TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization

TAILOR

Grant Agreement Number 952215

D3.1 Research Challenges and Technological Gaps of Trustworthy AI Report (v1)

Document type (nature)	Report
Deliverable No	3.1
Work package number(s)	3
Date	Due M22, 30 June 2022
Responsible Beneficiary	CNR, ID 2
Responsible Author(s)	Umberto Straccia, Francesca Pratesi
Publicity level	Public
Short description	Research Challenges and Technological Gaps of Trustworthy AI v.1

History			
Revision	**Date**	**Modification**	**Author**
1.0	2022-6-30	Initial version	Umberto Straccia

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Ana Paiva	8 / IST	30/06/2022
Koen van der Blom	7 / LEU	30/06/2022

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

Table of Contents

1 Summary

This deliverable illustrates the main research challenges the TAILOR project foresees for the near future to make AI systems trustworthy. To do so, we describe the challenges along the six well known dimensions of trustworthy AI:

1. Explainability;
2. Safety and robustness;
3. Fairness, equity and justice;
4. Accountability and reproducibility;
5. Respect for privacy;
6. Sustainability.

2 Contributors

The following people have been involved in the Deliverable:

Partner ID / Acronym	Name	Role

2 / CNR	Umberto Straccia	Coordination, contributor
2 / CNR	Francesca Pratesi	Contributor to the dimension “Explainable AI Systems”
43 / UPV	Jose Hernandez-Orallo	Contributor to the dimension “Safety and Robustness”
40 / UniPI	Salvatore Ruggieri	Contributor to the dimension “Fairness, Equity, and Justice by Design”
25 / TUD 41 / UGA	Luciano Cavalcante Siebert Marie-Christine Rousset	Contributor to the dimension “Accountability and Reproducibility by design” Contributor to the dimension “Respect for Privacy”
4 / UCC	Andrea Visentin	Contributor to the dimension “Sustainability”

3 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 an intense research activity. In this deliverable, version 1, we point out to some challenges we believe to be of fundamental importance towards the development of AI systems that are perceived by an agent, being it human or just another artificial system, as “trustworthy”[1].

4 Trustworthy AI Systems: Challenges

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)[2] 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”[3]. 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[4].

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 identified dimensions that concur to create a trustworthy AI system, like *the capability of being explainable, safe and robust, able*

[1] cf. trustworthy - worthy of confidence, Merriam Webster Dictionary, - that you can rely on to be good, honest, sincere, etc., Oxford Dictionary.

[2] https://artificialintelligenceact.eu

[3] Lewis, Michael, Katia Sycara, and Phillip Walker. "The role of trust in human-robot interaction." Foundations of trusted autonomy. Springer, Cham, 2018. 135-159

[4] Falcone, R., & Castelfranchi, C. (2001). Social trust: A cognitive approach. In Trust and deception in virtual societies (pp. 55-90). Springer, Dordrecht.

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).

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:

1. Assessing if an existing AI system is compliant with the guidelines
2. Repairing it in case it is not
3. 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 we believe need in-depth 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, work package meetings and, ultimately, from the SRIR deliverable D2.1.

4.1 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. The situation becomes even worse in case of so-called neuro-symbolic systems in which e.g., explanations about both a reasoning and learning decision process should be illustrated in a human interpretable way.

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 along the following 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 AI-based decision model (black box) aims to reconstruct its logic either by mimicking the opaque model with a transparent one (global approaches)[5] or by concentrating on the construction of a useful explanation (e.g., reasoning steps, feature relevance, factual and counterfactual) for a specific instance (local)[6].

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.);
- Metrics to quantify the grade of comprehensibility of an explanation for

[5] 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).

[6] 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)

humans (e.g., Fidelity, Stability, Minimality, Plausibility, Faithfulness, Actionability). These may need to take into account the research results from the HCI, DataVis, and Cognitive Sciences communities: e.g.,

  - benchmarking platforms (datasets, metrics and methods etc.) for creating a common ground for researchers on explanation from different domains;

- Investigate methods to design, develop, assess and enhance systems with the ultimate goal to create explainable neuro-symbolic AI systems, i.e. systems that are able to explain, in a human, or machine understandable way, the results of inference (e.g., deduction, abduction, induction, argumentation, causal, non-monotone, conditional, uncertain and vague reasoning, etc.) and learning for the integrated representations of symbolic and neural systems. The goal here is to provide explanations of learning-based decisions as well as the progressive acquisition of knowledge. A fundamental step is that of developing also knowledge representation formalisms that can naturally be coupled together with learning processes.

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 (see later on). In neuro-symbolic systems causality is mixed-up with the notion of causality coming from the knowledge representation and reasoning research area.

Beside the above-mentioned research topics, that are fundamental cornerstones to be addressed by the research community, we have identified open areas that may be important 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.

4.2 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[7]. 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)[8]. This is particularly relevant in safety-critical AI systems, such as those appearing in automotive and avionics; and
- Develop metrics to quantify the degree of saftyness and robustness.

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.

[7] J. Burden, J. Hernàndez-Orallo, S. Ó hÉigeartaigh, Negative Side Effects and AI Agent Indicators: Experiments in SafeLife, SafeAI@AAAI (2021)

[8] D. Hicks, Lessons from Philosophy of Science, IEEE Technology and Society Magazine (2018)

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 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 time-scales 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.

4.3 Fairness, equity and justice

Decisions are being increasingly (partly or fully) delegated to AI algorithms for a wide range of socially sensitive tasks. While the benefits of algorithmic-based decision making cannot be neglected, e.g., procedural regularity -same procedure applied to each data subject, automated decisions based on profiling or social sorting may be biased[9] for several reasons. Historical data may contain human (cognitive) bias and discriminatory practices that are endemic, to which the algorithms assign the status of general rules. AI algorithms may introduce new forms of bias[10], or reinforce existing biases because data about model's decisions become inputs in subsequent model construction (feedback loops). For instance, AI algorithms may wrongly interpret spurious correlations in data as causation, making predictions based on ungrounded reasons. Moreover, they pursue the optimization of quality metrics, such as accuracy of predictions, that favour precision over the majority of people against small groups. These risks are exacerbated by the fact that the AI/ML models are complex for human understanding, or not even intelligible (see Sect. 4.1).

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). Auditing AI aims to identify and address possible risks and impacts while ensuring robust and trustworthy accountability. Methods for auditing AI-based systems[11] 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. Specifically, understanding causal influences among variables is a fundamental tool for dealing with bias. Moreover, the choice of a quantitative measure of discrimination/fairness is a critical issue in this context, as more than 20 metrics have been proposed in the literature, and incompatibility results are established. Auditing is being addressed also at the levels

[9] E. Ntouts et al. Bias in data-driven artificial intelligence systems - an introductory survey. WIREs Data Mining Knowl. Discov., 10 (3), 2020.

[10] G. Alves, M. Amblard, F. Bernier, M. Couceiro, A. Napoli, Reducing Unintended Bias of ML Models on Tabular and Textual Data, DSAA 2021

[11] 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)

of internal governance mechanisms[12], conformity assessment[13], standardisation[14] and certification.

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 amplified when AI is deployed in complex sociotechnical systems. Mis-representation in the data, and how to address it, is still under-investigated in the scientific community. For instance, if gender is coded with a binary feature (male/female), then any further discrimination analysis is limited to contrasting only such two groups, excluding non-binary people. There is then the need for a more elaborate representation of human identity in raw data, e.g, using ontologies for concept reasoning[15]. The issue of *source criticism*[16], which is central to historical and humanistic disciplines, is still in its infancy in the area of big data and AI. Source criticism attains the provenance, authenticity, and completeness of data collected, especially in social media platforms.

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

[12] J. Metcalf, E. Moss, E.A. Watkins, R. Singh, M. C. Elish. Algorithmic impact assessments and accountability:

The co-construction of impacts. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 735–746, 2021.

[13] J. M okander, M.Axente, F. Casolari, Luciano Floridi. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. Minds and Machines, pages 1–28, 2021.

[14] S. Nativi, S. De Nigris. AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework, EUR 30772 EN, Publications Office of the European Union, Luxembourg, 2021.

[15] C. A. Kronk, J. W. Dexheimer. Development of the gender, sex, and sexual orientation ontology: Evaluation and workflow. J. Am. Medical Informatics Assoc., 27(7):1110–1115, 2020.

[16] G. Koch, K. Kinder-Kurlanda. Source criticism of data platform logics on the internet. Historical Social Research, 45(3):270–287, 2020.

design some fairness constraints for a variety of tasks such as classification, recommendation, resource allocation or matching is missing. Very few scientific works[17], however, attempt at investigating the practical applicability of fairness in AI. This issue is challenging, and likely to require domain-specific approaches.

For what concerns the industrial and service sector, we have to consider the legal framework that has been put in place by the European 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

[17] K.Makhlouf, S. Zhioua, C. Palamidessi. On the applicability of machine learning fairness notions. SIGKDD Explor., 23(1):14–23, 2021.

an engineering problem. It requires a multi-disciplinary approach and critical viewpoints that AI professionals have not been taught yet.

4.4 Accountability and reproducibility

Whenever something goes wrong, there is often a call to define who is responsible for this wrongdoing. Responsibility in this sense refers to one's obligation to render an account of your actions and the consequences of these, i.e., accountability[18]. AI systems bring particular concerns with respect to accountability, since their applications can conceal broader organisational and societal processes, and the black box nature of many learning algorithms complicates the situation.

The governance of the design, development, and deployment of algorithmic systems takes into consideration all stakeholders and interactions with socio-technical systems. To enable accountability, it should acknowledge the fact that some societal problems are wicked, i.e. their formulation and possible solution depends on the viewpoint of those presenting them, and that when many stakeholders co-design a solution it might be often difficult, if not impossible, to pinpoint responsibility as afterthought. To develop accountable AI systems should be a design issue tackled from early on. Further, AI systems should be auditable and traceable. In other words, mechanisms must be put in place to ensure that AI systems and their outcomes, both before and after their development, deployment and use, can be observed and analysed.

In this context, we see accountability and reproducibility as interrelated concepts. Developing reproducible AI systems can enable accountability over AI systems. On the other hand, the process of record-tracking and logging for accountability can support an increasing level of reproducibility.

[18] 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

Reproducibility[19] 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 in 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. Further, as AI systems are deployed “in the wild”, human control and prediction over their behaviour are very difficult if not impossible, leading to so-called accountability gaps. Industry, governments, and academia need to work together to avoid AI-related harms to a human or group of humans, in the first instance, and the possibility to account for and attribute responsibility in any situations.

[19] 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)

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. A particular challenge is the trade-off between making every system that processes personal data accountable, while at the same time empowering individuals with private rights of action and other rights, like access and the right to object.

4.5 Respect for Privacy

Privacy is one of the first human rights that has been considered in legal frameworks for AI regulation. The General Data Protection Regulation (GDPR), in its Article 5, promotes *privacy by design* in the form of a certain number of general principles for ensuring *privacy as the default* in the whole chain of data processing for a given task. However, implementing those high-level principles raises scientific and technical challenges. In particular, we have to investigate new methodologies and approaches for:

- Defining formally and detecting automatically privacy risks raised by AI systems handling different kinds of personal data[20];
- Designing data anonymisation and attribute hiding algorithms that are robust to sophisticated attacks[21];
- Designing AI algorithms that respect by design privacy constraints[22].

When protecting personal data, we are faced with the dilemma of disclosing no sensitive data while learning useful information about a population. The way to handle this tension between privacy and utility differs according to the privacy models.

[20] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership Inference Attacks Against Machine Learning Models. IEEE Symposium on Security and Privacy (2017)

[21] F. Pratesi, L. Gabrielli, P. Cintia, A. Monreale, F. Giannotti, PRIMULE: Privacy risk mitigation for user profiles, Data & Knowledge Engineering 125 (2020)

[22] 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)

Differential privacy[23] is a prominent family of privacy-preserving data publishing models. A differentially private computation (of statistics or of machine learning models) limits the impact of any individual data on its output. Designing a computational function that satisfies differential privacy consists in carefully combining basic perturbation mechanisms and in demonstrating formally some privacy guarantees[24]. Differential privacy models all share this common intuitive goal but they differ in the way they formalize it - for example, on the quantification of the impact of an individual or on the tolerance to possible failures of the guarantees.

K-anonymity models[25] aim to make each individual so similar as to be indistinguishable from at least K-1 others. In *k*-anonymity techniques, strategies such as generalization and suppression are usually applied to reduce the granularity of representation of quasi-identifiers. Quasi-identifiers are attributes that can be linked to external information to re-identify the individual to whom the sensitive attributes refer. These methods guarantee privacy but also reduce the accuracy of applications on the transformed data. Indeed, one of the main challenges of *k*-anonymity is to find the minimum level of changes (in terms of generalization or suppression) that allows us to guarantee high privacy and good data precision. *K*-anonymity models can also be vulnerable in some cases. In particular, it is not safe against homogeneity attack and background knowledge attack. The homogeneity attack exploits a possible lack of variety in the sensitive attributes. In a background knowledge attack, an attacker knows information useful to associate some quasi-identifiers with some sensitive attributes, for example that certain diseases are more frequent in a specific gender.

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

[23] Cynthia Dwork. Differential privacy. In *ICALP*. 2006.

[24] Bing-Rong Lin and Daniel Kifer. Towards a systematic analysis of privacy definitions. *J. Priv. Confidentiality*, 2014.

[25] Latanya Sweeney. K-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, oct 2002

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

4.6 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*[26].

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. 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.

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

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 techniques for energy reduction and scale on edge devices, those models that can and should be run close to the data sources. In summary, the main challenges concern:

- the need of large EU infrastructures both for industries as well as for the research community
- the development of energy efficiency metrics to optimise AI system

for economic, environmental and scalability reasons.

5 Towards Trustworthy AI

To conclude, 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, measure, enhance systems that fully implement Trustworthy AI with the ultimate goal to create AI systems that incorporate trustworthiness by-design. The basic question is how to instil all these principles by-design and develop measures to quantify the degree of trustworthiness into the basic research themes to the aim of defining methodologies for designing and assessing Trustworthy AI.

6 Appendix: Questionnaire

To gather some feedback from TAILOR participants related to the topic of this deliverable we set-up a simple questionnaire with the following questions:

1. Did you address Trustworthy AI in your activity somewhat? Yes/No

2. With respect to the 6 dimensions of Trustworthy AI, which are the ones in which your work did contribute? (multiple selections allowed)

    a. Explainable AI Systems

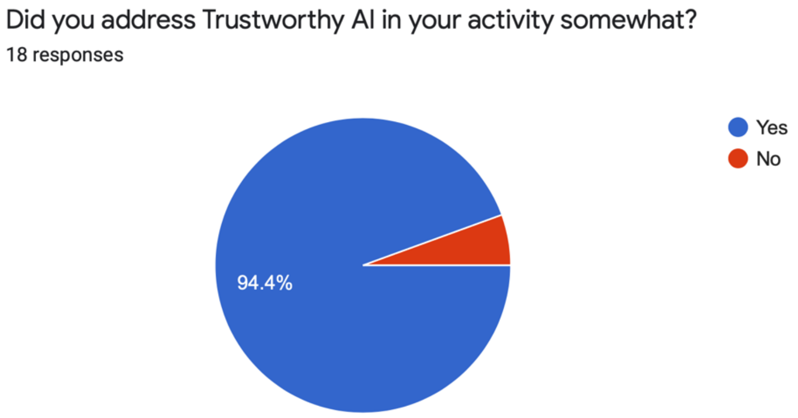
   b. Safety and Robustness
   c. Fairness, Equity, and Justice by Design
   d. Accountability and Reproducibility by Design
   e. Respect for Privacy
   f. Sustainability

3. If yes, could you please briefly describe the contribution(s)?

4. Are there open research problems or technological gaps related to Trustworthy AI that your work is going to address or you believe should be addressed in the near future?

However, we only got 18 answers at the time of writing.

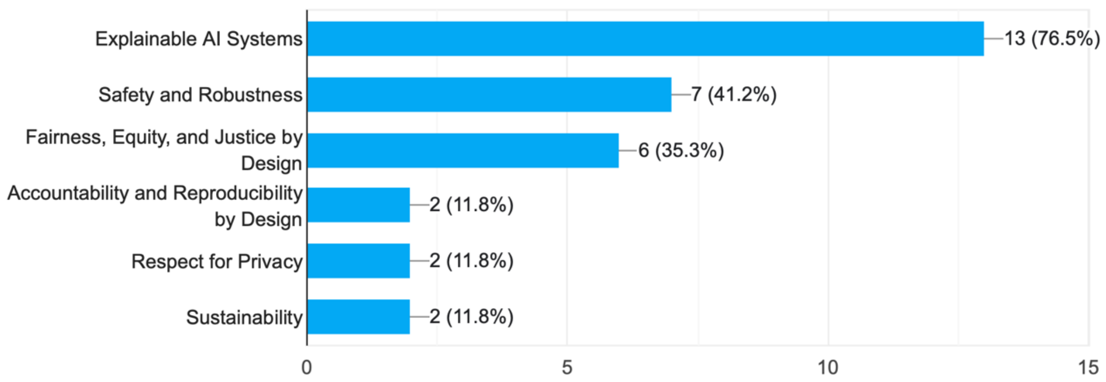
Concerning question 1, all, except one, responded yes.



Did you address Trustworthy AI in your activity somewhat?

18 responses

Concerning question 2, the distribution statistics is as follows.



Respect to the 6 dimensions of Trustworthy AI, which are the ones in which your work did contribute?

17 responses

Among those that responded, Explainable AI Systems is the more investigated one, which is not surprising.

Concerning question 3, the list of answers is as follows:

- complexity of explaining inconsistencies in combinatorial problems;
- Work on explainable agency; work on evaluation of trust.
- Producing appropriate explanations in the context of (neuro)symbolic AI.
- Design of Explainable AI approaches and of Interpretable by design Models. Usage of XAI approaches for bias detection and debiasing.
- We are working on certifying planning systems whose correct computations can be automatically independently validated.
- Explainable XGBoost; adversarial robustness; confidential privacy & homomorphic encryption
- We are currently working on the design of visual part detectors associated with a confidence measure that can be used to build more transparent classifiers.

- Explainable algorithms
- Applied research, tool development, dissemination
- We developed a framework for mitigating unintended biases in classifiers and masked language models.
- Explored explainability techniques in the context of Predictive Process Monitoring techniques and the usage of LTL based data encodings.
- I work on symbolic machine learning with Inductive Logic Programming. The models that are learned are explainable by construction.
- Combination of explainability and respect for privacy (given examples as explanation is not possible when there is potentially sensitive data in the training set). Methods to detect and mitigate biases, specifically in textual data.
- Research on techniques to forecast data centres workload to reduce overallocation (waste) of resources. Contribution to the handbook of trustworthy AI.
- Safe RL, value elicitation, value alignment.
- Creation CA on workload prediction. Research and supervision in Cloud Workload forecast.
- Applied research, tool development, dissemination.
- We provide a class of neural networks, called Logic Explained Networks (LENs), yielding a set of FOL rules as explanation for a classification task. LENs can be used either as a classifier or to explain the predictions of another black-box model.

Concerning question 4, the list of answers is as follows:

- We plan to include fairness into our prediction and accordingly into the resource allocation techniques.
- wide audience acceptability of AI
- Can we measure the level of explainability? Can we establish what mental advantages a good explanation has for the user? What do different forms of

explainability contribute to trustworthiness, and in what situation should you use which type of XAI?

- Make explainability real and not just a buzzword, make explainable techniques (1) solid, (2) robust, (3) scalable, (4) insightful, for a wide variety of data. Include Fairness, Equity, and Justice in recommender systems, especially the ones involving complex domains.
- Co-consideration of multiple TAI aspects.
- A better knowledge of which concepts can be explained in a compact/human-friendly way and which ones not.
- Yes, and we are currently developing two platforms: bias mitigation and analogical inference
- There are several open challenges in certifying planning systems, including certifying optimality of solutions/solution bounds and making certification more general and efficient.
- How to improve prototype-based approaches to reach performance on par with non-interpretable methods?
- Lifelong safe and robust learning.
- I believe Safety and Robustness are very important and should be addressed in the near future.
- Providing explanations for graph structured data and on text.
- Transdisciplinary research. How to better integrate different disciplines e.g. engineering, design, philosophy, psychology, social sciences and AI.
- How to measure trust. How to guarantee trust in H-AI interaction. How does explainability contribute to trust?
- Considering equality and fairness in workload prediction.
- Yes
- Heuristics to guide Formal methods tools, dedicated abstract domains, meaningful explanations (going beyond surface tools such as heat maps)

The number of answers is rather small. So, hardly we may infer some general conclusions from them. Nevertheless, we will devote to the questionnaire aspect much more dissemination effort in the final version 2 of this Deliverable (non-TAILOR mailing lists, TAILOR conferences, workshops ect.), with the hope to get some more useful insights from the answers: in fact, we believe that the questionnaire may be an additional means to gather feedback both from TAILOR and non-TAILOR participants.