PUTTING SCIENCE INTO STANDARDS

Workshop on Data quality requirements for inclusive, non-biased and trustworthy AI

Online on 8 and 9 June 2022

CEN CENELEC

European Commission

PUTTING SCIENCE INTO STANDARDS

Workshop on Data quality requirements for inclusive, non-biased and trustworthy AI

Online on 8 and 9 June 2022

The European Commission’s Joint Research Centre (JRC) and the European Standards Organizations, CEN and CENELEC , carry out an annual ‘foresight on standardisation’ exercise under the Putting Science into Standards (PSIS) initiative.

The PSIS initiative aims at identifying emerging science and technology areas that could benefit from standardisation activities to enable innovation and enhance industrial competitiveness. Every year, CEN and CENELEC and JRC select a topic for a PSIS workshop from a variety of proposals made by JRC scientists.

PSIS workshops bring together regulators, scientific communities, industry partners and the standardisation community to map the standardisation needs arising from European and international initiatives and to translate them into proposed actions for the technical committees.

This year, we invite you to participate in the 2022 PSIS workshop, which will focus on the topic of “Data quality requirements for inclusive, non-biased, and trustworthy artificial intelligence”. The workshop will take place online on 8 and 9 June 2022.

WHAT IS THE ROLE OF ETHICS IN AI?

The EU’s approach to artificial intelligence (AI) focuses on excellence and trust, aiming to boost research and industrial capacity and ensure fundamental rights. In April 2021, the European Commission presented its proposal for a Regulation laying down harmonised rules on Artificial Intelligence (COM/2021/206 final, AI Act), including a set of requirements (including on data quality and bias) for high-risk AI systems. The high-level requirements of the AI Act will be supported by standards developed by the European Standardisation Organisations. Bias in existing state-of-the-art AI models has been widely proven, raising concerns on societal consequences. Researchers in academia and industry have proposed different methods to evaluate and mitigate bias present in the different AI components. However, to date there is still not a common agreed methodology. Given the relevance of data in AI, it is of crucial relevance to establish standardised mechanisms to measure and mitigate the different biases present in data.

The objectives of the workshop are:

- Presenting current and future needs and recommendations to address data biases and related ethical concerns in the context of AI and the future AI Act;
- Mapping of existing and missing standardization efforts;
- Developing guidelines in view of data quality standards for AI models; and
- Proposing and recommending steps to start or complement the process of drafting standards.

RELEVANCE OF STANDARDISATION OF ETHICAL ASPECTS IN ARTIFICIAL INTELLIGENCE

Standards and guidelines are needed as to what type of data must be used in creating AI models to ensure that potential biases are detected and mitigated. Ensuring that data used in AI models uphold quality standards that result in non-biased, inclusive AI systems will provide the right foundation onto which trustworthy AI can be further developed and employed to improve EU citizens' lives.

STATE OF PLAY

Technological advances in digital transformation have created a situation where the volume of information generated and shared is outpacing the ability of humans to review and use such information. Novel artificial intelligence (AI) technologies, such as machine learning models and big data analytical tools are making sense of this information and providing insights.

Technological developments in computing infrastructures and algorithms have led to ever-more powerful AI algorithms and AI applications, capable of revolutionising virtually every aspect of our lives. In several areas, better processing of information has enabled big advantages, including healthcare, law enforcement, finance, media and education.

These developments have led to a situation where AI models are becoming more complex, more accurate and more widely used than ever before. Developments in storage and processing capabilities provide the base for deep neural network models that are trained on vast data, reaching unprecedented accuracy on many automated tasks. Applications of AI models today range from well-known fields, such as recommender systems and search result ranking, to sensitive use cases, such as medical diagnostics, bank loan approval or CV filtering.

Concerns on potential biases in AI are growing, with important and immediate effects in our lives. Biases are prevalent in existing state-of-the-art models, raising concerns on societal consequences, in particular for AI-informed decisions in the fields of health, security, finance, recruitment and education. The High-Level Expert Group on AI, appointed by the European Commission in 2018, stated "Diversity, non-discrimination and fairness" as one of the seven key requirements for trustworthy AI systems. This includes the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.

While the complexity of AI deep learning models makes their inner workings challenging to explain, the scientific community and tech companies are responding to the criticisms and expanding efforts to alleviate the root causes of AI biases.

Proving that the data on which models are trained fulfil quality standards is of paramount importance to ensure inclusiveness, ethical use of AI and absence of bias.

With the Commission's proposal of the EU artificial intelligence act, there is a need to translate requirements into norms and guidance documents, with the expectation that data quality standards will lead to:

1. An appropriate assessment and consideration of data shortcomings, biases, their impact on AI systems' performance and consequent mitigation actions to take;
2. Standardised practices for data privacy and governance, favouring data sharing and data ecosystems;
3. Guidance on concrete processes and transparency methodologies to address ethical concerns related to data;

In response to the proposed AI act that calls for ensuring trustworthiness and mitigating ethical risks, the regulation of AI technologies is gaining momentum, along with demands on standardisation efforts. As part of the "Europe fit for the digital age" priority, the European Commission published a review of the coordinated plan on AI in 2021 (AI Act), aiming for Europe to become a hub of trustworthy, excellent AI. At the same time, other initiatives complement the building of solid rules for the digital age, such as the EU Cybersecurity Strategy, the Digital Services Act, the Digital Markets Act, and the Data Governance Act. The European Regulation on Data Governance proposes to boost data sharing and availability.

The European Standardisation Organisations CEN and CENELEC recognised the rapidly growing standardisation need and launched last year its Joint Technical Committee 21 'artificial intelligence' (JTC 21), responsible for the development and adoption of standards for AI and related data, as well as providing guidance to other technical committees concerned with AI. JTC 21 identifies and adopts international standards already available or under development from other organizations such as ISO/IEC JTC 1 and its subcommittees, e.g. subcommittee 42 on artificial intelligence established in 2017. Furthermore, JTC 21 focuses on producing standardization deliverables that address European market and societal needs, as well as underpinning EU legislation, policies, principles and values.

First steps towards developing a standardisation roadmap to support the rules of the future AI Act have been undertaken and mapping standardisation needs are underway.

STRUCTURE OF THE WORKSHOP

The workshop will take place during two days. On Day 1 it will start with a plenary session, with a presentation of the political and technical implications of the AI Act as well as an overview of other international initiatives, e.g. from UNESCO and OECD, focusing on data and AI, and its link with trustworthiness, human rights and democratic values.

Following the plenary the workshop will break in parallel sessions held in two blocks, the first block held on Day 1, and the second block held on Day 2. The parallel sessions have a technical character, involving prominent AI practitioners and researchers, who will summarise the current state of the art on data-driven techniques and tools to ensure the trustworthiness of AI systems, with a focus on fairness and transparency.

The summary from the parallel sessions will be presented in plenary, followed by a panel discussion on the way forward.

BLOCK I: Horizontal initiatives for data quality assessment and bias mitigation in research and industry

The first block focuses on horizontal data quality and transparency approaches coming from industry and academia, and covering different phases of an AI system's lifecycle (e.g. dataset building, model training, system deployment and post-market monitoring). These horizontal initiatives encompass data biases with respect to different groups considering gender, nationality, culture, language and disciplines. The two horizontal approaches addressed by the first block are:

- Creating and documenting datasets for AI

At present, with the imminent adoption of the AI Act, there is a need to bridge the gap between existing voluntary practices in terms of dataset creation and documentation and the requirements defined in the legal text. In this context, this parallel session aims at presenting current state-of-the-art approaches for trustworthy dataset documentation and reflecting on whether they could be leveraged for future standards development.

Data is the raw material needed to train, test and validate AI systems. Furthermore, data is central throughout the AI development lifecycle, including, among others, steps devoted to data preparation, curation, annotation, cleaning and sharing. The creation of high-quality datasets in terms of completeness, correctness, representativeness and preservation of privacy have a direct impact in the development of trustworthy AI systems. Recently, important initiatives for the creation and comprehensive documentation of datasets have shed light in the form of checklists (e.g. HLEG's ALTAI, Microsoft), methodologies and templates (e.g. "Datasheets for Datasets", "Data Nutrition Label") or software tools (e.g. IBM's "AI Fairness 360").

- Data quality and bias examination and mitigation in AI

This session will present the latest research on methodologies, tools and best practices in the area of trustworthy AI, with a focus on those in the area of data quality as well as bias examination and mitigation.

In scope are, for example, concrete techniques that concern data quality, such as data augmentation, weighting loss functions (e.g. based on demographics), blinding methods (e.g. from a protected variable), feedback loops, fine-tuning, federated learning, transfer learning, robustness, evaluation techniques and benchmarks, adversarial attacks, explainability, human oversight and post-market monitoring.

Even in the presence of strong data quality measures, unwanted biases can still be present at the training and deployment stages of the AI life-cycle, causing unintended and possibly unexpected and harmful outcomes. It is therefore important to remain vigilant about best practices for continuous bias examination and mitigation during algorithm training, deployment and operation.

At the training stage, biases may feed into the system through data manipulation and augmentation techniques, fine tuning steps, the definition of objective functions or the use of specific algorithmic techniques. Even at the deployment stage, AI systems remain vulnerable to biases through potential feedback loops, improper evaluation techniques and benchmarks, or adversarial attacks. Detecting and addressing bias and other data quality issues throughout the entire AI system lifecycle requires the adoption of specific best practices, e.g. related to fairness, transparency, explainability, robustness and the specification of sustainable monitoring and evaluation techniques.

BLOCK II: Data quality needs and practice in selected sectors

In order to avoid fragmentation of the Digital Single Market and ensure harmonisation of provisions on AI across different sectors, the AI Act sets outs a horizontal framework. This choice is thoroughly analysed in the accompanying Impact Assessment and it is justified, since typically the same technology is characterized by the same problems and risks to fundamental rights (e.g. autonomy, data dependency, opacity etc.), irrespective of whether the AI system is developed by a public or private entity and irrespective of the sector where the system is deployed. Following this approach, horizontal standards are needed to ensure a level playing field and to avoid inconsistencies in how the same AI applications are regulated. Within this horizontal infrastructure, it is important that certain risks which may be specific to sectors could be properly considered in the context of the standardization process supporting the future AI Act, so as to ensure that the development of horizontal standards serves well the needs of the different sectors. . Hence, the second block of parallel sessions of the workshop will focus on certain different sectors which are considered as being at high-risk under the AI Act:

- Medicine and healthcare
- Law enforcement
- Finance
- Education and employment
- Industrial automation and robotics

While not being classified as high-risk under the AI Act, the sector of media, including social media, content moderation and recommender systems will be also looked at, taking into account its societal relevance and the fact that it is being made object of some obligations under the AI Act and the Digital Service Act.

Advisory board members

- **Werner BAILER** (Austrian Standards, ASI)
- **Arne BERRE** (Standards Norway, SN)
- **Patrick BEZOMBES** (ISO/IEC, JTC 1/SC 42/AHG 5 AI standardization landscape and roadmap)
- **Thierry BOULANCE** (European Commission, DG CNECT)
- **Tatjana EVAS** (European Commission, DG CNECT)
- **Anders FRIIS-CHRISTENSEN** (European Commission, JRC)
- **Ashok GANESH** (CEN-CENELEC)
- **Chiara GIOVANNINI** (ANEC)
- **Emilia GOMEZ** (European Commission, JRC)
- **Sebastian HALLENSLEBEN** (CEN-CENELEC, JTC 21 AI)
- **Laurens HERNALSTEEN** (CEN-CENELEC, JTC 21 AI)
- **Kim Skov HILDING** (CEN-CENELEC, JTC 21 AI)
- **Filipe JONES MOURAO** (European Commission, DG CNECT)
- **Marijan KAMENJAŠEVIĆ** (Croatian Standards Institute, HZN)
- **Matthew KING** (European Commission, JRC)
- **Irina ORSSICH** (European Commission, DG CNECT)
- **Andrea RAFFAELLI** (Small Business Standards, SBS)
- **Salvatore SCALZO** (European Commission, CNECT)
- **Emilia TANTAR** (Small Business Standards, SBS)
- **Fabio TAUCER** (European Commission, JRC)

Management team

- **Alexandra BALAHUR** (JRC I.2 Foresight, Modelling, Behavioural Insights & Design for Policy)
- **Isabelle HUPONT TORRES** (JRC B.6 Digital Economy)
- **Andreas JENET** (JRC A.5 Scientific Development)
- **Philip MAURER** (CEN-CENELEC, Market Perspectives & Innovation)
- **Livia MIAN** (CEN-CENELEC, Market Perspectives & Innovation)
- **Maurizio SALVI** (JRC I.2 Foresight, Modelling, Behavioural Insights & Design for Policy)
- **Josep SOLER GARRIDO** (JRC B.6 Digital Economy
- **Songul TOLAN** (JRC B.6 Digital Economy)