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	D6.4
Work package number(s)	WP6
Date	Due M40, December 2023
Responsible Beneficiary	IST, ID #8
Author(s)	Wico Mulder, TNO
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 WP6 (Learning and Reasoning in Social Contexts) on the other side.

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

Document Review		
Reviewer	**Partner ID / 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	UNIROMA	2024-02-25

Giacomo		
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 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).

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”.

This report describes the synergy in terms of relevance and impact between on the one hand the WP2 and WP8 ‘industry, challenges and roadmap’ activities, and on the other hand the research topics of WP6.

The deliverable 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.

Part II summarises the research topics and activities regarding WP6, Social AI.

This report, D6.4 is about the synergies between the scientific work on Social AI and the industrial work, data challenges and roadmap work in TAILOR.

Introduction to the Deliverable

This deliverable is part of a group of five deliverables entitled “Synergies Industry, Challenges, Roadmap”. Each of the five TAILOR scientific work packages (WP 3,4,5,6,and 7) has a similar type of deliverable (D3.6, D4.6, D5.4, D6.4, D7.4) entitled “Synergies Industry, Challenges and Roadmap”.

Each of the 5 deliverables reflects on the relation, relevance and impact between its scientific work in the particular work package and the industrial challenges, the data challenges, and the roadmap activities in TAILOR WP 2 and WP 8.

This report, TAILOR D6.4, is one of those five deliverables. It consists of two parts.

The first part is similar in all the five deliverables. It is a common part that summarises the TAILOR industry activities, challenges and roadmap.

The second part is written from the perspective of the specific scientific work package. It describes how the research topics and scientific activities address the elements of the first part.

In order to make each deliverable self-contained, the common part is copied into each of the five deliverables. This report, D6.4, is about the activities and research in WP6.

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.

The common part was developed in joint efforts of participants of all the involved WPs. The WP-specific parts were developed within each WP and reviewed by the participants of the other WPs.

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

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.

TAILOR Data Challenges

Within the context of the TAILOR project, computational competitions (‘challenges’) were organised aiming to tackle techniques, foster collaboration and address issues related to trustworthy AI. In order to overcome the ambiguity related to the word ‘challenges’ , we refer to these activities as ‘Tailor-data challenges’.

TAILOR scientists have 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:

Smarter Mobility Data Challenge, EDF + Manifest AI (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.

This challenge was run on Codalab. 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.

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

The “Learning to run a power network challenge 2022” is concerned with AI for smart grids, and it has been built by RTE, the French Power Grid operator, and the TAU team, 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 Reinforcement Learning, that performed worse. 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.

MetaLearn 2022, Inria partner et al., TAILOR-sponsored prizes (Summer 2022)

Meta-learning from learning curves is an important yet often neglected research area in the Machine Learning community. We introduce a series of Reinforcement Learning-based meta-learning challenges, in which an agent searches for the best suited algorithm for a given dataset, based on feedback of learning curves from the environment. Agents interact back and forth with an “environment”, similarly to a Reinforcement Learning setting. Meta-learning aims to leverage the experience from previous tasks to solve new tasks using only little training data, train faster and/or get better performance.

There were two challenge cases:

- Learning from Learning Curves - perf. w.r.t. dataset size (AutoML)

In this challenge series, we want to search for agents with high “any-time learning” capacities, which means the ability to have good performances if they were to be stopped at any point in time. Hence, the agent is evaluated by the Area under the agents’ 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).

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

The goal is to meta-learn a good model that can 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). We carve such tasks from various “mother datasets” selected from diverse domains, such as healthcare, ecology, biology, manufacturing, and others. By using mother datasets from these practical domains, we aim to maximise the humanitarian and societal impact.

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 winner is 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.

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.

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)

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 will be 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 will have access for their simulations to the LIPS (Learning Industrial Physical Simulation) platform described in the NeurIPS (dataset and benchmarks track) paper. The task will be to build surrogate models of these 5 fields for new airfoils, including Out-ot-Distribution cases. The evaluation will be 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 there are already 114 participants and 190 submissions!

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 triggered to use more advanced approaches on data modelling, and are now 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 the building itself. 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 price 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[1], 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.

[1] 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 WP 6 with Industry, Challenges, and Roadmap in TAILOR

This part *describes how the research activities in WP6 address the topics of part I, i.e. the challenges in industry (discussed in the Theme Development Workshops), the* TAILOR Data Challenges *and the TAILOR Roadmap.*

About WP6

WP6 is about Social AI, the field that focuses on the fundamental question: *How do we build agents that are able to communicate, cooperate and make decisions with other agents in a hybrid population?* Social AI studies the foundations, techniques and algorithms that allow autonomous AI agents to interact, negotiate, learn, and act in societies.

AI has become increasingly more present in our daily lives, especially during the last year, where Generative AI has been dominating the news and invades our lives. Such recent developments of AI, on the one hand, lead to a technological evolution resulting in more advanced applications, and, on the other hand, a change from isolated tool-based interaction towards more complex and collaborative role-based forms of interaction, where AI can become a real “partner”. This allows for a reiterated vision on the design and development of AI based systems, one that regards components of an AI system to be designed and realised in the form of agents that act in social context. “Agents” are designed to act in a society, interacting with other agents as well as with humans. WP 6 studies the mechanisms that allow such agents to cooperate, negotiate and learn from others (agents or humans).

Home assistant chatbots, self-driving cars, drones, and automated negotiation systems are among the various autonomous artificial agents integrated into our society. These agents streamline numerous tasks, conserving time and human labour. Yet, their integration into social contexts prompts a call for deeper comprehension of their interaction and on the impact they generate at a social level.

Agents should not reason, learn and act in isolation. They will need to do it with others and among others. It is therefore important to explore the foundation of social intelligence and social behaviour, on how AI systems should communicate, negotiate and reach agreements with other AI systems and humans within a multi-agent system.

So, it is of paramount that “trust”and “trust relationships” in such human-agent networks shall emerge as a result of their interaction, interpretations of their observations, or by explicitly given explanations.

Research topics

To address these challenges, the research activities in WP6 are grouped into four coherent topic lines:

- Modelling social cognition, collaboration and teamwork (Task 6.1)
- Theoretical models for cooperation between agents (Task 6.2)
- Learning from others (Task 6.3)

- Emergent Behaviour, agent societies and social networks (Task 6.4)

Below, we briefly mention the key topics for each of them and then relate these topics to the items we mentioned in part I concerning the synergies established in the TAILOR network.

Modelling social cognition, collaboration and teamwork (Task 6.1)

In order to make agents act in social settings and be able to interact with other agents in a social manner, we need to model an agent's cognitive and social capabilities. This entails integrating an individual's knowledge and behaviour with knowledge shared among various other agents, which can be acquired at different times and from diverse perspectives.

We, humans, are able to integrate the knowledge and suggestions from others depending on how trustworthy we consider the others. In that sense, humans base their decisions on the recommendations of others, often seeking explanations and motivations behind such suggestions, as well as the context and rationale. Building trust with an agent is essential in this interaction, as it directly impacts the user's likelihood of accepting the recommendations.

So, as we build on this emergence of trust from others, we need to consider the differences that are inherent to groups and seek to achieve the best, most trustworthy group of agents. Team formation algorithms enable the dynamic assembly of teams to address tasks requested over time. The process of team formation involves decision-making on diversity, skills, cognitive abilities and personality traits. When the team is formed, the same mechanisms hold for maintaining the team in its position, its dynamics, the preferences of their members, and the perceptions of one team member towards another. Our research attention is on the mechanisms that explain the suggestions for team formation.

Whereas the research on explainable AI is commonly addressed at the level of individual systems, we have addressed that topic in the context of multi-agent systems. We do so in combination with our research on real-time team formation. We have developed a mechanism to use contrastive explanations that motivate choices for team configuration. But in order to be able to reason about others, our agents must be endowed with the capabilities of forming mental models about others: Theory of Mind Mechanisms. In TAILOR, we have combined abductive reasoning with the concept of Theory of Mind (this concept and models is also explored in T6.2). Agents can use other agents as "sensors" with the purpose of being in a more informed position when it comes to their own decision making. We develop domain-independent models that describe mechanisms on how agents should observe the actions of others, adopt their perspective and generate explanations that justify their choice of action. The mechanism allows the agents to use the capacity to comprehend others, engage in reasoning about them. It allows an agent to perceive the state of the system through the eyes of the other, and infer the beliefs that account for their most recent action.

We studied the mechanisms in game-theoretic settings and applied the results for example in the formation of robot teams for search and rescue missions.

Normative systems and value alignment mechanisms ensure that the interactions at the heart of a society of agents are ethically appropriate. We make use of social simulations to investigate the use of prescriptive norms, also referred to as social laws. Prescriptive norms

provide guidance on the behaviour of agents. They consist of regulations, constraints and directives on the behaviour of agents, possibly accompanied by monitoring and sanctioning provisions for detected violations.

The purpose of prescriptive norms is to ensure conflict-free, coordinated operation of a team of agents. Contemporary solutions to achieving coordination through rules include online design that modifies and refines the norms in place at run-time in an open multi-agent system, and guarantees on the evolutionary stability of the resulting normative system.

We have developed a framework that allows communities of agents to perform what-if analysis on a given rule configuration and steer the group of agents towards more desirable end states. One of the applied fields in which we apply this work is in social engineering, e.g. on the challenge of poverty and equal opportunities for humans in our society.

Theoretical models for cooperation between agents (Task 6.2)

Collaborative decision making in groups of agents is a broad topic that has been addressed since the dawn discipline of multi agent systems. In the context of social AI it comes with a reasonable set of new challenges. We develop and employ economic frameworks to explore and enhance the principles, methodologies, algorithms, and instruments involved in collaborative decision-making among social agents.

Division of tasks

We model and elicit individual preferences and aggregate and mediate those preferences of multiple stakeholders in a fair manner. Fair division of tasks involves partitioning or allocating a set of resources to a set of agents each having diverse and heterogeneous preferences over these resources. It has been the focus of applied research in economics, mathematics and computer science for years. Contemporary research is on the fair division of indivisible resources. Common approaches are based on interpretations of proportionality, randomization and envy-freeness (i.e. when no agent believes that another agent was given a better bundle of resources). Some examples of fair division in the real world include course allocation at schools and in websites that facilitate humans in splitting rental costs or taxi’s fares.

Theory of mind

In accordance with the concept of Theory of Mind (ToM) we extend contemporary information-sharing with mechanisms that use the knowledge about the state of others. In order to interact within a team it is important to be able to understand each others' mental states and how those states might influence the actions of the other. Rather than reasoning only with one's own beliefs, desires, intentions, emotions, and thoughts, a person or agent with the awareness of others' states of mind can consider different and mindful acts, depending on a perceived context. In order to be able to do so, it requires the entities to use a form of social interaction, i.e. some form of communication with each other and learn from observations through that communication.

We also recognize limitations of ToM, as human reasoning is inherently imperfect. The design of artificial systems that interact with humans must consider the human perspective. Should agents prioritise qualities such as honesty, impartiality, and transparency in their

reasoning and decision-making processes when interacting with humans? Striking the right balance between ethical considerations and the functional aspects of AI systems is a pressing concern.

ToM mechanisms are expected to improve efficient decision-making, resource allocation, and coordination in complex, dynamic environments across a wide range of applications. Typical scenarios where ToM can be of value involve negotiations or where one has to deal with incomplete information. For example in Autonomous Driving Situations systems where multiple vehicles have to take into account the behaviour of others sharing the road. On a higher level, multi-agent systems could also negotiate on routes to minimise travel time or energy consumption while considering real-time traffic conditions and environmental factors.

In the field of energy-related negotiations, multi-agent systems can be employed to optimise energy distribution, consumption, and resource allocation. Negotiation might better ensure availability of energy and minimises wastage. One of our industrial use cases is about buildings, controlled by agents that team up with human actors to minimise their energy consumption, while maintaining the required levels of comfort.

In warehouse automation, multi-agent systems are used to optimise logistics and coordinate the actions of robots. Agents and humans work together to manage inventory levels, ensuring that products are restocked when needed and that warehouse space is efficiently utilised. In healthcare one can think of collaboration between healthcare professionals and AI agents in advisory roles. For example in decision support when it comes to tracing tumours, personalised medicine or treatment plans.

Cooperative game theory

Our study is on the underlying mechanisms and computational models of team formation and cooperation. Enforcing agents to act in a social context comes with the insight that they cannot be regarded purely from a computational view, based on altruistic agents using perfect calculating mechanisms. Instead, like humans, decision making in a social context requires agents to be self-interested. They have preferences, goals, and desires and will act to bring those attributes as best as they can. We study these mechanisms through the field of cooperative game theory.

We regard incentives to encourage self-interested agents to faithfully carry out their tasks. Automated mechanism design involves designing the rules of a mechanism so that a desirable outcome is reached, despite the fact that agents have self interests. An example area is voting.

Applications of this work can be found in the domains of communication networks, electricity grids, and auctions. Communication networks increasingly rely on wireless devices acting as nodes coalescing together to allow for the routing of packets on an adhoc basis. Strategic interactions between them allow the network to maintain communication in an efficient and autonomous manner as well as to defend itself against attackers. In electricity grids, cooperative game theory is applied in simulations and planning, where coalition making is used to best achieve the utilisation of distributed and renewable energy resources and foster reliability and resilience. Auctions are among the most important economic mechanisms used to allocate goods and procure services. It is difficult to find good mechanisms for so-called combinatorial settings, i.e. where a bidders valuation for an item may depend on

which other items are being received. Other examples of application areas in real life settings are electronic marketplaces, virtual organisations, policy making, surveillance networks and multi-agent task-allocation scenarios.

Learning from others (Task 6.3)

How can agents in a social context be efficiently guided in their joint learning process? The scenario of a single learning agent being guided by other agents or humans has undergone extensive examination in recent years. This guidance can take various forms, such as learning from demonstrations, receiving advice, or engaging in imitation learning. We address the key question of who should learn from whom, and what should be learned.

Interactive learning

Interaction enables agents to learn from each other. A system is said to learn from experience with respect to some class of tasks and a performance measure, if its performance at those tasks, as measured by that performance measure, improves with that experience. Typical research questions we address are :How to enable meaningful interaction and control among humans and agents (hybrid human-agent decision-making)? How to create awareness of each others' intentions? and How to detect each others' intentions from monitoring each others' behaviour? We study various forms of federated learning, where agents collaborate to learn a joint model while safeguarding the privacy of their individual data.

Existing approaches primarily focus on efficiently identifying the most qualified expert, but they tend to struggle when experts are either unqualified or exhibit consistent biases. This can potentially hinder the decision-making process. We propose a novel algorithmic method based on contextual multi-armed bandit problems to detect and rectify such biassed expertise. We investigate various expert group compositions (homogeneous, heterogeneous, and polarised) and demonstrate that this approach effectively leverages collective expertise, surpassing state-of-the-art methods, particularly in cases where the quality of provided expertise is subpar.

Multi Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) empowers us to develop adaptive agents that thrive in demanding environments, even when their observations are limited. Contemporary MARL techniques have concentrated on identifying factorised value functions, which have proven successful but often result in complex network structures. In contrast, we adopt a different approach by leveraging the structure of independent learners. Our algorithm employs a duelling architecture to represent decentralised policies as distinct individual advantage functions relative to a centralised critic, which is subsequently discarded after training. This critic serves as a stabilising agent, coordinating learning and formulating learning targets.

Industrial robots tasked with assembling customised products in small batches often require extensive reprogramming. Our work aims to simplify programming complexity by autonomously identifying efficient assembly plans. Initially, a digital twin of the robot uses a simulation environment to learn which assembly techniques (programmed through demonstration) are physically viable (i.e., without collisions). Experimental results confirm

that the system consistently converges to the swiftest assembly plans. Furthermore, pre-training in simulation significantly reduces the number of interactions needed before convergence compared to direct learning on the physical robot. This two-step process empowers the robot to independently discover accurate and rapid assembly sequences, without the need for additional human input or the risk of producing faulty products.

Emergent Behaviour, agent societies and social networks (Task 6.4)

AI is making its impact on various types of multi-agent systems in industrial fields, e.g. in logistics, urban settings and in cyber-physical systems. These fields come with challenges regarding understanding, control and maintenance, as they are characterised by large populations of both natural and artificial entities.

We focus on the societal level and explore the principles, methodologies, algorithms, and tools for modelling and crafting social structures, organisations, and institutions. We study various approaches to modelling social AI systems, including self-organisation, evolutionary game theory paradigms, and agent-based simulations.

Swarm robotics

Swarm robotics involves the development, assembly, and deployment of extensive robot collectives that collaborate to address challenges or complete tasks. It draws inspiration from self-organising phenomena observed in nature, such as the individual behaviour of social insects, schools of fish, or flocks of birds, which exhibit collective behaviour emerging from simple local interactions. Typically, the research field of swarm robotics extracts engineering principles from these natural systems to equip multi-robot systems with comparable capabilities. The goal of some of the work here reported was to create systems (swarms) that outperform single robots in terms of resilience, fault tolerance, and adaptability to environmental changes. Swarms are being applied across diverse domains, including rescue missions, environmental monitoring, logistics optimization, and traffic management in urban settings.

The interaction between a human and a robotic swarm opens completely new avenues, and trust, again plays a fundamental role. The main difficulty is given by the fact that, with the swarm being self-organised, there is no clear entity with which a human could establish communication. In a social environment human–swarm interaction provides the swarm with information about goals to be achieved or tasks to be performed. Swarms can be controlled indirectly by means of a few user-driven robots embedded within the swarm. Direct control of a swarm by a user is complicated by the fact that understanding what the swarm is doing can be very challenging due to the multitude of interactions happening within the swarm. This might be hard to “read” for a human observer, so therefore explainability is crucial. The design of swarm interaction solutions will require an understanding of the psychological effects induced on humans who interact with a robot swarm in order to favour interaction modalities that reduce stress and improve adoption, usability and trust.

Consensus algorithms

We studied the self-organisation of autonomous agents for collective action and have introduced a consensus problem in the context of the real-time Railway Traffic Management Problem, a challenging optimization problem in railway transportation that involves the

efficient management of train movements on a railway network while minimising delay propagation caused by unexpected events such as train breakdowns, signal failures, and other disruptions. In contrast with most of the optimization algorithms available in the literature, we proposed a decentralised approach, a consensus algorithm, in which trains are considered as intelligent agents able to self-organise and determine the best traffic management strategy without any central control.

In a different context, we studied strategies for decentralised multi-agent systems to estimate the existence of a quorum among the opinions promoted by each agent. We assume very limited agent capabilities to model minimalist agents like miniature robot swarms. We study how the ability to estimate the quorum is impacted by the motion abilities, memory and communication protocols, so as to determine under which condition the decentralised quorum sensing strategy provides the best results. In addition, we provide adaptive mechanisms to enable the identification of changing conditions by the swarms, in order to flexibly react to environmental dynamics.

Synergies with industry

Given these different areas of research and contribution to the research community, in this section we discuss how these research topics address the needs from industry mentioned in the TDW summary in part I.

Categories

Industrial needs challenges are commonly formulated in terms of generic value propositions and often expressed in business terms such as capacity planning, cost reduction etc. In order to be able to relate research topics to them properly, we grouped the industrial needs into a few categories. We then map the four WP6 topic tasks to each of these categories This allows us to relate the scientific work of to industrial topics mentioned in part I

The categorical themes, derived from the TDW overview in part I, are mentioned in the table below.

Resilience and robustness	Taking into account that algorithms are able to deal with situations that are different from the context in which they were designed or trained. Be able to continue somehow whilst recovering from an external surprise. Social AI can mitigate breaks in interaction and especially in trust, leading to better and more resilient systems. Furthermore, bringing the human in the loop in the creation of AI systems may lead to more resilient and robust systems that humans can trust.
Managing the quality of data	Data sharing is quintessential for all industrial fields. Technical data platforms have to ensure that data sources are secure, reliable, findable, and accessible (fair) in an interoperable manner for relevant parties. Industries struggle to keep pace with the dynamic nature of customer demands. The field of (social) AI may facilitate further development of European Data Spaces and allow for new types of data to be gathered.
Standardisation, verification, certification	How to empower individual AI agents to communicate with each other, collaborate, negotiate and reach agreements? How to deal with various forms of interoperability (technical, semantic, organisational, legal)?
Explainability and transparency of algorithms	Explainability is crucial for human interactions with AI systems. Whereas most Explainable AI (XAI) methods have focused primarily on algorithms there are also needs for approaches that considers both the sociotechnical aspects and organisational context when explaining AI-driven decisions.
Trusted decision making	For collaborative decision-making in cases where not everybody has complete and correct information, it is essential that each human and agent is aware of each others' points of view and understands that others

	possess mental states that might differ from one's own. The challenges and considerations revolve around the complex task of developing artificial systems that can effectively interact with humans, anticipate their behaviour, and foster trust.
Learning in federated context	How can we make agents learn from each other in a responsible and fair way, leading to more intelligent behaviour?
Responsibility	Designing, developing, and deploying AI with good intention to empower humans and operate with confidence. Responsible AI focuses on ethical principles guiding AI development and deployment, ensuring fairness, accountability, and transparency.
Education	The need for education, familiarity and adoption of AI driven approaches throughout the whole sector was expressed
Social interaction	The aim to empower our society by artificially intelligent systems becomes paramount when considering the collaboration in hybrid teams of humans and autonomous, agent based AI systems. Examples can be found in manufacturing, where multi-agent systems are involved in the planning of production tasks, maintenance tasks and order intake processes. Another example is in the domain of so-called adaptive operator support, where humans and machines (co-bots) cooperate in the production of goods. Challenges are in the field of adoption, proving the return on investment ROI in lowering production costs managing the innovation activities that rise after the first experiences with social AI systems.
Ethical and legal aspects	How can agents coordinate to fairly share common resources? How to create trustworthy hybrid human-AI societies that fulfil humans' expectations and follow their requirements?

Synergy table WP6 - WP8 Industry

Industrial Need	**Task 6.1**	**Task 6.2**	**Task 6.3**	**Task 6.4**
Resilience and Robustness.		x	x	x
Managing the quality of data	x			
Standardisation, verification, certification		x		x
Explainability and transparency of algorithms	x	x	x	x
Trusted Decision Making		x		x
Learning in federated contexts.			x	x
Responsibility	x	x		
Education	x	x	x	x
Social interaction	x	x	x	x
Ethical and legal aspects	x			

Synergies with and relevance to the (data) challenges

The table below relates the WP6 tasks with the individual data challenges mentioned in part I.

	Task 6.1	Task 6.2	Task 6.3	Task 6.4
Smarter Mobility Data Challenge. This challenge focuses on electric mobility. Capacity management and Sustainability were the theme-factors to be addressed. The focus of this data challenge is on the prediction of time series.		x	x	x
L2RPN II: Towards Carbon Neutrality. For this data challenge, main considerations were sustainability and robustness. 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.		x	x	

MetaLearn 2022. Various types of learners. Academic approaches.		x	x	x
Brain Age Prediction. In this challenge electroencephalogram (EEG) recordings are used to predict the age of individuals. Since the EEG recordings can be characterised by time series, this challenge relates to time series classification. Not much on the social AI field was addressed.				
Crossword puzzle. More on the NLP side. No interference with WP6.				
ML for Physical Simulations. Calculations in the field of fluid dynamics.				x
Mind your buildings. The focus of this challenge is to identify behavioural patterns classification of missing data and interaction with humans on finding patterns.	x	x		

Synergies with the TAILOR roadmap

Future Evolution regarding social AI.

There is no way back on the impact that AI is having on our society. A myriad of settings became the stage for AI applications, such as factories, roads, houses, hospitals and even schools. AI-powered machines must now place humans at the centre and are designed to interact with humans naturally.
And most importantly, ***AI is becoming social.*** We believe that there is now a place for a reiterated vision that regards AI situated in social contexts, and agents that are AI entities that cooperate and communicate in hybrid populations of humans and agents.

The evolution of AI entities regarding their interaction with humans can be seen from two perspectives; on the one hand we see a technological evolution resulting in more advanced individual AI entities, on the other hand we see an evolution from isolated tool-based expert systems towards interactive and collaborative systems. The latter has also led to systems in which AI entities interact with each other as well as with humans. The role of such agents is evolving from being merely task-oriented and assistive to being collaborative companions that care for each other and in some situations also for their surrounding environment.

Within WP6 we work on a maturity model that can be used to identify the maturity of AI entities in such human-AI ecosystems. In the future, we hope to be able to assess how social an AI system can be. Such models can be used to reflect on expected capacities, the role and the responsibilities of AI entities with respect to a human user, a team, and the society. It can be used in the process of planning and engineering the future AI entities that will act in a human AI ecosystem.

Preliminary TAILOR Roadmap elements

Since the due-date of this deliverable D6.4 is before the due date of the TAILOR roadmap, our reflections on the roadmap lean on a preliminary, perhaps even somehow a sketchy view, as was presented during a TAILOR meeting in November 2024. The figure below shows a preliminary view of the TAILOR Roadmap.

4 main directions that are currently being identified:

- Elsa & Governance
- TAI
- LOR
- Infrastructure

The roadmap contains many elements. In order to help with the process of structuring these elements, a generic use-case (here “decision support for lifestyle”) was chosen.

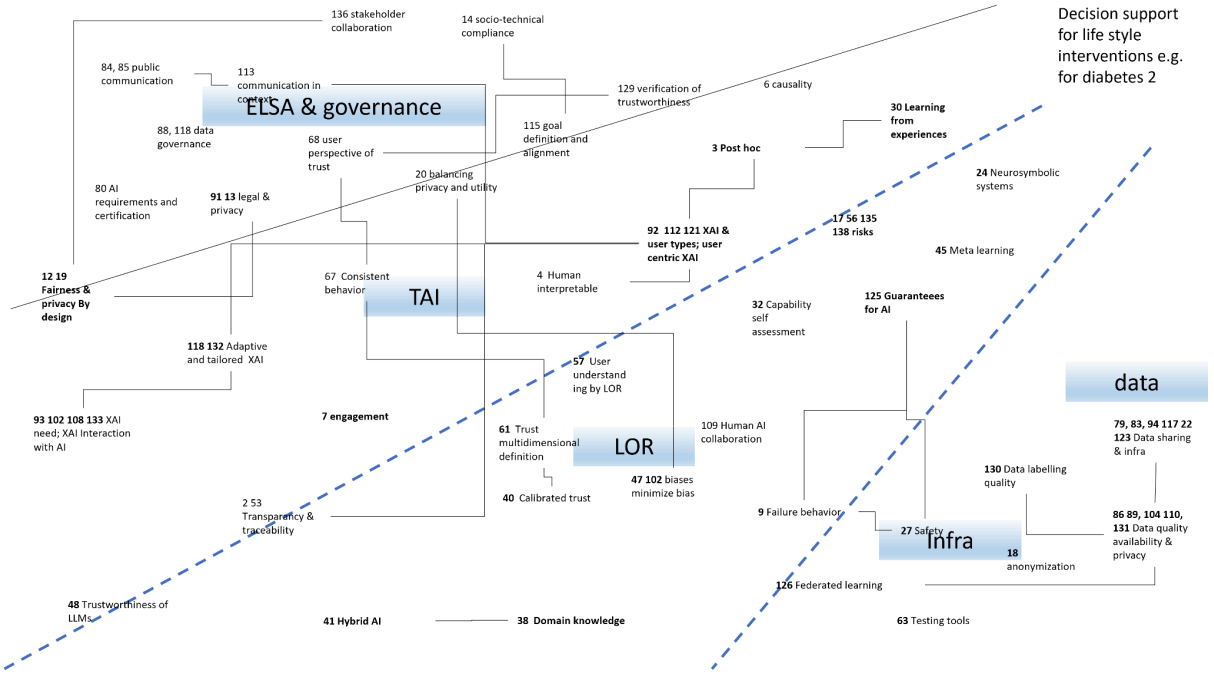

Figure 1: preliminary TAILOR roadmap overview

Synergies

WP6 synergies with the Tailor roadmap process can be expressed via a table. For each of the 4 main areas we picked a few relevant roadmap topics that are in line with the research activities in each of the four WP6 tasks.

All tasks within WP6 contribute to the future of social AI and are either directly or indirectly involved in the many topics that were sketched under the 4 main roadmap elements .

Conclusion

This deliverable has summarised the most important aspects of industrial needs, data challenges and roadmap elements of TAILOR. It has also listed the main topics that have been addressed in TAILOR WP6. The report has motivated the relevance of WP6 to the industry needs, data challenges and roadmap topics.

References

Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, Carles Sierra:Combining theory of mind and abductive reasoning in agent-oriented programming. Auton. Agents Multi Agent Syst. 37(2): 36 (2023)

Nieves Montes, Nardine Osman, Carles Sierra:A Computational Model of Ostrom's Institutional Analysis and Development Framework (Extended Abstract). IJCAI 2023: 6937-6941

Nieves Montes, Carles Sierra:Synthesis and Properties of Optimally Value-Aligned Normative Systems. J. Artif. Intell. Res. 74: 1739-1774 (2022)

Athina Georgara, Juan A. Rodríguez-Aguilar, Carles Sierra:Building Contrastive Explanations for Multi-Agent Team Formation. AAMAS 2022: 516-524

Athina Georgara, Juan A. Rodríguez-Aguilar, Carles Sierra, Ornella Mich, Raman Kazhamiakin, Alessio Palmero Aprosio, Jean-Christophe R. Pazzaglia:An Anytime Heuristic Algorithm for Allocating Many Teams to Many Tasks. AAMAS 2022: 1598-1600

Georgina Curto, Nieves Montes, Carles Sierra, Nardine Osman, Flavio Comim:A norm optimisation approach to SDGs: tackling poverty by acting on discrimination. IJCAI 2022: 5228-5235

Athina Georgara, Juan A. Rodríguez-Aguilar, Carles Sierra:Towards a Competence-Based Approach to Allocate Teams to Tasks. AAMAS 2021: 1504-1506

Georgios Amanatidis, Haris Aziz, Georgios Birmpas, Aris Filos-Ratsikas, Bo Li, Hervé Moulin, Alexandros A. Voudouris, Xiaowei Wu, Fair Division of Indivisible Goods: A Survey (2023)

Georgios Chalkiadakis , Edith Elkind , Michael Wooldridge, Computational Aspects of Cooperative Game Theory, (2012)

Mulder, W., Meyer-Vitali, A. . A Maturity Model for Collaborative Agents in Human-AI Ecosystems. In: Camarinha-Matos, L.M., Boucher, X., Ortiz, A. (eds) Collaborative Networks in Digitalization and Society 5.0. PRO-VE 2023. IFIP Advances in Information and Communication Technology, vol 688. Springer, Cham. (2023)