TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization

TAILOR

Grant Agreement Number 952215

D3.2 Research Challenges and Technological Gaps of Trustworthy AI Report, v.2

Document type (nature)	Report
Deliverable No	3.2
Work package number(s)	3
Date	Due M46, 30 June 2024
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.2

History			
Revision	**Date**	**Modification**	**Author**
1.0 (D3.1)	30 June 2022	-	Umberto Straccia, Francesca Pratesi
2.0 (D3.2)	29 October 2024	update	Umberto Straccia, Francesca Pratesi

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Luc de Raedt	5 / KUL	10 September, 2024
Marc Schoenauer	3 / INRIA	25 September 2024

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 various dimensions of trustworthy AI, which have been reformulated according to the EU AI act, specifically, the High-Level Expert Group on Artificial Intelligence (AI HLEG)[1], as follows:

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-Discrimination, and Fairness
6. Societal and Environmental Wellbeing
7. Accountability and reproducibility.

2 Contributors

The following people have been involved in the Deliverable:

Partner ID / Acronym	Name
2/CNR	Umberto Straccia
2/CNR	Francesca Pratesi
43/UPV	Jose Hernandez-Orallo
40/UniPI	Salvatore Ruggieri
25/TUD	Luciano Cavalcante Siebert
41/UGA	Marie-Christine Rousset
4/UCC	Andrea Visentini

[1] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

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. In particular, the last few years, significant progress has been made especially on generative AI, where the most significant development is probably the fast and growing adoption of *Large Language Models* (LLMs) by companies and organisations. There is initial research showing that some fields can significantly improve their productivity by using these tools effectively. However, the limitations of today's generative AI systems influence their reliability in deployment – with much recent attention on 'hallucinations' from such systems, for example – and further work is needed to ensure their safe, trustable and reliable use. Moreover, the use of generative AI to create misinformation and disinformation has been of widespread concern in a year where many countries across the world are convening elections. The prospect of rapid adoption of these technologies has led to a growing concern about the impact on the labour market, where significant changes may come. As AI becomes more usable, there have also been a range of concerns expressed by policymakers and the public about their security implications, their impact on privacy, their interaction with current data rights, and the impact of their use on marginalised communities.

Generally, 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 2, 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, be it human or just another artificial system, as

"trustworthy"[2]. This version updates some of the challenges foreseen in the previous version, considers the contribution made in the Handbook for Trustworthy AI[3], and takes into account the EU AI Act[4] about Trustworthy AI, as illustrated in the following.

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 to evaluate alternative scenarios. It is therefore crucial that in *high-risk* applications (as outlined in the AI Act)[5] AI systems possess features that make them trustworthy, where trust indeed is a 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"[6]. 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[7].

4.1 The three pillars

According to the Ethics Guidelines for Trustworthy AI by the AI HLEG,[8] Trustworthy Artificial Intelligence (Trustworthy AI) has three components, which should be met throughout the system's entire life cycle. Indeed, it should be:

---

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

[3] https://tailor-network.eu/handbook/

[4] Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf (visited on 2024-04-23).

[5] https://artificialintelligenceact.eu

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

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

[8] High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. See also http://tailor.isti.cnr.it/handbookTAI/main/Ethical_Legal_Framework/HLEG.html

- **Lawful**, complying with all applicable laws and regulations
    - AI systems do not operate in a lawless world. A number of legally binding rules at European, national, and international levels already apply or are relevant to the development, deployment, and use of AI systems today
- **Ethical**, ensuring adherence to ethical principles and values
    - Achieving Trustworthy AI requires not only compliance with the law, which is only one of its three components. Laws are not always up to speed with technological developments, can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms
- **Robust**, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm
    - Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. Such systems should perform in a safe, secure, and reliable manner. Moreover, safeguards should be foreseen to prevent any unintended adverse impacts. It is, therefore, important to ensure that AI systems are robust

Each component is necessary but not sufficient for the achievement of Trustworthy AI. Ideally, all three components work in harmony and overlap in their operation. If, in practice, tensions arise between these components, society should endeavour to align them.

4.2 The four ethical principles

AI systems should improve individual and collective wellbeing. The four ethical principles[9], rooted in fundamental rights, must be respected to ensure that AI systems are developed, deployed and used in a trustworthy manner. They are

[9] See footnote 7.

specified as ethical imperatives, such that AI practitioners should always strive to adhere to them. These are the principles of:

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability.

In the following, we are going to shortly detail those four principles.

The principle of respect for human autonomy. Humans interacting with AI systems must be able to maintain full and effective self-determination over themselves and partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement, and empower human cognitive, social, and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunities for human choice. This means securing human oversight over work processes in AI systems.

The principle of prevention of harm. AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings, where harms are intended to be both individual or collective, and can include intangible harm to the social, cultural and political environment. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. They must be technically robust and it should be ensured that they are not open to malicious use. Vulnerable persons should receive greater attention and be included in the development, deployment, and use of AI systems. Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and

consumers, or governments and citizens. Preventing harm also entails consideration of the natural environment and all living beings.

The principle of fairness. The development, deployment, and use of AI systems must be fair. There are, of course, many different interpretations of fairness, but according to the Ethics Guidelines for Trustworthy AI, this dimension implies a commitment to ensuring equal and just distribution of both benefits and costs and ensuring that individuals and groups are free from unfair bias, discrimination, and stigmatisation. Equal opportunity in terms of access to education, goods, services, and technology should also be fostered. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends and consider carefully how to balance competing interests and objectives.

The principle of explicability. Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as ''black box'' algorithms and require special attention. In those circumstances, other explicability measures (e.g., traceability, auditability, and transparent communication on system capabilities) may be required, provided that the system respects fundamental rights.

4.3 The seven key requirements

The principles outlined before must be translated into concrete requirements to achieve Trustworthy AI. These requirements are applicable to different stakeholders partaking in AI systems' life cycle: developers, deployers, and end-users, as well as the broader society. By developers, we refer to those who research, design and/or develop AI systems. By deployers, we refer to public or private organisations that

use AI systems within their business processes and offer products and services to others. End-users are those engaging with the AI system, directly or indirectly. Finally, the broader society encompasses all others that are directly or indirectly affected by AI systems.

Different groups of stakeholders have different roles to play in ensuring that the requirements are met:

- Developers should implement and apply the requirements to design and development processes
- Deployers should ensure that the systems they use and the products and services they offer meet the requirements
- End-users and the broader society should be informed about these requirements and able to request that they be upheld.

The Guidelines identified a list of seven requirements and the challenge is to develop frameworks that aim at implementing them. The seven requirements are the following:

1. **Human agency and oversight**. It covers fundamental rights, human agency, and human oversight
2. **Technical robustness and safety**. It covers resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. **Privacy and data governance**. It covers respect for privacy, quality and integrity of data, and access to data
4. **Transparency**. It covers traceability, explainability and communication
5. **Diversity, non-discrimination and fairness**. It covers the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing**. It covers sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability**. It covers auditability, minimization and reporting of negative impact, trade-offs, and redress.

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).

While technology alone cannot deliver on all these characteristics, advances in the technical capabilities of AI systems can contribute in each of these areas, providing the foundation for trustworthy AI.

For industry, it is essential to understand how these dimensions translate in practice and boil down to technical requirements.

Therefore, 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 these seven dimensions and highlight some research directions and areas that have been collected by (1) interacting with the scientific work packages of TAILOR, (2) the TAILOR Handbook of Trustworthy AI and (3) consolidating the input derived from the TAILOR Joint SRIR V.2 deliverable D2.5.

4.3.1 Human agency and oversight

AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights and allow for human oversight.

There are three different challenges to be considered when we talk about this ethical dimension:

Fundamental Rights. Like many technologies, AI systems can enable and hamper fundamental rights. That is, given the reach, capacity, and opacity of many AI systems, they can negatively affect fundamental rights. In situations where such risks exist, a fundamental rights impact assessment should be undertaken, which should be done a priori of system development.

Human agency. Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals.

Human oversight. Human oversight should ensure that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a “human-in-the-loop” (HITL), “human-on-the-loop” (HOTL), or “human-in-command” (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system’s operation.

The interplay between human decision-making and AI systems give rise to research challenges where progress is needed to create AI systems that enhance human agency, safety and oversight requiring various capabilities, such as:

- The ability to interrogate how and why a recommendation has been made

- The nature of the uncertainty connected to that recommendation or output, and how that might affect confidence in the system's workings
- The impact of the decision on different user groups and the operating environment.

These capabilities and their integration into system design presumably will be active areas of research.

4.3.2 Technical robustness and safety

The safety of an AI system refers to the extent the system meets its intended functionality without producing any physical or psychological harm, especially to human beings, and by extension to other material or immaterial elements that may be valuable for humans, including the system itself. Safety must also cover the way and conditions in which the system ceases its operation, and the consequences of stopping. The term robustness emphasises that safety and —conditionally to it— functionality, must be preserved under harsh conditions, including unanticipated errors, exceptional situations, unintended or intended damage, manipulation or catastrophic states.

Given the increasing capabilities and widespread use of AI Systems, there is a growing concern about its risks, as humans are progressively replaced or sidelined from the decision loop of such systems.

The field of AI safety and robustness can be organised into the following seven groups of thematic challenges:

- **AI Safety Foundations**: This category covers several foundational concepts, characteristics and problems related to AI safety that need special consideration from a theoretical perspective. This includes concepts such as uncertainty, generality or value alignment, as well as characteristics such autonomy levels, safety criticality, types of human-machine and environment-machine interaction

- **Specification and Modelling**: The main scope of this category is on how to describe needs, designs and actual operating AI systems from different perspectives (technical concerns) and abstraction levels. This includes the specification and modelling of risk management properties (e.g., hazards, failures modes, mitigation measures), as well as safety-related requirements, training, behaviour or quality attributes in AI-based systems
- **Verification and Validation**: This category concerns design and implementation-time approaches to ensure that an AI-based system meets its requirements (verification) and behaves as expected (validation). The range of techniques may cover any formal/mathematical, model-based simulation or testing approach that provides evidence that an AI-based system satisfies its defined (safety) requirements and does not deviate from its intended behaviour and causes unintended consequences, even in extreme and unanticipated situations (robustness)
- **Runtime Monitoring and Enforcement**: The growing autonomy and learning capabilities of AI systems present significant challenges for their Verification and Validation (V&V), as it is difficult to gather sufficient epistemological evidence to guarantee their correctness. Runtime monitoring is useful to cover the gaps of design-time V&V by observing the internal states of a given system and its interactions with external entities, with the aim of determining system behaviour correctness or predicting potential risks. Enforcement deals with runtime mechanisms to self-adapt, optimise or reconfigure system behaviour with the aim of supporting fallback to a safe system state from the (anomalous) current state
- **Human-Machine Interaction**: As autonomy progressively substitutes cognitive human tasks, some kind of human-machine interaction issues become more critical, such as the loss of situational awareness or overconfidence. Other issues include:
    - Collaborative missions that need unambiguous communication to manage self-initiative to start or transfer tasks
    - Safety-critical situations in which earning and maintaining trust is essential at operational phases

  - Cooperative human-machine decision tasks where understanding machine decisions are crucial to validate safe autonomous actions
- **Process Assurance and Certification**: Process Assurance is the planned and systematic activities that assure system lifecycle processes conform to its requirements (including safety) and quality procedures. In our context, it covers the management of the different phases of AI Systems, including training and operational phases, the traceability of data and artefacts, and people. Certification implies a (legal) recognition that a system or process complies with industry standards and regulations to ensure it delivers its intended functions safely. Certification is challenged by the inscrutability of AI-based systems and the inability to ensure functional safety under uncertain and exceptional situations prior to its operation
- **Safety-related Ethics, Security and Privacy**: While these are quite large fields, we are interested in their intersection and dependencies with safety and robustness. Ethics becomes increasingly important as autonomy (with learning and adaptive abilities) involves the transfer of safety risks, responsibility, and liability, among others. AI-specific security and privacy issues must be considered regarding its impact on safety and robustness. For example, malicious adversarial attacks can be studied with focus on situations that compromise systems towards a dangerous situation.

Towards this end, we consider of paramount importance the development of

- Metrics to quantify the degree of safeness and robustness of AI systems, inclusive the development of specific benchmarks
- Methods that precisely assess *how often* and *how much* the system may fail and *when*, leveraging both on formal methods for verification/validation and ML techniques.

4.3.3 Privacy and Data Governance

Publishing datasets plays an essential role in open data research and in promoting transparency of government agencies. Unfortunately, the process of data publication can be highly risky as it may disclose individuals' sensitive information. Hence, a first step before publishing datasets is to remove any uniquely identifiable information from them. A strength of AI technologies is the ability to combine multiple, complex data sources and process large amounts of data to identify insights that would not otherwise be available. To deliver this function, AI systems may require access to data about individuals that contains personal or sensitive data; they may also generate such sensitive data by analysing and combining datasets that may individually not appear to contain information that would cause concern. These complex patterns of data use contribute to a wider socio-technical environment in which it is challenging for individuals to understand or exert control over what data about them is used and for what purpose.

Assessing carefully privacy risks before the publication of datasets is crucial. Detection of privacy breaches should come with explanations that can then be used to guide the choice of the appropriate anonymization mechanisms to mitigate the detected privacy risks.

The European Data Protection Board[10] (EDPB) has published several guidelines. The EDPB Guidelines on Data Protection Impact Assessment[11] focus on determining whether a processing operation is likely to result in a high risk to the data subject or not. It provides guidance on how to assess data protection risks and how to carry out a data protection risk assessment. *Data minimisation* is a strong recommendation to limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose, and to retain the data only for as long as is necessary to fulfil that purpose.

[10] https://www.edpb.europa.eu/edpb_en
[11] https://ec.europa.eu/newsroom/article29/items/611236

Unlike in many other areas of trustworthy AI, there exist widely accepted frameworks, differential privacy and k-anonymity, that provide formal privacy guarantees.

Technical advances can help alleviate these concerns. In particular, the following challenges are foreseen:

- Progress in data-efficient AI that enables new methods that can deliver accurate results without access to large datasets
- Privacy-preserving AI methods that demonstrate the ability to process data without revealing personal data
- AI methods that process data locally, and share only the pieces of information needed for a given application instead of raw data (e.g. via federated learning)
- The generation of high-quality synthetic data that may offer an alternative to accessing personal information in the creation of AI systems and methods to measure the usefulness and accuracy of the AI models learned from synthetic data (e.g. via generative AI)
- The development of formal guarantees, which are an important component of sustainable privacy solutions as they enable long-term anonymity
- Innovations in data and AI governance are needed to empower individuals and communities to set boundaries on the use of data about them
- New approaches are required to ensure data integrity and quality, especially for self-learning systems, to avoid malfunction or malicious function of AI systems.

4.3.4 Transparency

The transparency requirement is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.

According to the AI HLEG report, the transparency dimension is related to three different but related aspects: *traceability*, *explainability*, and *communication* (as we report as follows) and the research challenges concern the development of AI systems helping to implement them.

Traceability. The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.

Explainability. Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g., application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g., layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

Communication. AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system (this transparency obligation is guaranteed by the EU AI Act). This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction

in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations.

We believe that 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*: given an opaque AI-based decision model (so called black-box) aims to reconstruct its logic either by mimicking the opaque model with a transparent one (global approaches) or by concentrating on the construction of a useful explanation (e.g., reasoning steps, feature relevance, factual and counterfactual) for a specific instance (local)
- *Human Interpretability*: Human interpretable formalisms to habilitate collaboration between humans and machine, capable to express high-level explanations (logical, causal, knowledge graph, Natural Language) for encoding domain knowledge (and, thus, investigating knowledge representation and reasoning formalisms that can naturally be coupled together with learning processes), causal relationships in the data and/or identified by learning models, and methods for generating multimodal explanations (cross-modal/cross-language, factual and counterfactual etc.)
- *Explainable neuro-symbolic AI systems*: 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

- *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. Moreover, in neuro-symbolic systems causality is mixed-up with the notion of causality coming from the knowledge representation and reasoning research area
- *Metrics*: Metrics to quantify the grade of comprehensibility of an explanation for humans (e.g., Fidelity, Stability, Minimality, Plausibility, Faithfulness, Actionability), inclusive benchmarking datasets.

4.3.5 Diversity, non-discrimination, and fairness

Increasingly sophisticated algorithms from AI and Machine Learning (ML) support knowledge discovery from big data of human activity. They enable the extraction of patterns and profiles of human behaviour which can make extremely accurate predictions. Decisions are then being partly or fully delegated to such algorithms for a wide range of socially sensitive tasks: personnel selection and wages, credit scoring, criminal justice, assisted diagnosis in medicine, personalization in schooling, sentiment analysis in texts and images, people monitoring through facial recognition,

news recommendation, community building in social networks, dynamic pricing of services and products.

The benefits of algorithmic-based decision making cannot be neglected, e.g., procedural regularity – same procedure applied to each data subject. However, automated decisions based on profiling or social sorting may be biassed 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. Also, the usage of AI/ML models reinforces such practices because data about model's decisions become inputs in subsequent model construction (feedback loops). Algorithms may wrongly interpret spurious correlations in data as causation, making predictions based on ungrounded reasons. Moreover, algorithms pursue the utilitarian optimization of quality metrics, such as accuracy of predictions, that favour precision over the majority of people against small groups. Finally, the technical process of designing and deploying algorithms is not yet mature and standardised. Rather, it is full of small and big decisions (sometimes, trial and error steps) that may hide bias, such as selecting non-representative data, performing overspecialization of the models, ignoring socio-technical impacts, or using models in deployment contexts they are not tested for. These risks are exacerbated by the fact that the AI/ML models are extremely large and complex for human understanding, or not even intelligible, sometimes they are based on randomness or time-dependent non-reproducible conditions.

Legal restrictions on automated decision-making are provided by the EU General Data Protection Regulation, which states (Article 22)[12] "the right not to be subject to a decision based solely on automated processing". Moreover, (Recital 71)[13] "in order to ensure fair and transparent processing in respect of the data subject […] the controller should use appropriate mathematical or statistical procedures […] to prevent, inter alia, discriminatory effects on natural persons".

The research challenges we ask for are to develop

[12] https://gdpr-info.eu/art-22-gdpr/
[13] https://gdpr-info.eu/recitals/no-71/

- Fair algorithms with the purpose of preventing biassed decisions in algorithmic decision making, possibly by adopting a Fairness-by-design approach
- Quantitative definitions (metrics) of fairness, by leveraging on those introduced in philosophy, economics, and machine learning, keeping in mind that the choice of a quantitative measure of discrimination/fairness is a critical issue (many metrics have already been proposed in the literature, and incompatibility results have been established among them). Hence, we need also methods that allow us to determine which fairness metrics are more appropriate for a given AI-System
- Auditing AI-based systems to discover cases of discrimination and to understand the reasons behind them and possible consequences, such as understanding causal influences among variables, inclusive methods that allow us to identify which segments of society the training data may reflect or exclude. This includes legal obligations for the large platforms to offer specific APIs with unlimited accesses for testing purposes, either by governmental agencies or by citizen-driven associations
- Methods that allow us to identify what are the main ethical harms or injustices that can be done in a particular context of an AI-System.

4.3.6 Societal and environmental wellbeing

Given the increasing capabilities and widespread use of artificial intelligence, there is a growing concern about its impact on the environment related to the carbon footprints and the power consumption needed for training, store and developing AI models and algorithms. There is a wide literature regarding the dangers of climate change and the need of modifying the habits of use of the technology by consumers and industries. Plans such as the European Green Deal promulgated by the European Commission has the aim to tackle climate change.

The challenge of AI research we demand for is

- To develop methods that may accelerate the efforts of protecting the planet with many applications such as the use of machine learning to optimise the energy consumption efficiency, reducing the CO2 emission, monitoring quality of the air, the water, the biodiversity changes, the vegetation, the forest cover, and preventing natural disasters, inclusive to control the environmental impact of the widening use of raw materials used for constructing digital devices. To this end, the development of energy efficiency metrics and benchmarks will certainly be beneficial
- To build models able to continuously monitor the effects of AI systems on individuals and groups (e.g. in terms of health, social relationships or agency) but also research on the impact of AI on society and democracy at large.
- To foster and leverage *Frugal AI*, which is about maximising efficiency while minimising resource consumption across all facets of AI systems (e.g., to develop increasingly complex AI systems with a smaller amount of data). It involves the design, development and deployment of AI systems that utilise minimal resources to efficiently achieve desired outcomes, such as environmental sustainability goals.

4.3.7 Accountability

Accountability and Reproducibility are two cornerstones of Trustworthy AI. Accountability requires mechanisms 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. This ability to review AI systems involves technical and organisational logging processes to enable investigators to draw the same conclusions from an experiment by following provided guidelines.

It is evident that accountability and reproducibility are 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.

Accountability requires the creation of mechanisms to assign responsibility for the use of AI in decision-making and to hold those responsible to account in the event of a failure of decision-making, especially where decisions have significant personal or social impacts.

Some challenges we believe that need to be addressed to develop accountable AI system include:

- *Auditability and decomposability of AI systems*: that is, the ability to interrogate how different sub-components, including physical components, in the system work, their contribution to a system output, and how different environments, models, data sources and sensing have influenced that output
- *Reproducibility of methods*: that is, 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*: that is, the production of corroborating results in a new study, having used the same experimental methods
- *Reproducibility of inference*: that is, the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study
- *Explainability and interpretability*: that is, the ability to generate accurate, reliable explanations of how and why different outputs have been produced, in line with the needs and interests of different domain users or AI experts
- *System design validation*: that is, methods to assess how AI methods are integrated into wider socio-technical systems for decision-making and physical interaction

- *Human-machine interactions*: that is, scrutiny of the interactions between human decision-making processes and AI systems, and how these are influenced by AI systems
- *Law and regulations*: that is, methods that allow us to address questions about assignment of liability and the regulation and certification of high-risk AI applications, alongside documentation to adhere to legal requirements, for example on key decisions throughout the system life cycle that can be used for auditing and assigning responsibility and liability.

4.4 Trade-offs and interactions

Trustworthy AI guidelines list several positive aims for AI systems, many of which are reproduced above. The different aims may be mutually contradictory and have a negative impact on system utility. For example, the transparency and explainability of an AI system may be in tension with attempts to increase the privacy of such systems.

More work is needed to understand conflicting interactions and learning costs to allow making informed decisions on which aims to prioritise to which degree in a given application.

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.