TAILOR

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

D3.6 Synergies Industry, Challenges, Roadmap concerning Trustworthy AI

Document type (nature)	Report
Deliverable No	D3.6
Work package number(s)	WP3
Date	Due M40, December 2023
Responsible Beneficiary	CNR, ID #2
Author(s)	Elisa Fromont (Inria), Francesca Pratesi (CNR)
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 WP3 (Trustworthy AI) on the other side.

History			
Revision	**Date**	**Modification**	**Author**
Version 1		-	-

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	UNIBRIS	2024-02-25
Joaquin Vanschoren	TUE	2024-02-25
Barry O’Sullivan	UCC	2024-02-25
Michela Milano	UNIBO	2024-02-25

This document is a public report. However, the information herein is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

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 WP3 (Trustworthy 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 WP3, Trustworthy AI, and the synergies with and relevance to industry and the (data) challenges, and synergies with the TAILOR roadmap.

Both are then related using a table overview.

This deliverable summarises crucial industrial requirements, data challenges, and roadmap components of the TAILOR project in alignment with the efforts undertaken in WP3, showing how the WP3 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
U Leiden	Annelot Bosman	WP7
LiU	Trine Platou	WP1, process support

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 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's 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 furter 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).
  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

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

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 WP3 with Industry, Challenges, and Roadmap in TAILOR

This part describes how the research activities in WP3 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 WP3

WP3 aims at investigating the methods and methodologies to design, develop, assess, enhance systems that fully implement Trustworthy AI with the ultimate goal to create AI systems that incorporate trustworthiness by design. This activity is organised along the six dimensions of Trustworthy AI: explainability, safety and robustness, fairness, accountability, privacy, and sustainability. Each task aims at advancing knowledge on a specific dimension and puts it in relationships with foundation themes. The overall mission for Trustworthy AI is to combine the various dimensions in the TAILOR research and innovation roadmap. The WP3 activities can be summarised as follows.

Task 3.1: Explainable AI Systems. This task focuses on the explainability and transparency of AI systems. The consortium launched two Coordinated Actions (CAs) on this task. The first CA is related to *explainability of medical images*. In particular, we defined a case-study (prostate cancer radiologies) where a convolutional neural network for the classification of multiparametric MRI images on unbalanced datasets is defined, and we aim to compare different methods for providing post-hoc explanations for the outcome of the black-box model. The second CA is about *explainable malware/security threat detection*. In particular, a comparison of methods for detection and prediction of malware/security attacks is on-going that are able to produce some kind of explanation or characterization of the attack. An ontology for malware detection, based on the EMBER dataset has been constructed.

Task 3.2: Safety and Robustness. This task tackles the challenge of bridging the gap between existing methodology for safety and critical systems based on software engineering, formal methods, and verification, and the current AI solutions. The contributions from TAILOR were divided into 4 coordinated actions. The first one explored the *analysis of what a truly adversarial example is*. The idea was using metrics of difficulty and comparing them to the result of different models to determine adversarial attacks. New articles in this direction were published at a NeurIPS workshop on Safe ML 2022, and more recently on issues of contamination at ECAI2023. The second CA explored ideas around *robust evaluation*. The third CA is represented by the *SafeAI and AISafety* workshops (IJCAI), where we had a special TAILOR session at IJCAI2022 . Finally, the fourth CA deals with *formal methods for verification in AI*.

Task 3.3: Fairness, Equity, and Justice by Design. This task addresses methods to enforce by design, values of fairness, equity, and justice within AI solutions from a multidisciplinary perspective. One Coordinated Action (CA) is on *Operationalizing Fairness Metrics: the case of credit scoring*. The CA aims at studying the impact, for the specific application context, of dealing with the values of fairness, equality, and equity in the design decisions of fair AI models, in particular in the selection and adoption of fairness metrics.

Task 3.4: Accountability and Reproducibility by design. This task is addressing accountability (it includes measures of governance of the design, development, and deployment of algorithmic systems required to prevent misuse of these systems) and reproducibility (methodological measures, quality standards, and scientific and technological procedures to better model the development of learning methods for AI). We have engaged in fruitful discussions exploring the *connections between* the sometimes ill-defined notions of *accountability and reproducibility in AI systems*, which has led to the emergence of a new CA.

Task 3.5: Respect for Privacy. This task investigates different dimensions of the problem of privacy protection when building AI systems. The task launched two CAs. The first CA is related to automatic tools for analysing and explaining privacy risks. We have surveyed and compared different complementary methods to analyse and quantify privacy risks raised by publishing aggregates of time series, depending on different publisher and attacker models (this is under review at ACM communication). The second CA plans to analyse the datasets' characteristics that impact the utility and/or efficiency of existing techniques such as Differential Privacy applied e.g. high-dimensional data such as multivariate time sequences (movements, speech, etc) whose private/non-private dimensions are strongly entangled.

Task 3.6: Sustainability. The task is about designing AI applications that improve sustainability. The goal is to reduce the impact of AI training and deployment and enhance sustainability solutions with AI. One of the topics we focus on is energy management algorithms and optimizations to reduce the energy consumption of data centers. The task launched one CA about Probabilistic Workload Forecasting in Cloud Computing using deep learning approaches. The CA aims to predict the future required workload of a data center to allow better resource scheduling and reduce overprovisioning. Moreover, the team worked on better simulators of data center resource usage. This is an example of an AI solution used to reduce a data center's energy consumption and emissions.

Regarding the AI approaches to enhance sustainability, we are working on electricity price forecasting and trading for battery systems. These allow the smoothing of energy production peaks of renewables, making their integration into the electric network easier.

Another project is the development of community-level road condition detection based on smartphone data. This system provides an updated map of the road condition without requiring specialised vehicles and physical inspection, reducing the impact of road condition detection and improving road safety and sustainability.

Task 3.7: Trustworthy AI as a whole. This task is related to use-cases which need all dimensions of trustworthy AI. It has been addressed jointly with the other tasks of WP3. For example, some works from the project address the particular connections existing between privacy, fairness and explainability.

Synergies with and relevance to industry

In this section we discuss how the research topics mentioned above address the challenges mentioned in the section on TDW in part I.

Industrial challenges are commonly formulated using generic terms, or generic value propositions expressed in business terms such as capacity planning, cost reduction etc. As a result, a mapping between relevant research topics and particular industrial challenges involves many-to-many relationships and contextual explanations on each individual level. To overcome this complexity we grouped the challenges and relate them to the six categories of WP3 activities.

Nukkai (https://nukk.ai/) is a French start-up that recently became a TAILOR network member. By leveraging the combined strength of symbolic AI and Machine Learning approaches, Nukkai is developing innovative AI solutions for a broad spectrum of applications that require processing incomplete data in a probabilistic universe while providing explanations to humans. Bridge is representative of what a multiagent system can be. It is a distributed game where each agent has only a partial view and has to build a strategy under uncertainty and to cooperate with another agent. By tuning the combination of AI paradigms implemented in the Nook bridge robot, NukkAI is deploying this hybrid approach to other domains such as cybersecurity, education, or transport. The problems tackled by the company fit well into Task 3.1 regarding Explainable AI Systems.

Besides, regarding the TDW listed above, the explainability and transparency principles fit with every topics, since they are both a law requirement and an important issue when dealing with self-driving cars (Future Mobility - Value of Data & Trust in AI), public sector (AI in the public Sector), data driven healthcare (AI for future healthcare), industry 4.0 (AI for Future Manufacturing), models for deepfake detection (AI Mitigating Bias & Disinformation), and predictive maintenance and optimization (AI for Future Energy & Sustainability). All these applications and domains require building reliable models that must be explained upon request of the data subjects involved, or of the final users of the system.

The study of adversarial examples is strictly related to images, and one of the most natural applications is the self-driving car. So, the Future Mobility - Value of Data & Trust in AI TDW was one of the occasions to explore a concrete application of the work in **Task 3.2**: Safety and Robustness.

The “Trusted AI: The Future of Creating Ethical & Responsible AI Systems” TDW is dealing with all the WP3 tasks and all the activities carried out within the TAILOR consortium (so fits into Task 3.7) since, as highlighted before, AI systems should cover strong assurance requirements in terms of robustness, fairness, accountability, transparency, privacy.

Here is the recap of how the TDW are related to WP3:

	Task 3.1	**Task 3.2**	**Task 3.3**	**Task 3.4**	**Task 3.5**	**Task 3.6**
Future Mobility - Value of Data & Trust in AI	X	X	-	-	-	-

AI in the public Sector	X	-	X	-	X	-
AI for future healthcare	X	-	-	-	X	-
AI for Future Manufacturing	X	-	-	-	-	-
AI Mitigating Bias & Disinformation	X	X	-	-	-	-
AI for Future Energy & Sustainability	X	-	-	-	-	X
Trusted AI: The Future of Creating Ethical & Responsible AI Systems	X	X	X	X	X	X

Synergies with and relevance to the (data) challenges

None of the proposed challenges were completely related or could be used as test-beds for the methods developed in WP3. We thus provide a description (as some kind of forward looking statement) of how these data challenges could be related to WP3 when relevant. The “Meta-learn” and “Crossword puzzle” challenges were considered too far from the considerations of WP3 and are not mentioned below.

Smarter Mobility Data Challenge. This challenge focuses on electric mobility, so the sustainability dimension is a core factor to be addressed. However, the actual studies within T3.6 are more focused on efficient server management and resource allocation, on edge computing, on efficient training of machine models, and on green design and manufacturing. Since the focus of this data challenge is on the prediction of time series, explainability may be important, for example to be able to understand the reason behind the optimisation choices suggested by the developed models. Within T3.1, we did some effort in the **explanation of time series**.

L2RPN II: Towards Carbon Neutrality. For this data challenge, the same consideration of the previous one still holds with regards to sustainability. However, as anticipated in the general description of the challenge, here the focus is more on the **robustness** than on the explainability. Indeed, agents must be robust to unexpected network events (such as both cyber-attacks or more physical threats) and maintain reliable electricity everywhere on the network, especially when the network is under stress from external events.

Brain Age Prediction EEG Challenge. In this challenge electroencephalogram (EEG) recordings are used to predict the age of individuals. The very sensitive domain implies that privacy is a key factor to be ensured since also small and hidden physiological damages or problems could be detected. **Fairness and equity** in the selection of participants and in the usage of both collected data and analysis results are important too. Finally, since the EEG recordings can be characterised by time series of EEG recording, this challenge is also related to **explainability** for the same reasons of the first two challenges.

ML for Physical Simulations. Since no actual individuals or personal data are involved in this challenge, the only ethical dimensions that need to be considered are **safety and robustness and accountability.** Indeed the domain considered in this challenge (aircraft modelling and aerodynamics in subsonic flights) could be sensitive considering these aspects.

Mind your buildings. The focus of this challenge is to identify behavioural patterns, so **privacy** plays an important role in protecting individuals' habits of participants. Being able to **explain** why a model took a decision instead of another could be relevant. Parts of the challenge focus on the reliability of missing or incomplete sensor data, so **robustness and accountability** are important aspects to be considered too.

	Task 3.1	**Task 3.2**	**Task 3.3**	**Task 3.4**	**Task 3.5**	**Task 3.6**
Smarter Mobility Data Challenge,	X	-	-	-	-	X
L2RPN II: Towards Carbon Neutrality,	-	X	-	-	-	X
MetaLearn 2022	-	-	-	-	-	-
Crossword puzzle	-	-	-	-	-	-
Brain Age Prediction	X	-	X	-	X	-
ML for Physical Simulations	-	X	-	X	-	-
Mind your buildings	X	X	-	X	X	-

Synergies with the tailor roadmap

In the recommendation of the TAILOR roadmap, there is a strong focus on how to "**measure and assess trustworthy AI systems**". This is, of course, relevant for WP3. However, some members of TAILOR are also part of an OECD.AI initiative and have proposed, in this context, a catalogue of tools and metrics for trustworthy artificial intelligence that can be found here https://oecd.ai/en/tools-report and https://oecd.ai/en/. We believe that TAILOR members should build on this initiative.

Concerning the "Scientific challenges" of the roadmap : one of the five short term scientific challenges that are identified in the SRIR is to develop "**human interpretable formalisms** to enable synergistic collaboration between humans and machines **with regards to the criteria of being explainable, safe, robust, fair, accountable**; and develop **standards and**

metrics to quantify the grade to which these criteria are satisfied.”. This point is clearly related to all tasks of WP3 and to the OECD initiative mentioned above. One of the five long term scientific challenges is to develop “the science, techniques and tools for adjustable autonomy for autonomous AI agents. In particular, equip autonomous agents with the ability to **understand when certain decisions** that it could take on its own **are questionable or unethical**, and human supervision should be required. This is one of the aims of Task 3.1.

Concerning the “Innovation” part of the roadmap: one of the two short term plans is to “consider that transparency (incl. explainability) targets different kinds of users: developers, domain experts, regulators, “users” (citizens, patients, etc.).”. This is also very much considered in Task 3.1. The long term innovation challenge is to “Implement Trust by Design: **Enable the design and verification of trusted AI systems** according to appropriate legal, social and technical criteria and aspects, focusing in particular on critical and risky applications.” This is also considered in all the tasks of WP3.

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 WP3.