadt, Darmstadt, Germany
[15] Amazon, Seattle, USA
[16] DISMI, University of Modena, Modena, Italy
[17] Reggio Emilia, Italy

Abstract. Neural-symbolic (NeSy) AI has gained a lot of popularity by enhancing learning models with explicit reasoning capabilities. Both new systems and new benchmarks are constantly introduced and used to evaluate learning and reasoning skills. The large variety of systems and benchmarks, however, makes it difficult to establish a fair comparison among the various frameworks, let alone a unifying set of benchmarking criteria. This paper analyzes the state-of-the-art in benchmarking NeSy systems, studies its limitations, and proposes ways to overcome them. We categorize popular neural-symbolic frameworks into three groups: model-theoretic, proof-theoretic fuzzy, and proof-theoretic probabilistic systems. We show how these three categories have distinct strengths and weaknesses, and how this is reflected in the type of tasks and benchmarks to which they are applied.

Keywords: Neural-Symbolic AI · Benchmarks · Evaluation of learning and reasoning

W.-Z. Dai (Ed.): ILP 2024, LNAI 16059, pp. 238–249, 2026.
https://doi.org/10.1007/978-3-032-09087-4_17

1 Introduction

Building systems that integrate learning, reasoning and optimization has long been a dream for artificial intelligence (AI). One of the major challenges within this context is evaluating novel ideas and frameworks on appropriate benchmarks. Too often, the tasks and the datasets that are considered and proposed for experimental evaluation are tailored to specific algorithms or methodologies and limited to ad-hoc scenarios and application domains. More generally, the neuro-symbolic (NeSy) community lacks a generally accepted perspective to test the existing approaches across a variety of different tasks and under different conditions, making deep experimental comparisons hard to obtain [44]. In addition, when a new system is proposed, it is often tested either on appositely introduced benchmarks, to emphasize the advantages of the novel approach (which is understandable and legitimate), or on well-known "old-fashioned" datasets and tasks. While a comparison on such classic benchmarks is useful to get an idea of the performance of an approach for some reference points, new challenges are necessary to drive the development of NeSy systems forward. For example, several well-known datasets in image classification such as MNIST or CIFAR have been used to design a variety of artificial tasks, each time with a specific goal: to propose a setting for continual learning or few-shot learning, to introduce explicit knowledge for reasoning, or to integrate rules and constraints for collective classification [4,24,25]. These datasets have become real benchmarking frameworks, but their environments are too limited for evaluating the development of systems integrating different paradigms.

To enable better benchmarking in NeSy AI we make the following contributions: (i) analyzing the current state of the art for what concerns the existing datasets and benchmarks at the intersection of learning, reasoning and optimization; (ii) studying their limitations; (iii) analyzing the existing systems that have been applied to such data; (iv) providing a list of the desiderata that new benchmarks should include; (v) proposing novel ideas for the evaluation and comparison of different approaches. This is all intended to provide insight into the abilities and limitations of current and future learning and reasoning systems.

2 A Categorization of Neural-Symbolic Benchmarking

Let us start off by taking a look at current benchmarking in NeSy AI. To do this, we use a categorization of the different types of neural-symbolic tasks due to Vermeulen et al. [44].

Distant Supervision. In this setting, we have a set of i.i.d. supervised examples, but supervision is not available at the level of the classifier that needs to be trained. Rather, it is available on the logic that is defined over the classifiers. A classic example is that of the MNIST Addition task [24], where the training examples are tuples of handwritten images, labelled with their sum. The learning task however is to learn to recognize individual digits.

Structured Prediction. In this setting, the system needs to classify entities connected by a logical structure (e.g., a graph). Often, a subset of the entities is labelled, while the label of the majority of entities needs to be inferred by taking into account entity-specific attributes, as well as the relational structure connecting them. This is typically done through logic knowledge. An example is the Citeseer [17] dataset. Here, individual entities are scientific papers, represented as a bag of the words present in the paper. The goal is to predict the domain to which the paper belongs. The citation graph is also provided, which includes an edge if one paper cites another. The background knowledge is that if one paper cites another, it likely belongs to the same domain.

Knowledge Base Completion. In this classic setting, we have a knowledge base with missing links between the entities that need to be completed. Although this is generally not a neural-symbolic setting, the idea is that the entities might have sub-symbolic attributes, or that a deep learning-based system might learn embedding representations and patterns that are not easily learned by a symbolic system. As a result, enhancing a symbolic data collection with a neural model may produce a significant improvement in solving the task, as shown, e.g., on the Countries [23,26] datasets

Learning to Optimize. In this setting, the goal is to generate solutions to computationally intractable tasks. The model is trained to approximate the optimal solution. The logic is used to make it more likely that a consistent solution is generated.

Overview of Popular Neural-Symbolic Benchmarks

Below is an overview of popular neural-symbolic benchmarks for each these category.

- Distant supervision (50): Add 2x2, Apply 2x2, BDD-OIA, CelebA, Chess, CLE4EVR, CLEVR, CLEVR-Hans, CLEVR-Math, Context-sensitive grammars, Crop yield prediction, CUB, DBA, DOT, Follow Suit Winner, Handwritten formulas, Hanoi, Indoor scene classification, Kandinsky patterns, Math, Member, MIMIC-II, MNIST Addition, MNIST AddMul, MNIST Even-Odd, MNIST Following Pairs, MNIST Half, MNIST Pairs, MNIST Sequential, MonumAI, Mutagenicity, Operator 2x2, Path, Predictive toxicology, RAVEN, ROAD-R, RPS, Shapeworld, Shortest path, Sudoku grid validity, Tic tac toe, Tic tac toe - next move, Trigonometry, vDEM, Visual Sudoku, V-LOL, VQAR, Well-formed parantheses, Word-algebra problems, XOR
- Structured prediction (5): AbstRCT, Arnetminer, CiteSeer, Cora, IPC
- Knowledge base completion (16): Countries, CQ2SPARQLOWL, EMBER/PE Malware Ontology, FB15k-237, Kinship, MedHop, MMKB, Nations, PharmKG, Pizza ontology, PubMed, Randomly generated KBs, UMLS, WebKB, WikiHop, WN18RR
- Learn to optimize (2): Hardware/Algorithm Dimensioning, Transprecision computing

3 A Comparison of Neural-Symbolic Systems

To discuss the state-of-the-art of neural-symbolic systems, we follow the categorization introduced in Marra et al. [27]. Here, NeSy systems were categorized along 6 dimensions. We exploit some of these dimensions below to provide a rough categorization of neural-symbolic systems. This is followed by an analysis of their capabilities. An up-to-date version of the tables are available at https://sites.google.com/view/benchmarking-in-nesy-ai.

3.1 Dimensions of Neural-Symbolic Systems

Proof- vs Model-Theoretic. The first dimension we select is the proof-theoretic vs model-theoretic dimension, as this property has a profound impact on the type of inference that is carried out by the systems. Proof-theoretic systems work by finding proofs for a query by chaining together several steps of logical inference using either backwards or forward reasoning. This type of inference thus has a defined direction and is strongly connected to (logic) programming. On the other hand, the model-theoretic approach considers the satisfying models for a given logical theory, which is related to SAT-solving. *Logical semantics.* Marra et al. [27] distinguish between three different levels of semantics: minimal, stable, and classical semantics. If the logical theory is limited to definite clauses, its semantics is generally defined in terms of the least Herbrand model. It is the unique *minimal* set of atoms that can be derived from the clauses. When relaxing this constraint allowing any type of clause, this minimal set might not be unique. Instead, the semantics is defined by all *stable* models. Finally, the semantics of arbitrary logical theories is defined by the *classic* semantic definition of First-Order Logic. These semantics can be extended by defining a probability distribution over models.

Structure vs Parameter Learning. For most machine learning models in a learning setting, the structure is fixed, but the parameters inside the structure have to be learned. For many logic-based methods, however, the structure itself defines the model as there are no other parameters. For neural-symbolic methods, this dichotomy becomes even more important, and whether structure learning is supported becomes a defining aspect of the system. In this paper, all systems support parameter learning, so we only indicate whether they support structure learning.

3.2 Overview

In Table 1, we give an overview of all neural-symbolic systems discussed in this paper. A lot of entries are reused from Marra et al. [27] with permission.

3.3 Categorization

Analyzing Table 1 from [27] only according to the dimensions mentioned above, we now identify three distinct groups of systems. In Table 2 we give a per-category overview of these properties with respect to the number of NeSy systems having them.

Group 1: Model-Theoretic Systems. The systems that only use a model-theoretic approach are quite uniform. They all use classical logical semantics, and almost none of them supports structure learning. The group is further divided into systems that use a fuzzy or a probabilistic interpretation on top of the classical semantics, with some systems offering (a mix of) both.

Group 2: Proof-theoretic fuzzy systems. Within the proof-theoretic dimension, we have a clear splitting between fuzzy and probabilistic systems. Most fuzzy systems use minimal semantics, and almost all of them support parameter learning.

Group 3: Proof-theoretic probabilistic systems. The probabilistic proof-theoretic systems are divided between minimal and stable model semantics. Furthermore, very few of them have support for structure learning. This is potentially due

Table 1. All neuro-symbolic systems considered in this survey.

System	Inference	Semantics	Learning	Benchmark type
	(P)roof (M)odel	(C)lassical (M)inimal (S)table (P)robability (F)uzzy	(P)arameters (S)tructure	(D)istant supervision (S)tructured prediction (K)B completion
αILP [35]	P+M	S+P	P+S	D
Concept Embedding Models (CEM) [12]	M	C+P+F	P	D
Deep Concept Reasoner (DCR) [4]	P	F	P+S	D
Deep Logic Models (DLM) [28]	M	C+P+F	P	D+K
DeepProbLog [24]	P+M	M+P	P+S	D+S
DeepStochLog [46]	P	M+P	P	D+S
Feed-Forward Neural-Symbolic Learner (FFNSL) [9]	P	S+F	P+S	D
Greedy Neural Theorem Provers (GNTP) [30]	P	M+F	P+S	K
Lifted Relational Neural Networks (LRNN) [37]	P	M+F	P+S	D+K
Logic Explained Networks (LEN) [7]	P+M	C+F	P+S	D
Logic Tensor Networks (LTN) [3]	M	C+F	P	D+K
NeurASP [49]	P+M	S+P	P	D
NeuralLP [47]	P	M+F	P	K
Neural Markov Logic Networks (NMLN) [29]	M	C+P	P+S	K
Neural Probabilistic Soft Logic (NeuPSL) [32]	M	C+F	P	D+S
NLog [42]	P	M+P	P	D
NLProlog [45]	P	M+P	P+S	K
Neural Theorem Prover (NTP) [33]	P	M+F	P+S	K
Reason-able Embeddings [1]	M	C+F	P	K
Relational-Concept Based Models (R-CBM) [5]	P+M	P+F	P+S	D+S+K
Relational Neural Machines (RNM) [26]	M	C+P	P	D+S
Relational Reasoning Networks (R2N) [26]	P+M	C+F	P	S+K
Semantic Based Regularization (SBR) [11]	M	C+F	P	S+K
Scallop [21]	P	M+P	P	D
SLASH [36]	P+M	S+P	P	D
TensorLog [8]	P	M+P	P	S+K

to the more expressive nature of probabilistic inference, making search over the space of rules expensive.

3.4 Capabilities

The category of a NeSy system has a large impact on what type of tasks it can be applied to. This is made clear by the benchmarks each system is generally evaluated on. In this section, we exemplify this by indicating which tasks the different categories of systems are evaluated on. The resulting signature is indicative of the capabilities of a system. We count these signatures for each system in the categories listed above ((D) Distant supervision, (S) Structured prediction, (K) Knowledge base completion). We then count how often each task type appears in these signatures. Here, we omitted the optimization-based tasks as they were not used in the systems used in this comparison.

The results of the categorization are shown in Table 3. From subtables (a)-(c) we can see that distant supervision tasks are common among all systems. Also, both model-theoretic and proof-theoretic systems are quite versatile. Proof-theoretic probabilistic systems seem to be mostly focused on distant supervision. This is probably due to the more expensive probabilistic inference that prevents them from being successfully applied to structured prediction and knowledge-base tasks.

Table 2. An overview of the properties of the systems in the different categories along the 3 dimensions.

Category	Proof- vs model-	Semantics	Fuzzy vs probability	Structure learning	#systems
	(P) Proof (M)odel	(C) Classical (M) Minimal (S) Stable	(P) Probability (F) Fuzzy	(✓) yes (X) no	
model	M	C	P + F	X	2
	M	C	F	✓	1
	M	C	F	X	4
	M	C	P	✓	1
	M	C	P	X	1
fuzzy	P	C	F	✓	1
	P	C	F	X	1
	P	M	F	✓	3
	P	M	F	X	1
	P	S	F	✓	2
probabibilistic	P	M	P	X	1
	P	S	P	✓	1
	P	S	P	X	2
	P	M	P	✓	1
	P	M	P	X	4

Table 3. Capabilities of each system category. We report the count of systems evaluated on a combination of (D)istant supervision, (S)tructured prediction, and (K)nowledge base completion tasks.

(a) Model-theoretic

D	S	K	# Systems
X	X		2
X		X	2
X			2
	X	X	1
		X	2
# Systems	6	3	5

(b) Proof-theoretic fuzzy

D	S	K	# Systems
X	X	X	1
X		X	1
X			2
	X	X	1
		X	3
# Systems	4	2	6

(c) Proof-theoretic probabilistic

D	S	K	# Systems
X	X		2
X			5
	X	X	1
		X	1
# Systems	7	3	2

4 Limitations of the State of the Art

We now discuss some limitations identified in the existing benchmarks, which have led to the recent introduction of novel benchmarks for the NeSy community.

Concerning Data. Combining data from different sources, and integrating low-level perceptual stimuli (images, videos, text, signals) with knowledge of any kind remains a cornerstone of most existing NeSy benchmarks. A large part of such benchmarks utilizes images as input, whereas text remains largely under-explored by the NeSy community [19]. The rise of Large Language Models (LLMs) has also rapidly changed the landscape, representing an additional element to account for. The integration of LLMs within NeSy approaches, to address reasoning and optimization tasks, seems a very promising though challenging research direction for the future. Images, instead, are the most frequently used category of input data, since they can be easily manipulated to create synthetic datasets with desired properties and characteristics. Moreover, they can be employed across a wide variety of applications like, e.g., game playing, as in Tic-Tac-Toe, constraint solving in Visual Sudoku, visual question answering as in CLEVR-Hans [39], plain classification as in Kandinsky Patterns [31]. Knowledge is usually implicit when dealing with certain input data categories, such as knowledge graphs, whereas the definition of specific tasks often requires the use of explicit knowledge, typically in the form of (soft or hard) logic rules: this is the case, for example, for the many benchmarks created, with different goals, from the MNIST dataset, or from CLEVR-based settings, such as CLE4EVR [25]. To mitigate the lack of explicit logic knowledge, several approaches have started co-learning a set of logic rules on knowledge graphs, or preprocess the graph with an external rule miner such as AMIE [14] or DRUM [34], and then use them for knowledge graph reasoning [10,18]. However, the extracted rules are often very different, and therefore it is difficult to build a comprehensive view of the capabilities enabled by exploiting this knowledge, even if evaluated on the same datasets.

Concerning Paradigms and Tasks. Besides traditional paradigms and tasks, such as classification and reasoning, interesting and novel research directions have

emerged, leading to the identification as well as the design, of novel benchmarks. This is the case, for example, for benchmarks inspired by an incremental or continuous learning process, such as MNIST Sequential, CLE4EVR, or KANDY. Yet, we believe that this direction is still largely under-explored, and it actually represents a true element of novelty that should be further considered by future benchmarks, and by systems as well. Moreover, existing benchmarks are often too specific and do not properly model complex and real-world interactions. In this regard, some interesting advances toward a more general perspective have been considered on the crossword application [50], where knowledge and constraints can be used to solve or generate thematic world puzzles. The possibility of having a human-in-the-loop is also a crucial ingredient to enhance explainability and trustworthiness in AI systems. Benchmarking the capability of a system to extract the correct explanation is quite challenging, some recent attempts, however, have been made on specific topics, such as universal algebra [16] and electrical power grids [43]. A novel task that has been recently addressed within the NeSy community is that of reasoning shortcuts (e.g., BDD-OIA on autonomous driving predictions, and MNIST Half or Sequential). Although some of the existing benchmarks allow for the definition of tasks in small-data regimes (i.e., few-shot learning), semi-supervised learning, or even unsupervised learning, we also consider this aspect as an open challenge for the design of NeSy benchmarks.

Concerning Performance. Measuring the performance of NeSy systems with metrics that can capture properties beyond plain accuracy in classification or pattern recognition still remains an open issue, and it is a highly relevant problem within the NeSy community [22]. Among the novel benchmarks proposed within the TAILOR project, there have been some attempts to include performance metrics that take into account properties like interpretability and trustworthiness. This is the case, for example, for the works that have been studying concept learning, as well as reasoning shortcuts [25]. In this case, beyond the accuracy of the classification task, the idea is to analyze to what extent the learned representations are aligned with a set of pre-defined concepts. Energy efficiency to reduce the carbon footprint is another dimension that is gaining relevance. In this context, some recently proposed benchmarks, related to hardware dimensioning and transprecision computing, are exploiting energy-related metrics [13,38].

Concerning Implementation. From a more practical perspective, we remark that the comparison of the same system across different benchmarks, or of different systems on the same benchmark, is made difficult by the heterogeneity of the formalisms used to represent data and to model background knowledge. A standardization of frameworks would represent a crucial step to improve such comparisons and to advance the state-of-the-art: this could be enabled by providing APIs to the systems, by providing knowledge in different formats, or by including benchmarks within existing platforms such as OpenML. Ongoing work is looking into creating a knowledge representation language for NeSy that could

be used to unambiguously and uniformly represent the knowledge in tasks and benchmarks [20].

Concerning Domains. Analysis of existing datasets may be very useful in highlighting how some domains are under-represented in the panorama of benchmarks usually considered by the NeSy community. Planning is an example of an under-represented domain, as it can easily provide both symbolic data, such as activity traces or maps, and numeric data, coming from perception. Novel benchmarks have been proposed within TAILOR for goal recognition and classic planning [6]. The medical and legal domains also represent two scenarios where background knowledge provided by experts could be a crucial element to boost the performance of purely data-driven systems: such knowledge could be provided in various formats, including knowledge graphs, ontologies, or even plain natural language. Biomedical data have been proposed as benchmarks for knowledge graph completion (e.g., the PharmKG benchmarks [10,51]), whereas legal documents (e.g., online terms of service) have been proposed for tasks related to distant supervision. Yet, more opportunities will likely emerge in these fields in the coming years. Regarding textual documents, computational argumentation and argument mining could be an additional research field where symbolic knowledge might be employed, for example, to encode argument models. Some preliminary works using NeSy systems for this kind of task have been proposed [15]. Finally, safety-critical applications have also been identified as a domain, where it is common to have hard and soft constraints that intelligent agents have to satisfy when interacting with the environment: even if some work in, e.g., autonomous driving [40], reinforcement learning [48] or malware detection [2,41], has been done in this context, a more extensive and systematic application of NeSy systems in this setting could also be an interesting research direction for the future.

5 Conclusions

In this paper, by borrowing relevant criteria from other work, we have given a categorization and overview of popular systems and benchmarks within neural-symbolic AI. We have categorized popular neural-symbolic frameworks into three categories: model-theoretic, proof-theoretic fuzzy, and proof-theoretic probabilistic systems. Our analysis shows that these three systems have distinct strengths and weaknesses, and this is reflected in the type of tasks to which they are applied. Going forward, we will further deepen our analysis of both systems and benchmarks with a more fine-grained analysis of the state-of-the-art.

References

1. Adamski, D.M., Potoniec, J.: Reasonable embeddings: learning concept embeddings with a transferable neural reasoner. Semantic Web (Preprint), 1–33 (2023)
2. Anthony, P., et al.: Explainable malware detection with tailored logic explained networks. arXiv preprint arXiv:2405.03009 (2024)

3. Badreddine, S., Garcez, A.d., Serafini, L., Spranger, M.: Logic tensor networks. Artif. Intell. **303**, 103649 (2022)
4. Barbiero, P., et al.: Interpretable neural-symbolic concept reasoning. In: International Conference on Machine Learning, pp. 1801–1825. PMLR (2023)
5. Barbiero, P., Giannini, F., Ciravegna, G., Diligenti, M., Marra, G.: Relational concept based models. arXiv preprint arXiv:2308.11991 (2023)
6. Chiari, M., Gerevini, A.E., Percassi, F., Putelli, L., Serina, I., Olivato, M.: Goal recognition as a deep learning task: the grnet approach. In: Proceedings of the International Conference on Automated Planning and Scheduling, vol. 33, pp. 560–568 (2023)
7. Ciravegna, G.: Logic explained networks. Artif. Intell. **314**, 103822 (2023)
8. Cohen, W.W., Yang, F., Mazaitis, K.: Tensorlog: a probabilistic database implemented using deep-learning infrastructure. J. Artif. Intell. Res. **67**, 285–325 (2020)
9. Cunnington, D., Law, M., Lobo, J., Russo, A.: FFNSL: feed-forward neural-symbolic learner. Mach. Learn. **112**(2), 515–569 (2023)
10. Diligenti, M., Giannini, F., Fioravanti, S., Graziani, C., Falaschi, M., Marra, G.: Enhancing embedding representations of biomedical data using logic knowledge. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2023)
11. Diligenti, M., Gori, M., Sacca, C.: Semantic-based regularization for learning and inference. Artif. Intell. **244**, 143–165 (2017)
12. Espinosa Zarlenga, M., et al.: Concept embedding models: beyond the accuracy-explainability trade-off. Adv. Neural. Inf. Process. Syst. **35**, 21400–21413 (2022)
13. Francobaldi, M., et al.: Tinderai: support system for matching AI algorithms and embedded devices. In: The International FLAIRS Conference Proceedings, vol. 36 (2023)
14. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: Amie: association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 413–422 (2013)
15. Galassi, A., Lippi, M., Torroni, P.: Investigating logic tensor networks for neural-symbolic argument mining. In: Proceedings of 1st International Joint Conference on Learning, Reasoning, pp. 1–7 (2021)
16. Giannini, F., et al.: Interpretable graph networks formulate universal algebra conjectures. Adv. Neural Inf. Process. Syst. **36** (2024)
17. Giles, C.L., Bollacker, K.D., Lawrence, S.: Citeseer: an automatic citation indexing system. In: Proceedings of the Third ACM Conference on Digital Libraries, pp. 89–98 (1998)
18. Guo, S., Wang, Q., Wang, L., Wang, B., Guo, L.: Knowledge graph embedding with iterative guidance from soft rules. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
19. Hamilton, K., Nayak, A., Božić, B., Longo, L.: Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. Semantic Web (Preprint), 1–42 (2022)
20. van Krieken, E., Badreddine, S., Manhaeve, R., Giunchiglia, E.: Uller: a unified language for learning and reasoning. arXiv preprint arXiv:2405.00532 (2024)
21. Li, Z., Huang, J., Naik, M.: Scallop: a language for neurosymbolic programming. Proc. ACM Program. Lang. **7**(PLDI), 1463–1487 (2023)
22. Lorello, L.S., Lippi, M.: The challenge of learning symbolic representations. In: d'Avila Garcez, A.S., Besold, T.R., Gori, M., Jiménez-Ruiz, E. (eds.) Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La

Certosa di Pontignano, Siena, Italy, 3–5 July 2023. CEUR Workshop Proceedings, vol. 3432, pp. 44–61. CEUR-WS.org (2023). https://ceur-ws.org/Vol-3432/paper4.pdf
23. Maene, J., De Raedt, L.: Soft-unification in deep probabilistic logic. Adv. Neural Inf. Process. Syst. **36** (2024)
24. Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., De Raedt, L.: Deepproblog: neural probabilistic logic programming. Adv. Neural Inf. Process. Syst. **31** (2018)
25. Marconato, E., Bontempo, G., Ficarra, E., Calderara, S., Passerini, A., Teso, S.: Neuro-symbolic continual learning: knowledge, reasoning shortcuts and concept rehearsal. In: Proceedings of the 40th International Conference on Machine Learning, pp. 23915–23936 (2023)
26. Marra, G., Diligenti, M., Giannini, F.: Relational reasoning networks. arXiv preprint arXiv:2106.00393 (2021)
27. Marra, G., Dumancic, S., Manhaeve, R., De Raedt, L.: From statistical relational to neurosymbolic artificial intelligence: a survey. Artif. Intell. **328**, 104062 (2024). https://doi.org/10.1016/J.ARTINT.2023.104062
28. Marra, G., Giannini, F., Diligenti, M., Gori, M.: Integrating learning and reasoning with deep logic models. In: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (eds.) ECML PKDD 2019. LNCS (LNAI), vol. 11907, pp. 517–532. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46147-8_31
29. Marra, G., Kuželka, O.: Neural markov logic networks. In: Uncertainty in Artificial Intelligence, pp. 908–917. PMLR (2021)
30. Minervini, P., Bosnjak, M., Rocktäschel, T., Riedel, S., Grefenstette, E.: Differentiable reasoning on large knowledge bases and natural language. In: AAAI, pp. 5182–5190. AAAI Press (2020)
31. Müller, H., Holzinger, A.: Kandinsky patterns. Artif. Intell. **300**, 103546 (2021)
32. Pryor, C., Dickens, C., Augustine, E., Albalak, A., Wang, W.Y., Getoor, L.: Neupsl: neural probabilistic soft logic. In: IJCAI, pp. 4145–4153. ijcai.org (2023)
33. Rocktäschel, T., Riedel, S.: End-to-end differentiable proving. Adv. Neural Inf. Process. Syst. **30** (2017)
34. Sadeghian, A., Armandpour, M., Ding, P., Wang, D.Z.: Drum: end-to-end differentiable rule mining on knowledge graphs. Adv. Neural Inf. Process. Syst. **32** (2019)
35. Shindo, H., Pfanschilling, V., Dhami, D.S., Kersting, K.: αilp: thinking visual scenes as differentiable logic programs. Mach. Learn. **112**(5), 1465–1497 (2023)
36. Skryagin, A., Stammer, W., Ochs, D., Dhami, D.S., Kersting, K.: Neural-probabilistic answer set programming. In: KR (2022)
37. Sourek, G., Aschenbrenner, V., Zelezny, F., Schockaert, S., Kuzelka, O.: Lifted relational neural networks: efficient learning of latent relational structures. J. Artif. Intell. Res. **62**, 69–100 (2018)
38. Spillo, G., De Filippo, A., Musto, C., Milano, M., Semeraro, G.: Towards sustainability-aware recommender systems: analyzing the trade-off between algorithms performance and carbon footprint. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 856–862 (2023)
39. Stammer, W., Schramowski, P., Kersting, K.: Right for the right concept: revising neuro-symbolic concepts by interacting with their explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3619–3629 (2021)
40. Stoian, M.C., Giunchiglia, E., Lukasiewicz, T.: Exploiting t-norms for deep learning in autonomous driving. arXiv preprint arXiv:2402.11362 (2024)

41. Švec, P., Balogh, Š., Homola, M., Kl'uka, J., Bisták, T.: Semantic data representation for explainable windows malware detection models. arXiv preprint arXiv:2403.11669 (2024)
42. Tsamoura, E., Hospedales, T.M., Michael, L.: Neural-symbolic integration: a compositional perspective. In: 35th Conference on Artificial Intelligence, AAAI 2021, Virtual Event, 2–9 February 2021, pp. 5051–5060. AAAI Press (2021). https://doi.org/10.1609/aaai.v35i6.16639
43. Varbella, A., Amara, K., Gjorgiev, B., Sansavini, G.: Powergraph: a power grid benchmark dataset for graph neural networks. In: NeurIPS 2023 Workshop: New Frontiers in Graph Learning (2023)
44. Vermeulen, A., Manhaeve, R., Marra, G.: An experimental overview of neural-symbolic systems. In: International Conference on Inductive Logic Programming, pp. 124–138. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-49299-0_9
45. Weber, L., Minervini, P., Münchmeyer, J., Leser, U., Rocktäschel, T.: Nlprolog: reasoning with weak unification for question answering in natural language. In: ACL (1), pp. 6151–6161. Association for Computational Linguistics (2019)
46. Winters, T., Marra, G., Manhaeve, R., De Raedt, L.: Deepstochlog: neural stochastic logic programming. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 10090–10100 (2022)
47. Yang, F., Yang, Z., Cohen, W.W.: Differentiable learning of logical rules for knowledge base reasoning. Adv. Neural Inf. Process. Syst. **30** (2017)
48. Yang, W.C., Marra, G., Rens, G., De Raedt, L.: Safe reinforcement learning via probabilistic logic shields. In: Proceedings IJCAI (2023)
49. Yang, Z., Ishay, A., Lee, J.: Neurasp: embracing neural networks into answer set programming. In: IJCAI, pp. 1755–1762. ijcai.org (2020)
50. Zeinalipour, K., Iaquinta, T., Angelini, G., Rigutini, L., Maggini, M., Gori, M.: Building bridges of knowledge: innovating education with automated crossword generation. In: 2023 International Conference on Machine Learning and Applications (ICMLA), pp. 1228–1236. IEEE (2023)
51. Zheng, S., et al.: Pharmkg: a dedicated knowledge graph benchmark for biomedical data mining. Brief. Bioinf. **22**(4), bbaa344 (2021)