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

Synergies Industry, Challenges, Roadmap for Social AI

Document type (nature)	Report
Deliverable No	D7.4
Work package number(s)	WP7
Date	Due M40, December 2023
Responsible Beneficiary	LEU, ID #7
Author(s)	Joaquin Vanschoren (TUE), Annelot Bosman (LEU), Holger Hoos (LEU)
Publicity level	Public
Short description	This deliverable is dedicated to the synergies between the industry and the data challenges tackled in TAILOR on one side, and the academic work explored in WP7 (Automated AI) on the other side.

History			
Revision	**Date**	**Modification**	**Author**
Version 1	16-02-2024	-	Joaquin Vanschoren (TUE), Annelot Bosman (LEU), Holger Hoos (LEU)

Document Review		
Reviewer	**Partner Acronym**	**Date of report approval**
Fredrik Heintz	LiU	2024-02-25
Marc Schoenauer	Inria	2024-02-25
Umberto Straccia	CNR	2024-02-25
Luc De Raedt	KUL	2024-02-25
Giuseppe De Giacomo	UNIROMA	2024-02-25

Ana Paiva	IST	2024-02-25
Holger Hoos	ULEI	2024-02-25
Philipp Slusallek	DFKI	2024-02-25
Peter Flach	UNIBRISTOL	2024-02-25
Joaquin Vanschoren	TUE	2024-02-25
Barry O'Sullivan	UCC	2024-02-25
Michela Milano	UNIBO	2024-02-25

Table of Contents

Summary of the report

This deliverable is dedicated to the synergies between the industry and the data challenges tackled in TAILOR on one side, and the academic work explored in WP7 (Automated AI) on the other side.

The document consists of two parts. Part I summarises the TAILOR activities and results regarding the Industry, mainly covered by the Theme Development Workshops, the Data-challenges, and the activities around the roadmap of trustworthy AI, such as (i) future mobility & Trust in AI; (ii) AI in the public sector; (iii) AI for future manufacturing; (iv) AI mitigating bias & disinformation; (v) AI for future energy & sustainability; and (vi) Trusted AI: The Future of Creating Ethical & Responsible AI Systems.

Part II summarises the research topics and activities regarding WP7, Automated AI, and the synergies with and relevance to industry and the (data) challenges, and synergies with the tailor roadmap.

This deliverable summarises crucial industrial requirements, data challenges, and roadmap components of the TAILOR project in alignment with the efforts undertaken in WP7, showing how the WP7 topics are central in the development of trustworthy AI systems.

Introduction to the Deliverable

This report is one in a group of five Synergies-deliverables in TAILOR, each pertaining to one of the five TAILOR scientific work packages (WPs 3-7), as shown in the table below. Each of the five Synergies-deliverables reflects on synergies between the scientific work done, and the work of WPs 2 “Strategic Research and Innovation Roadmap” which also includes data-Challenges, and 8 “Industry, Innovation and Transfer program”.

Scientific WP	Title
WP3	Trustworthy AI
WP4	Integrating AI Paradigms and Representations
WP5	Deciding and Learning How to Act
WP6	Learning and Reasoning in Social Contexts
WP7	Automated AI

Each of the five deliverables has two parts:

- Part 1 is introducing the work in WPs 2 and 8 and is the same in all the reports:
  - summarises the TAILOR industry activities, challenges and roadmap and was developed in joint efforts of participants of all the involved WPs. It is included here in order to make the deliverable self-contained.
- Part 2 is proper to the WP.
  - developed within each WP and positions the WP w.r.t. the first part.

This report, D4.6, is about the synergies between the scientific work on Paradigms and Representation and the data-challenges, industry efforts and roadmap work.

Process and people

All five scientific WPs have been represented in the joint working group for the first, common part. This joint working group was led by TNO with support from the project management office at LiU.

Table 1 below lists the people involved in writing the common part.

The project industry partners have all been engaged in WP2 (Roadmaps and Challenges) and WP8 (Industry).

Partner ID / Acronym	Name	Role
TNO	Wico Mulder	WP6, process lead
INRIA	Marc Schoenauer	WP2
DFKI	Janina Hoppstaedter	WP8
CNRS-IRIT	Andreas Herzig	WP5
CNR	Francesca Pratesi	WP3
Inria	Elisa Fromont	WP3
KU Leuven	Robin Manhaeve	WP4
TU/e	Joaquin Vanschoren	WP7, main author
U Leiden	Annelot Bosman	WP7, main author
LiU	Trine Platou	WP1, process support

In addition, the people listed in the table below are engaged in WP7 and have contributed to writing the parts of this report specifically related to WP7 AutoAI.

Partner Acronym	Name
TUE	Joaquin Vanschoren
LEU	Annelot Bosman
LEU	Julia Wasala
TUE	Pieter Gijsbers
LEU	Holger Hoos
LEU	Jan van Rijn

Part I: Industry, Challenges, and Roadmap in TAILOR

(To jump to the WP-specific part, click here)

Industry

Theme Development Workshops (TDWs)

TAILOR has organised so-called Theme Development Workshops (TDWs) during which players from industry and academia discuss challenges and key AI research topics in a certain area or in a specific industry sector. In total, seven workshops have been organised. This section provides a brief summary of the industrial challenges obtained from the outcome of those TDWs. Full reports can be retrieved from the Tailor website.

Future Mobility - Value of Data & Trust in AI (October-2021)

DFKI and ZF Group presented on AI techniques related to self-driving cars. An overarching challenge is to deal with safety and security. There is a strong need for robust metrics and automated checking of the quality of data and labels. Furthermore, robustness of algorithms to unforeseen environmental changes and adversarial attacks is something to work on, as well as topics related to explainability. Also privacy was discussed, pointing to the need for safe and controllable forms of data sharing, learning from anonymized and encrypted data and forms of federated learning. Volkswagen AG stressed the difference between invention and innovation. There is an overarching need for valorisation of research results and a data driven approach to innovation. Also understanding (getting grip on) the aspects of trust is a major concern since this is in the end what will define the success of innovative AI solutions in the eyes of end-users.

During the workshops it was discussed on how AI algorithms could monitor and detect situations to decide when it is necessary to hand over control to a human. The need for education, familiarity and adoption of AI driven approaches throughout the whole sector was expressed. It was also perceived that the act of estimating the business value of data for

different types of users was found to be complex. Also the difference between explainability and trust was found to be complex and hard to generalise across different domains.

AI in the public Sector (November 2021)

Upcoming technological solutions and adoption of transformation processes in the context of cities and municipalities, urges the need for urban labs. Education and methods that foster the growth of startups and scaleups, which are booming in the overall domain of AI, are important for economic growth. There is also a need to keep a grip on the lawful and ethical aspects of AI. Upcoming Data and AI-ACTs were discussed. Since the rise of AI application comes with an increasing number and type of risks and societal threats, opinions were discussed on the leading role of the public sector in how it should address the various aspects of trustworthy AI.

The breakout sessions addressed fairness, accountability, transparency, explainability which are generic concepts that underlie the overall need for guaranteeing safety of AI systems. The challenge is to allow technology to evolve from within a human-centric paradigm. Reliability plays a crucial role in this. Attention for education and career development was conceived as very relevant for further adoption of AI in our society. There is also still a strong need for techniques that can better deal with the timeliness, complexity, availability and quality of data.

AI for Future Healthcare (January 2022)

The Luxembourg Institute of Health presented on the role of AI in healthcare using data driven methods in numerous fields, e.g. efficiency in diagnostics and precision medicine. These methods aim for economic savings, prevention and better patient care. Barcelona Supercomputing Center explained the field of genomic data science. Both organisations stressed the urge for quality standards, common analysis standards and pipelines as well as data sharing in terms of federated access, discovery systems and federated learning. Some of the key technology areas with applications in the healthcare domain are Natural Language Processing, deep learning for imaging and detection, and tools for adequate decision making.

Philips Research stressed the importance of responsible usage of data and recent developments of using AI techniques in the field of MRI scanning. The fourth presentation, held by NTT Data, was about healthcare systems which make estimations and predictions about population health, care needs, healthcare professionals' decision-making and direct healthcare to persons using data centric approaches. It is challenging to guarantee the sustainability of the healthcare system to be resilient and flexible when facing threats.

In the workshops, the following needs for AI (research) were identified: 1) standards on frameworks that can support AI trustworthiness, including data quality, privacy enhancing technologies and data sovereignty. 2) explainability of AI models for trust as well as regulatory compliance. 3) the availability of adequate infrastructure for the conception, development, and validation of AI systems. 4) to understand how decision making and practitioners' behaviour are affected when AI detection and decision making systems will get more and more into play. It was also concluded that support for education and career development is needed.

Solutions that involve the monitoring of patients through daily interaction, stress the attention for further inclusion of social psychology and related disciplines into the field of AI research and innovation. Like persuasive technologies in various marketing domains, nudging and learning in social contexts were found to be crucial in advanced advisory and coaching systems. In the context of dialogue based interaction, dealing with ambiguity was mentioned as one of the key areas to improve upon.

AI for Future Manufacturing (October 2022)

DFKI started with a keynote on the topic of industrial AI across industry 4.0 and how it encompasses competitive manufacturing processes. Examples include predictive maintenance, planning, zero-error production and quality monitoring. Directions go in using cyber-physical systems and hybrid-ai solutions. The ZF group continued and highlighted the need for explainable AI. The third talk was given by CIIRC on robotics and edge-computing and ABB concluded the series of presentations. Both urged for higher quality of data in order to reach the required levels of reliability of AI solutions.

In the breakout sessions it was discussed to what extent an industry can give guarantees on AI trustworthiness of its products. E.g. how to verify that a solution is trustworthy, and the question who takes responsibility during deployment: supplier(s) or customer? A different group discussed the challenges around training AI models without giving up data sovereignty. Approaches to share models instead of data were addressed. The application areas of design and assembly demand for richer and transferable models and machine learning techniques for running simulations and algorithms that are robust to different types of sensors. In manufacturing for the space-industry the challenge of energy-efficient AI methods was mentioned. In the session about zero defect production and the session synthetic data generation the challenges were identified: the need for formal representation of data, ageing of models, lack of training data, and dealing with false alarms.

AI Mitigating Bias & Disinformation (November 2022)

The participating organisations discussed the difference between *mis*information, which is understood to be false or incorrect information, and *dis*information which describes false information that has been purposefully spread to deceive others. The idea of psychological inoculation functions similar to vaccines, as it may be possible to protect people from misinformation by either warning them of the fact that they are about to be misled or by pre-emptively providing them with the correct information, if false information about an issue is currently being spread. However, just with fact-checking, there are issues of scaling this solution, as anticipating each new misinformation trend is incredibly difficult.

Main concerns mentioned were on misusing AI technology and the increasing speed at which disinformation evolved and spread. Deepfake generation and detection methods deserve serious attention. On the front of deepfake generation models, it appears that diffusion-based models are now surpassing GAN-based methods in terms of realism and quality. In terms of detection approaches, a variety of approaches seem to be necessary, including for instance fingerprinting approaches, data augmentation (for more robust training), and person-specific biometric/semantic approaches. It was discussed whether neuro symbolic approaches could help in addressing these challenges.

From an AI perspective, a big challenge is how to build tools that help AI systems to "understand" human social rules, that recognize potential social biases, and possibly correct their effect on the system. On the topic of generative models, the evaluation of the performance of large language models was identified as challenging. It is important to know the quality and fit of generated text regarding the content and the message conveyed in it. Last but not least, AI-driven social media is found to be the key arenas for shaping public opinion, political controls. Many challenges lie here, and regulations might be a necessary measure, as they play a central role in our society and thereby in every industrial domain.

AI for Future Energy & Sustainability (May 2023)

ABB explained in their keynote how AI contributes to the integration of renewable energies, supply forecasts and monitoring & prevention. They mentioned the importance of balancing the potential benefits of AI with its environmental impact. Sharing best practices and leveraging collective intelligence is a key step in creating sustainable solutions, as it enables organisations to learn from each other and work together towards a common goal.

The keynote of ETH Zurich was about disruption of the legacy energy system from fossil fuels towards renewable energy sources. AI is playing a pivotal role in smart grid management, predictive maintenance, energy storage, and optimization. However, this comes with challenges on adoption to new power demand patterns and controlling new sources of flexibility. Other challenges associated with the use of AI in the energy system that were identified are: data privacy, data security, explainability, transparency, and accountability. Each of these challenges needs to be addressed to ensure that AI is used responsibly and effectively in the transition towards a sustainable and intelligent energy future.

The third presentation, given by TNO stressed the increasing role of AI as asset moderator, and discussed concerns about feasibility and safety due to the high responsibility involved. While implicit competition, especially price-based, could align well with AI, there are still many unanswered questions. Moreover, market-based competition, which is prevalent in the energy system, poses its own challenges. Additionally, the governance of distributed energy system operation needs to be better defined. It should be treated as an organisational challenge where AI handles responsibilities. Interoperability is also essential; designs should contribute to the broader picture instead of focusing on isolated systems.

This was also the message in the closing keynote given by EDF. Two challenges were mentioned: first, how to build a generic and trustworthy AI model for time series data, which is useful for several applications such as peak load estimation, flexibility management, network balancing and customer consumption analysis. Besides the operational constraints of data quality, an important regulatory constraint is the European GDPR, as individual load curves are classified as private information. Therefore, it is imperative to build models that are generic, privacy preserving, and robust against attacks all while maintaining good performance levels. The second challenge was how to build explainable AI models when dealing with multimodal data. The data collected can be either structured (tables, time series, contract information) or unstructured (emails, audio transcriptions, power plant photos, drone photos, etc.). The goal is to build an AI model that can handle all this variety of data, while being able to explain how the output is obtained.

The breakout sessions discussed various examples of domain related problems such as addressing responsibility, as well as complexity in operational management. It was stressed the importance of approaching the challenges in a multi disciplinary approach. The promises of AI in the energy sector are manyfold; on improving energy production (nuclear, hydro, renewable) by monitoring, fault detection and diagnosis, uncertainty quantification, etc. and in the operation of distribution networks via forecasting models for load, demand and prices can be realised, including its role in getting knowledge of consumer behaviour to help reduce electricity consumption and prepare for e-mobility and interaction using tools for customer relationship management, text and voice processing etc .

Trusted AI: The Future of Creating Ethical & Responsible AI Systems (September 2023)

DFKI offered a comprehensive insight into the European Commission's initiatives in the field of Artificial Intelligence (AI) and provided an overview of the forthcoming AI Act. It also delved into the Commission's strategies to ensure the effective implementation of AI legislation. The presentation outlined various areas where harmonised standards would be developed to operationalize the AI Act's requirements. These areas encompassed cybersecurity, transparency, robustness, accuracy, and the need for advanced explainability methods to generate explanations that are accurate and informative.

The challenges surrounding generative AI encompass a wide array of ethical, societal, and technical considerations. Addressing these challenges requires collaboration among various stakeholders, a commitment to ethical design, and ongoing efforts to ensure the responsible and equitable use of generative AI technology.

The challenges and considerations discussed in the breakout session revolve around the complex task of developing artificial systems that can effectively interact with humans, anticipate their behaviour, and foster trust. It was also suggested that having many different ethical AI frameworks may be beneficial because of the variety of orientations they apply to. However, to be meaningful, they should be industry and/or use-case specific.

The participants indicated that in the last decade we observe a massive imbalance in resources and talent between private and public sector, aggregated by the fact that currently, 70% of individuals with PhDs in AI find employment in the private sector. To this end, it is a private sector-centred logic that drives what we, as a society, focus on. More funding is needed to develop technology which prioritises public, and not private, values.

An argument was made that the principle-based approach to AI ethics has failed. That is because it is unclear how to evaluate and balance values against each other, how to implement them in technical systems, and how to enforce them in practice. There is a need for a novel set of interdisciplinary skills and on-going governance required to embed ethics in the entire cycle of AI development: from concept development to evaluation. Responsible development of technology requires groundwork, implementation of the processes, documentation, multi-disciplinary collaboration, stakeholder convening, a skills set different from what most academics, ethicists and philosophers traditionally do.

The participants also discussed a regulatory approach to AI ethics through the lens of the AI Act proposal. It was pointed out that the AI Act proposal has two main aims when it comes to

AI ethics: i) harmonisation of the vocabulary; ii) making principles enforceable. Experts pointed out that the AI Act does not contain a specific list of ethical principles, but rather requirements which are based on ethical principles. To illustrate, a human agency and oversight principle translates into auditing and impact assessments requirements. Similarly, a transparency principle translates into a requirement of the disclosure of the datasets for the foundation models.

Other challenging aspects that were discussed were a) finding effective control strategies in the interaction between intelligent machines and human agents. For instance, traded control (where a human agent completely relinquishes control at some point in time) might offer advantages in certain cases, while a symbiotic, dynamic interaction (where the amount of contribution may e.g. dynamically and continuously vary) might be recommendable in other cases and b) defining effective mechanisms of responsibility attribution through forms of control that can grant a meaningful (self-)attribution of responsibility across the different controllers and agents that populate a sociotechnical system. This is a challenge that touches many factors affecting human-AI interaction, such as opacity, unpredictability, delusions of agency and so on. A key point is the study of how trust naturally emerges in systems that incorporate the concepts of Theory of Mind (ToM) within their negotiation mechanisms. We have to bridge the gap between theoretical insights, particularly from game theory, and their practical application in real-world scenarios containing human-agent interactions. A crucial caveat is recognizing the limitations of ToM, as human reasoning is inherently imperfect. This exploration is essential for building trust in AI systems that can collaborate effectively with humans.

Categories

Indicatively, in very generic way, one can group the industrial challenges as follows:

- Robustness of algorithms
- Managing the quality of data
- Standardisation, verification, certification
- Explainability and transparency of algorithms
- Learning in federated context
- Responsibility
- Education on data driven and algorithmic processes
- Interaction on social level
- Ethical and legal aspects

TAILOR Data Challenges

Within the context of the TAILOR project, computational competitions (originally named as 'challenges') were organised aiming to tackle techniques, foster collaboration and address issues related to trustworthiness.

In order to overcome the ambiguity here, we refer to these activities as 'Tailor-data challenges.

TAILOR scientists have co-organised data-challenges together with leading industrial groups to create data challenges and hackathons for Trustworthy AI. The ambition is to jointly identify data sets that are suitable for advancing science, in a real-world industrial application setting. The following challenges have been organised in the context of TAILOR, but note that several of these challenges were presented in detail in Deliverable D2.3, "Foundational benchmarks and challenges Report" delivered in August 2022. The ones that were run later (or are still being run now) will be similarly presented in the Version 2 of this Deliverable, D2.6, due at Month 46. Furthermore, all Challenges will be thoroughly analysed in Deliverable 2.4, "Lessons learned from TAILOR Challenges", also due at Month 46.

Smarter Mobility Data Challenge, EDF + Manifest AI + Inria (Oct. Dec. 2022)

The Smarter Mobility Data Challenge aimed at testing statistical and machine learning forecasting models to forecast the states of a set of charging stations in Paris at different geographical resolutions. Transport represents almost a quarter of Europe greenhouse gas emissions.

Electric mobility development entails new needs for energy providers and consumers. Businesses and researchers are proposing solutions including pricing strategies and smart charging. The goal of these solutions is to avoid dramatically shifting EV users' behaviours and power plants production schedules. However, their implementation requires a precise understanding of charging behaviours. Thus, EV load models are necessary in order to better understand the impacts of EVs on the grid. With this information, the merit of EV charging strategies can be realistically assessed.

Forecasting occupation of a charging station can be a crucial need for utilities to optimise their production units in accordance with charging needs. On the user side, having information about when and where a charging station will be available is of course of interest.

The Dataset consisted of time based status data of 91 charging stations and was posed as a clustering and time series prediction problem. A detailed description of this challenge was provided in Deliverable D2.3 in August 2022, i.e., before the actual start of the challenge: As said above, the results will be described in Deliverable D2.6, and analysed together with the results of all TAILOR challenges in Deliverable D2.4, and we only present them rapidly here.

This challenge was run on Codalab, from October to December 2022. Twenty-eight teams participated in the Development phase, for a total of 296 submissions. However, only eight submitted their best solution to the final phase, and there were three clear winners, well above the others – the first two being very close, clearly above the third one. The winners used CatBoost, an Open Source implementation of Gradient Boosting chosen after some algorithm selection method (pertaining to AutoML). The second team used a weighted average of tree-based regression, tree-based classification (after discretization) and classical ARIMA method. Interestingly, these two teams obtained very close scores (206 vs 209, to compare to 220 for the third one and 255 for the fourth) though using very different approaches. The third team used different CatBoost models.

TAILOR was involved in this challenge through EDF, who was the most pro-active partner in the organisers (together with Air Liquide), providing and cleaning the data, and Inria: Sébastien Treguer participated to the preparation of the data and the design of the scoring function; Marc Schoenauer was member of the jury, chaired by Cédric Villani, the well-known Mathematician (2010 Field Medal) and Member of French Parliament. A jury was mandatory as the elegance of the solution was one of the criteria.

L2RPN II: Towards Carbon Neutrality, RTE and Inria (June-Sept. 2022)

The “Learning to run a power network challenge 2022” is concerned with AI for smart grids, and is the last of a long series of challenges. All have been built by RTE, the French Power Grid operator, and the Inria TAU team (Isabelle Guyon, Sébastien Tréguer), in collaboration with EPRI, CHA Learn, Google research, UCL and IQT labs.

Power networks (“grids”) transport electricity across regions, countries and even continents. They are the backbone of power distribution, playing a central economical and societal role by supplying reliable power to industry, services, and consumers. Their importance appears even more critical today as we transition towards a more sustainable world within a carbon-free economy and concentrate energy distribution in the form of electricity. Problems that arise within the power grid range from transient brownouts to complete electrical blackouts which can create significant economic and social perturbations.

Grid operators are still responsible for ensuring that a reliable supply of electricity is provided everywhere, at all times. With the advent of renewable energy, electric mobility, and limitations placed on engaging in new grid infrastructure projects, the task of controlling existing grids is becoming increasingly difficult, forcing grid operators to do “more with less”.

This challenge aimed at testing the potential of AI to address this important real-world problem to anticipate future scenarios of supply and demand of electricity at horizon 2050, aiming to maximally use renewable energies to eventually reach carbon neutrality. The challenge was intended to simulate a 2050 power system. One is

expected to develop the agent to be robust to unexpected network events and maintain reliable electricity everywhere on the network, especially when the network is under stress from external events. An opponent, which will be disclosed, will attack in an adversarial fashion some lines of the grid everyday at different times (as an example, you can think of lightning strikes or cyber-attacks). One has also to overcome the opponents' attacks and ensure the grid is operated safely and reliably (with no overloads).

Like the previous ones, this challenge is run on Codalab. A total of 16 participating teams made an entry on the final phase of the competition, among which only 5 were ranked above the baseline. The winner used an AlphaZero-based grid topology optimization. However, it should be noted that they had prior domain knowledge, as they are working on a congestion management solution for the energy sector, based on their topology optimization methodology. The second team used a single-step agent based on brute-force search and optimization tuned on the offline test set. Note that they did try PPO, a popular and usually powerful Reinforcement Learning algorithm, that performed worse here. Interestingly, the third team used no training at all. They choose the best action among 1000 randomly chosen ones, however with bells and whistles here and there. Again, a detailed description of this challenge was provided in Deliverable D2.3 in August 2022, i.e., while the challenge was still running, and further details on the results will be given in Deliverable D2.6.

MetaLearn 2022, Inria, Leiden U., and TU Eindhoven (Summer 2022)

Meta-learning is the field of research that deals with learning across datasets. While Machine Learning has solved with success many mono-task problems, though at the expense of long wasteful training times, Meta-learning promises to leverage the experience gained on previous tasks to train models on new datasets faster, with fewer examples, and possibly better performance. Such challenges obviously pertain to AutoAI (TAILOR WP7). But though grounded on learning, they also imply approaches from Unifying paradigms (WP4), depending on the solutions used by the candidates, and greatly improve the generalisation capabilities (e.g., across domain, see below) of the trained models, thus increasing the trustworthiness of the results (WP3).

Two series of challenges were organised under Isabelle Guyon's (Inria partner) scientific supervision, Meta-Learning from Learning Curves, and Cross-Domain MetaDL. Beyond Isabelle's role, TAILOR participated to the second rounds of both series, by sponsoring the winners' prizes and also through other TAILOR partners than Inria, namely Leiden University (partner #7) and TU Eindhoven (partner #12). All details regarding the datasets and the ranking measures have been given in Deliverable D2.3, but the results were not yet available at the time of writing D2.3, and will be detailed, as said in the introduction of this Section, in both Deliverables D2.6 and D2.4. We are only providing a bird's eye view here.

- **Meta-Learning from Learning Curves. Round 2: performance. w.r.t. dataset size**

  In this challenge series, the goal is to train a Reinforcement Learning agent that will choose the algorithm (with its hyperparameters) to use during the optimization. The training is made on meta-examples that are the learning curves obtained on some meta-datasets by some algorithms and given hyperparameters. The agent is evaluated by the Area under the Learning Curve (ALC) which is constructed using the learning curves of the best algorithms chosen at each time step (validation learning curves in the Development phase, and the test learning curves in the Final phase). While, in round 1, the meta-examples were 'performance vs time' curves, in round 2 they were 'performance vs dataset size'. The final score of the submitted algorithm was the worst one obtained out of 3 independent runs with different random seeds.
  The results of the challenge were officially announced during the AutoML conference in Potsdam, September 12th – 15th 2023. Ten teams only had submitted entries to the final phase. The winning team used a kind of Direct Policy Search approach, directly aiming at maximising the ALC (thus mixing Optimisation and Learning). They reached an ALC score of 0.39, remarkably stable across the random seeds. The second best score (0.35) was obtained by … the provided DDQN (Double Deep Q-learning Network[1]). It was however less stable than the winning DPS, with a maximum of 0.37. The next two scores (second and third prizes) obtained 0.32 and 0.31 respectively, though they both reached 0.36 as their maximum over the three random seeds. The second team trained an ensemble of models to predict both the performance and the CPU cost of a given algorithm from meta-data, that was used online during the run on the test examples. The third-prized team trained an algorithm comparator using embeddings of both algorithms and datasets, using end-to-end learning on the meta-training datasets.

- **Cross-domain MetaDL - Any way/any shot meta learning**

  The goal is to meta-learn a good model that can later quickly learn tasks from a variety of domains, with any number of classes (also called "ways") within the range 2-20, and any number of training examples per class (also called "shots"), within the range 1-20. All tasks were taken from various "mother datasets" selected from diverse domains, such as healthcare, ecology, biology, manufacturing, and others with the long-term goal to maximise the human and societal impact of the challenge. The average normalised classification accuracy over all meta-test tasks is used as the ranking metric, and the lowest of three independent runs is used for the final ranking (again, all details are given in Deliverable D2.3).
  Different "leagues" were proposed, with corresponding prized. The two main leagues were the **Free-style** league, in which pre-trained models were allowed, and the **Meta-learning** league, where no pre-training is allowed. A **New-in-ML** league, a **Women** league, and a **Participant of a rarely represented country** league were also given prizes, selected from the participants of the main two leagues (several teams won two prizes, one in the main leagues and one in some under-represented leagues).

[1] van Hasselt et al., Deep Reinforcement Learning with Double Q-learning, AAAI 2016.

The competition started July 1. for the main Development phase, and ended October 31. About 100 teams participated in the Development phase, with almost 400 submissions, 200 being valid. All winners used variations of Deep Learning techniques with specific bells and whistles. Note that the winner of the Meta-Learning league (and also of the New-in-ML league) is the only team which used attention mechanisms.

Brain Age Prediction from EEG Challenge, NeuroTechX (Nov. 2022)

In this challenge, participants were invited to use AI to predict the age of an individual from an electroencephalogram (EEG) recording time series. Such age predictions can be an important path to the development of computational psychiatry diagnosis methods. Computational psychiatry is a new approach in which algorithms are not only used to manage and organise data but also to understand hidden physiological and behavioural signals from the patient. This computational discrimination allows for both computer aided diagnosis (CAD) as well as a deeper understanding of the condition itself through generative models. By inferring the subject's age from their neuroimaging data one can then use the discrepancy between their biological age and estimated age to gather some insight into their individual developmental trajectory. The problem was posed as a regression problem. Each subject was characterised by time-series of EEG recording, with eyes opened and eyes closed. One had to predict the age of the individual.

This challenge was run on Codalab and was organised by the NeuroTechX company together with TAILOR partner Inria (Sébastien Tréguer). It attracted 36 competitors and more than 500 submissions for the development phase, and 20 made it to the final phase. The winners came way above the other teams, reaching 1.15 prediction score, while teams 2 and 3 were only separated by 3.10^{-3} around 1.6. Interestingly, they used a mix of expert hand-designed features and classical learning: an Empirical Wavelet Transform was used to extract 3 Intrinsic Mode Functions, obtaining a hybrid time-frequency representation, to which they added classical statistics for brain signal (variance, skewness, kurtosis, Point to point range, Root mean square, Standard deviation, number of zero crossings, Hjorth mobility and Hjroth complexity, Petrosian Fractal Dimension), leading to 3*11 features in time-frequency space. They also computed the so-called Power Spectral Density function through several frequency windows, together with the ratios of power across bands, leading to 9 features in the frequency domain. They then tried several learning algorithms, and found out that RandomForest gave the best results. All their code uses standard Python libraries (generic Scikit-Learn and neurophysiologically -specialised MNE).

Crossword puzzle

Organised by Prof. Marco Gori's WebCrow team at U. of Siena, this challenge has two phases, addressing automated crossword solving and generation, based on common modules hybridising Natural Language Understanding (NLU), Machine

Learning and constraint satisfaction, while gathering knowledge and data from several sources (web search, dictionaries, specialised multilingual schools curricula). Understanding crossword definition goes beyond NLU: Understanding clues requires several logical steps in Language Analysis.

The challenge was about solving and creating crossword puzzles. Crossword solving involves gradual tasks, from traditional clue answering and grid filling to integrated approaches for constrained clue answering, crossword correction, and end-to-end Neuro-Symbolic models. Crossword generation is about finding topic-relevant terms and clues/definitions, and involves the design (or fine-tuning) of some LLM for direct generation of clues/answers.

ML for Physical Simulations (aka Scientific Machine Learning – SciML)

Organised by IRT-SystemX, and co-organised by TAILOR (through its Inria partner) and several industrial partners (including NVIDIA, RTE and Criteo), this challenge intends to promote the use of Machine Learning based surrogate models to numerically solve physical problems, through a task addressing a Computational Fluid Dynamics (CFD) use case related to airfoil modelling. The challenge is held on the Codalab platform (maintained by the Inria partner), from Nov. 16. 2023 to end February 2024. The public training dataset is the AirFrans dataset described in the NeurIPS (dataset and benchmarks track) paper, made of 1000 CFD simulations of steady-state aerodynamics over two dimensions airfoils in a subsonic flight regime (5 real values at every point of the point cloud defined by the mesh on the simulation domain), and the participants have access for their simulations to the LIPS (Learning Industrial Physical Simulation) platform described in the NeurIPS (dataset and benchmarks track) paper. The task is to build surrogate models of these 5 fields for new airfoils, including Out-ot-Distribution cases, and the evaluation is a mix of accuracy (MSE), computational cost, and, last but not least, respect of the physical constraints (Navier-Stokes equations).

This challenge is run on Codabench (the new version of Codalab), and is still in its Development Phase, but at the time of writing, there are already 114 participants and 190 submissions, from both academia and industry.

Mind your buildings (feb 2023)

The challenge was about identifying behavioural patterns related to building occupancy using sensor data coming from a multi tenant building. In the period from January to March 2023, a group of 25 people worked on data science problems in the context of urban energy sustainability. It was organised by TNO and DFKI, in collaboration with the Hanze university of applied sciences in the Netherlands and the company AIMZ. The groups developed algorithms that could pinpoint and repair missing data in incomplete sensor data and/or floor plans of buildings. Models for prediction of occupancy were retrieved from the sensor data.

The organisers were thinking of organising a follow up (intended name ‘mind the avatars’ mind) in which they would like to study various implementations of using Theory of Mind.

The challenge was organised in the form of a ‘dilated three day hackathon’ by TNO in collaboration with the Hanze university of applied sciences, DFKI, and the company AIMZ.

20 people in three groups worked on questions related to energy management of a multi-tenant building. The evenings were organised in that particular building. The challenge involved mixed mode competition where discussions and presentations were plenary with all the teams, whereas there was a competitive element in the form of a prize for the best individual team. Various data science approaches were used to cluster data and learn predictive models.

Roadmap

Roadmapping aims at supporting strategic and long-range planning. It is referred to as the process that provides structured (and often graphical) means for exploring and communicating the relationships between evolving and developing research topics, technologies, and products.

The process of roadmapping involves the identification and the prioritisation, usually in time, of different elements in order to understand and steer the direction of research, technologies and product evolution. The process of developing a roadmap is as important as the final roadmap-document itself, as it requires researchers and stakeholders to think in terms of relationships and to work together to develop a plan to achieve common goals and objectives.

From a research perspective, a roadmap contains topics that show the evolution from a research content. The milestones cover the steps of their evolutionary paths, and address how the topic is related to a particular field of research. The research perspective provides insights in common planning horizons and might support funding decisions for European research programs that foster the economic strength of organisations and research institutes in Europe.

An industrial perspective on a roadmap captures stakeholder interests from a business perspective in various markets and industrial domains. Industrial roadmaps help to ensure that existing and potential technology can get aligned with economic and societal objectives and with the needs of end users. Both perspectives can be combined in order to provide insights into how important problems for society can be addressed, and highlights how to pursue important future research.

The first version of TAILOR roadmap was written following the structure of the scientific Work Packages of the network[2], WP3-7: one Chapter per WP, only with one additional Chapter dedicated to the Foundation models and the rising LLMs. The resulting document was written in a collaborative manner within each WP, after a series of discussions led by the WP2 and Task 2.2 leaders during the respective WP internal meetings during spring and summer 2021. All important aspects of Trustworthy AI were present in the different Chapters, but two main ingredients needed to be added: the links between the different WPs, i.e., between the Learning, the Optimization and the Reasoning aspects of AI (the L, O, and R), and some prioritisation among the objectives that had been identified. The Version 2 of the SRIR, due on month 44 (April 2024) will correct this. After fetching feedback from the whole consortium, a “Spring Camp” is being organised on April 8-9 to spread the collaborative work among the partners for the fine-tuning of this final phase. In particular, cross-WP discussions will take place in breakout sessions, in order to favour a more coherent topic-oriented organisation of the SRIR and ensure completeness and quality of the final document.

[2] After a totally unsuccessful attempt, via some poll sent to all partners, to adopt a different structure, oriented toward hybridization of AI – from hand-in-hand LOR, as in WP4, to much wider hybridization with other domains, of Computer Science and beyond.

Part II: Synergies WP7 with Industry, Challenges, and Roadmap in TAILOR

This part describes how the research activities in WP7 address the topics part I, i.e. the challenges in industry (discussed in the Theme Development Workshops), the data oriented hackathons (originally denoted as Challenges), and the TAILOR roadmap.

About WP7

The main goal of WP7 is to leverage research on AI methods at the “meta-level” to ensure that AI tools and systems, especially when built, deployed, maintained and monitored by people with limited AI expertise, are performant, robust and trustworthy. This includes both work on expanding the scope of current Automated ML (AutoML) methods, with the goal of covering machine learning settings currently not satisfactorily addressed by AutoML, as well as automatically detecting when AI systems (especially automatically constructed or configured ones) 'get off-track', automatically finding good tradeoffs between performance and the various dimensions of trustworthiness via multi-objective optimization, and ensuring that the costly AutoML process does not have to start from scratch whenever the use case slightly changes. As such, this WP is of direct relevance for a variety of industrial use cases, to optimise machine learning models while solving real-world constraints, such as resilience against noise, undesired effects in real-world data, reduce computational costs, or reduce the footprint of ML models (e.g. to make them fit in embedded devices). The WP7 activities can be summarised as follows.

Task 7.1 AutoML in the wild*:* Facilitate the **usability** of machine learning by non-machine learning experts who have data and a clear target to predict, but who are not familiar enough with machine learning to know which neural architecture or machine learning pipeline to use, and how to set its hyperparameters. Efforts are concentrated on developing neural architecture search (NAS) methods that are usable in the wild. This task aims to design robust and efficient methods for determining strong neural architectures and their hyperparameter settings for previously unseen datasets. Secondly, it aims to design methods to automatically handle messy real-world data in AutoML. This includes developing new methods for feature preprocessing and feature selection, as well as methods for handling unstructured data, data with many defects, missing data, concept drift, and data wrangling.

Task 7.2 Beyond standard supervised learning: Expand the scope of AutoML to diverse and rich learning settings. Work in this task will bring together AutoML researchers with experts from multi-target regression, unsupervised learning, semi-supervised learning and learning on spatio-temporal data, such as time series, location traces and trajectory data, with the goal of designing rich, highly flexible algorithmic frameworks for these settings, as well as adapting powerful general-purpose algorithm configuration techniques for configuring these automatically. Moreover, it aims to assemble collections of benchmark data and scenarios for these settings, drawing from existing public libraries and repositories and elicit novel real-world datasets in partnership with industry.

Task 7.3 Self-monitoring AI systems: Automatically detect when an AI system (such as a classifier, predictor or reasoning engine obtained from an AutoAI system) gets 'off-track' and

can no longer be used safely and reliably. AI systems should be robust against changes in the environment in which they operate, whether the changes are due to natural drifts, the system actions, or adversarial attacks. A key step towards this goal, especially when using AutoAI for automated customisation and performance optimisation, is to be able to recognise when an AI system begins to show non-robust behaviour -- and, in a second step, automatically corrects itself. We refer to this capability as self-monitoring. This task aims to produce general-purpose methods and techniques for achieving self-calibration, not only for machine learning, but also for systems or components for other AI problems, such as reasoning, planning and optimisation problems. It will also develop metrics for critically assessing such techniques, and collect and design suitable benchmarks.

Task 7.4 Multi-objective AutoAI: Develop multi-objective AutoAI methods to automatically determine the best tradeoffs between performance and other objectives, e.g., derived from the six dimensions of trustworthiness. All AI systems are designed and implemented to fulfil some operational goal. Hence, any dimension of trustworthiness that we want the system to meet can only be one additional objective. And more often than not, it is necessary to give up some efficiency on the main goal in order to improve along the trustworthiness dimension(s). Using AutoAI to design and optimise trustworthy AI systems is hence de facto a multi-objective problem. Building on existing multi-objective general-purpose algorithm configuration, this task brings together researchers in optimization and ML to adapt and develop, in the AutoML context, general-purpose multi-objective algorithms, and assemble data sets for evaluating multi-objective AutoAI methods.

Task 7.5 Ever-learning AutoAI: Ensure that AutoAI gets better over time, producing better models with less data, and avoids the computational overhead of starting from scratch for any new use case, or change in scenario. Modern AI systems are often much less resource and sample-efficient than humans, requiring much more data and computation to learn a given task. One reason for this is that humans never really learn from scratch, they can build on a lifetime of experience on other related tasks. Research into meta-learning can give us AI systems that follow a similar approach. These will be much more efficient to train (**sustainability**) and will be **more robust** since they reuse previously learned patterns and representations, and built on data from many prior learning episodes, not just data from the task at hand. Current successes in this field (e.g. in transfer learning), still require manual trial-and-error. In this task we aim to set up infrastructure, collect (meta)data, and develop techniques that leverage this metadata to automatically design efficient open-ended AI systems.

Synergies with and relevance to industry

In this section we discuss how the research topics mentioned above address the challenges mentioned in part I. AutoAI enables us to build tailored, customised and performance-optimised AI solutions to problems from diverse domains. WP7 aims to expand the scope of AutoML and AutoAI (beyond ML) to diverse learning settings and ensure AI systems' trustworthiness, industries across various sectors will benefit. Whether it's manufacturing, healthcare, finance, or retail, companies can deploy tailored AI models without extensive in-house expertise.

Addressing Real-World Challenges

Industries often grapple with messy and unstructured real-world data, as well as data for which few (or no) labels are available. ***Task 7.1 (AutoML in the wild) and Task 7.2 (Beyond standard supervised learning)*** have an emphasis on handling such data using AutoML techniques, ensuring that industries can derive actionable insights without extensive and labour-intense data cleaning and processing. We have worked with a variety of companies, allowing the project to address specific, tangible challenges faced by these entities, ensuring the solutions developed are not just theoretical but practically viable.

One real world challenge that is prevalent in industry is finding the right (hyper-)parameters for a process, AI method or machine learning pipeline. Often practitioners have some idea of a good initial setting and wish to further improve which is previously quite difficult to incorporate into (hyper)parameter optimisation tools. The work of Mallik et al. (PriorBand[3]), sponsored by TAILOR, helps to alleviate this issue where practitioners can inform the optimisation system of what they believe to be a good initial setting to dramatically speed up the time taken to find the optimal settings for their problem. This work was done in collaboration with Lund University, Sweden and Hannover, Germany and the University of Freiburg.

One question that the industry is beginning to face is “Is your system fair?”, with non-discrimination legislation being published by the EU as well as the US. AutoAI, and more particularly AutoML methods, can be employed to help an industry improve efficiency through automated means, yet these automated systems are often critically un-aware of any fairness constraints that an industry may have when building predictive models. A recent TAILOR publication “Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML[4] attempts to address these issues, acting primarily as a guide for developers of AutoML systems to be aware of both the potential and challenges of utilising these systems in practice. This piece sets forth a path for industry to deploy AutoML systems and be aware of their utility and limitations, such that industry can meet governmental legislation. This paper was published jointly with LMU Munich, Eindhoven University of Technology and the University of Freiburg.

We also applied AutoML techniques in direct collaboration with companies. For instance, in the context of self-driving cars, we collaborated with NXP to leverage AutoML to design novel radar-based object detection models that are simultaneously much smaller and more accurate.[5] In the context of semiconductors, we worked with AMSP to produce novel machine learning methods to detect anomalies that can be used to detect defects in semiconductor chips.[6]

Constant interaction with industry is critical to shape research priorities based on real-world challenges. This ensures the outcomes of the research are always relevant and can lead to

[3] Neeratyoy Mallik, Carl Hvarfner, Edward Bergman, Danny Stoll, Maciej Janowski, Marius Lindauer, Luigi Nardi, Frank Hutter. PriorBand: Practical Hyperparameter Optimization in the Age of Deep Learning, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA

[4] Hilde Weerts, Florian Pfisterer, Matthias Feurer, Katharina Eggensperger, Edward Bergman, Noor Awad, Joaquin Vanschoren, Mykola Pechenizkiy, Bernd Bischl, Frank Hutter, “Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML”. Journal of Artificial Intelligence Research (JAIR), 2024.

[5] Boot, T., Cazin, N., Sanberg, W., & Vanschoren, J. (2023). Efficient-DASH: Automated Radar Neural Network Design Across Tasks and Datasets. In *2023 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1-7). IEEE.

[6] Kerssies, T. & Vanschoren, J. (2023). Neural architecture search for visual anomaly segmentation. *Automated Machine Learning Conference (AutoML 2023).*

tangible improvements in industrial operations. The second AutoML conference[7] was successfully hosted in Berlin, 12-15 September, 2023 with ~300 participants from across both Academia and Industry. Notably, large names such as Amazon and Google but also startups and other businesses connecting with research. The conference featured a dedicated “Industry Meets Academia” day with a focus on connecting research with industry and the problems they face in practice. This included speeches from industrial researchers but also academics who collaborate frequently with industry.

Robust, self-monitoring AI systems

Before AutoAI can be adopted effectively in industry, it must be very robust and able to optimise very complex multi-dimensional objectives. ***Task 7.3 (Self-monitoring AI systems)*** aims to ensure that AI systems are robust against changes in the environment in which they operate, while ***Task 7.4 (Multi-objective AutoAI)*** ensures that industries make decisions that are not just performance-driven but also account for other factors such as sustainability, ethics, and cost-efficiency. This holistic decision-making can lead to long-term growth and sustainability.

A recent research line on robustness of deep neural networks, involving Leiden University and RWTH Aachen and partially funded by TAILOR, has introduced novel approaches for increasing the efficiency of computationally challenging neural network robustness verification problems and for assessing neural network verification in unprecedented detail.[8] [9] [10]. While this work is not yet at the stage where it can be directly applied in real-world applications, the aim and the potential for such applications is an important driving factor.

Work carried out by researchers at Leiden University, the University of Twente, the University of Münster and the University of Hannover and partially funded by TAILOR has recently extended one of the state-of-the-art automated algorithm configuration techniques underlying several AutoML systems based on Bayesian optimisation to multiple optimisation objectives, as often encountered in real-world applications. We fully expect this work, once published, to have significant potential for industrial applications of AutoML.

Very recently, a large cross-academia-industry collaboration has been started on AI Safety[11], in which TUE is playing a leading role, including key AI companies such as Google, Meta, Anthropic, Cohere, Coactive, as well as Stanford University and stakeholders from the public sector. It focuses on the design and implementation of a platform for the systematic evaluation of large language models (LLMs). It will evaluate LLMs according to a wide range of safety metrics, and it will be entirely open so that new tests and metrics can be added at any time. This will allow researchers and companies that develop or use LLMs to thoroughly test them and detect when they are out of line.

---

[7] AutoML Conference Industry Day. https://2023.automl.cc/

[8] Matthias Konig, Holger H. Hoos, Jan N. van Rijn. "Speeding up neural network robustness verification via algorithm configuration and an optimised mixed integer linear programming solver portfolio". In: Machine Learning 2022. paged 1–20.

[9] Matthias Konig, Annelot W. Bosman, Holger H. Hoos, Jan N. van Rijn. “Critically Assessing the State of the Art in CPU-based Local Robustness Verification**”.**In *Proceedings of the Workshop on Artificial Intelligence Safety 2023 (SafeAI 2023) co-located with the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI2023)*. 2023. Best paper award. Also under review at Journal of Machine Learning Research.

[10] Annelot W. Bosman, Holger H. Hoos, Jan N. van Rijn. A Preliminary Study of Critical Robustness Distributions in Neural Network Verification.In: 6th Workshop on Formal Methods for ML-Enabled Autonomous Systems (FoMLAS), 2023.

[11] https://mlcommons.org/working-groups/ai-safety

Enhancing Industrial Efficiency

Finally, AutoAI methods should be able to adapt quickly to new tasks to be useful for industry applications which emerge quickly or have very little data available. ***Task 7.5 Ever-learning AutoAI*** can revolutionise industrial operations by ensuring that their AI systems adapt to evolving challenges without the need for constant oversight or extensive reconfigurations. For a more complete discussion on the many application areas of meta-learning and open challenges, please see a recent survey that we did on this topic.[12]

A frequent problem in industry is that data evolves over time, also known as drift, and AutoML pipelines that worked well in the past may degrade suddenly or gradually. To address this problem, we developed a novel *online* AutoML framework[13], sponsored by TAILOR, that automatically detects drift and appropriately responds, by partially or entirely re-optimizing the pipeline, switching to methods that are more robust to drift, ensembling with other pipelines, or, when the drift is cyclic, switching back to earlier pipelines. As such, the AutoML system is also learning continually over a stream of real-world data as it is gathered.

A very common industry scenario is that there is plenty of data, but none (or very little) of it is labelled. Unsupervised problems, such as clustering and outlier detection, are therefore critical in many industrial applications. Still, they have long eluded AutoML since there is no ground truth to optimise towards. However, human experts can still be very effective at this task based on their experience on *similar* problems. This inspired us to use a meta-learning approach, LOTUS[14], in which a large number of proxy tasks (with ground truth) are created so that we can learn which unsupervised techniques work well (or not) on different problems. When given a new unsupervised problem, it finds the most similar prior problem using Optimal Transport, a measure of similarity between two data distributions, and consistently recommends the best clustering or outlier detection techniques.

Another frequent problem with real-world data that is felt across industries is that it contains many errors, due to human error or data gathering issues, that severely limit the effectiveness of machine learning solutions. In data-centric AI, the goal is to automatically improve data quality. Together with a large cross-academia-industry collaboration, including key AI players such as Google, Meta, Cohere, Coactive, as well as Harvard, Stanford, and Carnegie-Mellon University, we created a platform, DataPerf[15], to develop and benchmark automated techniques for data acquisition, augmentation, selection, quality assessment, and cleaning. This is stimulating the creation of novel data-centric AI techniques that will likely benefit many industries.

Another fruitful collaboration was achieved with Level 42 Inc., a US-based manufacturer of advanced wide-band e-stethoscopes, resulted in a novel method of pulmonary anomaly detection[16], an adaptive machine learning system capable of learning from small data (only

[12] A. Vettoruzzo, M. -R. Bouguelia, J. Vanschoren, T. Rognvaldsson and K. Santosh, "Advances and Challenges in Meta-Learning: A Technical Review," in *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.*

[13] Celik, B., Singh, P., & Vanschoren, J. (2023). Online automl: An adaptive automl framework for online learning. *Machine Learning*, *112*(6), 1897-1921.

[14] Prabhant Singh & Joaquin Vanschoren, AutoML for Outlier Detection with Optimal Transport Distances. Proceedings of IJCAI 2023.

[15] Mazumder, M., et al. Dataperf: Benchmarks for Data-Centric AI development. *Advances in Neural Information Processing Systems 36, NeurIPS 2023*, New Orleans, LA, USA

[16] Piotr Kaszuba, Andrew Turner, Bartosz Mikulski, Nl Shasha Jumbe, Andreas Schuh, Michael Morimoto, Peter Rexelius, Ryan Hafen, Ron Deiotte, Kevin Hammond, Jerry Swan, and Krzysztof Krawiec. “Synthesizing Effective Diagnostic Models from Small Samples Using Structural Machine

~20 subjects were engaged in the associated clinical trial conducted at the Johns Hopkins University). This capability makes it prospectively possible to adapt quickly to changing circumstances, (e.g. strains of viruses/infections varying temporally or/and geographically. The learning algorithm is also multi-objective — in addition to minimising the empirical loss, it attempts to limit the complexity of synthesised models in order to reduce the risk of overfitting. The method is a part of a recently granted patent.[17]

Here is the recap of how the TDW are related to WP7:

	Task 7.1	**Task 7.2**	**Task 7.3**	**Task 7.4**	**Task 7.5**
Future Mobility - Value of Data & Trust in AI	X	-	-	-	X
AI in the public Sector	X	-	-	-	-
AI for future healthcare	-	-	-	-	X
AI for Future Manufacturing	-	-	-	-	-
AI Mitigating Bias & Disinformation	-	-	X	X	-
AI for Future Energy & Sustainability	X	-	-	-	-
Trusted AI: The Future of Creating Ethical & Responsible AI Systems	X	X	X	X	X

Learning: A Case Study in Automating COVID-19 Diagnosis". In: Proceedings of the Companion Conference on Genetic and Evolutionary Computation. GECCO '23 Companion.

[17] Sensor systems and methods for characterizing health conditions, USPTO Patent US20210345939A1, Inventors: Nelson L. Jumbe, Andreas Schuh, Peter Rexelius, Michael Morimoto, Dimosthenis Katsis, Nikola Knezevic, Steve Krawczyk, Kevin Hammond, Krzysztof Krawiec, Gregory A. Kirkos. https://patents.google.com/patent/US20210345939A1

Synergies with and relevance to the (data) challenges

Of the proposed challenges discussed in Part 1, only the MetaLearn 2022 is directly related to the work done in Task 7.5. That said, WP7 has been very active in creating its own datasets and benchmarks with industry and societal impact. Indeed, benchmarks play a key role in testing and developing AutoAI applications for real-world problems. For AutoAI, multiple datasets with varying degrees of difficulty are essential: performance comparisons on single datasets do not give a full picture.

First, to obtain a clear view on the robustness and limitations of today's AutoML systems, we have created the AutoML benchmark[18], supported by TAILOR, that very quickly became the standard framework to evaluate AutoML techniques from both academia and industry. It is already adopted by most major AutoML systems, and provides a very detailed view on the real-world capabilities of AutoML systems. It automates the installation and configuration of the AutoML frameworks, downloading the relevant data, and processing the framework predictions. This software has been used by industry parties, such as Microsoft and Amazon AWS, to evaluate their frameworks in paper (e.g. AutoGluon[19] and FLAML[20]) but also in their development environment (in continuous integration). It is also designed to be extensible and will be expanded to include more and more real-world tasks as AutoML systems emerge that can deal with them.

The last years have seen remarkable progress in few-shot learning (learning from very few examples based on past experience), and continual learning (in which a single neural network is learned and continually adapted as it encounters new tasks, much as humans do). This rapid progress has led to an explosion of techniques and very little guidance on which techniques to use. The Meta-Album project[21], partially funded by TAILOR, is a novel benchmark to systematically evaluate meta-learning, few-shot learning, and continual learning techniques, and is seeing rapid adoption by researchers. This paper was published jointly with Eindhoven University of Technology and Leiden University.

It is challenging to compile benchmark suites for real-world applications such as climate modelling and problems from medicine[22], because they use a variety of data formats that seem very obscure to most AI researchers, and that need to be unified to be able to be used to train machine learning models. The benchmark also has to be diverse, which introduces the problem that different tasks often require different types of data. Real-world problems like sea ice detection (e.g., AutoICE) are very often multimodal, which only expands the number of possibilities (e.g., a combination of satellite imagery and time series of ground-based measurements).

One initiative to address these challenges, is the creation of a new Datasets & Benchmarks track[23] at NeurIPS, a top machine learning conference. This was co-initiated by TUE. It provides a new and strong incentive for domain scientists and practitioners to create new datasets and benchmarks that are of high quality and easy to use by the machine learning

[18] Gijsbers, P., Bueno, M., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B. & Vanschoren, J. AMLB: an AutoML Benchmark., Journal of Machine Learning Research (JMLR), 2024.
[19] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*
[20] Wang, C., Wu, Q., Weimer, M., & Zhu, E. (2021). Flaml: A fast and lightweight automl library. *Proceedings of Machine Learning and Systems*, *3*, 434-447
[21] Ullah, Ihsan and Carrion, Dustin and Escalera, Sergio and Guyon, Isabelle M and Huisman, Mike and Mohr, Felix and van Rijn, Jan N and Sun, Haozhe and Vanschoren, Joaquin and Vu, Phan Anh. Meta-Album: Multi-domain Meta-Dataset for Few-Shot Image Classification. 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Datasets and Benchmarks Trak, 2022.
[22] https://www.thelancet.com/pdfs/journals/landig/PIIS2589-7500(19)30108-6.pdf
[23] https://nips.cc/Conferences/2023/CallForDatasetsBenchmarks

community. This track has triggered an immense interest - e.g. in 2023, almost 1000 new datasets and benchmarks were submitted, and 322 accepted - and is turning the attention of the AI community away from 'toy' datasets and towards large real-world datasets that have real industry or societal impact, such as medicine, climate change, and generative AI.

A second initiative, also co-initiated by TUE, is the creation of a new journal, the Journal for Data-Centric Machine Learning Research (DMLR)[24], part of the JMLR family, that provides similar new incentives to create novel high-quality data and benchmarks.

Synergies with the tailor roadmap

AutoAI is highlighted in both the short version and the full version of the TAILOR roadmap. The overarching short term challenges are the creation of AutoAI benchmarks, which is of importance to all WP7 tasks and expanding existing techniques for application to real-world problems, which is particularly relevant for task 7.1 (AutoML in the wild) and Task 7.2 (Beyond standard supervised learning). The benchmarks that are created to date are discussed in the previous section.

The short-term goal of integrating learning, reasoning and optimisation (LOR) is particularly relevant for Task 7.3 (Self-monitoring AI systems), as we see the use of LOR in verification techniques and model checking for safety, bias and robustness. This is discussed above in the section on 'Robust, self-monitoring AI systems'.

The five tasks that are part of WP7 are all carefully constructed to benefit the long term goal of providing broad, safe and efficient use of AutoAI techniques for all sectors of industry and society. AutoAI has the potential to guarantee safe use of AI in sectors where AI expertise is not widely available. Task 7.4, multi-objective AI, for instance is necessary to not only make accurate AI systems, but also integrate different objectives, such as safety, fairness and sustainability. In general, good AutoAI systems are necessary to increase accessibility to state-of-the-art methods for practitioners, without having to reimplement, recalibrate or even search for them. This is an important step towards TAILORs second main objective.

Conclusion

This deliverable has summarised the most important aspects of industrial needs, data challenges and roadmap elements of TAILOR in synergies with what had been done in WP7.

[24] https://data.mlr.press/