TAILOR

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

Integrated Learning, Reasoning and Optimisation in Practice v.1 Report

Document type (nature)	Report
Deliverable No	4.3
Work package number(s)	4
Date	23 June 2022
Responsible Beneficiary	P 5 - KU Leuven
Editor(s)	Luc De Raedt
Publicity level	Public
Short description (Please insert the text in the Description of Deliverables in the Appendix 1.)	Integrated Learning, Reasoning and Optimisation in Practice v.1

History			
Revision	**Date**	**Modification**	**Author**
1.0	30 June 2022	Minor	Luc de Raedt

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Philipp Slusalek	DFKI	25 June, 2022
Michela Milano	UNIBO	25 June, 2022

Table of Contents

Summary of the report

Objectives: 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. It will also study applications in two different domains. The WP is divided into four main Tasks, and is connected to other WPs by two tasks.

Introduction to the Deliverable

There are two deliverables for WP 4, that are both divided into an intermediate report (v1 M22) and a final report (v2 due at the end of the project).

Deliverables

D4.1: **Foundations**, techniques, algorithms and tools for integrating learning, reasoning and optimisation. (report) Report on the scientific challenges tasks T4.1 & T4.2.

D4.3: Integrated learning, reasoning and optimisation **in practice** (report). Report on the scientific challenges tasks T4.3 & T4.4.

This TAILOR WP has largely focused on two types of meetings and workshops. In the first type, there has been an emphasis on foundations, techniques, and tools for integrating learning, reasoning and optimization. In this type of workshop, the four scientific topics that characterize the first four tasks of WP4 within TAILOR have been covered. This has not only provided us with insight into the foundations and challenges connected to this WP, it has also delivered a number of interesting tutorials and survey papers, that have partly or fully been inspired by TAILOR and that led to novel insights and often also collaborations.

Deliverable 4.1 starts with these results, and then outlines the other results obtained within the WP. The second type of meeting was connected to the important taskforce of WP4 around benchmarks, datasets and systems. Given the plethora of different systems, representations and datasets, it is not easy to see the general picture in this diverse landscape. Therefore, we decided to start up a taskforce that would collect existing data, systems and study and compare them in order to get insight into the current and future abilities of integrated learning, reasoning and optimization approaches. This is the topic of Deliverable 4.3 and promises to result in publications summarizing useful observations and insights about the **practice** of *integrated learning, reasoning and optimisation* approaches.

Thus rather than dividing the deliverables along the task dimensions T4.1 / 2 vs T4.3 / 4 we found it more appropriate to report on the foundational issues in D4.1 and focus on the results of the taskforce in D4.3 as this is related to the potential and practice of WP 4 techniques.

Organisation

Main Contributors
Marco Lippi (UNIMORE)
Francesco Giannini (CINI)
Andrea Passerini (UNITN)
Emanuele Sansone (KUL)
Luc De Raedt (KUL)

Other People Involved
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 certainly to evaluate novel ideas and frameworks on appropriate benchmarks. Too often, in fact, the tasks and the datasets that are considered and proposed for experimental evaluation are tailored to some algorithms or methodologies, and limited to ad-hoc scenarios and application domains. More in general, 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, too often novel systems that aim to integrate learning, reasoning and optimization still rely on old-fashioned data and tasks: while a comparison with standard benchmarks is always useful to have an idea of the performance of an approach with respect to some reference point, we argue that the time is ripe for considering new challenges, which can drive the development of new integrated systems. To make an example, several classic 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 playground for the development of systems integrating different paradigms (e.g. MNIST is limited both from the learning/perceptual perspective, as it is mainly devised to solve simple digit recognition tasks, and also from the reasoning perspective, as enabling forms of reasoning restricted to operations on natural numbers).
Consequently, can we define a set of requirements for a challenge that goes beyond those currently available?
Can we do this with the goal of obtaining a benchmarking framework that meets these requirements and that can still be implemented in a reasonable time? Possibly building on top of existing ones?

To address these questions, 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 **expected outcomes** at the end of the project are:

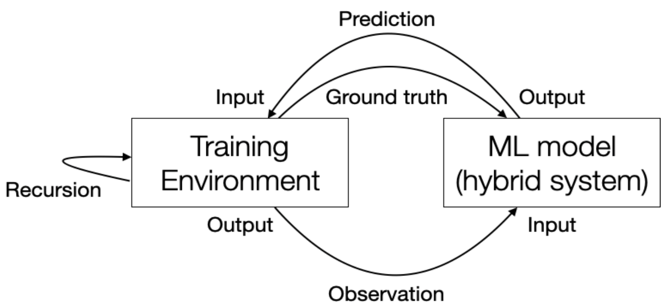
- insights into the abilities and limitations of the hybrid systems we study
- a number of publications comparing such systems, both theoretically and empirically
- a number of new challenges for hybrid systems

In the final deliverable, we also intend to go one step further and highlight some actual applications developed with integrated learning, reasoning and optimisation techniques.

Description of Tables - Datasets and Systems

The taskforce focused on contexts where low-level data is combined with knowledge, so that perception and reasoning skills need to be integrated in order to solve the tasks at hand. Knowledge could be either implicit (i.e., specific of the domain, derived from commonsense, encoded into data structures) or explicit (i.e., made available in the form of logic predicates and rules, or as constraints); either exact or uncertain; features could be just numeric or symbolic, or a combination of the two; examples could be either independent or involved in a variety of relations. All these characteristics have a strong impact on the categories of systems that can handle the corresponding datasets and which benchmarks can be defined upon.

To handle the complexity of the problem, the taskforce has produced two tables, one for datasets/benchmarks and one for systems. The general structure of the tables is based on a reinforcement learning setting, where the machine learning model (namely our hypothetical hybrid system) and the training environment interact with each other. The machine learning model provides a prediction for each observation. The training environment provides ground truth feedback to the machine learning model based on the received prediction and generates new observations optionally using historical information (i.e. in the form of a recursion). The overall setting is depicted in the following figure.



Importantly, the two tables take two opposite perspectives of the same setting. Indeed, the table about datasets/benchmarks focuses on the training environment, hence the input and the output are the prediction of the ML model and the new observation, respectively. The table about the ML model swaps the input and the output. In other words, the input consists of the observation provided by the training environment, whereas the output consists of the prediction for the observation.

The ground truth feedback from the training environment to the ML model represents the supervisory information which can be used to drive the learning, reasoning and optimization of the ML model.

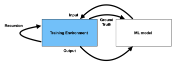
Table about Datasets/Benchmarks

The table contains a list of datasets for each task of TAILOR WP 4:

1. **Task 4.1, Learning and Reasoning**: List of all datasets including **explicit knowledge**.
2. **Task 4.2, Learning and Optimisation**: List of all datasets related to **constraints and optimization problems**.
3. **Task 4.3, Knowledge Graphs, Embeddings, Ontologies:** List of all datasets relying on **relational knowledge** and **embedding representation**, such as **KG.**
4. **Task 4.4, Perception, Spatial Reasoning, and Vision**: List of all datasets with connection to learning and reasoning with **implicit knowledge**.

Information is structured according to general content (grey columns, such as URL of dataset, license, brief description etc.), the training environment (blue columns, defining the input, the output, the ground truth feedback and whether recursion is used or not) and the evaluation procedure (green columns, such task, metrics and baselines).

The table about datasets/benchmarks is shown in the next page.

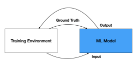


		GENERAL							TRAINING ENVIRONMENT					EVALUATION		
WP 4 Tasks	Name of dataset	**License** (Yes, which? No, collect/reuse data?)	**URL**	**Challenge** (Which task/setting can be formulated? Describe the dataset and the possible tasks/settings in one sentence)	**Domain** (Synthetic, Real. If real indicate the domain. Es. medicine, finance)	**Number of training samples**	**Nature** (Symbolic/subsymbolic/both)	**Domain knowledge** (explicit/implicit, exact/uncertain)	**Input** (the output of the ML model) **Type** (integer/real/vector/tensor/graph/...)	**Output** (the input to the ML model) **Type** (integer/real/vector/tensor/graph/...)	**Interactive** (does the input affect the next output? Yes/No)	**Ground truth** (supervised feedback provided to the ML model) **Kind** (noisy/missing/delayed/sparse/...)	**Recursion** (does the output depend on previous output? Yes/No)	**Setting** (Same as training? If no, describe)	**Metrics**	**Baselines**
Task 4.1 -- Learning and Reasoning	MNIST database	No, but collect and modify	http://yann.lecun.com/exdb/mnist/	Handwritten digits: any prediction task based on numbers	Synthetic (even though is based onn real handwritten digits	60000	both	The dataset is simple enough to create explicit and exact domain knowledge	integer (predicted class label), vector (image)	vector (image)	No	Direct (correct label for each input)	No (iid setting)	Yes	Accuracy (+ traditional metrics for classification) Reconstruction error FID (Frechet Inception Distance) (for generation)	Neural networks (https://paperswithcode.com/sota/image-classification-on-mnist) Neuro-symbolic (DeepProbLog, DeepStochLog) Generative models (VAEs, GANs, etc.)
	Beta-residue partners in proteins	No	https://www.ics.uci.edu/~baldig/betasheet_d	Link prediction within graphs (proteins)	Real (bioinformatics)	916 proteins	both	Explicit, could be modeled as soft rules	graph (binary labels)	protein sequence (string)	Proteins are independent, links in the same protein are correlated	Direct (correct labels)	In the same protein, yes	Yes	Accuracy	Betapro, ground-specific Markov Logic Networks, neural networks
	Inductive General Game Playing	No license	https://github.com/andrewcropper/mlj19-iggp	Collection of inductive logic programming tasks Reference http://andrewcropper.com/pubs/mlj19-iggp.pdf	Synthetic	Small data regime (only data, no generating scripts)	symbolic	explicit, exact	binary (ground atom)	binary (ground atoms)	No	Direct (correct assigment to atom)	No	Yes	Accuracy	Inductive logic programming systems, Aleph, ASPAL, Metagol, ILASP
	Handwritten Formula Recognition	CC BY-NC-SA 3.0	https://github.com/liqing-ustc/NGS https://www.cs.rit.edu/~crohme2019/task.htr	Solving handwritten expression	Synthetic (even though is based onn real handwritten	10000	both	Arithmetic	Real value (the value of the expression)	list of images representing symbols of the expression	No	Direct (correct label for each input) - Distant (on the elements of the list)	No	Yes	Accuracy (computed with a eps-tolerant distance)	Neural networks, Neural Symbolic (DeepStochLog, Neural Symbolic Grammars)
	Multi-Object Datasets (multi d-sprites, Objects Room, CLEVR, Tetrominoes)	Apache-2.0	https://github.com/deepmind/multi_object_d	Collection of image datasets including multiple objects with different color and shape. Challenge: prediction tasks based on objects - object segmentation or query based prediction	Synthetic	Depends on dataset 1 M for multi d-sprites and Objects Room infinite for CLEVR	both	The dataset is simple enough to create explicit and exact domain knowledge	integer (predicted class label), vector (image)	vector (image)	No	Direct (correct label for each input)	No (iid setting)	No, in the sense that you have control on the setting, to test ood and task generalization	Accuracy (for classification) Adjusted Rand Index (for segmentation masks) Reconstruction error FID (Frechet Inception Distance) (for generation)	Unsupervised scene decomposition and generation Check https://paperswithcode.com/dataset/multi-dsprites
Task 4.2 -- Learning and Optimisation	Hardware and Algorithm Dimensioning	CC BY 4.0	https://zenodo.org/record/5795216	Challenge: calibrate parameters of an online anticipatory algorithm and determine suitable hardware configurations	Energy	30000	both	Basic domain knowledge available (expected monotonicity)	PV and load curves, algorithn parameters	solution quality, solution time, memory	No	Direct (correct measurements -- inputs and output can be swapped in this dataset)	No	Yes	Accuracy, expected values for metrics (solution quality, solution time...)	
	Amazon Last Mile Routing Challenge		https://dataverse.harvard.edu/dataset.xhtml	Challenge: predict 'best' routing of drivers (TSP problem)	Real, transportation		both	explicit (distances, zones, travel time)	sequence (routes)	stops data	no	direct	no	yes	custom routing quality	Distance-based TSP
	Ireland energy + scheduling	No	https://github.com/CryoCardiogram/ijcai-cac	Challenge: half hour price predictions + do energy-aware scheduling of tasks on machines	Real: energy prices	38016	Both	Explicit (task scheduling specification)	real vector (multi-output)	real valued	No	direct	Yes, but typically trained as no	Yes	Regret, MSE	regression, smart-predict-then-optimize
	Warcraft Shortest Path		https://sites.google.com/view/combinatorialg	Challenge: tile cost prediction + find shortest path on a grid of tiles	Synthethic	10000 grids, 12x12 tiles	Both	Explicit grid size and tile dimension, image content is implicit	real vector (multi-output)	tensor (image)	No	Direct	No	Yes	Accuracy, solution quality	CNN+Shortest Path solver, combinatorial gradients
	Bipartite Cora		https://github.com/bwilder0/aaai_melding_c	Edge prediction within graph + find maximum matching in bipartite graph	Real, citations	27 subgraphs, 2500 pairs of node per grap	Both	Explicit (bipartite matching problem, citation network)	real vector (multi-output edge probabilities)	vector (concatenated nodes features)	No	Direct	No	Yes	Regret, Log-loss, MSE	classification, smart-predict-then-optimize
	Transprecision Computing (Micro-benchn	CC BY 4.0	https://zenodo.org/record/5831793	Challenge: minimze the number of bits assigned to floating-point variables within a micro-benchmark while respecting a constraint on the maximum accuracy error on the output (w.r.t. to non-reduced precision)	Transprecision computing	5 Micro-benchmarks with 10K samples; 2 with ~4K samples	Both	Basic domain knowledge available (DAG describing how FP variables are related)	one integer value for each FP variable (the number of bits assigned to it)	The error associated to the reduced-preci	No	Direct (correct measurements for each precision configuration)	No	Yes	MAE, MSE, MAPE (regression task)	
Task 4.3 -- Knowledge graphs, Embeddings, Ontologies	Terms of Service for unfair clause detecti	No	https://github.com/federicoruggeri/Memnet_	Sentence classification: predict whether a clause is potentially unfair for the consumer	Real (law)	21063	both?	Explicit, modeled as a collection of legal rationales (a set of sentences)	integer (binary labels)	text	Not necessarily, but it could be possible to perform collective classification across subsequent document sentences	Direct for final label, plus weak or strong supervisions for legal rationales	Not necessarily, but it could be possible to perform collective classification across subsequent document sentences	Yes, except for supervisions on rationales	Accuracy + metrics for rationales matching	Memory netwoks, BERT, MemBERT
	Argument Mining (various)	?	http://argumentationmining.disi.unibo.it/reso	Sentence classification and link prediction between setences in a document	Real	Various sizes	subsymbolic (text) with also constraints	Explicit, modeled as a set of constraints (rules)	In general, a graph	text	No	Direct	No	Yes	Accuracy, F1	Transformers, Integer Linear Programming, Attention-based models
	ImageGraph	BSD-3-Clause	https://github.com/mniepert/mmkb	KB completion	Real (several)	1,330 relation types, 14,870 entities, 829,931 images	both	Explicit	Entity / link	graph / image	No	Direct	No	Yes	Rank	VGG-16 + Distmult
	AGENT	No	https://github.com/MicroSTM/AGENT-synthe	Goal preferences, action efficiency, unobserved constraints, cost-reward trade-off	Synthetic (inspired by infant cognition)	3360 trials	subsymbolic	Implicit	Vector (video)	binary	Not necessarily	Direct	No	Possibly a novel physical situation can be provided	Accuracy	Bayesian Inverse Planning and Core Knowledge, ToMnet-G
	PharmKG	No	https://github.com/MindRank-Biotech/Pharm	Link prediction	Synthetic (corrected using real bioinformatics data)	500348 triplets	symbolic (potentially extendable with feature representations)	Explicit relations among data, implicit logiral rules	Entity / relation	binary	No	Direct	No	Yes	Accuracy	ConvE, ComplEx
	Countries KB	ODbL	https://github.com/mledoze/countries	Link prediction within graphs (cities located in continents)	Synthetic (geographical)	244 countries, 5 regions, 23 subregions	Symbolic	Explicit, modeled with one hard rule and one soft rule	integer (binary labels)	vector (embeddings)	No	Direct (correct labels), with missing facts and noisy rules	No	Yes	AUC-PR	ComplEx, NTP, MINERVA
	Nations	CC0	https://alchemy.cs.washington.edu/data/nati	Link prediction within graphs	Real (politics)	2565 true triplets	Symbolic	Implicit (or extracted with tools)	binary labels (missing links)	binary labels (known links)	No	Direct, sparse	No	Possible, but not always	Ranking (MRR, HITS@n)	ComplEx, NTP, MINERVA, Neural MLN
	UMLS	CC0	https://alchemy.cs.washington.edu/data/uml	Link prediction within graphs	Real (medicine)	6529 true triplets	Symbolic	Implicit (or extracted with tools)	binary labels (missing links)	binary labels (known links)	No	Direct, sparse	No	Possible, but not always	Ranking (MRR, HITS@n)	ComplEx, NTP, MINERVA, Neural MLN
		CC0	https://alchemy.cs.washington.edu/data/kins	Link prediction within graphs	Real (kinship)	10686 true triplets	Symbolic	Implicit (or extracted with tools)	binary labels (missing links)	binary labels (known links)	No	Direct, sparse	No	Possible, but not always	Ranking (MRR, HITS@n)	ComplEx, NTP, MINERVA, Neural MLN
	Cora	CC0	http://www.cs.utexas.edu/users/ml/riddle/dat	Knowledge base completion: de-duplicate entities	Real (citations)	2708	both	Explicit hard rules	integer (binary labels)	real-valued vectors	No	Direct (correct labels), with noisy rules	No	Yes	AUC-PR	Neural networks (https://paperswithcode.com/sota/node-classification-on-cora), ACORA
	Yago		https://yago-knowledge.org/	Knowledge base completion	Real (Wikipedia, WordNet)	6M facts, 3M entities (v3.0)	Symbolic	Implicit	multilabel (multiple classes)	multiple labels (known links)	No	Direct, sparse	No	Yes	Ranking (MRR, HITS@n)	DistMult, ComplEx, ConvE, ...
	oFB15k-237		https://paperswithcode.com/dataset/fb15k	Knowledge base completion	Real (FreeBase)	592213 triplets	Symbolic	Implicit	binary labels (missing links)	binary labels (known links)	No	Direct, sparse	No	Yes	Ranking (MRR, HITS@n)	RGCN
	oWN18RR		https://paperswithcode.com/dataset/wn18rr	Knowledge base completion	Real (WordNet)	93003 triplets	Symbolic	Implicit	binary labels (missing links)	binary labels (known links)	No	Direct, sparse	No	Yes	Ranking (MRR, HITS@n)	RGCN
	MovieLens 1M		https://github.com/Mehran-k/RelNN/tree/ma	Node classification within graphs	Real (movies)		Symbolic	Implicit	Integer (binary labels, for gender) or real-value (for age)	Logic predicates	No	Direct, sparse	No	Yes	Accuracy, Log-loss, MSE	Matrix Factorization, Collaborative Filtering, RDN-Boost, RLR/MLN, RelNN
	PAKDD 2015		https://github.com/Mehran-k/RelNN/tree/ma	Node classification within graphs	Real (commerce)		Symbolic	Implicit	Integer (binary labels)	Logic predicates	No	Direct, sparse	No	Yes	Accuracy, Log-loss, MSE	Matrix Factorization, Collaborative Filtering, RDN-Boost, RLR/MLN, RelNN
	NELL	open	https://github.com/huggingface/datasets/tree	Instantiation of new entities and relation triples	Real (web pages)	more than 100M	Symbolic	Implicit	Ontology + relations + web pages	Knowledge graph	yes	Sparse and noisy + unsupervised data	Yes	Possible, but not always	Accuracy	GCN, GraphVAT, DFNet-ATT
	DBpedia	GNU GPL	https://github.com/dbpedia/	Text Classification and Entity Retrieval on Wikipedia	Real (web pages)	9.5 billion triplets	Symbolic	Implicit	text and graph	Knowledge graph	yes	Direct, Sparse + unsupervised data	yes	Possible, but not always	Accuracy	XLNet, ColBERT
	Yelp		https://github.com/Mehran-k/RelNN/tree/ma	Node classification within graphs	Real (food)		Symbolic	Implicit	Integer (binary labels)	Logic predicates	No	Direct, sparse	No	Yes	Accuracy	Matrix Factorization, Collaborative Filtering, RDN-Boost, RLR/MLN, RelNN
Task 4.4 -- Perception, Spatial Reasoning, and Vision	Hi-Phy	MIT licence	https://github.com/Cheng-Xue/Hi-Phy	Learning about physics forces from images	Synthetic	10 tasks with 65 templates	both	atomic actions	images + symbolic representation	actions	yes	missing	no	Yes	reward	various reinforcement learning agents
	PTR	MIT licence	http://ptr.csail.mit.edu/	Part-based visual scene understanding	Synthetic	70 000 examples	both	explicit	images + symbolic representation	objects/binary	no	Direct	no	Yes	Accuracy	varia
	CLEVRER			VQA from videos	Synthetic	20.000 videos, 300.000 questions	subsymbolic	implicit, but can be extracted	integer	video	no	direct	No	Yes	Accuracy	LSTM, CNN+LSTM, ...
	Visual genome	CC BY-NC-SA 4.0	http://visualgenome.org/	VQA from images	Real	108K images	both	dense object annotation, including relationships		image	no	direct	No	Yes	Accuracy	CNN, CNN+LSTM, ...
	Home Action Genome (HOMAGE)	Data is owned by Panasonic, access	https://homeactiongenome.org/	Recognition of indoor daily activities from multi-view videos	Real (recorder participants)	1,752 synchronized sequences (5,700 videos in total)	both	atomic actions + compositional actions, relationship implicit	integer (predicted class label)	vector (video)	no	direct	No	Yes	Accuracy, maP	unimodal prediction models
	Virtual tools	?	https://sites.google.com/view/virtualtoolsgan	Slving physical puzzles and learning how tool impact the environment	Synthetic	30 different games	effectively symbolic but does have visual input	needs to be learned	positions and types of objects	actions	yes	missing	possibly	yes	number of solved tasks	custom sampling-based agent
	Alchemy	Apache-2.0	https://github.com/deepmind/dm_alchemy	Learning how components interact across environments, in which some features are shared and some differ	Synthetic	infinite?	both	needs to be learned	graph identifying interaction	image and symbolic representation	yes	Direct	no	no, the underlying structure if different	reward	reinforcement learning agents: IMPALA, VMPO
	Abstraction and Reasoning Corpus (ARC	Apache-2.0	https://github.com/fchollet/ARC	Series of image tasks based on Raven's Progressive Matrices	Synthetic	Training: 400 tasks * 4	both	implicit, exact	vector (image)	set of vectors (set of image pairs)	No	Direct (correct vector/image)	No	Yes	Generalization difficulty Intelligence of system Source: https://arxiv.org/pdf/1911.01547.pdf	Dreaming with ARC (system based on DreamCoder) Paper: https://openreview.net/pdf?id=-gjy2V1ko6t
	BABI Tasks	No	https://research.facebook.com/downloads/b	Various: reasoning, question answering, reading comprehension, etc.	Synthetic and real (text)	Different size for different tasks	both (mostly subsymbolc, being text)	Mostly implicit	typically strings (QA)	text	No	Can be both weak and strong	No	Yes	Accuracy	Memory networks
	ClevrTex	BSD-3.0-Clause	https://github.com/karazijal/clevrtex-generat	Extension of CLEVR with more realistic (texture-based) background	Synthetic	infinite	both	The dataset is simple enough to create explicit and exact domain knowledge	integer (predicted class label), vector (image)	vector (image)	No	Direct (correct label for each input)	No (iid setting)	No, in the sense that you have control on the setting, to test ood and task generalization	Accuracy (for classification) Adjusted Rand Index (for segmentation masks) Reconstruction error FID (Frechet Inception Distance) (for generation)	Unsupervised scence decomposition and generation Check https://paperswithcode.com/dataset/clevrtex
	Multi-Object Datasets (CATER)	Apache-2.0	https://github.com/deepmind/multi_object_d	Collection of videos including multiple moving objects with different color and shape. Challenge: prediction tasks based on objects, motion, occlusions	Synthetic	infinite	both	The dataset is simple enough to create explicit and exact domain knowledge	integer (predicted class label), vector (video)	vector (video)	No	Direct (correct label for each input)	Yes	No, in the sense that you have control on the setting, to test ood and task generalization	Accuracy (for classification) Adjusted Rand Index (for segmentation masks) Reconstruction error FID (Frechet Inception Distance) (for generation)	Object, part detection, scene parsing Check https://paperswithcode.com/dataset/pascal-person-part
	Pascal-Part	Flickr terms of use see http://host.robots.ox.ac.uk/pasca	http://roozbehm.info/pascal-parts/pascal-par	Collection of images of real-world objects Challenges: object detection, part detection	Real (ground truth symmetric)	~20000	both	implicit, uncertain	integer (predicted class label), vector (image)	vector (image)	No	Direct (correct label for each input)	No (iid setting)	Yes	Accuracy (for classification) Intersection over Union	Object, part detection
	CLUTRR	CC-BY-NC 4.0	https://github.com/facebookresearch/clutrr	Reason about family relations mentioned in text.	Synthetic	10k-15k	symbolic	implicit but can be provided	integer	text	no	direct	no	similar, but with noise or different reasoning "depth"	accuracy	LSTM, Bert, ...

Table about Systems

The table contains a list of systems for each task of TAILOR WP 4. As mentioned earlier, the table takes the perspective of the ML model.
Information is structured according to learning systems (blue columns, related to learning component dealing with perception tasks), reasoning systems (green columns, related to the reasoning/optimization component dealing with high-level tasks) and their integration (orange columns)

The table about datasets/benchmarks is shown in the next page.



README
The TRAINING ENVIRONMENT generates the data, receives the prediction from the machine learning (ML) model (called Output), provides the feedback to ML model (called GROUND TRUTH) and generates the next data item (called INPUT)

As an example The 7 dimensions from "Statistical Relational to Neuro-Symbolic AI" :
Reasoning System
Dimension 1: Underlying structure (directed/undirected)
Dimension 2: Inference (grounding (based on SAT solvers) versus proof-oriented (based on Prolog))
Dimension 4: Discrete/Continuous (Boolean vs Fuzzy)
Dimension 6: Learning (Parameter versus Structure)
Dimension 7: Type of Logic (Propositional, First Order Logic or other forms)

Integration
Dimension 3: Paradigms (Logic, Probability, Neural)
Dimension 4: Representations (Symbolic, SubSymbolic)

WP 4 Tasks	**Name of system**	**General Description**	**URL**	**Challenges/Tasks**	**Experimented Datasets**	**Learning System** **Input** (the input of the ML model) **Type**	**Output** (the output of System 1) **Type**	**Function description (map relating input to output)**	**Reasoning System** **Input** (the input of System 2) **Type**	**Output** (the output of the ML model) **Type**	**Function description (map relating input to output)** See for example above (ROW K1)	**Integration** **End-to-end** (Yes, if no provide description)	**Paradigms** (e.g. Logic (L), Probability (P), Neural (N))	**Representations** (symbolic (S), subsymbolic (Sub))
Task 4.1 -- Learning and Reasoning	DeepProbLog	Unified framework for deep learning and probabilistic logic programming	https://github.com/ML-KULeuven/	Mathematical operations on MNIST, program induction (sketching) on word algebra problems, neuro-symbolic on Coin-Urn MNIST images	MNIST; Coin-Urn MNIST images	Image	Probabilities for boolean variables (facts) or categorical variables (annotated disjunctions)	any kind of neural network. Experimented with CNN and linear dense layers	Neural predicates	Probability on logical query	directed proof-oriented discrete parameter learning definite clauses (logic programs)	Yes	L+P+N	S+Sub
	Semantic-based regularization / Lyrics	Integration of logic rules converted as fuzzy logic constraints on top of deep learners	https://github.com/GiuseppeMarra	Collective classification; integration of learning and reasoning	Cora, CIFAR 10, Citeseer; MNIST; ConLL 2000; CelebA	Images, feature-based	True-values of logic predicates	any kind of neural network. Experimented with MLP, CNN, GAN	Neural predicates	Rules' satisfaction	undirected grounding fuzzy parameter learning arbitrary FOL	Yes	L+N	S+Sub
	Markov Logic Networks													
	Probabilistic Soft Logic													
	Neural Logic Programming													
	Neural Logic Machines													
	Semantic Loss													
	Logic Tensor Networks	Integration of logic rules converted as fuzzy logic constraints on top of deep learners	https://github.com/logictensornetw	Querying, learning, reasoning, clustering	MNIST, Iris, Pascal-Part,	Images, feature-based	True-values of logic predicates	any kind of neural network	Neural predicates	Rules' satisfaction	undirected grounding fuzzy parameter learning arbitrary FOL	Yes	L+N	S+Sub
	Tensorlog	Framework compiling probabilistic FOL queries into a differentiable neural network	https://github.com/TeamCohen/Te	Querying, learning, reasoning	Citeseer, Cora, UMLS, Wordnet, WikiMovies	Feature-based; embeddings	Probabilities for boolean variables (facts)	Tensor Calculus	Real-valued tensors	Probability on logical query	directed proof-oriented discrete parameter learning Horn clauses	Yes	L+N	S+Sub
	Deep Logic Models	Unified framework to integrate logical reasoning and deep learning into a PGM	https://arxiv.org/pdf/1901.04195.p	Learning and Inference to KBC; Application of partially violated rules; Link prediction on graphs	MNIST following pairs; Countries;	Images; Embeddings	Probabilities for boolean variables (facts)	any kind of neural network; experimented with CNNs and MLPs	Random variables of atomic formulas	Probability of logic facts	undirected grounding discrete parameter learning arbitrary FOL	Yes	L+P+N	S+Sub
	Relational Neural Machines	Unified framework to integrate logical reasoning and deep learning into a PGM	https://www.ai4europe.eu/researc	Learning and Inference to KBC; Application of partially violated rules; Link prediction on graphs; Recovering supervised learning and MLN as special cases	MNIST following pairs; Citeseer;	Images; Feature-based	Probabilities for boolean variables (facts)	any kind of neural network; experimented with CNNs and MLPs	Random variables of atomic formulas	Probability of logic facts	undirected grounding discrete parameter learning arbitrary FOL	Yes	L+P+N	S+Sub
	Neural Markov Logic Networks													
	DiffLog	Structure learning system based differentiable	https://arxiv.org/pdf/1906.00163.p	Structure learning	https://github.com/XujieSi/fse18-artifact-183				Templates of rules, facts	scores of rules	fuzzy circuit	Yes	L+N	S
Task 4.2 -- Learning and Optimisation	Empirical Model Learning	Library, tutorial, paper	https://emlopt.github.io/											
	Smart Predict + Optimize	paper with code												
	CVXlayers	paper with code												
	SATNET	paper with code												
	Blackbox solver differentiation	paper with code												
	Learning Modulo Theory	paper with code												
	NCE solution cache	paper with code												
Task 4.3 -- Knowledge graphs, Embeddings, Ontologies	Relational Reasoning Networks	Models performing sub-symbolic reasoning steps in a latent space	https://arxiv.org/pdf/2106.00393.p	Exploiting relational knowledge to refine embeddings on top of KGE representations; Link prediction;	Countries; UMLS, Nations; Kinship; Cora	Embeddings; Feature-based	Embeddings of facts	any KGE and/or neural model returning an embedding for logic atoms	Embedding of facts	Embedding of facts	undirected grounding continuous embeddings parameter (and possibly structure) learning arbitrary FOL or list of atoms with not-explicitely known relationships	Yes	L+N	S+Sub
	Express GNN	Model to approximate inference in MLNs with GNNs	https://github.com/expressGNN/E	Link prediction	Cora, FB15k,UW-CSE	Embeddings; Feature-based	Embeddings of facts	GNN	Embedding of facts	Embedding of facts	undirected grounding continuous embeddings parameter learning arbitrary FOL	Yes	L+P+N	S+Sub
	KGEs: like Transe, Distmult, NTN	Methods to provide fact embeddings according to relational knowledge	example of Python library for KGE https://pykeen.readthedocs.io/en/	Link prediction; graph classification; node classification	Any kind of KG/ontology/relational database, e.g. Freebase, FB15k, WordNet, WN18, NELL, DBpedia, Yago	Embeddings of constant and relation	Embeddings of facts	KGEs			directed and undirected grounding continuous embeddings no logic knowledge	yes	N	S+Sub
	Lifted Relational Neural Networks	Compiling of definite clauses into a neural network architecture to perform forward chaining	https://github.com/GustikS/GNNw	Link prediction	Mutagenesis, NCI-GI,							Yes	L+N	S+Sub
	PyKEEN	Python package to train and evaluate knowledge graph embedding models	https://github.com/pykeen/pykeen	(Inductive) Link Prediction	Kinships, FB15K-237, WN18RR, YAGO3-10	Triples	Embeddings of entities and relations & plausability scores for triples	Knowledge graph embedding model				Yes	N	S+Sub
	Neural Theorem Provers	End-to-end differentiable provers exploiting embedding representations of predicates and entities	https://github.com/uclnlp/ntp	Theorem Provers	Countries, Kinship, Nations, UMLS							Yes	L+N	S+Sub
Task 4.4 -- Perception, Spatial Reasoning, and Vision	Neuro-Symbolic Dynamic Reasoning													
	ToMnet													
	BlPaCK													

Analysis of the Tables

From the analysis of the tables, the taskforce collected a list of desiderata describing some properties that potential new datasets and benchmarking frameworks should include. We hereby describe such properties, trying to focus on different aspects: the nature of data; paradigms and tasks for learning, optimization, and reasoning; novel metrics to measure performance; practical issues dealing with software platforms, tools, and implementation; novel domains of interest that have been seldom investigated within this context.

Concerning data. By considering the data level only, a clear starting point is to combine low-level data (images, videos, text, signals) with knowledge of some kind. This knowledge could be implicit or explicit, exact or uncertain. One desirable feature would be to enable the possibility to ask different questions within the same dataset, thus by exploiting different sets or types of knowledge across different tasks. For this reason, considering multiple data sources (e.g., multimodal data) could be an interesting additional feature, as well as to include a dynamic dimension to tackle evolving data. That of temporal data is indeed a challenging domain that has seldom been considered, and which would need to rethink paradigms and tasks for experimental evaluation.

Concerning paradigms and tasks. Regarding the tasks and the paradigms for learning, reasoning or optimization, the taskforce identified a crucial element of novelty in interactive learning, where humans could interact with the systems, by providing various forms of feedback, from simple labels to critiques, explanations and arguments. This will habilitate interactive debugging during learning and foster interpretability and trustworthiness, and it will be especially relevant in systems that consider lifelong or continual learning and again the temporal dimension, allowing them to adapt to distribution and knowledge drifts and to small-data regimes (i.e., few-shot learning). Additionally, it would also be interesting to jointly consider multiple learning tasks within a single benchmark, since this would allow testing multiple skills at once of the systems.

Concerning performance. Another point that was raised by the analysis of the tables is how performance should be measured. Besides considering classic metrics that essentially focus on accuracy, benchmarks that aim to include and exploit background knowledge should also measure the interpretability of the results (following the recent trends in eXplainable AI) and possibly the coherence of the predictions with the available knowledge. Energy efficiency to reduce the carbon footprint is yet another dimension to consider.

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.

Concerning domains. Finally, the analysis of the datasets table was very useful in highlighting how some domains are under-represented in the panorama of benchmarks that are usually considered. Planning is a clear example of an application domain that would be perfectly suitable for testing the integration of learning, reasoning and optimization, as it can easily provide both symbolic data, such as activity traces or maps, and numeric data, coming from perception. 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. Visual question answering and conversational agents are instead two candidate applications to allow interaction with users and knowledge integration in the fields of computer vision and natural language processing: in the latter case, computational argumentation and argumentation mining could be an additional research field where symbolic knowledge is typically employed to encode argument models. 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.

Towards Creating the Next Generation of 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.

The taskforce organized 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).

Appendix: Program of a WP 4 Workshop on This Deliverable

What Are the Next Measurable Challenges in AI? (March 3, 2022)

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 certainly to evaluate novel ideas and frameworks on appropriate benchmarks. Too often, in fact, the tasks and the datasets that are considered and proposed for experimental evaluation are tailored to some algorithms or methodologies, and limited to ad-hoc scenarios and application domains. More in general, 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.

Can we define a set of requirements for a challenge/benchmark that goes beyond those currently available?
Can we do it with the goal of having a benchmark (or rather a benchmarking framework maybe) that meets these requirements and can still be implemented in a reasonable time? possibly building on top of existing ones?

Program

13:00-13:15 Doors open

Introduction

13:15-13:30 Introduction & Expectations - Luc De Raedt

13:30-14:00 Invited Talk: Lessons Learned at NeurIPS 2021 Datasets and Benchmarks - Joaquin Vanschoren

PART I (grounding the discussion in the literature)

14:00-14:30 Presentation Datasets/Systems Tables - Marco Lippi/Francesco Giannini

14:30-15:30 Discussion on Tables - Working groups

15:30-15:45 Break

PART II (widening the perspective)

15:45-16:45 Panel on Limitations of Existing Benchmarks and New Challenges - Andrea Passerini

- Fosca Giannotti
- Marco Gori
- Kristian Kersting
- Michele Sebag
- Joaquin Vanschoren

16:45-18:00 Discussion on Panel - Working groups

18:00-18:15 What's Next? - Luc De Raedt