TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

Integrated learning, reasoning and optimisation in practice v.2 Report

Document type (nature)	Report
Deliverable No	4.4
Work package number(s)	4
Date	Due M44, April 2024
Responsible Beneficiary	P 5 - KU Leuven
Author(s)	Robin Manhaeve, Francesco Giannini, Marco Lippi, Luc De Raedt
Publicity level	Public
Short description	Integrated learning, reasoning and optimisation in practice v.2

History			
Revision	**Date**	**Modification**	**Author**
v.1	03/05/2024	First version	Robin Manhaeve, Francesco Giannini, Marco Lippi, Luc De Raedt

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Fredrik Heintz	LiU	2024-05-13
Nic Wilson and Steve Prestwich	UCC	2024-05-14

Table of Contents

Introduction to the Deliverable

The question addressed in this WP is how to integrate learning, reasoning and optimisation, that is, how to computationally and mathematically integrate different AI paradigms. The most apparent difference between paradigms lies in the representations that are used and so an operational way to answer the question is to tightly integrate different representations as to offer both learning, reasoning and/or optimisation in common frameworks. This theme will therefore design representational systems with accompanying inference, learning and optimisation algorithms that can support trustworthy artificial intelligence.

In this deliverable, we focus on systems that integrate learning, reasoning and optimisation in practice. More specifically, we will look at the tasks, benchmarks and applications that such systems are applied to. We first look at what types of existing datasets are good candidates to be turned into novel benchmarks. Next, we give a classification of the state of the art for both systems and tasks. We then give an overview of which class of system is best / most applied to which type of task. After this, we look at the state-of-the-art in applications. We finish this report by listing challenges and future work for the field.

Organisation

This is the second version of this deliverable. The first version is available at: https://tailor-network.eu/wp-content/uploads/2022/07/D4.3-Integrated-learning-reasoning-and-optimisation-in-practice-v.1.pdf.

The following people have been involved in the Deliverable:

Partner ID / Acronym	**Name**	**Role**
KU Leuven	Robin Manhaeve	Author
UNIFI	Marco Lippi	Author
CINI	Francesco Giannini	Author
KU Leuven	Luc De Raedt	Author
KU Leuven	Emanuele Sansone	Author v.1
UNITN	Andrea Passerini	Author v.1

Other contributors:
Fabrizio Riguzzi (UNIFE), Neil Yorke-Smith (TU Delft), Sebastijan Dumancic (TU Delft), Tias Guns (KUL), Michele Lombardi (UNIBO), Debjit Paul (EPFL), Boi Faltings (EPFL), Kristian Kersting (TUDA), Devendra Dhami (TUDA), Mehdi Ali (FhG), Jens Lehmann (FhG), Michele Lombardi (UNIBO), Andrea Borghesi (UNIBO)

Motivation and Aim

Building systems that can integrate learning, reasoning and optimization has long been a dream for artificial intelligence. One of the major challenges, within this context, is to evaluate 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, they lack an open and wider perspective to test the considered approaches across a variety of different tasks and under different conditions, making experimental comparisons hard to obtain. In addition, novel systems that aim to integrate learning, reasoning and optimization often rely on old-fashioned data and tasks. While a comparison with standard benchmarks is useful to get a rough idea of the performance of an approach with respect to some reference point, we argue that it is now necessary to consider new challenges to drive the development of new integrated systems.
For example, several well-known datasets in image classification, such as MNIST or CIFAR, have been used for a wide variety of artificial tasks, each time with a specific goal: to propose a setting for few-shot learning, to introduce explicit knowledge for reasoning, to integrate rules and constraints for collective classification. ... In this sense, they have nowadays become real benchmarking frameworks. However, these datasets offer a limited environment for the development of systems integrating different paradigms.

To address these issues, the TAILOR project has established a taskforce working across the different tasks of WP 4, identifying the following phases: (i) to analyze the current state-of-the-art for what concerns the existing datasets and corpora at the intersection of learning, reasoning and optimization; (ii) to study their limitations; (iii) to analyze the existing systems that have been applied to such data; (iv) to provide a list of the desiderata that new benchmarks should include; (v) to propose 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.

It is worth mentioning that the goal is not just to list data collections, but especially to highlight which tasks can be applied to such data (i.e., in the form of benchmarks), and how a more extensive benchmarking framework could be designed, by unifying and composing a variety of heterogeneous tasks, working on the same original data collection. As a consequence, the ultimate goal of the taskforce is to provide a suite of benchmarks which enable the creation of new tasks at a minimal cost and also provide a methodological evaluation to assess the performance of hybrid systems, which integrate the paradigms of learning, reasoning and optimization, thus providing insight into the practice and driving also future research.

The outcomes at the end of the project are:

- a number of publications comparing such systems, both theoretically and empirically
- a number of new challenges and benchmarks for hybrid systems
- a categorization of state-of-the-art hybrid systems
- a categorization of state-of-the-art hybrid benchmarks
- a highlight of applications developed with integrated learning, reasoning and optimisation techniques

Evolution from D 4.3

The first version of this deliverable introduced two tables, one concerning benchmarks and one on neuro-symbolic (NeSy) systems. These were formulated in terms of a reinforcement learning view which modeled the interaction between a system and its learning environment (i.e. the task).
In this version of the deliverable, we have stepped away from this view as it was unnecessarily complex. We still have a table for systems and benchmarks, but they have been simplified, reducing the number of columns for each. The goal behind this simplification was to give a clearer, higher-level overview of both, that can further develop into a usable categorization and comparison.
In the first version, one of the original goals was to list existing data collections that can be used to create new benchmarks. We have decided to move away from this goal, since such data collections are ubiquitous, and many more novel benchmarks have been introduced in the past few years. The current issue does not seem to be a lack of interesting benchmarks, but rather a lack of a consistent comparison between systems on a shared suite of benchmarks. That is why this final version has shifted towards the latter goal.

Datasets for NeSy benchmarks

As previously mentioned, the work of the taskforce had initially focused on listing the existing datasets in machine learning (i.e., not necessarily those used within the NeSy community) which could represent good candidates to create benchmarks for the NeSy community. While in the second part of the project we stepped away from that view, we hereby find it useful to remark which properties of such datasets are relevant to make them suitable for an extended use within the NeSy community. In particular, the ideal NeSy benchmark should consist in a setting where data are complemented with, or integrated by, some kind of knowledge. This knowledge could be either explicit (e.g., soft rules in probabilistic logic) or implicit (e.g., links in a knowledge graph), and data could be in either sub-symbolic form (e.g., images) or symbolic form (e.g., facts in the form of logic predicates). Also, we especially focus on problems and tasks where both neural approaches struggle and purely symbolic methods are unable to handle data uncertainty: in such cases, the integration of the two worlds should indeed provide an added value.

Categorization of Neuro-Symbolic Systems

To discuss the state-of-the-art of neuro-symbolic systems, we follow the categorization as introduced in [Marra et al. 2024]. Here, NeSy systems were categorized along 6 dimensions. From these 6, we select the following dimensions to create a categorization.

Proof- vs model-theoretic

Proof-theoretic approaches work by finding proofs for a query in a logic theory. A proof for a query is a sequence of logical inference steps that demonstrates the truth of that query based on the given program. Typically, forward or backward chaining inference is used to search for proofs for queries. Conversely, the model theoretic perspective on logic is to find a model or truth assignment to the logical atoms that satisfy a given logic theory. An interpretation, or possible world, is a truth-assignment to the propositions (or ground atoms) of the language, and can be uniquely identified with the set of propositions it assigns True (thus considering all the other False). In the model-theoretic perspective, one uses the logic theory as a set of constraints on the propositions, that is, the propositions are related to one another, without imposing a directed inference relationship between them as in forward or backward chaining.

Logical semantics

We can distinguish three different levels of semantics, which are also closely tied to the used syntax of the underlying logic. First, when the logical theory consists of definite clauses only, the semantics is given by the least Herbrand model. The least Herbrand model of a definite clause theory is unique and it is the minimal w.r.t. set inclusion. It contains all ground facts (from the Herbrand domain) that are logically entailed by the theory. Second, when the logical theory can contain any set of clauses, the semantics is given by the set of all stable Herbrand models. Third, while Horn-clauses are the basis for "pure" Prolog and logic programs, there exist several extensions to this formalism to accommodate negated literals in the condition part of rules or disjunction in the head. A popular framework in this regard is

answer set programming (ASP). The previous three levels of semantics are based on Boolean models, i.e. models where each atom is either present (i.e. True) or absent (i.e. False). Differently, fuzzy logic, and in particular t-norm fuzzy logic, assigns a truth value to atoms in the continuous real interval [0, 1]. Logical operators are then turned into real-valued functions, mathematically grounded in the t-norm theory.
The previous semantics have been extended by defining probability distributions over models, or possible worlds. In particular, the probability that a certain formula holds is computed as the sum of the probabilities of the possible worlds that are models of that formula.

Structure vs parameter learning

Learning approaches are usually distinguished by whether the structure or the parameters of the model are learned. In structure learning, the learning task is to discover the logical theory, i.e., a set of logical clauses and their corresponding probabilities or weights that reliably explains the examples. What explaining the examples exactly means depends on the learning setting. In contrast to structure learning, parameter learning starts with a given logical theory, and only learns the corresponding probabilities or weights. In our selection, all systems support parameter learning, so we only indicate whether a system supports structure learning.

Categorization

We now identify three groups of systems:

Group 1: Model-theoretic system

The systems that only use a model-theoretic approach are quite uniform. They all use a classical logical semantics, and almost none of them support 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 system, we have a clear splitting between fuzzy and probabilistic systems. The fuzzy systems (nearly) all use minimal semantics, and they almost all 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 to the more expressive nature of probabilistic inference, preventing efficient search over the space of rules.

An overview of a representative group of state-of-the-art NeSy systems is given in Table 1. Proof- and model-theoretic are indicated as P and M respectively in the first column. Classical, minimal, and stable semantics are indiciated as C, M and S respectively in the second column. This column also indicates (P)robabilistic and (F)uzzy extensions of these

semantics. The third column indicates whether a system supports structure learning. The final column indicates the category of the system: (M)odel theoretic, proof-theoretic (F)uzzy, or proof-theoretic (P)robabilistic. Many entries originate from [Marra et al. 2024].

Table 1: An overview and categorization of state-of-the-art NeSy systems

System	P vs M	Semantics	Structure	Category
DCR	P	C+F	X	F
NeuralLP	P	M+F		F
dILP	P	M+F	X	F
DiffLog	P	M+F	X	F
LRNN	P	M+F	X	F
NLM	P	M+F	X	F
NTP	P	M+F	X	F
R2N	P+M	C+F		F
R-CBM	P+M	C+F	X	F
FFNSL	P	S+F	X	F
GNTP	P	M+F	X	F
DI2	M	C+F		M
LTN	M	C+F		M
SBR	M	C+F		M
Reason-able Embeddings	M	C+F		M
LEN	M	C+F	X	M
CEM	M	C+F+P		M
DLM	M	C+F+P		M
RNM	M	C+P		M
SL	M	C+P		M
NMLN	M	C+P	X	M
NeuPSL	M	C+F		M
DeepStochLog	P	M+P		P
NLog	P	M+P		P
Scallop	P	M+P		P
TensorLog	P	M+P		P
NLProlog	P	M+P	X	P
DeepProbLog	P+M	M+P		P
NeurASP	P+M	S+P		P
Slash	P+M	S+P		P
aILP	P+M	S+P	X	P

The State-of-the-Art in Neuro-Symbolic Benchmarking

We will now take a look at the benchmarking in the field of NeSy AI. For this, we borrow from [Vermeulen et al. 2023] the following categorization.

Distant supervision (D)
In this setting, we have supervised examples, but the 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. These types of tasks are thus strongly linked to task 4.1 of this WP.

Structured prediction (S)
In this setting, the system needs to make a prediction over structured objects. This means that the entities in the data can no longer be considered separately, but the structure that connects them needs to be considered as well. This is typically done through logic. These types of tasks are thus strongly linked to task 4.1 of this WP.

Learning to optimize (O)
Optimization problems are often very hard to solve exactly. However, in many cases, a good approximation to the exact solution is sufficient, and a lot easier to find. In this setting, the system is expected to learn how to generate an (approximately) optimal solution to the input problem. By using a neural-symbolic approach, neural networks can be used to predict an (approximate) solution, while logic can incorporate the knowledge of the problem (i.e. what is a valid route).These types of tasks are thus strongly linked to task 4.2 of this WP.

Knowledge base completion (K)
A knowledge base is a collection of entities, and the relations between the entities. However, the relations between the entities are often incomplete. In this setting, the system needs to generate the missing relations from the incomplete knowledge base.These types of tasks are thus strongly linked to task 4.3 of this WP.

Below, we give an overview of popular neuro-symbolic benchmarks along the categories defined above.

- 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 EvenOdd, 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

Benchmarking coverage

The category of the NeSy system has a large impact on what type of tasks it is suitable for. This is made clear by the benchmarks each system is generally evaluated on. We look at each system and see what type of task they have been evaluated on. The resulting signature is indicative of the capabilities of a system. We count these signatures for each system in the categories as listed above ((D) Distant supervision, (S) Structured prediction, (K) Knowledge base completion). By counting how often each task type appears in these signatures, we can also see what the most frequent tasks are for each system category. Here, we omitted the optimization-based tasks as they were not used in the systems used in this comparison.

Model-theoretic systems			
D	**S**	**K**	**# Systems**
X	X		2
X		X	1
X			3
	X	X	1
		X	2
6	3	4	

Proof-theoretic fuzzy systems			
D	**S**	**K**	**# Systems**
X	X	X	1
X			3
	X	X	1
		X	3
4	2	5	

Proof-theoretic probabilistic systems			
D	**S**	**K**	**# Systems**
X	X		1
X			6
	X	X	1
		X	1
7	2	2	

From these tables 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 preventing them from being successfully applied on structured prediction and knowledge-base tasks.

Lessons Learned, Desiderata and Limitations

The analysis of the existing benchmarks developed by and used within the NeSy community allowed us to refine the list of desiderata that was initially proposed within D4.3, and to identify limitations that led to the introduction of novel benchmarks for the community.

Concerning data. Combining data from different sources, and integrating low-level perceptual stimuli (images, videos, text, signals) with knowledge of any kind still remains a cornerstone of most existing NeSy benchmarks. The analysis conducted throughout the TAILOR project shows that a large part of such benchmarks utilize images as input, whereas text still remains largely underexplored by the NeSy community [Hamilton et al., 2022]. 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. On 19/04/24, a workshop on “LLM and Learning, reasoning and optimization” was jointly organized by WP4 and 5. It explored the fusion of large language models and reasoning with invited talks by Guy Van den Broeck (UCLA) and Scott Sanner (University of Toronto).
Images, instead, are the most easy-to-use category of input data, since they can be easily manipulated, to create synthetic data sets with desired properties and characteristics. Moreover, they can be employed across a wide variety of applications (e.g., game playing, as in Tic-Tac-Toe, constraint solving in Visual Sudoku, visual question answering as in CLEVR-Hans, plain classification as in Kandinsky Patterns). 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, of the many benchmarks created, with different goals, from the MNIST data set, or from CLEVR-based settings, such as CLE4EVR.

Concerning paradigms and tasks. Besides traditional paradigms and tasks, such as classification and reasoning, interesting and novel research directions have emerged within the taskforce, leading to the identification as well as the design of novel benchmarks. This is the case, for example, of 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. The possibility to have human-in-the-loop is also a crucial ingredient to enhance explainability and trustworthiness in AI systems. Similarly, a novel task that has been recently addressed within the NeSy community is that of reasoning shortcuts (e.g., BDD-OIA data set 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 [Lorello and Lippi, 2023]. Anyhow, among the novel benchmarks proposed within the TAILOR project, there have been some attempts to include performance metrics that could take into account properties like interpretability and trustworthiness. This is the case, for example, of the works that have been studying concept learning as well as reasoning shortcuts [Marconato et al., 2023]. 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. Another additional metric, that is becoming more and more important nowadays, is energy efficiency to reduce the carbon footprint, that is yet another dimension to consider. Some benchmarks proposed within the TAILOR project related to hardware dimensioning and transprecision computing are exploiting such metrics.

Concerning implementation. From a more practical perspective, it has been noted that the comparison of the same system across different benchmarks, or of different systems on the same benchmark, is made difficult by the heterogeneity in 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.

Concerning domains. The initial analysis of datasets was very useful in highlighting how some domains were under-represented in the panorama of benchmarks usually considered by the NeSy community. The work carried out by the taskforce has helped in producing more benchmarks in such domains, as well as to highlight novel domains where NeSy could find application in the future. 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 [Chiari et al., 2024]. The medical and legal domains represent as well two scenarios where background knowledge provided by experts could be a crucial element to boost 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 [Zheng et al. 2021,Diligenti et al. 2023]) whereas legal documents (e.g., online terms of service) have been proposed for tasks related to distant supervision. Yet, more opportunities are likely to be explored 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 within the TAILOR project [Galassi et al., 2021]. Finally, safety-critical applications have also been identified as a domain where it is quite usual to have hard and soft constraints that intelligent agents have to satisfy when interacting with the environment: this could also be an interesting research direction for the future.

New benchmarks / datasets

Within the work package, several datasets and benchmarks have been introduced. We give an overview below.

In [Lorello et al., 2024], the authors introduce Kandy. It is a framework for generating curricula of tasks in the style of Kandinsky patterns, with a focus on continual and semi-supervised learning. Researchers at UNIFE have introduced multiple benchmarks, including the Tic Tac Toe dataset that includes two tasks: one on predicting the winner for a given board, and one for predicting the optimal next move. They also proposed to use within the NeSy community the RAVEN dataset [Zhang et al. 2019], consisting of Raven Progressive Matrices. These are IQ tests where the player is presented with a 3x3 array of panels with geometric figures that are arranged according to unknown rules. Of the 9 panels, the last one is masked and is to be selected among 8 possible answer panels.

While the task of theorem proving has been extensively investigated within the AI literature, the problem of providing novel mathematical conjectures is still barely explored. For instance, in the field of universal algebra (UA), a prominent line of work aims at studying the correspondence between equational properties and topological structures on graph-based data called *algebraic lattices*. This kind of problem can be suitably investigated by NeSy systems, however there were no existing datasets regarding such a benchmark. More recently, [Keskin et al. 2023] and [Giannini et al. 2023] have introduced a set of datasets with lattices up to size of 100 nodes, an algorithm to generate more AI-ready datasets based on UA's conjectures, and a novel neural layer to build fully interpretable GNNs. Within this frame, there were considered two main tasks: strong generalization capability of the systems on graph classification (i.e. training on lattices of size up to 8-9 nodes and testing on larger lattices), and explanation capability (i.e. the interpretable GNN is required to point out the correct explanation as a subgraph of the input graph for the class predicted for each graph).

Several benchmarks have been introduced to investigate the problem of reasoning shortcuts and continual learning. In [Marconato et al. 2023], the authors introduce several modifications of the original MNIST Addition dataset, as well as adaptations to BDD-OIA, a real-world and high-stakes dataset for autonomous driving, and CLE4EVR, a continual learning adaptation of the well-known synthetic VQA dataset CLEVR. Finally, they have also introduced a dataset based on Kandinsky patterns.

The State-of-the-Art in Neuro-Symbolic Applications

Malware Detection. The rapid development of new attack techniques make Malware Detection (MD) a constant challenge in cybersecurity. While machine learning approaches offer a promising solution, they lack an explanation for the instances identified as malware. On the other hand, white models are interpretable but often at the price of very low performances. As common MD datasets are often missing a clear interpretation of data features, [Švec et al. 2022] extracted a knowledge based dataset from the EMBER dataset, available at https://github.com/orbis-security/pe-malware-ontology, enabling this type of data to be exploited in semantic-enabled tools (either symbolic or hybrid). For instance, this dataset has been recently investigated by [Anthony et al. 2024] by using Logic Explained Networks (LENs), which are a recently proposed class of interpretable neural networks providing explanations in the form of First-Order Logic (FOL) rules. The results show that LENs achieve robustness that exceeds traditional interpretable methods and that are rivaling black-box models.

Mathematical Olympiads. AlphaGeometry [Trinh et al. 2024] is introduced as a groundbreaking theorem prover tailored for Euclidean plane geometry. Unlike traditional methods reliant on human demonstrations, AlphaGeometry leverages a neuro-symbolic approach, synthesizing millions of theorems and proofs across various complexity levels. This system integrates a neural language model, trained from scratch on extensive synthetic data, to guide a symbolic deduction engine through intricate problem-solving processes. In evaluations against a set of 30 challenging olympiad-level problems, AlphaGeometry showcases exceptional performance by solving 25 problems, surpassing the previous best method that could only handle ten. Its efficacy nears that of an average International Mathematical Olympiad (IMO) gold medallist.

Card Games. The integration of games into the domain of AI represents a relatively recent challenge. In fact, several factors contribute to this trend, like the diversity of the complexity among the different games, ranging from simple board games like Tic-Tac-Toe to complex strategy games like Go or StarCraft, and the demanding aspect of real-time decision-making under uncertainty, possibly in presence of incomplete information. Moreover, games have served as intricate representations of complex social dynamics, revealing that individual players exhibit diverse behaviors, like investigated by [Sauvain et al. 2022] for the game of Bridge with a data-driven approach. Then further studies carried out by the NukkAI group have shown the capabilities of using a relational and interpretable model, possibly exploiting an interaction between a domain expert and a data scientist, for both the game of Bridge [Ventos et al. 2024] and more card games such as poker [Li et al. 2024]. These applications have shown impressive results, computing optimal and/or robust strategies in games with incomplete information, given various types of knowledge about opponent models.

Recognition of Chemical Compounds. In the realm of drug discovery, understanding the graph structure of chemical compounds holds significant importance. Despite the extensive research in chemistry and pharmaceutical sciences, a considerable number of scientific articles only provide the structure of these compounds in the form of images. Previous research, as highlighted by [Oldenhof et al. 2020], addressed this challenge by employing deep neural network models for optical compound recognition. These models automatically

analyze the images, converting them into a chemical graph structure. However, such models typically necessitate detailed annotations at the instance level, including the positions and labels of all objects within each image. Addressing this limitation, recent work by [Oldenhof et al. 2023] introduced ProbKT, a novel framework grounded in probabilistic logical reasoning. This framework enables the training of object detection models using various forms of weak supervision, leading to notable advancements across different application domains.

Visual Reasoning. Recent techniques aim to boost the performance of deep learning models for Scene Graph Generation (SGG) by incorporating background knowledge. These approaches fall into two categories: sub-symbolic integration within the model and symbolic maintenance. However, both encounter drawbacks. The sub-symbolic method requires complex neural architectures, increasing costs, while the symbolic approach struggles with scalability relative to background knowledge size. More recently, [Buffelli et al. 2023] introduced Neural-Guided Projection (NGP), a neuro-symbolic regularization technique for injecting symbolic background knowledge into neural SGG models that overcomes the limitations of prior art. NGP is model-agnostic, does not incur any cost at inference time, and scales to previously unmanageable background knowledge sizes.

Self-Driving Cars. Despite the impressive results of deep learning models, their application in safety critical domains is still prevented by multiple issues. For instance, machine learning models for autonomous driving face significant challenges in ensuring safety and interpretability. Safety concerns arise due to vulnerabilities to adversarial attacks, sensor failures, and unpredictable human behavior. Interpretability issues stem from the lack of transparency in model decision-making, raising questions about trust and regulatory compliance. Addressing these challenges is crucial for the reliable and ethical deployment of autonomous driving systems. A possible solution to model the problem of pedestrian collision-free navigation of self-driving cars is given by using a partially observable Markov decision process, and then to rely on deep reinforcement learning or approximate planning. In this regard, some advancements have been produced by hybridizing these two approaches in the systems HyLEAP [Pusse et al. 2019] and HyLEAR [Gupta and Klusch 2023], which have shown to outperform previous state-of-the-art models in most accident scenarios regarding safety, while appearing equally competitive regarding smoothness of driving and time to goal on average. An alternative approach considers a class of models capable of learning from a set of requirements expressing a background knowledge about the problem and guaranteeing to be compliant with these requirements. For instance, the requirements can be expressed as logical constraints to be integrated into the learning objective of an optimization problem. According to this formulation, [Giunchiglia et al. 2023] has introduced the ROad event Awareness Dataset with logical Requirements (ROAD-R), showing that current state-of-the-art models often violate these logical constraints. This dataset has also been tested by [Stoian et al. 2024], which defines a memory-efficient logic-based loss function, allowing for exploiting t-norms fuzzy logics for the task of event detection in autonomous driving. The experimental results show that using t-norm-based losses improve the performances, especially when the available labeled data is scarce, or when a corpus of unlabeled data is available.

Challenges

While benchmarks are clearly extremely important in providing a common ground to quantitative evaluate the performance of different solutions, in modern research on AI there is a concrete risk of benchmark *hyperspecialization* and *overfitting*, in which the goal of research becomes beating the state-of-the-art on a specific benchmark (or group of closely related benchmarks), and the longer-term objective of which the benchmark is an initial and very partial proxy is lost.

At the “What are the Next Measurable Challenges in AI?” workshop, organized on 03/03/2022, the taskforce held a panel discussing these topics, and how to create novel challenges that allow to overcome the limitations of existing benchmarks and encourage the exploration of radically new ideas, in particular involving the combination of learning, reasoning and optimization. The panelists were Fosca Giannotti, Marco Gori, Kristian Kersting, Michèle Sebag and Joaquin Vanschoren, and the panel was moderated by Andrea Passerini.

A first critical aspect was identified in the obsolescence of benchmarks, which is especially important when talking about standard, static benchmarks, and calls for solutions involving evaluation of benchmark overfitting, benchmark evolution, dynamic benchmarking and the relation with lifelong and continual learning tasks. A major requirement for long-term challenges was identified in the possibility of having a diverse set of tasks to be accomplished. This calls for solutions relying on interactive learning environments, most likely virtual ones, where a combination of broad perceptual and reasoning abilities are needed in order to successfully accomplish the tasks. A second major requirement concerns the need to have the human in-the-loop of the process. This is in-line with the human-centric and trustworthy perspective on AI fostered by the EC, and poses a number of new challenges in how to make this interaction efficient and effective. Finally, the evaluation metrics and process for these systems should be substantially revised. Standard measures like accuracy are clearly insufficient and need to be complemented with aspects involving energy efficiency, interpretability, reliability, but most importantly the utility of the *joint* system that combines machine(s) and human(s).

Another key challenge for benchmarking is the systematic and consistent comparison on a representative set of benchmarks. It is not uncommon for a new system to introduce its own task on which it performs well, but a systematic comparison with other systems is lacking. Overall, we are currently lacking a good overview of NeSy system performance, both in terms of accuracy and efficiency. A large part of this is due to the fact that each NeSy system introduces its own way of integrating / representing knowledge, requiring a large effort from the researcher to port existing benchmarks to their system. This is in contrast to the more uniform way of representing data as is common in deep learning and other ML fields.

References

[Anthony et al. 2024], Peter Anthony,Francesco Giannini, Michelangelo Diligenti, Martin Homola, Marco Gori, Stefan Balogh and Jan Mojzis (2024). Explainable Malware Detection with Tailored Logic Explained Networks. arXiv preprint arXiv:2405.03009.

[Buffelli et al. 2023], Buffelli, D., & Tsamoura, E. (2023, June). Scalable theory-driven regularization of scene graph generation models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 6, pp. 6850-6859).

[Chiari et al., 2023], Chiari M., Gerevini A.E., Percassi F., Putelli L., Serina I., Olivato M., Goal Recognition as a Deep Learning Task: The GRNet Approach. ICAPS 2023

[Diligenti et al. 2023], Diligenti, M., Giannini, F., Fioravanti, S., Graziani, C., Falaschi, M., & Marra, G. (2023, June). Enhancing Embedding Representations of Biomedical Data using Logic Knowledge. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[Galassi et al. 2021], Galassi, A., Lippi, M., Torroni, P., Investigating logic tensor networks for neural-symbolic argument mining, IJCLR 2021.

[Giannini et al. 2023], Giannini, F., Fioravanti, S., Keskin, O., Lupidi, A., Magister, L. C., Lió, P., & Barbiero, P. (2023). Interpretable Graph Networks Formulate Universal Algebra Conjectures. Advances in Neural Information Processing Systems, 36.

[Gupta and Klusch 2023], Gupta, D., & Klusch, M. (2023, June). Hylear: Hybrid deep reinforcement learning and planning for safe and comfortable automated driving. In 2023 IEEE Intelligent Vehicles Symposium (IV) (pp. 1-8). IEEE.

[Hamilton et al. 2022], Is neuro-symbolic AI meeting its promises in natural language processing? a structured review, Semantic Web, 2022

[Keskin et al. 2023], Keskin, O., Lupidi, A. M., Fioravanti, S., Magister, L. C., Barbiero, P., Lio, P., & Giannini, F. (2023, September). Bridging Equational Properties and Patterns on Graphs: an AI-Based Approach. In Topological, Algebraic and Geometric Learning Workshops 2023 (pp. 156-168). PMLR.

[Li et al. 2024], Li, J., Zanuttini, B., & Ventos, V. (2024, March). Opponent-model search in games with incomplete information. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 9, pp. 9840-9847).

[Lorello and Lippi 2023], Lorello, L.S., Lippi, M., The Challenge of Learning Symbolic Representations, NeSy, 2023

[Lorello et al. 2024], Lorello L.S., Lippi, M., Melacci, S., The KANDY Benchmark: Incremental Neuro-Symbolic Learning and Reasoning with Kandinsky Patterns. Available online at https://arxiv.org/pdf/2402.17431.pdf

[Marconato et al., 2023], Marconato E., Teso, S., Vergari, A., Passerini, A., Not All Neuro-Symbolic Concepts Are Created Equal: Analysis and Mitigation of Reasoning Shortcuts, NeurIPS 2023.

[Marra et al. 2024], Marra, G., Dumančić, S., Manhaeve, R. and De Raedt, L., 2024. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, p.104062.

[Pusse et al. 2019], Pusse, F., & Klusch, M. (2019, June). Hybrid online pomdp planning and deep reinforcement learning for safer self-driving cars. In 2019 ieee intelligent vehicles symposium (iv) (pp. 1013-1020). IEEE.

[Sauvain et al. 2022], Sauvain Camille, Véronique Ventos, and Jérôme Sackur. "A Variety of Players Type: A Data-Driven Approach Applied to the Game of Bridge." Available at SSRN 4203174 (2022).

[Stoian et al. 2024], Stoian, M. C., Giunchiglia, E., & Lukasiewicz, T. (2024). Exploiting t-norms for deep learning in autonomous driving. arXiv preprint arXiv:2402.11362.

[Švec et al. 2022], Švec, P., Balogh, Š., Homola, M., & Kľuka, J. (2022). Knowledge-based dataset for training PE malware detection models. arXiv preprint arXiv:2301.00153.

[Trinh et al. 2024], Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. Nature, 625(7995), 476-482.

[Ventos et al. 2024], Ventos, V., Braun, D., Deheeger, C., Desmoulins, J. P., Fantun, J. B., Legras, S., ... & Thépaut, S. (2024). Construction and Elicitation of a Black Box Model in the Game of Bridge. In Advances in Knowledge Discovery and Management (pp. 29-53). Springer, Cham.

[Vermeulen et al. 2023], Vermeulen, Arne, Robin Manhaeve, and Giuseppe Marra. "An Experimental Overview of Neural-Symbolic Systems." *International Conference on Inductive Logic Programming*. Cham: Springer Nature Switzerland, 2023.

[Zhang et al. 2019], Zhang, C., Gao F., Jia B., Zhu Y., Zhu S.-C., RAVEN: A Dataset for Relational and Analogical Visual rEasoNing, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

[Zheng et al. 2021], Zheng, S., Rao, J., Song, Y., Zhang, J., Xiao, X., Fang, E. F., ... & Niu, Z. (2021). PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Briefings in bioinformatics, 22(4), bbaa344.