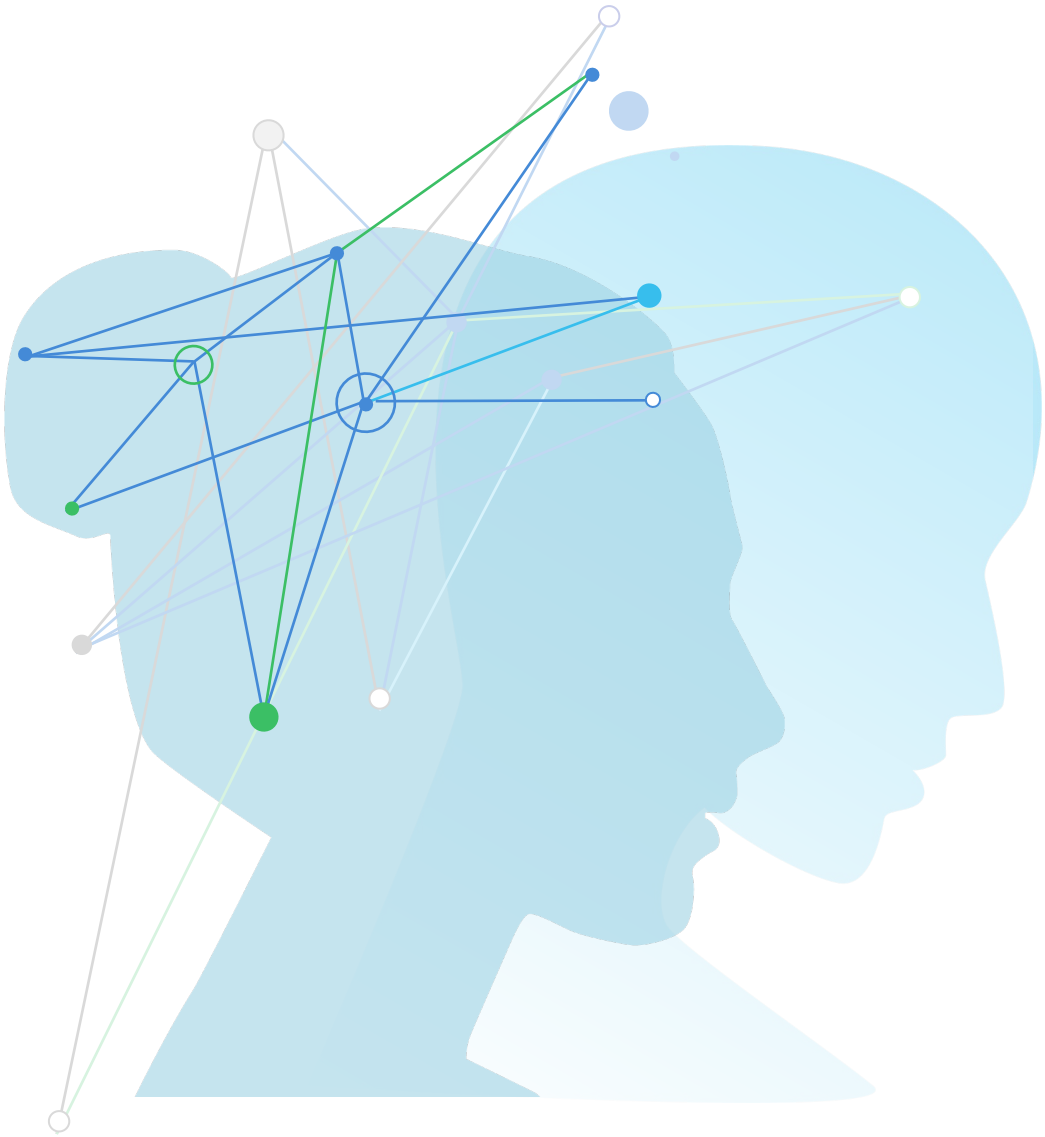
TAILOR

STRATEGIC RESEARCH & INNOVATION ROADMAP OF TRUSTWORTHY AI

The Scientific Foundations of Trustworthy AI in Europe for the Years 2022-2030



Funded by the European Union

Editors:

Michela Milano

Department of Computer Science and Engineering
Bologna University

Contact: michela.milano@unibo.it

Marc Schoenauer

INRIA Saclay
Université Paris Sud

Contact: marc.schoenauer@inria.fr

Fredrik Heintz

Department of Computer and Information Science
Linköping University

Contact: Fredrik.heintz@liu.se

Publication date: 2022-04-13

This document is an partly updated version of the TAILOR deliverable 2.1: Strategic Research and Innovation Roadmap (SRIR) for Trustworthy AI for the years 2022-2030

TAILOR is an ICT-48 Network of AI Research Excellence Centers funded by EU Horizon 2020 research and innovation programme under grant agreement No 952215

Foreword

TAILOR is a European network developing the scientific foundations for realising the European vision of human-centred Trustworthy AI. To achieve this, we strongly believe that it is necessary to combine and integrate learning, optimisation and reasoning. TAILOR is one of four EU-funded networks of AI research excellence centers. One of our most important tools is the TAILOR Strategic Research and Innovation Roadmap.

It is our ambition that the roadmap will identify major scientific research challenges and push the scientific progress towards the most pressing challenges. It is based on input from both industry and academia, especially working groups related to the five research themes of TAILOR and a series of theme development workshops organised in close collaboration with industry.

This document is the first version of the roadmap. It should be considered a living document that acts as a starting point for this important initiative. Even though it is a snapshot of the current situation, it also contains short- and long-term recommendations for the direction of future European AI research.

We hope that the roadmap will inspire funding agencies and researchers to start addressing these major research challenges so that we together can boost research on Trustworthy AI in Europe and globally.

It is our desire to engage with the broader research community to further discuss this important topic, develop the roadmap, and to support the European AI ecosystem in achieving the scientific foundations of Trustworthy AI. We therefore welcome comments, constructive criticism, and suggestions.

Fredrik Heintz

Coordinator of TAILOR and Professor of Computer Science, Linköping University, Sweden

Table of contents

Executive Summary

This document is the first version of the Strategic Research and Innovation Roadmap of the TAILOR project, focussed on Trustworthy Artificial Intelligence through Learning, Optimization and Reasoning. The project objectives are extremely ambitious, and address topics that are currently very actively investigated. Therefore, defining a roadmap is itself an ambitious goal. We have started analysing many documents containing Roadmaps and Research and Innovation agendas of AI related initiatives (in particular we have analysed the AI4EU Strategic Research and Innovation Agenda and the AI, Data and Robotics PPP Strategic Research Innovation and Deployment Agenda and the AI Watch Index 2021). Also, strategic and roadmapping documents of initiatives from connected fields (e.g., HPC, IoT, Cybersecurity) have been evaluated to find connections and synergies.

As in the Ethical Guidelines for Trustworthy Artificial Intelligence document released in 2019 by the High-Level Expert Group on AI, we need to consolidate ongoing research activities, solid foundational theories, and methodological guidelines that are not yet common in neither industry nor academia. To this end, we have consolidated input coming from scientific and innovation work packages of the TAILOR Network of Excellence, that have released impressive scientific results in one and a half year, but these results still need to be conceptualised, organised, and classified in a rationale shaping future avenues.

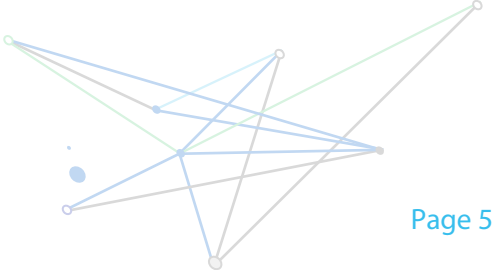
Still, in the limited time passed from the project start, the TAILOR consortium has identified interesting research directions and urgent industrial needs. Prioritisation of actions and their timing is not yet perfect, but we are confident that a clear plan will be available for the second and final version of the SRIR.

The document is organised with a short snapshot of the state of European research and innovation landscape. We then define the challenges related to the dimensions of trustworthy AI, namely explainability, safety, robustness, fairness, accountability, privacy and sustainability.

Following TAILOR work packages, learning, optimization and reasoning are considered and several aspects of their integration are analysed: unifying formalisms for integrating reasoning and learning, learning and reasoning on how to act, social perspectives, and AutoAI. A last section is devoted to Foundation models that have been gaining momentum since the TAILOR proposal was written.

In addition, the industrial and service sectors along with the public sector are considered in the report based on three Theme Development Workshops organised in the context of WP8. Important priorities are identified and gaps needing to be filled in are outlined.

Clearly, all these areas have huge gaps and research questions that need to be addressed in the short and long term. Some recommendations are proposed before the concluding section of this document.



Introduction

Artificial Intelligence has grown in the last ten years at an unprecedented pace. It has been applied to many industrial and service sectors, becoming ubiquitous in our everyday life. More and more often, AI systems are used to suggest decisions to human experts, to propose scenarios, and to provide predictions. Because these systems might influence our life and have a significant impact on the way we decide, they need to be trustworthy. How can a radiologist trust an AI system analysing medical images? How can a financial broker trust an AI system providing stock price predictions? How can a passenger trust a self-driving car?

These are fundamental questions that deserve deep analysis and intense research activity as well as a **new generation of AI talents** who are skilled in the scientific foundations of Trustworthy Artificial Intelligence, who know how to assess and how to design trustworthy systems. Some of the current issues related to lack of trust in AI systems are a direct consequence of the massive use of black-box methods relying only on data. We need to define the foundations of a **new generation of AI systems** not only relying on data-driven approaches, but also on the whole set of AI techniques, including symbolic AI methods, optimization, reasoning, and planning.

Europe is ready for such a challenge, as its research landscape is strong and well structured, even if quite fragmented. Initiatives such as TAILOR and the other Networks of Excellence on AI, recently funded by the European Commission, as well as the AI-on-demand platform aggregating all AI stakeholders represent a very good strategy to reduce fragmentation of the European scientific community around AI. In fact, AI research in Europe has a long and successful history dating back at least to the creation of the ECAI[1] archival conference in 1974. Bibliometric data[2] show that Europe ranks first worldwide in terms of published AI research papers (Europe: 170,800; China: 135,000, US: 106,600). Many of today most used AI methods and tools originated in European universities and research institutes, for example, in areas such as constraint programming (Gecode[3]), logic programming (SICtus Prolog[4]), Semantic Web (OWL, RDFS, Reasoners), LSTM[5], and Evolution Strategies[6].

The European education system is also very well positioned and structured, even though the **talent retention** rate needs to be improved through a flourishing and stimulating ecosystem.

---

[1] European Conference on Artificial Intelligence, https://www.eurai.org/activities/ECAI_conferences
[2] Data from 1998 to 2017, source: Elsevier AI Report 2018, Scopus.
[3] https://www.gecode.org, https://choco-solver.org
[4] https://sicstus.sics.se
[5] S. Hochreiter and J. Schmidhuber, Long Short-term Memory. Neural Computation 9(8):1735-1780, 1997.
[6] I. Rechenberg, Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Frommann-Holzboog Verlag, 1973.

To set the foundation of Trustworthy AI principles that adhere to European values, the European Commission (EC) selected a High-Level Expert Group (HLEG) that published the Ethical Guidelines for Trustworthy AI[7] in April 2019. Then in February 2020, the EC released a White Paper on AI[8] defining a set of features a trustworthy AI system should have. More recently the EC has proposed an AI regulation, the AI Act[9], a first-of-a-kind regulation forbidding certain uses of AI and defining high risk applications where AI systems need to be carefully assessed.

From the White paper on AI, building on the Ethics Guidelines for Trustworthy AI developed by the HLEG[10], the main features that a Trustworthy AI system should be compliant with are:

- Human agency and oversight
- Technical robustness and safety
- Privacy and Data governance
- Transparency and explainability
- Diversity, non-discrimination, and fairness
- Societal and environmental well-being
- Accountability

These features will be discussed in turn in the next section, as they represent fundamental pillars for building trust and fostering AI uptake at large scale.

The purpose of the European Project TAILOR is to build the capacity of providing the **scientific foundations for Trustworthy AI in Europe** by developing a network of research excellence centres leveraging and **combining learning, optimisation, and reasoning**. These systems are meant to provide descriptive, predictive and prescriptive systems glueing data-driven and knowledge-based components.

TAILOR is based on four powerful instruments: this roadmap, a basic research program to address grand challenges, a connectivity fund for active inclusion of the larger AI community, and network collaboration activities promoting research exchanges, training materials and events, and joint PhD supervision.

TAILOR is mainly research oriented. However, it keeps an eye on the **industrial perspective** toward trustworthy AI systems, trying to translate into technical requirements the features identified by the Ethical Guidelines and the White paper, assessing the impact of these requirements on the entire value-chain and guiding them into the new regulatory framework.

[7] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
[8] https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[9] https://artificialintelligenceact.eu/
[10] https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Society and the public sector are also considered, as AI systems could play an important role in Europe to solve or at least support the solution of important societal challenges like climate change, the pandemics, education, immigration and digital divide. Having systems that suggest decisions and provide indications in these settings inevitably requires an approach heavily based on trust.

Challenges ahead are extremely important and to face them with an effective approach we need to identify concrete research avenues and technology gaps that need to be filled to achieve a Trustworthy AI culture. The current document is a first step in this direction. In the next sections, we first outline the objectives of the Strategic Research and Innovation Roadmap and then we identify research topics that need to be investigated to improve the foundations of Trustworthy AI and to properly transfer this knowledge to the industrial and service sectors both in private and public bodies.

The process: Writing this Strategic Research and Innovation Roadmap is the purpose of Task 2.2, and UNIBO, as Task leader, and Inria, as WP leader, were the main contributors. Several other TAILOR members also participated. A Roadmap Editorial Board (REB[11]) was created (the goal of Task 2.1). Furthermore, an Extended Roadmap Editorial Board (EREB[12]) was set up, including the leaders of all research and innovation WPs (3-8), plus the leaders of the tasks dedicated to contributions to the roadmap (as well as links with industry) present in these WPs.

The process started with a questionnaire, sent to the EREB, but the contributions received were too scarce to build a complete roadmap. A more interactive process was then run in each of the scientific WPs (WPs 3-7), during one of the workshops each one of them involving partners of the scientific WP. In addition, as planned in the original plan, outputs were collected from the Theme Development Workshops organised within WP8. This document is a synthesis of these different contributions, consolidated by the Task leader, and by further contributions of the whole EREB.

[11] Please see REB members in Appendix 1.
[12] Please see EREB members in Appendix 1.

Objective 1

Providing guidelines for strengthening and enlarging the pan-European network of research excellence centres on the Foundations of Trustworthy AI

Objectives of the TAILOR SRIR

Objective 2

Defining paths for advancing the scientific foundations for Trustworthy AI and translating them into technical requirements to be adopted broadly by industry

Objective 3

Identifying directions for fostering collaborations between academic, industrial, governmental, and community stakeholders on the Foundations of Trustworthy AI

Objectives

DEFINING SCIENTIFIC RESEARCH DIRECTIONS IN TRUSTWORTHY AI, ENLARGING AND STRENGTHENING THE COMMUNITY AND FOSTERING COLLABORATIONS BETWEEN ALL STAKEHOLDERS ARE THE CORE OBJECTIVES OF THE STRATEGIC RESEARCH AND INNOVATION ROADMAP.

In this section, we state the objectives of this Strategic Research and Innovation Roadmap (SRIR) that are built upon the objectives of the TAILOR project. The Strategic Research and Innovation Roadmap not only shapes the objectives of the project but goes beyond the project duration with the purpose of influencing and providing guidelines for the European Commission in their future work programmes. The roadmap should define the foundations of Trustworthy AI for the **years 2022-2030**. Seeded with existing Roadmaps/Agendas worldwide and coordinated with the VISION roadmap, the TAILOR roadmap will actively pursue alignment with European Commission Work Programmes and EU-wide relevant initiatives such as the AI Watch[13], the European AI-on-demand platform[14], the AI Alliance[15], and the AI, Data and Robotics PPP[16].

The SRIR is strongly focusing on AI research covering both curiosity-driven and application-driven research topics. For curiosity-driven research the roadmap will target grand challenges with long-term impact to ensure excellent research and help train the best AI talents in Europe. For application-driven research, the future avenues will be identified by combining extensive requirements collection from vertical domains, with horizontal cross-pollination and leverage, and its transfer to industry. In both cases, the TAILOR roadmap will be coordinated with the CSA and the other networks.

In the following we define the three main objectives of the current SRIR:

[13] https://knowledge4policy.ec.europa.eu/ai-watch_en
[14] https://www.ai4europe.eu/
[15] https://futurium.ec.europa.eu/en/european-ai-alliance
[16] https://ai-data-robotics-partnership.eu/

1 Providing guidelines for strengthening and enlarging the pan-European network of research excellence centres on the Foundations of Trustworthy AI

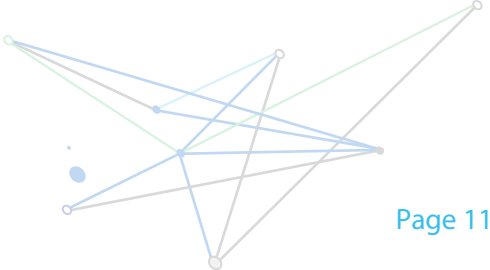
While the TAILOR consortium has built a fully functional network of centres of research excellence on the foundations of Trustworthy AI covering all of Europe, defining the governance structure of the network, leveraging existing expertise and platforms ecosystem, the SRIR aims at further strengthening and enlarging this network and pushing the research frontier on Trustworthy AI, during the network lifetime and, more importantly, after it has ended.

2 Defining paths for advancing the scientific foundations for Trustworthy AI and translating them into technical requirements to be adopted broadly by industry.

The objective of the TAILOR project is to enable faster research progress by providing a common resource for Europe and the world with easy access to state-of-the-art knowledge and expertise in the foundations of Trustworthy AI. The objective of the SRIR is to strengthen the foundation of Trustworthy AI and guide the transfer of these principles to practical applications. In particular, translating high level principles defined in the Ethical guidelines of Trustworthy AI to technical requirements is crucial for research and innovation. In addition, assessment methodologies for measuring the compliance to trustworthy AI principles of AI systems merging learning, optimization and reasoning are of paramount importance for auditing purposes.

3 Identifying directions for fostering collaborations between academic, industrial, governmental, and community stakeholders on the Foundations of Trustworthy AI

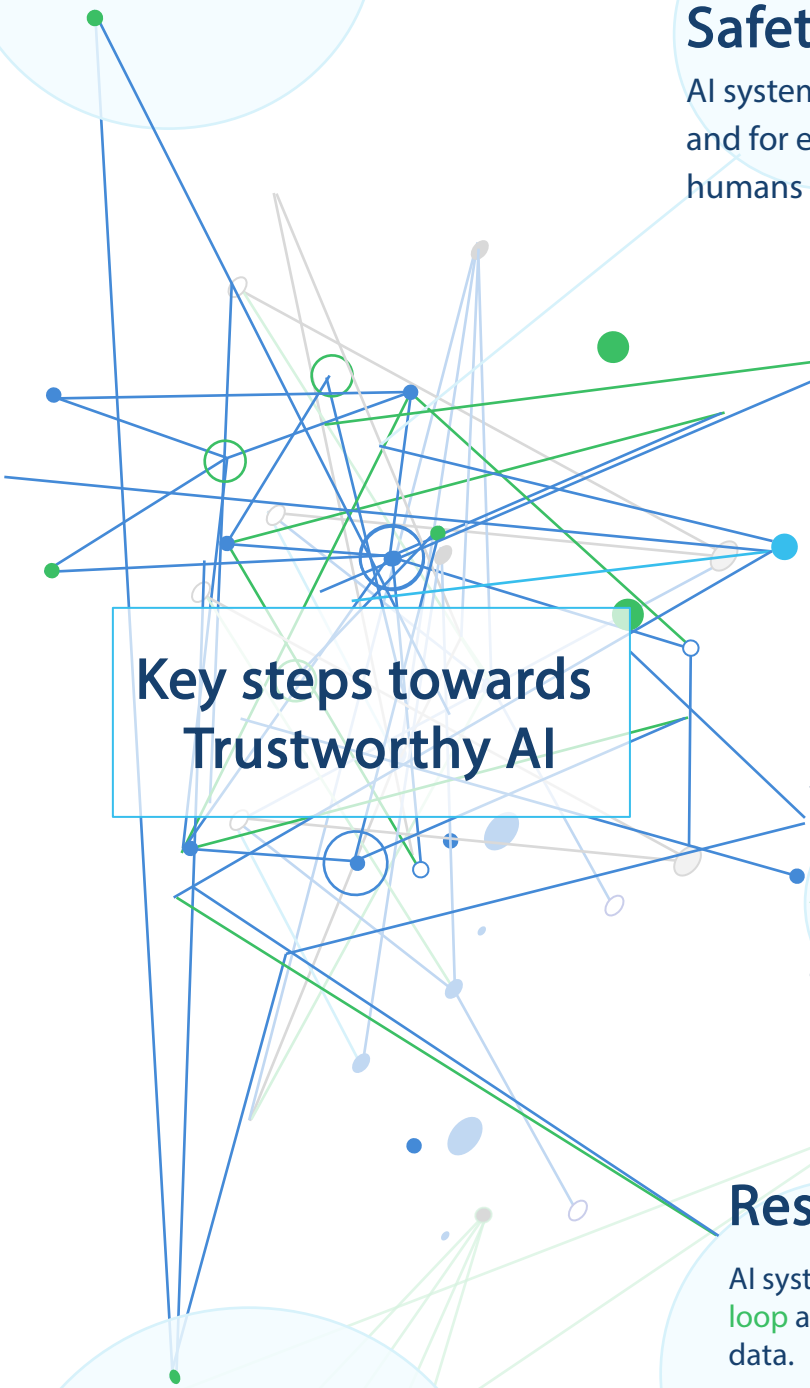
The objective of the TAILOR project is to develop an Innovation and Transfer Program to develop synergies and cross-fertilisation between industry and the TAILOR network as well as foster innovation and exploit new ideas. The objective of the SRIR is to strengthen collaboration between academics, industrials, public bodies, and citizens to create an ecosystem of stakeholders on Trustworthy AI. Stakeholders can take on several roles, sometimes in conflict of interest: as researchers, educators, developers, service and data providers, (legal) consultants, problem owners, regulators, trusted validation parties, corporate and individual private end-users. The connections with the future AI-on-demand platform will serve as a glue for strengthening these collaborations.



Detailed priorities, actions, and timing needs further discussion within and outside the TAILOR project and will be detailed in the second version of the SRIR. For the time being, both short- and long-term recommendations are proposed.

AI systems are more and more often used in critical sectors, to support the decision-making process, to provide accurate predictions, and evaluate alternative scenarios. It is therefore crucial that in high-risk applications, as outlined in the AI Act, AI systems exhibit features that make them trustworthy. Trust indeed is a more complex concept. Trust can be conceptualised as "a multidimensional psychological attitude involving beliefs and expectations by a trustor about a trustee, derived from experience and interactions with that trustee in situations involving uncertainty and risk"[17]. This commonly agreed conceptualization of trust, coming from human-human and human-machine literature, considers several ingredients of trust: beliefs about the trustee's capabilities; expectations; and some degree of risk associated with the possibility that the expectations will not be met[18].

[17] Lewis, Michael, Katia Sycara, and Phillip Walker. "The role of trust in human-robot interaction." Foundations of trusted autonomy. Springer, Cham, 2018. 135-159
[18] Falcone, R., & Castelfranchi, C. (2001). Social trust: A cognitive approach. In Trust and deception in virtual societies (pp. 55-90). Springer, Dordrecht.



Key steps towards Trustworthy AI

Explainable AI systems

An AI system should be able to explain its results and justify its decisions

Safety and Robustness

AI systems should be safe for humans, and for everything that is valuable to humans

Fairness, equity, and justice

Fairness should be embedded by design in all AI systems

Accountability and Reproducibility

Accountable and reproducable AI systems in design, development, and deployment

Respect for privacy

AI systems with humans in the loop and in control of their own data.

Sustainability

Large data storage and computing infrastrucftures are needed to deliver good AI solutions for the planet, and AI solutions with reduced energy consumption.

Trustworthy AI

SEVERAL DIMENSIONS CONCUR TO CREATE A TRUSTWORTHY AI SYSTEM, LIKE THE CAPABILITY OF BEING EXPLAINABLE, SAFE AND ROBUST, ABLE TO PROMOTE FAIRNESS, EQUITY AND JUSTICE, ACCOUNTABLE AND REPRODUCIBLE, RESPECTFUL FOR PRIVACY, AND SUSTAINABLE. FOR INDUSTRY, IT IS ESSENTIAL TO MEASURE THEIR DIMENSION AND TRANSLATE THEM IN TECHNICAL REQUIREMENTS.

AI systems are more and more often used in critical sectors, to support the decision-making process, to provide accurate predictions, and evaluate alternative scenarios. It is therefore crucial that in high-risk applications, as outlined in the AI Act, AI systems exhibit features that make them trustworthy. Trust indeed is a more complex concept. Trust can be conceptualised as "a multidimensional psychological attitude involving beliefs and expectations by a trustor about a trustee, derived from experience and interactions with that trustee in situations involving uncertainty and risk"[19]. This commonly agreed conceptualization of trust, coming from human-human and human-machine literature, considers several ingredients of trust: beliefs about the trustee's capabilities; expectations; and some degree of risk associated with the possibility that the expectations will not be met[20].

Even if trust is a complex psychological attitude, and often not rational, it is important to identify clear indications for AI system developers to try to achieve trustworthiness. There are several dimensions that concur to create a trustworthy AI system, like the capability of being explainable, safe and robust, able to promote fairness, equity and justice, accountable and reproducible, respectful for privacy, and sustainable.

The combination of all these dimensions, together with research directions for supporting them, is a long-term research objective and is also likely to cope with properties and tensions among conflicting goals (e.g, accuracy vs. fairness).

[19] Lewis, Michael, Katia Sycara, and Phillip Walker. "The role of trust in human-robot interaction." Foundations of trusted autonomy. Springer, Cham, 2018. 135-159
[20] Falcone, R., & Castelfranchi, C. (2001). Social trust: A cognitive approach. In Trust and deception in virtual societies (pp. 55-90). Springer, Dordrecht.

For industry, it is essential to understand how these dimensions translate in practice and boil down to technical requirements. There is a need for each dimension to create methodologies for:

- Assessing if an existing AI system is compliant with the guidelines
- Repairing it in case it is not
- Designing a new AI system compliant with the guidelines.

In the following we dive into several directions, and we outline the main research directions that need investigation and also impactful areas for the industrial and service sector. These research directions and areas have been collected by (1) interacting with the scientific work packages of TAILOR and (2) consolidating the input derived by the SRIR workshop and work package meetings. They will now be discussed in turn at a general level in the rest of this chapter, while the following chapter will detail the specifics of each scientific Work Package.

Explainable AI systems

Explainability in AI systems concerns the capability of a system to explain its results, to justify its decisions and to bring evidence about the choices made and to debug it to understand when, where and why a mistake was made. This aspect is exacerbated by the intense development of deep neural networks that are black boxes providing no human-understandable clue about their results.

Explainable-AI explores and investigates methods to produce or complement AI models to make accessible and interpretable the internal logic and the outcome of the algorithms, making such processes understandable by humans.

In this field, it is important to push forward the research, for example by proposing new Explainability methods in both directions:

- **Transparent-by-design**: AI tools, methods and processes that are explainable on their own, following a transparent by design approach also capable of incorporating existing background knowledge ;
- **Post-hoc explanations** that given an opaque ML model (black box) aims to reconstruct its logic either by mimicking the opaque model with a transparent one (global approaches[21]) or by concentrating on the construction of a useful explanation (feature relevance, factual and counterfactual) for a specific instance (local)[22].

[21] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, GLocalX - From Local to Global Explanations of Black Box AI Models, Artificial Intelligence, Volume 294 (2021).
[22] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and Counterfactual Explanations for Black Box Decision Making, IEEE Intell. Syst. 34(6): 14-23 (2019)

An important aspect concerns the trade-off between accuracy and interpretability, and the ambitious challenge to propose innovative models that strive to achieve both.

In addition a number of fundamental challenges are still open , such as:

- Human interpretable formalisms to habilitate synergistic collaboration between humans and machine, capable to express high-level explanations (logical, causal, knowledge graph) for encoding domain knowledge, and/or taking into account causal relationships in the data and/or identified by learning models;
- methods for generating multimodal explanations (cross-modal/cross-language, factual and counterfactual etc.). standards and metrics to quantify the grade of comprehensibility of an explanation for humans (e.g., Fidelity, Stability, Minimality, Plausibility, Faithfulness, Actionability) . Those standards need to take into account the research results from the HCI, DataVis, and Cognitive Sciences communities.
- benchmarking platforms (datasets, metrics and methods etc.) for creating a common ground for researchers on explanation from different domains.

Last but not least, an important aspect of explainability has to do with causality. Supervised learning techniques today only learn correlations, whereas causality is necessary when it comes to decisions. In many application domains, causal links are implicit, known from past scientific corpus or simply common sense. However, when it is not the case, being able to learn causal links from data can become crucial, and add a layer of explainability to the learned model: in health, finance, environments for instance. Several approaches have been proposed, and their main limitations are the scale-up to thousands of variables, and the detection of hidden confounders, that hinder the identification of true causal dependencies. Note that causality is also important when it comes to fairness and accountability.

Beside the above mentioned research topics, that are fundamental cornerstones to be addressed by the research community, we have identified open areas that are crucial for the industrial uptake of trustworthy AI. These are of course also research areas, but they are driven by applications.

One important challenge concerns the development of AI systems aimed at empowering and engaging people, across multiple scientific disciplines and industry sectors. Beyond the specific challenges that each discipline or application generates, a general problem requires our attention, i.e., finding a right trade-off in the provided explanations.

Indeed, in multiple practical decision making scenarios, human-machine collaboration and argumentation is needed, with humans keeping the responsibility for the decisions, but relying on machine aids. A human expert is more likely to rely on AI systems when she (or someone we can trust, somewhere) understands the reasons for the behaviour observed or the decision suggested. Even in the extreme case of statistical validation, there should exist some logical and rational hints that support the statistics.

Essentially, the explanation problem for a decision support system can be understood as "where" to place boundaries between the algorithmic details to be delivered. We must define what details the decision maker can safely ignore and, on the contrary, what meaningful information the decision maker should absolutely know to make an informed decision. Therefore, the explanation is intertwined with trustworthiness (what to safely ignore), comprehensibility (meaningfulness of the explanations), and accountability (humans keeping the ultimate responsibility for the decision).

The challenge is hard, as explanations should be sound and complete in statistical and causal terms, and yet should be able to adapt the level of explanations to all the involved stakeholders, such as the users subject to decisions, the developers of the AI system, researchers, data scientists and policymakers, authorities and auditors, etc.

Safety and Robustness

AI systems should be conceived and engineered to be safe for humans, and for everything that is valuable to humans, with their cultural biases. They should also be robust against perturbations, varying contexts and malicious attacks. In safety critical domains, these features are of paramount importance and need to be addressed with special care[23]. In particular, as AI systems become more complex, in order to achieve safety and robustness, we need to re-understand their evaluation to

- verify and validate a system under acceptable assumptions whenever possible (verifiability);
- precisely assess *how often* and *how much* the system may fail (calibration) and *when* (capability profiling, context-dependent evaluation)[24]. This is particularly relevant in safety-critical AI systems, such as those appearing in automotive and avionics.

The technical foundations and assumptions on which traditional safety engineering principles are based are inadequate to ensure safety and robustness of systems in which AI/ML algorithms are interacting with people and the environment at increasingly higher levels of autonomy, even more so in case of continuous online/real-time adaptation, subject to concept drift. Specifically:

- The perspective from AI/ML evaluation has focused on performance on specific benchmarks and distributions, but not on safety or robustness, originating problems such as adversarial attacks or data/concept shift;
- We need to reinforce the emergent links from safety engineering, formal methods and verification to the way AI/ML systems are conceived and evaluated.

Also in this setting, the TAILOR consortium has identified areas that might be important for the industrial and service sector. All stakeholders in AI (users, industry, governments) will not put a

[23] J. Burden, J. Hernàndez-Orallo, S. hÉigeartaigh, Negative Side Effects and AI Agent Indicators: Experiments in SafeLife, SafeAI@AAAI (2021)
[24] D. Hicks, Lessons from Philosophy of Science, IEEE Technology and Society Magazine (2018)

system in operation (or will remove it soon after use) if they do not trust its behaviour in terms of safety and robustness. This is a principle that holds for every engineering discipline, for every technology, and very much so for AI. Even if the benefits compensate for the risks, any safety backlash (e.g., an accident) will have an important effect on the penetration of the technology and on the reputation of companies using AI.

There has been significant involvement of industry in some activities for which the TAILOR network has been associated, such as the significant participation of papers and speakers from industry in the SafeAI@AAAI and AISafety@IJCAI workshops. There is also an important activity from industry in the debate about regulation and certification of AI systems, especially after the new drafts on AI regulation from the EU. There seems to be independent entities to certify the capabilities, safety and robustness of AI systems, and even the creation of evaluation sites (e.g., for self-driving cars, for drones, etc.). The evaluation of AI systems goes much beyond the research-oriented measurement and testing of scientific papers, but has to consider a context-oriented, user-oriented, on-the-ground evaluation in real environments.

Academia can also help anticipate risks and contingencies that industry is not able to visualise, given the shorter timescales of their R&D cycles. This is especially relevant for general-purpose technologies, recently exemplified with a new generation of systems that are built once, but repurposed for many different applications, such as language models.

Fairness, equity, and justice

AI-based systems may produce decisions or have impacts that are unfair, or even discriminatory, both under a legal or an ethical perspective[25]. In this context, auditing AI-based systems is essential to discover cases of discrimination and to understand the reasons behind them and possible consequences (e.g., segregation).

Methods for auditing AI-based systems[26] for discrimination discovery typically investigate how decisions vary between social groups that differ w.r.t. sensitive variables. The perils of correlation analysis have been pointed out only recently. Specifically, understanding causal influences among variables is a fundamental tool for dealing with bias.

It is important to notice that bias can come from the training data, from the algorithm used to interpret the data or from the human interpretation of results. Therefore, all these dimensions should be considered and measured. AI relies heavily on human-generated data, whose biases can be

[25] G. Alves, M. Amblard, F. Bernier, M. Couceiro, A. Napoli, Reducing Unintended Bias of ML Models on Tabular and Textual Data, DSAA 2021
[26] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, D. Pedreschi, FairLens: Auditing black-box clinical decision support systems, Information Processing & Management, volume 58 (5) Elsevier (2021)

amplified when AI is deployed in complex sociotechnical systems. Misrepresentation in the data, and how to address it, is still under-investigated in the scientific community.

The objective of equity can be achieved by embedding the fairness value in the design of such systems (Fairness-by-design) and by upholding that value (justice). A systematic approach that investigates how to build AI systems that respect by design some fairness constraints for a variety of tasks such as classification, recommendation, resource allocation or matching is missing.

For what concerns the industrial and service sector, we have to consider the legal framework that has been put in place by the commission. Provisions on equality or non-discrimination are firmly embedded within the key Human Rights treaties. In the European Union, there is a harmonised framework established by Directive 2000/43 on "Implementing the Principle of Equal Treatment between Persons Irrespective of Racial or Ethnic Origin", and the Directive 2000/78 on "Establishing a General Framework for Equal Treatment in Employment and Occupation". The GDPR established the principle that personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject. Finally, the proposal of EU regulation on AI "complements EU law on non-discrimination with specific requirements that aim to minimise the risk of algorithmic discrimination, in particular in relation to the design and the quality of data sets used for the development of AI systems complemented with obligations for testing, risk management, documentation and human oversight throughout the AI systems' lifecycle".

In this legal context, industrial applications of AI that impact individuals and groups must be designed or tested for non-discrimination. Embedding fairness, equity and justice by-design requires re-thinking the AI-development cycle, taking those values already into account at design time: What are the main ethical harms or injustices that can be done in this context of the application? What segments of society does the training data reflect or exclude? Which fairness metrics are more appropriate? How to monitor compliance of the socio-technical system to fairness? How to prevent feedback loops? Tackling these questions in an industrial setting is not only an engineering problem. It requires a multi-disciplinary approach and critical viewpoints that AI professionals have not been taught yet.

Accountability and reproducibility

Accountability[27] regards the governance of the design, development, and deployment of algorithmic systems, which takes into consideration all stakeholders and interactions with socio-technical systems. More specifically, bias mitigation includes introducing techniques for data collection and analysis and processing that

[27] European Commission, Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*, Publications Office, 2019, https://data.europa.eu/doi/10.2759/177365

- acknowledge the systemic bias and discrimination that may be present in datasets and models;
- formalise fairness objectives based on notions from the social sciences, law, and humanistic studies;
- build socio-technical systems incorporating these insights to minimise harm on historically disadvantaged communities and empower them;
- introducing methods for decision validation, correction and participation in co-designing algorithmic systems.

Reproducibility[28] is the ability to consistently obtain commensurate results from an experimental setting. It is an important factor to build trust in a result or a specific method that is not supported by a strong theory. Ensuring the reproducibility of learning methods can be difficult, especially when dealing with data science and machine learning (ML), due to the complexity of ML methods in terms of the number of parameters, the optimization strategies needed to make them perform as expected, and the availability and inner peculiarities of the data used to their development. Specifically, reproducibility can be addressed at different levels:

- reproducibility of methods: the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results;
- reproducibility of results: the production of corroborating results in a new study, having used the same experimental methods;
- reproducibility of inference: the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study.

In summary we need to define scientific and methodological measures, quality standards and procedures to better model the development process of learning methods.

From an industrial perspective, an important goal of the accountability task is to uncover and explore available legal answers to tackle bias and unfairness in algorithmic decision-making as well as the accountability gap, i.e., the capacity to attribute AI-related harm to a human or group of humans in the first instance.

It is important to investigate which are the best available solutions or highlight which are the missing parts in existing guidelines, and suggest new possibilities. Another goal is making every system that processes personal data accountable, while at the same time empowering individuals with private rights of action and other subjective rights, like access and the right to object.

[28] O. Gundersen, Y. Gil, D. Aha, On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. AI Magazine, 39(3) (2018)

Respect for privacy

Privacy is one of the first human rights that has been considered in legal frameworks for AI regulation. Nevertheless, there is still the need to investigate new methodologies and approaches for:

- defining formally and detecting automatically privacy risks raised by AI systems handling different kinds of personal data[29]
- designing data anonymisation and attribute hiding algorithms that are robust to sophisticated attacks[30]
- designing AI algorithms that respect by design privacy constraints[31]

Respect for privacy is in tension with other properties that are required for trustworthy AI such as fairness, explainability and transparency. It is therefore very important to investigate the interplay with other aspects and human values, in particular to study and measure the impact of techniques of anonymization, encryption, noise injection on:

- the usefulness and accuracy of the AI models learned from sanitised data
- the fairness of decisions or recommendations computed on the transformed data
- the understandability and interpretability of the results returned by AI systems in the setting of varied tasks handling personal data such as classification, recommendation, resource allocation or matching.

From an industrial perspective, more and more applications use AI techniques that apply to personal data for developing decision making applications that directly impact humans. European industry should promote the development of AI products for the benefit of European citizens, with strong guarantees of their compliance with GDPR. This requires collaborative projects between academia and industry for a continuous transfer of robust anonymization techniques and of novel algorithms that respect by design privacy constraints.

In many applications, humans are the data providers and it is very important to put humans in the loop so that users keep the control on the data they accept to transmit according to their own privacy policy. This requires developers of applications to explain the services offered to the end-users in exchange of their data and to justify precisely which personal data are needed. Therefore, privacy

[29] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership Inference Attacks Against Machine Learning Models. IEEE Symposium on Security and Privacy (2017)
[30] F. Pratesi, L. Gabrielli, P. Cintia, A. Monreale, F. Giannotti, PRIMULE: Privacy risk mitigation for user profiles, Data & Knowledge Engineering 125 (2020)
[31] H. Asghar, C. Bobineau, M.-C. Rousset. Compatibility Checking Between Privacy and Utility Policies: A Query-Based Approach. INP; Laboratoire d'informatique de Grenoble (2021)

cannot be considered in isolation and has to be handled in its interplay with explainability, accountability and fairness.

Sustainability

The position of AI w.r.t. sustainability, and more particularly environmental issues, is ambiguous. On the one hand, AI can bring (and is already bringing) beneficial solutions to many problems related to climate change, global warming and human carbon footprint. On the other hand, AI-based computations are responsible for a large part of the carbon emissions in ICT, which are one important cause of global warming. Such ambiguity has been clearly highlighted in the recent GPAI report *Climate change and AI*[32].

As of today, indeed, beneficial results can only be obtained at a cost in terms of carbon emissions, as many fields of AI research (e.g., deep learning, integration of AI paradigms, auto AI) require both a considerable amount of data and large computing and storage infrastructure. We thus need such large infrastructures to deliver the promises of "good AI" for the planet, at least in the short and medium term. This raises another issue for the academic research community, as such infrastructure is usually not available in academic contexts. This is a crucial issue: Even in the US, researchers are asking for the creation of a National Resource Infrastructure for AI, claiming that suitable computing resources for AI are only available to companies, which invest on short term industrial goals. The lack of sufficient resources for basic AI research has led NSF to a call for proposals for hosting such a national centre that has already received 80 applications from various US academic institutions. Europe is lagging behind in this perspective, but it is important to change this trend in the short term.

But at the same time, research is needed to investigate how to reduce energy consumption and the carbon footprint of AI solutions, be they centralised or distributed, in particular in the field of Deep Learning, where networks have reached such huge sizes (see e.g. Foundation Models, Section 4.e). Improved algorithmic approaches, including symbolic constraints from background knowledge, network quantization and data compression as well as incremental learning and scarce data situations (up to one- and zero-shot learning) should be considered during learning; network reduction and distillation, and local symbolic models for frugal inference. More generally, as suggested for the dimension of explainability, energy efficiency should be another metric to be considered in the design of AI systems and models in order to achieve

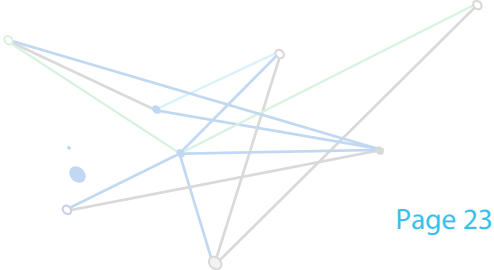
Clearly, taking sustainability into account is crucial also in the industrial and service sector, for economic reasons, and, more and more, in terms of reputation. Data centres, industries that make intensive use of AI algorithms need to take a close look at the sustainability aspect for providing

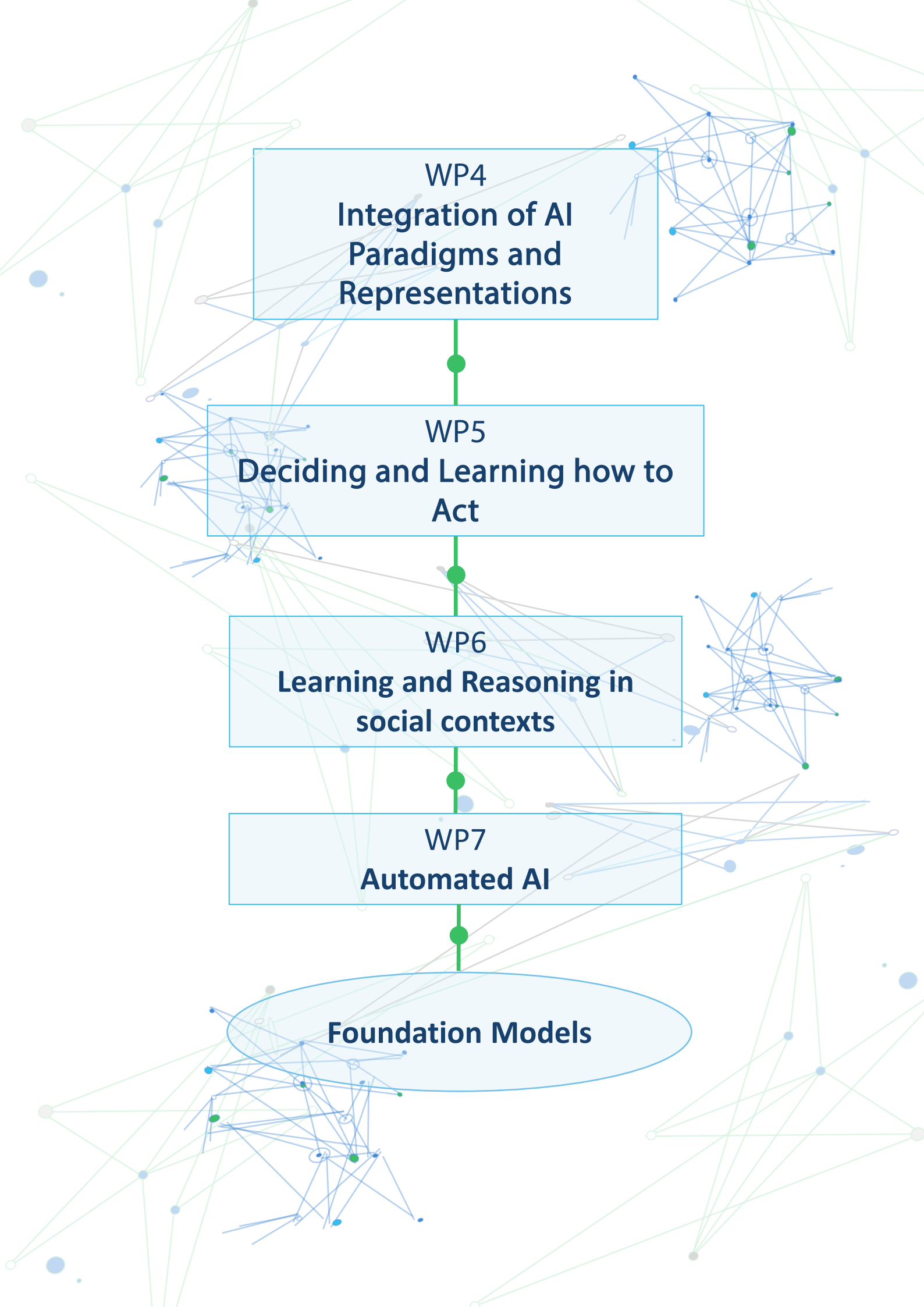
[32] https://www.gpai.ai/projects/climate-change-and-ai.pdf , Nov. 2021

techniques for energy reduction and scale on edge devices, those models that can and should be run close to the data sources.

Towards Trustworthy AI

To conclude this section, the ultimate goal of trustworthy AI research and innovation is to establish a continuous interdisciplinary dialogue for investigating the methods and methodologies to design, develop, assess, enhance systems that fully implement Trustworthy AI with the ultimate goal is to create AI systems that incorporate trustworthiness by-design. The basic question is how to instil all these principles by-design into the basic research themes to the aim of defining methodologies for designing and assessing Trustworthy AI.




WP4
Integration of AI Paradigms and Representations
WP5
Deciding and Learning how to Act
WP6
Learning and Reasoning in social contexts
WP7
Automated AI
Foundation Models

Learning, Optimization, and Reasoning

LEARNING, OPTIMIZATION AND REASONING HAVE REPRESENTED CORE AI TECHNIQUES, OFTEN STUDIED IN ISOLATION. THE TAILOR PROJECT IS FOUNDED ON THESE PILLARS, BUT GOES BEYOND THEIR DEVELOPMENT IN AN INDEPENDENT FASHION. RATHER, ITS PURPOSE IS TO INTEGRATE THEM TO CREATE REALISTIC MODELS AND PROVIDE DECISION SUPPORT IN A COMPREHENSIVE WAY.

In this section, we outline the consolidated result obtained by the interactions with the scientific work packages of the TAILOR project, i.e., WP4-7, devoted to important aspects of the integration of Learning, Optimization and Reasoning. These areas represent the mathematical and algorithmic foundations on which Artificial Intelligence and its applications rest. However, they have so far been tackled mostly independently of one another, giving rise to quite different models studied in fragmented communities.

AI has focussed on reasoning and optimization for a very long time, and has contributed numerous effective techniques and formalisms for representing knowledge and inference. Recent breakthroughs in machine learning, and in particular in deep learning, have, however, revolutionised AI and provide solutions to many hard problems in perception and beyond. However, this has also created the false impression that AI is just learning, not to say Deep learning, and that data is all one needs to solve problems pertaining to AI.

The TAILOR project is founded on these pillars, but goes beyond their development in an independent fashion. Rather, its purpose is to integrate them to create realistic models and provide decision support in a comprehensive way. The four work packages devoted to this aspect are WP4 that pertains to the Integration of AI Paradigms and Representations, WP5 that concerns Deciding and Learning How to Act, WP6 that is devoted to Learning and Reasoning in Social Contexts, integrating agency and autonomy of, with and within AI systems, and WP7 that studies Automated AI, namely the automatic creation and deployment of AI systems without the need of AI experts. For each of these fields, we have collected structured inputs from the various work packages in terms of general description of the context, relevant scientific areas to be covered, also mentioning connections with other disciplines and scientific communities and industrial needs. In the following subsections we outline these topics for the four above mentioned areas. The section ends with addressing the pressing issues in Trustworthiness raised by the so-called Foundation Models, which

were not so prominent at the time TAILOR was conceived, but have become very visible in many application areas since then.

Integration of AI paradigms and representations

History and context: Recently we have witnessed a growing interest from the scientific community toward the combinations of Learning, Optimisation and Reasoning (LOR). Europe has been the cradle of logic programming, inductive programming, constraint programming and logical and relational learning and has a very strong tradition and excellent researchers. Many results come from European universities and research centres. These are a number of surveys[33] [34] [35] that mention important European results and success stories.

Clearly to achieve important results in this field, it is today often important to have domain experts that provide their knowledge, considerable amounts of data, and large computing and storage infrastructure, even more so when addressing different paradigms together. Fighting this trend, moving towards frugal AI, should also be included in the research goal of all responsible researchers.

Scientific promising research areas: While from the literature we can conclude that many approaches have already been proposed and studied, it is clear that there are many open directions that deserve further investigation. In techniques that merge neural approaches with symbolic models, it is important to understand how to balance the models, how to craft knowledge in the symbolic and in the neural part. Also, it is important to understand if a universal representation is needed or if, instead, distinct representations are more effective. Also to deal with scalability issues, it is important to define tractable languages able to incorporate both symbolic and sub-symbolic representations.

There are at least three different types of neural symbolic AI systems that are very promising today but not yet well understood. First, there are those where the symbols refer to logical representations and rules. These neural-symbolic systems aim at easily incorporating constraints and rules as background knowledge in the neural networks as to allow classical reasoning[36]. Second, there are the constraint programming and optimisation systems where approaches such as empirical model learning[37] and smart predict and optimise[38] indicate that solvers and learners should be tightly integrated to get good performance. Third, there are many challenges connected to constructing

[33] https://web.ecs.syr.edu/~ffiorett/files/papers/arxiv21b.pdf
[34] https://www.sciencedirect.com/science/article/abs/pii/S0377221720306895?via%3Dihub
[35] https://www.ijcai.org/proceedings/2018/772
[36] For instance, https://arxiv.org/abs/2003.08316 and Hitzler, P., Sarker, M.K. (Eds) Neuro-Symbolic Artificial Intelligence: The State of the Art, IOS Press, 2022
[37] Lombardi, Michele, Michela Milano, and Andrea Bartolini. "Empirical decision model learning." *Artificial Intelligence* 244 (2017): 343-367.
[38] https://arxiv.org/abs/1911.10092

knowledge graphs from text with neural symbolic methods, as well as a high potential for applying neural symbolic AI to multimodal intelligence, e.g., combining language and vision.

Integrating symbolic and sub-symbolic systems can strongly impact explainability and interpretability of the models, pursuing one of the dimensions of trustworthy AI, as explained in section 3. There are many applications where this feature assumes a paramount importance, for instance in health applications where trust in the decisions or predictions coming from an AI system is crucial for promoting acceptance and adoption of AI techniques. In addition, the improved comprehensibility of machine learned representations might help to improve the communication, the interaction and the collaboration between machines and humans.

Impactful areas for European industry: Many industrial and service sectors might be impacted by systems merging reasoning, optimization and machine learning. For instance, the integration of dynamic systems (ODEs) and ML approaches are of interest in engineering predictive maintenance, and population dynamics (ecology, epidemics...). Robotics of course is an area where it is essential to seamlessly glue perception, planning, actions, learning and reasoning. Other sectors that might be impacted positively by systems that merge learning with optimization/reasoning are resilient large-scale scheduling/planning under uncertainty, logistics, transportation, energy distribution and smart manufacturing. Also the social context might be affected by this integration like climate change, health management, personalised computing services, smart personal assistant, domotics, and so on.

Now, despite the undoubtful advantage that these techniques could bring to the industrial and service sector, many barriers still counteract the wide uptake of these technologies in European industries. The limited knowledge of the potential of AI techniques, especially in SMEs and startups, the lack of off-the-shelf solutions, the lack of trust in automated solutions that could be mitigated by working on trustworthy AI dimensions, the scarcity of good quality data (especially for small businesses), the fact that high quality industrial data are often not available to academia, the lack of scalability of many approaches and the fact that most software systems are still academic prototypes that are not robust enough for applications in industry.

Deciding and learning how to act

History and context:

One of the key aspects of scientific work in TAILOR, carried out within work package 5, is to focus on the fundamental question: ***how an AI agent decides and learn on how to act.*** In particular, research in this area aims at empowering the agent with the ability of deliberating on how to act in the world in an autonomous fashion without the direct intervention of humans. Crucially, empowering an AI agent with the ability to self-deliberate its own behaviour carries significant risks of the agent getting out-of-control; hence this ability must be balanced with safety. Autonomous behaviours must indeed be guided by human specifications, guarded by human oversight, verifiable and

comprehensible in human terms, and ultimately trustworthy. Assessing safety is essential, and formal verification, model checking and automated synthesis to guarantee safety specifications is central to this effort. This line of research involves several fields of AI, including planning, knowledge representation, logics in AI and probabilistic reasoning as well as verification and automated synthesis in Formal Methods. As outlined in the introduction, Europe has a strong tradition and a leading position worldwide in all these fields, as witnessed by the presence of European researchers in the top-rate venues of these areas[39] as well as of AI as a whole[40]. One related area that has been underrepresented in Europe is deep reinforcement learning and reinforcement learning in general, not considering DeepMind in the UK. Another important gap to be addressed is the one between theoretical research in planning and industry applications. Interestingly recently, the European Commission has approved and funded, among the ICT49 projects, the AIPlan4EU[41] project focusing on deploying planning technologies to industry.

Promising scientific research areas

Many interesting research avenues have been identified by WP 5 partners during their work, workshops and meetings, including

- Reasoning and planning for acting[42]
- Learning strategies/plans from data
- Learning heuristics for planning[43]
- Learning models from data, and then do reasoning and planning
- Learning from past experiences and simulations, for refining strategies/plans or models
- Monitoring the actual outcome of actions
- Recognizing possibly unexpected outcomes
- Reasoning, planning and learning how to deal with unexpected outcomes

This work isolated an important research direction that concerns the integration and development of model-based and model-free approaches for learning and planning. In particular, the following areas are considered important:

- Learning action models (related to WP4);

[39] ICAPS <https://www.icaps-conference.org>, KR <https://kr.org>, AAMAS <https://www.ifaamas.org>, HIghlights <https://highlights-conference.org>.
[40] IJCAI <https://www.ijcai.org>, AAAI <https://aaai.org>, ECAI <https://www.eurai.org>.
[41] https://aiplan4eu.fbk.eu
[42] Ivan D. Rodriguez, Blai Bonet, Sebastian Sardiña, Hector Geffner: Flexible FOND Planning with Explicit Fairness Assumptions. ICAPS 2021: 290-298 - best paper awards at ICAPS 2021.
[43] Simon Ståhlberg, Guillem Francès, Jendrik Seipp: Learning Generalized Unsolvability Heuristics for Classical Planning. IJCAI 2021: 4175-4181 - Distinguished paper at IJCAI 2021

- Non-Markovian reinforcement learning (e.g. reward machines, temporally extended rewards and dynamics);
- Integrating logic-based reasoning about actions and data-driven learning;
- Learning and acting in robotics (behaviour trees);
- Theory of mind in order to reason about beliefs, capabilities and goals, when deliberating and executing actions (related to WP6);
- Connections and synergies with formal methods;
- Goal reasoning and formation;
- Learning and exploiting automata/goal structure;
- Considering multiple models to handle various levels of contingencies.

Understanding and regulating the action autonomy of AI agents is considered very important. Ideally one would like to have AI agents that can assess on their own their ability of taking certain decisions autonomously and be ready to ask for human supervision if this is not the case. One example is when an autonomous car gives back control to the human, or a robot that asks for help to humans when unable to perform an action, say press the button of an elevator. But we foresee forms of "adjustable autonomy" that go much beyond these cases studied today, for example asking for human supervision not because the agent cannot do something, but because it considers questionable, or unethical, taking a certain decision.

This area also has important connections with the trustworthy AI dimensions that need to be further investigated: for example inductive biases need to be understood and their foundation need to be developed (WP3).

The creation of benchmarks for these augmented capacities of deliberation and learning on how to act is also important, as demonstrated by classical planning in the last decade when established benchmarks and competitions have had a primary role in the advancement of the field. This is in particular the case especially for deep reinforcement learning, which could leverage ideas from knowledge representation and planning.

Learning and reasoning on how to act is strongly connected with other scientific disciplines, where AI for deciding and learning how to act can boost research and technological development. For example, the connection of planning and formal methods, enable to plan and reason about actions, MDPs, best-effort synthesis, focus on finite traces and not only infinite traces (intelligent agents do not work for a task forever). There are also significant connections with Operations Research, psychology, human science and cybersecurity.

Impactful areas for European industry

Deciding and learning how to act is important in several contexts: mobility, production, interacting with humans, fintech, entertainment, and many others. For example, autonomous mobile robot platforms are focusing less on hardware aspects and more on organisation and software, to

automate warehouses and logistics. This shift is an opportunity for introducing advanced forms of autonomy based on the kind of work done in WP5. Smart manufacturing could benefit from research in learning and reasoning on how to act by automated program-synthesis and learning how to handle unexpected exceptions. Interaction with humans requires autonomous capability in acting in order not to be too annoying to the humans themselves. FinTech is interested in creating autonomous agents that can act rationally while learning from actual data during operation. Also video games, augmented reality, interactive entertainment is heavily relying on these techniques for improving the interaction and the behaviour of avatars. A further application of learning and reasoning on how to act is the education field, for instance, to plan individual curricula for students of online classes.

Learning and reasoning in social contexts

The full deployment of artificial intelligence devices in our society, be they robots, chatbots or IoT systems, makes it a distributed socio-technical system. Such a system can only work, and be trusted, if it is aware of its social context. This is considered here from a multi-agent standpoint.

History and context

Many AI approaches rely on a multitude of agents. Distributed AI, such as in multi-agent systems, swarm intelligence and robotics, is based on a number of agents interacting and communicating with each other and in a virtual or physical context. According to social interaction paradigms, agents should not reason, learn and act in isolation, but with and among others. Therefore, it is important to explore the foundations of social intelligence and social behaviour of how AI systems should communicate, collaborate, negotiate and reach agreements with other AI and (eventually) human agents within a multi-agent system (MAS).

Nowadays, computation is increasingly distributed and the IoT will enable devices to become more intelligent, to communicate, and in the end to socialise. Social AI will be observable within Massive Multi-Agent Systems (MMAS), which will include all sorts of devices and different interaction modes with people, organisations and institutions. Important research challenges are open and need to be investigated in TAILOR and beyond. One research question concerns how we empower individual AI agents to communicate with each other, collaborate, negotiate and reach agreements/consensus and how they coordinate to fairly share common resources, and how they differentiate to accomplish collaborative tasks together.

When interacting with one or more human agents, software agents or robots need to explain their motives and intended actions and to understand those of their human partners. Models that capture the other agents' actions and reasoning, like theory of mind models, need to be included in each agent, which allows reflection about one's own intentions, those of others and the effects of one's actions on others.

Multi-agent systems are also used for studying and simulating human behaviour, such as the behaviour of individuals in crowds, spreading diseases in pandemics or financial market dynamics. In agent-based social simulation, individual (group) behaviour is modelled in order to allow for observing emerging global behaviour in more detail and precision than using models that use population averages.

Another important aspect concerns how to make agents learn from each other in a responsible and fair way, leading to more intelligent and fair collective behaviour (e.g., multi-agent reinforcement learning, MARL). And finally how to create trustworthy hybrid human-AI societies that fulfil humans' expectations and follow their requirements.

Preserving privacy for federated and social learning is of course of paramount importance for ensuring privacy within the planning and coordination activities carried out within the MAS.

A number of issues are still open. For example, modelling domain knowledge is labour intensive in general, and it is even harder in social domains featuring human factors and social dynamics. We may need to revive old models and methods as important results have already been achieved in the field of social agents[44]. As an example, domain knowledge can be applied towards increasing explainability. Automatically injecting domain knowledge into agents would be a step towards improving transparency and explainability. Furthermore, using embodiment features and social clues can also add to the richness of the interaction making social AI more understandable[45]. Domain knowledge can also be represented in the form of "digital twins", for instance in the field of transport or logistics. Updating these simulators with online observations would be extremely helpful also in industry.

Important connections between social AI and trustworthiness need to be further investigated at the moment. Not only humans need to trust the agents but also agents need to trust the other agents that they are placed together with, in particular if they plan to learn from them. Thus, finding trustworthy equilibria would be important. For large scale multi agent systems, equilibria are more subtle, calling for a connection of classic MAS methods with population dynamics or dynamical methods.

Another dimension of trustworthy social AI that pertains to social and multi-agent systems is that explainability should be directed towards what are the goals of the AI system, which solution concepts it will reach and how to explain them appropriately to the specific target audience.

---

[44] Lugrin, B., Pelachaud, C., & Traum, D. (Eds.). (2021). The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition.

[45] Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., & Chetouani, M. (2021). Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction (THRI)*, *10*(3), 1-24.

Impactful areas for European industry With the emergence of Hydrid AI systems, the impact of Social AI systems will affect all industrial domains. Already in healthcare we see that AI systems are in dialogue with humans to detect and analyse cancer cells, as well as systems that suggest diagnoses. In addition, social AI is more and more supporting humans in self-care and prevention.

Another example is application in precision agriculture and dairy farming. Humans collaborate with machines by tuning the model parameters in the AI systems that are used for crop production and cattle management. Also in the traffic and transport sector uses interaction/dialogue based mechanisms in their traffic management systems. Like in the TV entertainment sector, humans pass preferences and systems classify and personalise their interaction.

Social AI systems are also used in the energy sector, e.g. citizens optimise the energy consumption in buildings. In near future buildings do share information between each other to learn and collaborate in energy management towards the local power grids. This will be extended at smart city level, and efficient building occupancy management. Experiments have started in the field of law enforcement, for example to use federated reasoning mechanisms to gain a better understanding of debt problems or resolve cold cases.

One of the generic areas of application of social AI, useful in all industrial domains, is modelling and simulation. It concentrates on observing the behaviour of humans or systems of agents in order to better understand that behaviour in terms of derived rules and patterns in the underlying mechanisms. E.g. modelling negotiation mechanisms, decision making or group formation.

The insights obtained, i.e. the retrieved rules and models, can be used in simulations and implemented in real-life applications. One sees this already in e.g. multi robot task allocation in search and rescue contexts, traffic management and control in smart cities or advanced planning in digital manufacturing factories.

Hybrid team interactions for multi-party decision-making are being explored in simulated environments where agents are represented as active digital twins and humans participate either interactively or by modelling their (social) behaviour. Nevertheless, the deployment in critical trustworthy real-world use cases is the ultimate goal for hybrid intelligent systems. (see also section \ref{Projects})Real life situations sometimes require a coordinated and combined use of different approaches, E.g. Shortcomings in the AI have to be solved by designing teamwork that allows the human to take over.

Automated AI

History and Context: The substantial and fast progress of AI seen in recent years often comes at the price of a paradoxical increase in complexity. The many choices of approaches, algorithms, configurations and (hyper-)parameter settings also create challenges for deployment, and despite more and more off-the-shelf tools being available, non-AI experts still struggle to get the best out of

them. Automated AI aims to bridge this gap, by enabling people without specialised training in AI (e.g., researchers, engineers or students) to benefit from the best AI techniques, with limited help from AI experts. At the same time, AI experts can make faster progress and obtain better results using AutoAI methods and tools.

Historically, AutoAI can be traced back to work in the AI and Optimization communities in the area of Algorithm Selection and Configuration (ASC), and mainly concerned the data-driven selection of the best algorithm for solving a given problem as well as the configuration of its parameters. This work ASC has had, and continues to have, major impact, notably in its most general form, known as Programming by Optimization (PbO)[46]. In the Machine Learning community, the concept of AutoML first emerged from directly building on earlier work on ASC. Its coming of age was boosted by the series of AutoML challenges[47], and gradually AutoML included design choices across entire ML pipelines. One particularly prominent area in AutoML is AutoDL, aka Neural Architecture Search (NAS), which aims to optimise the architecture of the deep neural network on a given dataset or class of datasets.

Today, AutoAI is a fast-growing research area, with many promising directions for further exploration; WP7 of TAILOR (and associated activities) aim to facilitate that exploration. It is an essential part of TAILOR strategy to reach the goals of Trustworthy AI, combining Learning, Optimization and Reasoning. AutoAI contributes to these goals by automatically (a) detecting an AI system is no longer reliable for its originally intended task, (b) adapting to new conditions, and (c) making trade-offs between aspects such as explainability and simplicity.

AutoAI for Hybrid AI Systems: Purely data-driven AutoAI has limitations, and domain knowledge needs to be included in AutoAI systems to design better pipelines (e.g., to preprocess messy data and include model preferences in AutoML and AutoDS[48]) and to make better decisions (e.g., choice of algorithm type in ASC). Similarly, combining AutoAI techniques with HCI approaches could allow users ranges of automation, between full automation and completely human-driven processes, and to gradually improve their skills. Furthermore, we see much promise in expanding the scope of AutoAI approaches to include algorithmic approaches beyond AI, such as simulation based on physical models.

Another type of hybridization combines optimization and reasoning techniques for ensuring "trustworthiness by design", provided some meaningful and reliable metrics of such requirements are made available. Similarly, multi-agent systems (encompassing both human and artificial agents)

[46] H.H. Hoos. Programming by optimization. Commun. ACM, 55(2):70–80, 2012
[47] F. Hutter, L. Kotthoff and J. Vanschoren, *AutoML - Methods, Systems, Challenges*, Springer Verlag, 2019.
[48] T. De Bie et al. "Automating data science." Communications of the ACM 65.3 (2022): 76-87.

would benefit from AutoAI approach, e.g., in the context of timetabling tasks. Due to the strong interaction with humans, trustworthiness issues are especially relevant here.

The holy grail of task similarity: To efficiently explore which algorithms may work well on a new dataset (or task), it is often useful to assess whether the new task is similar to previous tasks, so that algorithms that worked well before can be tried again. One option is the use of meta-features, a set of properties that identifies a specific dataset/task. They allow quick assessment of similarities and differences between two datasets. This allows much more targeted selection of algorithms and hyperparameters. Good meta-features exist for various optimization problems.

However, in AutoML, high-quality meta-features that are both useful (to allow clear choices of algorithmic bricks and their configuration) and meaningful (to contribute to the trustworthiness of the whole process) as descriptions of datasets have remained elusive. Recently, progress has been made especially for specific types of datasets, such as task embeddings for image or tabular datasets[49]. Still, much work remains for effective adoption. Further research is needed into task similarity measures that transfer to practical and effective use. An alternative is to not assess task similarity beforehand, but by interacting with the new task (e.g., by active testing[50]): after evaluating some algorithms on the new task, better similarity assessment to previous tasks may be possible.

Robust, efficient and multi-objective AutoAI is needed to enable real-world AI applications: The real world is dynamic, and the mid- to long-term deployment of AI systems requires them to adapt to change. This is true for learning (incoming data may change) and optimization (the objective function changes with the environment) algorithms. These changes may require changes in (hyper-)parameter settings or even the choice of model or algorithm to be used. An operational AI system could be monitored by an AutoAI "supervisor" that responds to change by adjusting the system. When the AutoAI supervisor cannot adapt to guarantee safe use of the AI system, it should warn the user to call in a human expert (see also the previously mentioned link to HCI).

Broader adoption of AutoAI tools and systems in real-world settings critically hinges on improved accessibility and usability. Current AutoAI systems are often designed for research and can require substantial AutoAI expertise to use. Designing AutoAI systems for non-experts requires addressing multiple challenges, including the ease of setting up and applying the system to a problem, and the explainability of the results and process (cf. explainable AI section). Once developed, easy-to-use and explainable AutoAI systems would also aid sustainability by, e.g., reducing computationally costly mistakes. Especially in the area of AutoML, broad use in real-world applications will require not only improved accessibility (e.g., in terms of the computation cost and complexity of use of neural architecture search), but also a much improved ability to handle messy real-world data[51].

---

[49] J. Vanschoren, Meta-Learning. In: *AutoML - Methods, Systems, Challenges*, Springer Verlag, 2019.
[50] R. Leite et al., Selecting Classification Algorithms with Active Testing. MLDM'2012
[51] John W. van Lith and Joaquin Vanschoren: https://doi.org/10.48550/arXiv.2111.01868

Real-world problems often involve complex and conflicting sets of requirements, e.g., accuracy *vs* explainability vs robustness vs fairness (detection or removal of biases in the data), accuracy *vs* complexity (to address environmental concerns), accuracy *vs* resource requirements (e.g., amount of training time/data), and tradeoffs between different solution quality objectives (e.g., false positives *vs* false negatives in binary classification). Multi-objective AutoAI methods, currently still in their infancy, are therefore of key importance to real-world uses of AutoAI.

Leveraging multi-task- and meta-learning: As mentioned earlier, the space of possible algorithms, algorithm configurations, and pipelines of multiple algorithms, is huge, and searching this space "from scratch" is very inefficient. Practical AutoAI tools tend to make many assumptions about future tasks to drastically constrict the search space (e.g., by imposing a semi-fixed pipeline). This may lead to suboptimal results when the optimal solution is outside the preselected space, and making good preselections for unfamiliar application domains is difficult. One promising way of addressing this issue is based on learning which algorithms and configurations work well for an application domain, with the goal of creating domain-specific AutoAI tools focussed on selecting and fine-tuning AI algorithms within that domain. This meta-knowledge could also be combined with knowledge from domain experts, resulting in better generalisation to new use cases (e.g., involving different types of images or text) and reduced need for training data.

Benchmarking AutoAI: AutoAI research is largely empirical, and new approaches are validated by benchmarking against state-of-the-art methods. Therefore, carefully constructed, widely available benchmarks are of crucial importance. Existing AutoAI benchmarks are too limited, and the development of AutoAI tools for more diverse problems critically depends on high-quality benchmarks for those settings.

Widely used datasets also exist in subfields of AI related to combinatorial decision and optimization problems, as well as for continuous optimization problems, but AutoAI extensions are needed. For AutoML, the OpenML platform gave rise to thousands of useful ML datasets. For reinforcement learning, platforms exist with an open interface that should allow to implement reproducible benchmarks[52], but no clear set of environments and tasks appears to have emerged as a recognized benchmark. Finally, the huge computational costs for AutoDL gave rise to tabular and surrogate NAS benchmarks, culminating in the current NAS-Bench-Suite[53].

In many cases, it remains an open issue which metrics to use to compare different approaches, and for noisy or dynamic problems, the type of noise or change also has to be defined.

Impactful areas for European industry: AutoAI is critical for all companies and organisations that are too small to afford hiring highly qualified AI experts, including SMEs and public institutions.

[52] https://gym.openai.com/
[53] https://openreview.net/forum?id=0DLwqQLmqV

Although superficial expertise in AI is becoming more commonly available, AutoAI is expected to yield better results in terms of performance and trustworthiness. Furthermore, even in situations where specialised AI expertise is available, AutoAI has an important role to play: when the problems to be solved evolve or vary over time, the AI methods used to tackle them also need to be adapted (examples are found in the logistics, manufacturing and agri-food sectors), and AutoAI techniques are key to achieving this adaptation in a timely, cost-effective and trusted manner.

Foundation Models

The term Foundation Models was coined in Summer 2021 by researchers from Stanford[54] to include these huge models that are trained on huge amounts of data on no specific task, but can be "easily" fine-tuned into a number of particular tasks. The first Foundation Models were designed in the domain of Natural Language Processing (NLP), but the concept was soon applied to other domains, like speech, image, robotics, … In the following, NLP Foundation Models, aka Large Language Models (LLMs) will be used to illustrate the opportunities and the risks of Foundation Models at large.

History and Context: Language Models (LMs) are the basis for many natural language processing (NLP) tasks, and in the recent years, we could observe major improvement across several NLP tasks using LLMs (like e.g., GPT3, Gopher, RETRO). A (statistical) LM represents a probability distribution that specifies the probabilities of sequences over tokens. Different objectives such as predicting the next token in a sequence, or predicting missing tokens in a sequence, have been proposed to train LMs. The goal is to learn token representations that capture their meanings, so that tokens with similar meaning are represented by similar distributed representations.

Vector representations of words have been used to model the semantics of natural languages based on the lexical co-occurrences in large corpora[55]. Deep neural network architectures for NLP, which usually represent words as vectors, have naturally incorporated these word-embedding models as pre-trained representations for downstream tasks (e.g., sentiment analysis)[56]. Since 2018, pre-trained word-embeddings have been replaced by LLMs, as they produce contextualised representations which improve the performance on most tasks[57]. Nowadays, one of the main paradigms in NLP consists of using a pre-trained LLM (e.g., BERT, XLNet, RETRO) and fine-tune it to a specific task[58]. Very

[54] Rishi Bommasani et al. (66 authors), On the Opportunities and Risks of Foundation Models}, arxiv 2108.07258, 2021.
[55] Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. Language and Linguistics Compass, 6(10):635–653.
[56] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In Proc. Conf. on Empirical Methods in NLP, pages 298–307. ACL.
[57] Matthew E. Peters, Mark Neumann et al. 2018. Deep Contextualized Word Representations. In Proc. Conf. of the North American Chapter of the ACL: Human Language Technologies, Volume 1, pages 2227–2237. ACL.
[58] Jacob Devlinet al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. Conf. of the North American Chapter of the ACL: Human Language Technologies, Volume 1, pages 4171–4186. ACL.

well-known examples of the effectiveness and high quality results of LLM-based tools, applications and services are, for example, the advances in the machine translation field, among others.

Despite (or because of) their impressive capabilities, pre-trained LLMs raise severe concerns, such as the incorporation of various biases present in the training data, including among others, the overrepresentation of hegemonic viewpoints or the potential damaging of marginalized groups[59].

Due to their huge complexity (billions of weights), LLMs suffer from all issues mentioned previously regarding lack of trustworthiness, and what emergent properties they present. There are also worrying shortcomings in the text corpora used to train these anglo-centric models, like the lack of representation of low-resource languages. For instance, in accepting large Internet-based datasets as models training data, we risk perpetuating dominant viewpoints, increasing power imbalances and further reifying inequality, and introducing stereotypical and derogatory associations along gender, race, ethnicity or disability status presented in them. These shortcomings are general to all models and imply that they could return incoherent output, a critical issue when the models are used for questions-answering, advice-giving or in dialog systems in general. Some of the sources of incoherency in its outputs can be explained by a lack of integration with formal knowledge representation models, which does not allow these systems to perform logical reasoning or even very simple deductions, the shortness of its internal memory of recent facts, lack of clear understanding of temporal reasoning, and its general impossibility to appropriately manage and deal with imprecise terms expressions, which are so common in human natural language. An identified scientific challenge is the integration of these models with knowledge representation and reasoning models, enriched with relevant elements which are necessary for dealing appropriately and naturally with human language, such as temporal concepts, or vague terms and expressions.

AI & Data Sovereignty: LLMs have rapidly become ubiquitous in technologies that we use in everyday life, such as chatbots and digital agents (Siri, Alexa), or automatic translators, that are critical to Europe unity. However, these models are developed primarily outside Europe by Tech giants like OpenAI. To retain Europe's AI sovereignty, increased attention to European efforts to train such models should be devoted. European efforts would also ensure that trustworthy AI is addressed within a European framework. For instance, the following questions that address the basic components "lawful," "ethical," and "robust" of trustworthy AI should be addressed within a European framework when LMs are used within industrial applications: who is accountable for the output of generative language models (*lawful*)? How do we address bias in the data and the models (*ethical*)? How can we improve factual correctness of LMs (*robust*)?

[59] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency (pp. 610-623).

Furthermore, a related aspect is trustworthy data infrastructure. Where should training data and inference data be saved? One European initiative that addresses this and related questions is the "GAIA-X" initiative, where a trustworthy data infrastructure is developed.

In the recently started project "OpenGPT-X" the concerns mentioned above are addressed. Within OpenGPT-X, LMs *"Made in Europe"* are developed relying on a GAIA-X data infrastructure. The LMs will be provided to the public as well as smart services relying on large LMs such as question answering, machine reading comprehension, and dialogue systems are provided to the public and European industry.

Biases/trustworthy/explainable: Large language models are trained on massively huge textual corpora, which include substantial amounts of non-curated data from the web. As these models learn statistical word patterns, they incorporate information which humans can associate with biases and harmful attitudes. In fact, several studies have pointed out how LMs encode stereotypical associations about race, gender, disability status, or ethnicity[60] [61] [62]. When used for text generation, LLMs can produce highly toxic language utterances which incorporate some of the mentioned biases. Moreover, current LLMs do not provide insights on their internal structure, so that they are extreme black-box models with a very limited explainability.

Fairness/equity/justice: Since the emergence of LLMs such as ELMo and BERT, institutions and companies seem to compete to produce ever larger language models, increasing their training data and number of parameters aimed at improving the results on various NLP tasks. Although some of these models contain multilingual representations, most of them are trained only for English. Furthermore, multilingual models include at most the top 100 languages in Wikipedia (out of the approximately 7,000 world languages) and behave notoriously worse than monolingual ones. On the one hand, this prevents many citizens from interacting with AI systems using their own languages, thus increasing their potential exclusion, and undermining language and cultural diversity. On the other hand, accepting that the web data used to train LMs is representative of all humanity may involve the perpetuation of dominant viewpoints and power imbalances as well as the disregard of views coming from communities which are underrepresented on the Internet.

Sustainability: Training foundation models based on deep neural networks need large computational infrastructures with multiple instances of GPUs or TPUs, therefore limiting access to these models on an economic basis. For instance, the training of GPT-3 reportedly cost $12 Million

[60] Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In Proc. First Workshop on Gender Bias in NLP. 33–39.
[61] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In Proc. First Workshop on Gender Bias in NLP. 166–172.
[62] Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In NeurIPS. 13230–13241.

for just one run[63]. Besides, the training process of a single model usually needs weeks or months to be completed, involving substantial costs to the environment due to the energy required to power this hardware. As an example, it was estimated that a single deep learning model for NLP can produce more than 626k pounds of CO2 emissions, which can be equivalent to the total lifetime carbon footprint of five cars.

Reproducibility and validation: Use of automatic general metrics has been widely extended as the usual validation methodology in all LLM applications. In some areas, such as machine translation, well-known valid metrics such as BLEU have been established for confronting the translation models results with the golden standards. Nevertheless, these metrics have been transferred to and adopted by other NLP areas in which their validity is under discussion by the scientific community. Another major problem in this context is the under-reporting of errors, since most of the evaluations, either manual or automatic, only produce a final score and do not report any details about the actual errors that systems make and their outreach. Because of this, validation is in general an unsolved open issue within these contexts, which demands research to be conducted on the appropriateness of automatic metric validation, as well as feasible and sustainable methods for human validation of LLM-based systems.

Bundle Forces: Training large language models requires, despite the scientific expertise, immense hardware resources, and the expertise to efficiently use such resources. Usually, a single European academic institution does not have the resources to train such models. However, scientific expertise and hardware resources are distributed across European institutions. For instance, the supercomputer JUWELS has more than 4000 GPUs and is among the fastest supercomputers in the world. Bundling forces across Europe would enable the development of large-scale language models for Europe. The recently started initiative GPT-X is a first step in that direction.

Impactful areas for European industry: As researched by market research firm Statista[64], the growth of the NLP industry in Europe has steadily increased since 2017, with a turnover of $208.7m to a projected $2.2bn by 2025. Other studies[65] have also pointed to this growth.

In both studies, different challenges for the European NLP industry are observed, such as the high dependence on the large North American NLP industries, the shortage of specialised technicians or the lack of highly competitive technological infrastructures, in a context of political and, above all, energy volatility, both related to the COVID-19 pandemic and climate change.

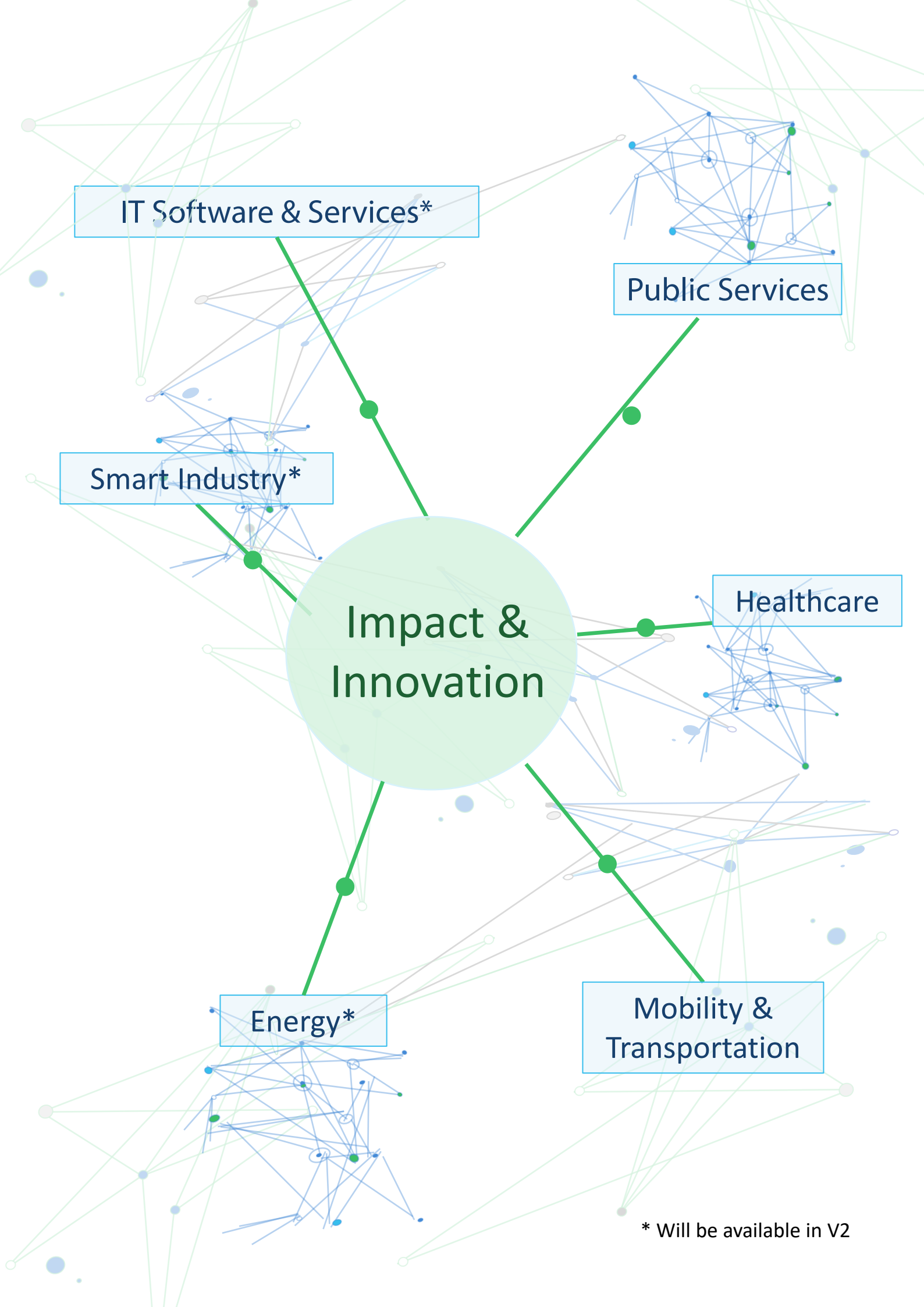
63 https://venturebeat.com/2020/06/01/ai-machine-learning-openai-gpt-3-size-isnt-everything/
64 https://www.statista.com/statistics/1042474/europe-natural-language-processing-market-revenues/
65 https://www.kbvresearch.com/europe-natural-language-processing-market/

Thus, LM and corpuses are being developed and built by the GAFAMs, although some European companies have been playing a cutting-edge role in specific topics (e.g., DeepL in the machine translation area), but usually the major players in NLP are companies based outside of Europe.

At the same time, Europe, with its enormous diversity of languages and linguistic varieties, has a superior knowledge of the perspective that users expect from such NLP resources, tools, and products than other parts of the world.

For all these reasons, there are great opportunities for growing a truly European industry on language technologies, which can address the challenges of language diversity and equality across Europe. Some areas which can benefit from this are:

- Voice technologies, such as Automatic Speech Recognition (ASR) and Text-to-Speech (TTS)
- Dialogue systems, including chatbots and conversational assistants
- Machine translation
- Linguistic correction and assessment
- Information extraction, Opinion mining, sentiment analysis and fact-checking/verification
- Natural language generation, including Text-To-Text and summarization or Data-To-Text systems.



* Will be available in V2

Impact and innovation

While the TAILOR project is mainly concerned with research areas, both fundamental and applied, the network activities, through WP8, and hence the Strategic Research and Innovation Roadmap, are also concerned with the impact this research will bring on the industry, on the public sector and the society at large, beyond the project boundaries.

AI could bring important breakthroughs in the definition of regulation, democracy and the common good, as well as facing global social challenges such as health care and personalised medicine, well-being, pandemic response, climate change, poverty, equality, and inclusion.

Three are the main objectives of the Roadmap for the industrial sector:

1. Raising awareness in Trustworthy AI
2. Supporting companies in assessing/repairing/designing trustworthy AI systems
3. Identifying which sectors/applications are more affected by Trustworthy AI guidelines, by analysing the high-risk systems identified in the regulation AI Act.

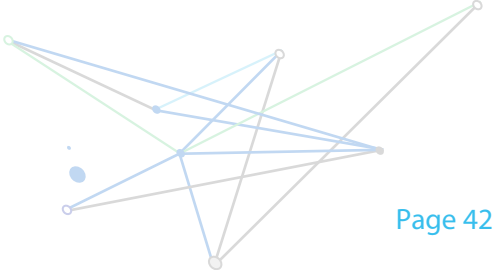
To achieve these objectives and obtain a substantial impact on the economy, the society, and the environment, it is important first to identify which sectors would benefit most from the broad uptake of AI techniques.

Theme Development Workshops

Theme Development Workshops (TDWs) are an innovative format bringing together key players from industry, academia and politics to jointly identify the key AI research topics and challenges in a certain area or for a specific industry sector. In December 2020, an agreement was made between the respective coordinators and leadership teams of TAILOR, VISION, HumanE-AI-Net and CLAIRE to plan and execute a series of Joint (co-organised) Theme Development Workshops, starting in 2021. In the following subsections, we summarise key observations from the first three Joint Theme Development Workshops (TDW) organised under the lead of TAILOR.

In total, the TDWs brought together 180 experts from academia, industry and politics to jointly identify and discuss the most promising and emerging AI topics in the public sector, mobility & transportation sector as well as in the healthcare sector. The full reports are available for download via the TAILOR website:

- Theme Development Workshop "AI in the Public Sector"
- Theme Development Workshop "Future Mobility - Value of Data & Trust in AI"
- Theme Development Workshop "AI for Future Healthcare"



In this roadmap document, we specifically point out concerns, problems and/or bottlenecks that need attention and answers for Trustworthy AI to become reality, as well as some first ideas on how to address them. Accordingly, these topics should be reflected in the TAILOR roadmap for guiding research and development towards providing solutions.

Focus on six industry domains

One objective of TAILOR is to develop synergies and cross-fertilization between industry and the TAILOR network of excellence centres to provide the basis for Trustworthy AI in Europe. Industrial participation is ensured by leveraging the strong connections to many different application domains and big industry networks via our large network of partners, including some selected multinationals and major companies, representing and providing links to their specific industrial sectors. In particular, the following industry sectors are focussed in TAILOR: Smart industry, IT Software & Services, Public Services, Mobility & Transportation, Energy, and Healthcare, as illustrated in Figure 1.

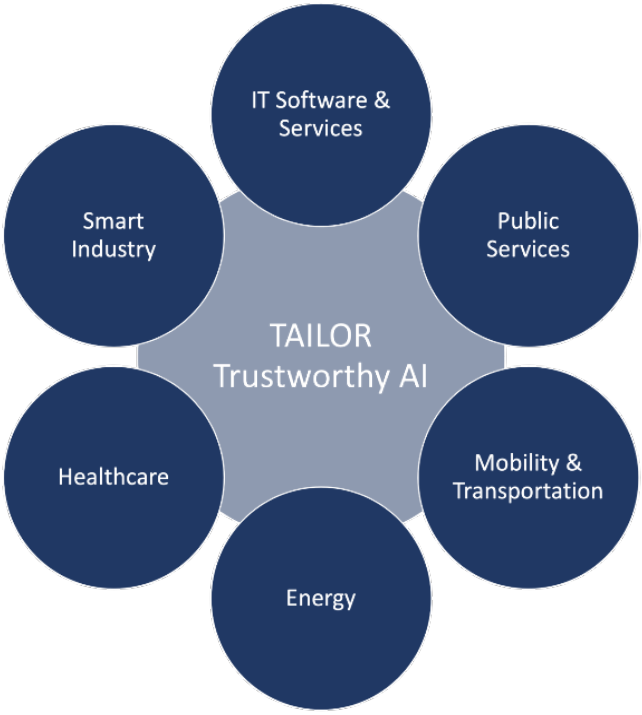


Figure 1: The six industry sectors represented by partners in TAILOR: Smart industry, IT Software & Services, Public Services, Mobility & Transportation, Energy, and Healthcare.

The following section will present the outputs of the first three TDWs: AI in the Public Sector, AI for Future Mobility, and AI for Future Healthcare.

AI in the Public Sector

The TDW was focussed around application areas for AI in the public sector, as well as on more horizontal topics of AI spanning many application areas with a particular focus on different aspects of Trustworthy AI. The following challenges and ideas were identified during the worksop, which are of specific interest for the TAILOR Roadmap:

Building trust in AI systems

Building trust in AI systems is essential to strengthen the application of AI in the public sector on both the public and private side. Thus, it is important to involve humans more closely in the development of AI systems and to guarantee humans the possibility to intervene in AI applications at any time. The following ideas were mentioned in order to address this challenge:

- Could an independent and objective institution guarantee the safety of AI systems, and if yes, should it be on a local, regional, national or European level?
- Can we develop ecosystems of trust for enabling and safeguarding AI systems?
- In order to achieve public trust in AI systems, the user of an AI system needs to know when a system is operating out of bounds, and this also needs to be clearly communicated. So a main point in this context is transparency. How can this be achieved? Education can be considered as a preparatory action, in particular in the context of public administrations, where a clear understanding of the overall framework for the potential introduction of AI in the processes is needed, as well as a maturing awareness concerning the limits and capabilities of AI.

Reliability and accountability of AI-based systems and solutions

Especially in the discussions with statistics institutes involved, some interesting key elements were identified in order to improve **reliability** and **accountability** of AI-based systems and solutions. In particular, experts addressed and suggested the following ideas and challenges:

- A **certification approach** (including governance) was identified as crucial for the public acceptance of AI.
- The idea of an **algorithm registry** for the use of AI.
- The possibility of making software code (or parts of it) available to the public (e.g., open source) or to public administrations, balancing the need for a minimum level of openness (from the perspective of the public administrations) with the need of protecting strategic technological assets (from the perspective of solution providers), in order to **improve accountability and trust** individually addressed per project.
- How to **measure the performance of AI ecosystems**? The performance is more than financial or operational performance; impacts and systemic change should also be considered.

- AI is built on data, but many times public administrations suffer from internal data silos; unavailability of data could result in not-suited solutions for the needs of the public administration or to solutions that are affected by bias. To obtain reliable and accountable AI-based systems in public administrations, these considerations highlight the need to overcome (internal) information silos in different public organisations.

Socio-technical context and systemic effects

Another key insight from the workshop was that AI-based systems should be understood, studied, developed and implemented in their **socio-technical context**. Considering AI technology in isolation is not sufficient, because system safety can then not be guaranteed for example. Accordingly, the experts suggest:

- A paradigm shift from technology-centred to a more system- and human-centred approach.
- The further development of mechanisms like sandboxes to reproduce some algorithmic decisions, and approaches how to **measure systemic effects**.

Closely connected to this was also the discussion, that currently dominant topics like fairness, accountability, transparency or explainability are all contextual. Although **the requirements for AI systems as identified by the High-level Expert Group on AI** (HLEG-AI) are still considered to be very valid and relevant, the experts pointed out that it would be beneficial to:

- Further define and clarify trustworthy AI in general, and its requirements (as defined by the HLEG-AI) in particular.
- Focus on a more socio-technical specification of AI systems, including a broad and integrated view on different aspects and the context.
- Development of **standardisation and validation processes for AI components**: The objective is to bring these to the market while dealing with aspects of safety, security and privacy.

AI for Future Mobility

During the TDW a broad range of topics were discussed, among them challenges for validation and verification of AI systems, safety and security in the automotive industry including robustness to unforeseen changes, as well as the many aspects trustworthy AI encompasses in the different phases of development of products incorporating AI. Within these discussions, the following sector-specific challenges in particular emerged:

Challenges of developing standardisation and validation processes for AI components

The experts identified these topics as important mechanisms to bring AI systems and components to the market while dealing with aspects of safety, security and privacy. In this context, also the

certification of AI systems was discussed in more detail, producing the following ideas and challenges:

- **Controllability by humans**, so that humans have control over the AI system at any time and can intervene if necessary. A new way of controllability might be the process of keeping the human in-the-loop.
- **Self-awareness of the systems**, meaning that the system can assist itself in case it is unfamiliar with the situation. This aspect could also address the safety argument of certification. Some possible solutions were highlighted by the participants, like Bayesian Inference Technique for handing over the control to humans or how to quantify the uncertainty or methods for quantification of uncertainty.
- The idea of **"Breaking the Rule" in Autonomous Driving**, which means that such systems should be allowed to break traffic rules in case of an emergency. However, this will make it much more difficult to get a certification of the AI-based system.

Explainable AI for time series and verification approaches

This topic, which was brought to the workshop by one of TAILOR's industrial partners, was considered as a very complex and relatively new one by the experts. In particular, the following ideas and challenges were mentioned:

- Discuss the **distinction of trustworthiness and explainability** in more detail, especially as many open questions of explainable methods exist in general.
- How could **generalisation** be established, not only from a local but also from a global approach. Therefore, multiple methods should be considered, including mechanisms **ensuring that the algorithms work properly and give reasonable results**.
- A **big data pool** would be beneficial in this context, so that everyone can **work and train on common datasets** for an easier understanding and to identify potential errors.
- Focus on creating an **Explainable AI rulebook**, maybe starting with a proper categorization of AI application areas.

Interdisciplinary nature of Trustworthy AI

Trustworthy AI has an interdisciplinary nature and therefore must cover a lot of aspects. Especially, better understanding of gaps in Trustworthy AI plays an important role in the mobility sector and should be developed further. In general, Trustworthy AI is considered as a difficult to grasp topic in this sector and therefore the experts suggest to perceive and approach it holistically, including the areas of "Robustness & Security", "Human-in-the Loop & Explainability", "Ethics, Privacy & Liability", "AI Governance & Monitoring", "Verification & Validation", "Data Availability/Quality", "Reliability & Safety". On a political level, it would be beneficial to consider using the same terms in ongoing and future initiatives, especially in the area of AI/digital Ethics.

Building trust in AI and bringing in data

It will be important to explain how data is being used to increase trust, to have more experts and resources in Europe, to invest in projects to educate people how to build good models and remove biases, as well as collecting data in the right way. Furthermore, the challenge is also to define and visibly present the value of the data and its release on the customer and provider side in order to motivate sharing data for specific applications.

Availability of and standards for data

While the international availability of, and standards for, shared data spaces are considered a prerequisite for trustworthy AI, there are conflicts between collecting and sharing enormous amounts of data and simultaneously protecting sensitive data. Especially in the mobility sector, significant problems related to the availability and use of data are slowing down the industry but could in part be addressed by:

- Creating and using easy to obtain data bases.
- Creating better standards for data-driven programming and increasing investments on the hardware side to have secure supply chains for a fully trustworthy system.
- Creating a big data pool to work with and train on common datasets for easier understanding and error identification.
- Redefining some rules regarding the certification of AI systems and specifically Autonomous Driving technologies.
- Ensuring that users are able to understand the general behaviour of AI systems, including their capabilities and limitations which requires increased data transparency.

Accessing expertise and attracting talents

Industry and academia should work together to identify the kind of expertise that is needed in a specific field, but also in order to define a basic level of AI-related knowledge needed for leadership positions as well as for the general public. In this context, knowledge management and especially sharing AI knowledge in the AI communities and initiatives, was identified as a key factor besides bridging the gap between the needs in industry and the training in academia through small courses and activities in universities. Furthermore, it was stated that it is becoming increasingly difficult to attract interdisciplinary AI researchers to industry and research across Europe's borders or to prevent them from leaving.

AI Training and Upskilling Programmes

The experts discussed specific needs for AI training and upskilling programmes and how these needs can be aligned with academic activities and doctoral programmes. To structure the discussion the topic was divided into three pillars: (1) AI technology for AI experts who build AI systems; (2) Other users who use AI systems to build other non-AI-systems; (3) People who simply use AI systems. The

discussion has shown that within the first pillar, scientists from various disciplines with problem-solving skills, for example physicists and engineers, start to work with AI, but they might need more training in AI. This is why AI or Data Science in general should be taught more in the various scientific courses in academic studies. This issue also became relevant in the second pillar where people need more skills on statistical knowledge and data handling. Therefore, Data Science and Statistics should play an important role in all curricula. To ease the use of AI for non-AI researchers, the solution could be to build (modular) frameworks in order to reduce the complexity of AI models. Although some frameworks exist that are widely used, they seem to be too complex for other disciplines and too much tailored to Computer Science. In general, technology needs to move towards the average user (pillar 3), and users need to be able to "understand" the general behaviour of an AI system, including the capabilities and limitations of those systems.

AI for Future Healthcare

Healthcare is an area with many facets and perspectives, accordingly the discussions in the TDW covered a broad range of topics from diagnostics to precision and personalised medicine, genomics, bioinformatics, as well as infodemics and more consumer-oriented healthcare solutions. Within these discussions, the following sector-specific challenges in particular emerged:

AI and genomics

Within this specific area, reliable AI techniques and their support to bioinformatics in clinical diagnosis were discussed. Some challenges and ideas that were specifically identified are:

- **New protocols and standards** are needed for the collection of information which also address the polyhedral aspects of health and its diseases.
- **Quality controls** should be implemented to give a minimum reliability of the information collected. This includes software as well, thus requiring AI and ML predictive approaches to adhere to some essential quality principles (e.g., DOME recommendations[66]).
- It is very important to also collect information about negative results on clinical experimentation in order to **avoid undesirable biases** in AI systems.
- **Transfer learning** technologies seem to be able to provide successful scenarios in multiple health application domains. These techniques together with a combination of data-driven and model-based approaches are seen as an adequate framework, especially in the field of genomic medicine.
- In the field of genomics and health research, systems with **black boxes** have proven their usefulness, however in the case of clinical practice, **fully explainable AI systems** are required for their application. Therefore, working in the field of genomics and health should tend to produce AI systems that explain the relationship between genotypes and phenotypes. The

[66] https://www.nature.com/articles/s41592-021-01205-4#Bib1

balance between usefulness and explainability, however, should be modulated depending on the task.

- Existence of biases in **AI systems applied to genomics and health**: Due to the biological nature of genomic information, some of these biases are far from being undesirable. Accordingly, there is a need to study and control the introduction of biases in AI systems in this application area.

Analysis of health data stored across different stakeholders and/or borders

This should, for instance, avoid the transfer or exchange of data and ensure increased security. In particular, **federated learning approaches** were discussed by the experts in order to address this challenge. In general, two main types of use cases were identified, which are vastly different in their requirements and therefore solution approaches:

- **Use cases featuring a few big data silos** (e.g., hospitals), which are always online and are endowed with sufficient computing power.
- **Use cases where many small devices** (like wearables or smart phones) **are the data sources** but are only sometimes online and with very limited computing capacity. The following challenges and ideas were outlined by the experts in order to address these two use cases:
- A fundamental issue for all use cases is the **legal question of data security and anonymity**. How can these be guaranteed when using federated learning? As a possible solution, **a binding legal framework** in which to conduct federated learning was discussed by the experts.
- For successful federated learning, a high communication load is required which is difficult for low battery distributed devices with low connectivity. How to choose the best trade-off leads to a **multi-objective optimization problem**, which is hard to solve practically. Making use of sparsification and quantification strategies or a combination thereof could be a way to handle this challenge.
- An additional point is the **problem of devices dropping on and off** due to individual user behaviour. This causes a **data bias**, because some devices deliver much more data than others, which in turn can skew the federated learning. A solution could be the use of **asynchronous training** when building the federated learning models.
- Assuming that such a model was trained successfully, how to evaluate it, since the original data on which it was trained is no longer available? This is especially difficult for use cases with many devices (wearables for example), where data is not stored for a longer period of time.

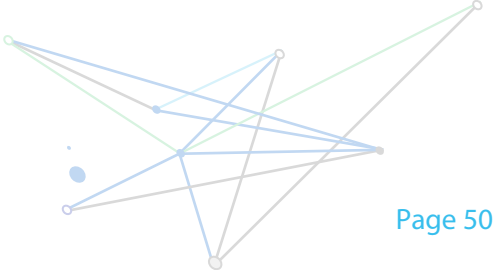
Learning models that are aware of, and consistent with, biomedical concepts and knowledge

This was identified as a key research challenge to enable a widespread adoption of AI-based solutions in the life sciences. Three main areas have been identified throughout the discussions, which are of specific interest for TAILOR:

- Learning-reasoning integration would **strengthen trust** of the life science community towards the use of data-driven methods and enhance self-explicability of the models, e.g., by having the model provide interpretations rooted on well-understood biomedical concepts.
- Learning-reasoning integration seems to be fundamental to surpass limitations of purely data-driven methods, such as machine learning and deep learning models, in **unfavourable conditions such as data scarcity**. In this respect, the experts have identified **rare diseases** as a relevant challenge which can highly benefit from an integrated approach capable of fusing symbolic knowledge, available under the form of knowledge graphs and interactomes, with high-dimensional / small-sample-size data.
- On a methodological level, the experts identified the research field of learning from complex data structures as a key enabler to effectively pursue the integration. Additionally, it is also advised to carefully investigate and consider the role of bias in knowledge representation, and how this can affect black-box systems that integrate such knowledge.

Apart from this, also some **key enabling factors for Trustworthy AI** in the healthcare sector were identified on a more general level:

- **Standards as one of the most important aspects for Trustworthy AI**. Increased efforts in the development of such standards were identified as necessary, including incentives that promote the involvement of all relevant stakeholders.
- **Explainability of the AI models** as a crucial element for gaining the trust of all stakeholders, as well as a likely requirement for regulatory compliance. In this respect, significant support is required in the development of methods that allow opening of the black-box. The experts pointed out that it is important to keep in mind that not all AI systems for healthcare rely on deep-learning techniques and approaches involving hybrid AI may play an important role in the future.
- **NLP** was identified **as an enabling technology**, meaning that it is at the core of a wide range of interaction scenarios within a healthcare environment. It can support making AI solutions explainable, e.g., through dialogue, although **ambiguity and context understanding** are some **very challenging aspects for NLP applications** in this sector.
- The availability of adequate infrastructure for the conception, development, and validation of AI systems.



- Medical data is very sensitive and exploiting that data to its maximum use will inevitably create tensions between values. **Privacy Enhancing Technologies** are a technological development that can alleviate some of the problems and these need to be developed further. Another important related concept is **sovereignty**, meaning that the involved persons should be given good opportunities for consent management.
- In **infodemics** – defined as overabundance of information, not necessarily reliable, circulating online and offline about an epidemic outbreak – it would be of great interest for future research to combine insights from **interdisciplinary disciplines** like Computational Social Science, Behavioural Neuroscience and Complexity Science to achieve better results and a broader insight. In this context, the experts stated that **Trustworthy AI** will play a major role in this area and will be key to fight infodemics.

Recommendations

Trust requires many system aspects to play together, such as transparency, fairness, accountability, robustness, and accuracy. These aspects need to be considered in a given socio-technical context and bound to a specific purpose. Trust emerges from experience in a human-centred ecosystem, which includes legal frameworks and the coordination among many stakeholders. As a concluding chapter of the SRIR we provide some recommendations concerning the most important research and innovation directions that need to be explored and investigated substantially in the short and long term.

Recommendations related to measure and assess Trustworthy AI dimensions

SHORT TERM

- Develop methods for measuring and evaluating the trustworthiness of AI systems.

LONG TERM

- Develop tools for continuously auditing and adapting Trustworthy AI systems: monitoring, dynamically identifying issues, and mitigating them.

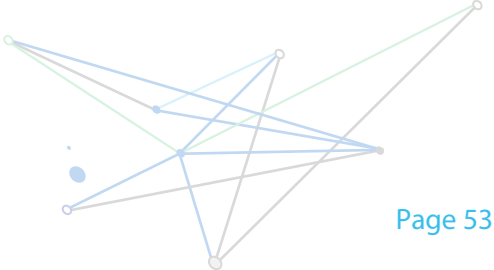
Recommendations related to scientific challenges

SHORT TERM

- Develop human interpretable formalisms to enable synergistic collaboration between humans and machines w.r.t. the criteria of being explainable, safe, robust, fair, accountable; and develop standards and metrics to quantify the grade to which these criteria are satisfied.
- Develop methods for integrating model-based and data-driven approaches to autonomous acting.
- Develop a broad range of AutoAI benchmarks to facilitate development and critical assessment of AutoAI techniques and systems.
- Expand current AutoAI techniques to better meet the demands of real-world applications, including multiple interacting design objectives (with aspects of trustworthiness), scalability, scope and ease of use.
- Develop integrated representations and frameworks for learning, reasoning and optimisation based on probability, logic, neural networks, ontologies, knowledge graphs and constraints.

LONG TERM

- 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.
- Develop a computational theory of mind that considers mental attitudes such as beliefs, knowledge, goals, intentions, capabilities, emotions, and integrates them in a computational effective fashion into autonomous acting.
- Enable the broad, safe, and efficient use of AutoAI techniques across all sectors of industry and society, especially in contexts where limited AI expertise is available (SMEs, public administration, …).
- Develop a unifying theory and framework of learning, reasoning and optimisation that that bridges the gap between the data- and knowledge-driven and the symbolic and subsymbolic approaches in AI.



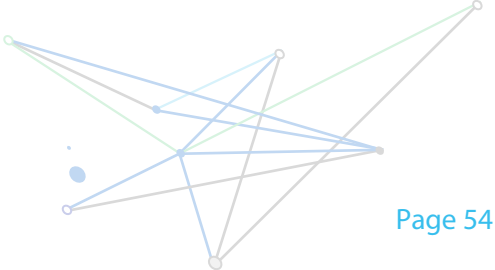
Recommendations related to innovation challenges

SHORT TERM

- Develop generic operational models of hybrid approaches allowing their reuse in various domains and propose metrics/benchmarks for validating these models.
- Consider that transparency (incl. explainability) targets different kinds of users: developers, domain experts, regulators, "users" (citizens, patients, etc.).

LONG TERM

- 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.



Conclusion

This is the first version of the living TAILOR Strategic Research and Innovation Roadmap. It should be seen as a snapshot of the current situation, focused on the integration of learning, optimization and reasoning to improve the trustworthiness of AI systems. It should, and will, be further developed throughout the project, and hopefully beyond.

AI is advancing fast, and some prospective directions sketched here will become state-of-the-art in the near future, raising more open questions. The whole European AI community, not only TAILOR partners, will be invited in the continued development of the roadmap. More Theme Development Workshops will be organised by TAILOR which will further enrich the Impact section – while also, though more indirectly, impacting its evolution by unveiling and highlighting research directions that need to be explored, in the virtuous cycle tightening the connection between Research and Innovation.

This document should thus be seen as a starting point.

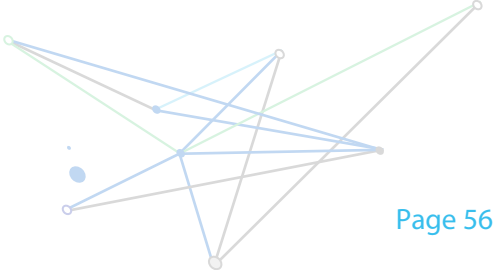
Appendix 1: REB & EREB members

The Roadmap Editorial Board members are:

- Michela Milano, Università di Bologna, Italy (Task 2.2 Leader)
- Marc Schoenauer, INRIA, France (WP2 Leader)
- Fredrik Heintz, Linköping University, Sweden (Project Coordinator)
- Silke Balzert-Walter, DFKI, Germany
- Kristian Kersting, Technische Universität Darmstadt, Germany
- Barry O'Sullivan, University College Cork, Ireland
- Philip Slusallek, DFKI, Germany

As each WP 3-8 involves activities related to the SRIR, an Extended REB (EREB) was formed, including the WP leaders and the task leaders of the SRIR-related tasks in these WPs

- Fosca Gianotti and Umberto Straccia, CNR, Italy
- Luis Galarraga, INRIA, France
- Luc De Raedt, KULeuven, Belgium
- Mehdi Ali and Jens Lehmann, Fraunhofer, Germany
- Giuseppe De Giacomo, Università degli Studi di Roma La Sapienza, Italy
- Andreas HERZIG, IRIT, CNRS, France
- Ana Paiva, IST-UL, Portugal
- Wico Mulder, TNO, Netherlands
- Hoger Hoos, U. Leiden, Netherlands
- Joaquin Vanschoren, TU/e, Eindhoven, Netherlands
- André Meyer-Vitali, DFKI, Germany



About the TAILOR project

The purpose of the EU Project TAILOR is to build the capacity to provide the scientific foundations for Trustworthy AI in Europe by developing a network of research excellence centres leveraging and combining learning, optimisation, and reasoning. These systems are meant to provide descriptive, predictive, and prescriptive systems integrating data-driven and knowledge-based approaches.

Artificial Intelligence (AI) has grown in the last ten years at an unprecedented pace. It has been applied to many industrial and service sectors, becoming ubiquitous in our everyday life. More and more often, AI systems are used to suggest decisions to human experts, to propose actions, and to provide predictions. Because these systems might influence our life and have a significant impact on the way we decide, they need to be trustworthy. How can a radiologist trust an AI system analysing medical images? How can a financial broker trust an AI system providing stock price predictions? How can a passenger trust a self-driving car?

These are fundamental questions that require deep analysis and fundamental research activity as well as a new generation of AI talents who are skilled in the scientific foundations of Trustworthy AI, who know how to assess, and how to design, trustworthy AI systems. Some of the current issues related to lack of trust in AI systems are a direct consequence of the massive use of black-box methods relying only on data. We need to define the foundations of a new generation of AI systems not only relying on data-driven approaches, but also on the whole set of AI techniques, including symbolic AI methods, optimization, reasoning, and planning.

Read more on www.tailor-network.eu