Towards International Collaboration for Mindful, Holistic and Inclusive Validation of Artificial Intelligence

Birgitta Dresp, Roberto Zicari, Paola Aurucci, Geneviève Fieux-Castagnet, Gérald Santucci

Towards International Collaboration for Mindful, Holistic and Inclusive Validation of Artificial Intelligence

Birgitta Dresp-Langley[*], Roberto Zicari[+], Giovanni Sartor[#], Paola Aurucci[#], Genevieve Castagnet[^] and Gerald Santucci[@]

[*]Centre National de la Recherche Scientifique UMR 7357 CNRS and Strasbourg University, France
[+]University of Seoul, South Korea and Arcadia University of Applied Sciences, Helsinki Finland
[#]University of Bologna Law Department, Bologna, Italy
[^]Human Resources Directorate, SNCF, France
[@]Global Forum Association, Paris, France and European Commission, DG Communication Networks, Content and Technology
Brussels, Belgium

Abstract

This paper summarizes some of the major contributions from speakers at Tutorial 4 of the PFIA Summer School in Strasbourg France, which took place at Strasbourg University between July 3 and 7, 2023. The tutorial was organized by the Trustworthy AI Lab of the CNRS and Strasbourg University https://trustworthyai-lab.icube.unistra.fr/ and the different contributions, placed in the context of the AI Act of the European Commission, adopted by the European Parliament in June 2023, illustrate why human-centered approaches to validating Artificial Intelligence and its applications in context are needed, and suggest how this can be made possible by including all actors, from citizens to domain experts, and viewpoints from society science and industry in a long-term assessment loop for trustworthiness evaluation. Rules of good procedure must be found and established within the specific framework of a given field of application. At the same time, domain-related limitations and obstacles must be considered. The potential as well as major challenges of implementing holistic and inclusive procedures for evaluating trustworthy AI are brought forward.

Introduction

Artificial intelligence (AI) represents opportunities and threats for society. The aim of this tutorial is to introduce an international approach, co-constructed by scientists and members of the public in cooperation with the European Commission and its expert groups on AI, towards best practices to ensure an environment ethical, responsible, conscious and sustainable ("trustworthy" deployment) of AI and AI-based applications. In December 2022, the European Commission (EC) ratified its first version of the EU AI law following several years of extensive work and discussions, with regular meetings in Brussels and online during the Covid19 pandemic of the High Level Expert Groups on AI, which includes the Subgroup on AI and Citizen/Consumer Security of the Directorate General of Justice (DG-JUST). EU AI law aims at ensuring that AI is human-centered and trustworthy. Such a goal is reflected in the European approach to excellence and trust through rules and concrete actions and research projects. The European AI Act is the first to propose an attempt towards a transnational legal framework on AI. The international Z-inspection© initiative and affiliated labs emerged from this work. Speakers will explain the state of the art in international AI ethics and policy following almost five years of intensive collaborative work in European Commission expert groups. What is meant by a "conscious", "holistic" and "inclusive" approach to artificial intelligence in this context will be clarified. Implications for data protection, clinical research,

applications and commercial innovation will be brought forward in light of fundamental human rights and due process rules as stipulated in the EC texts. The following paragraphs provide a brief summary overview of the different contributions.

Summary of Contributions

Setting the stage for the tutorial, the first presentation by **Birgitta Dresp-Langley** summarizes the essentials on and around the European AI Act. Adopted by the European Parliament in June 2023, the European AI Act **[1]** is the European Commission's proposal for an international legal framework for Artificial Intelligence (AI) and its application(s) as a result of seven years of work meetings, in Brussels and online during the years of the Covid19 pandemic, of the European Commission's High Level Expert Group on AI (HLEGAI) and its subgroup Consumer Safety Network, AI and connected technologies (CSN-AI) appointed under the umbrella of the Directorate General for Justice (DG-JUST). Closely linked to concurrent, similar efforts of the European Commission towards working out legislation on all matters pertaining to the digital society (General Data Protection Regulation, Digital Safety Act, Cyber Security Act), the European AI Act stipulates ground conditions for trustworthy Artificial Intelligence. Beyond technical requirements relative to the transparency, adequacy, and parsimony of input data structures, neural network architectures, learning algorithms, and output data, trustworthiness of AI involves conditions and criteria beyond purely technical aspects. Such additional conditions and criteria are mandatory under the light of problems relative to the different levels of autonomy of AI systems. These have been defined by the Stockholm International Peace Research Institute in 2017 **[2]** in terms of three distinct levels

1- Human on the loop (human experts have full control over the system at all steps)
2- Human in the loop (human experts have control over some steps)
3- Autonomous (no human control over the system at any step)

On the basis of this fundamental distinction between AI systems as a function of the extent of human control they allow for, the HLEGAI worked out ***seven criteria for trustworthy AI*** that were first published by the EC in April 2019 in terms of an Assessment List for Trustworthy AI (ALTAI). Therein it is stipulated further that prior to assessing an AI system for these seven criteria, a Fundamental Rights Impact Assessment (FRIA) is to be performed to ensure that the AI system does not violate any of the rights stipulated in the Universal Declaration of Human Rights of the United Nations. Under this premise, it appears quite clearly that AI systems of the third type listed here above (autonomous, no human control over any step) are by their nature incompatible with the fundamental human right to freedom, and potentially others, depending on the field of implementation/application. The European AI Act of 2023 is entirely based on the criteria formulated in the ALTAI, and is challenged by several problems, which will be difficult to resolve. A first problem is that the EC will have to ensure that all secondary EU legislation pertaining to data protection (General Data Protection Regulations-GDPR), product safety (General Product Safety Regulations-GPSR), and digital technologies in general will be modified and adapted accordingly to be fully consistent with the AI Act, which has teeth but will not be able to bite if this issue cannot be resolved. A second problem is that of a clear ***definition of risks*** represented by different types of AI. There is currently no fine distinction between risk levels apart from some a coarse division into "high" and "low" risks, with "high risk" AI requiring further distinction between "unacceptable" and "acceptable" levels. "Unacceptable" risk up to now is defined in terms of high and unacceptable risk of injury or grievous bodily harm; the mental health risks associated with some AI-driven technology (social media and other interactive platforms), which are tangibly high and definitely unacceptable, partly under the premise of the fundamental human right to pursuit of happiness and freedom of choice, partly because they result from manipulative technology ("brainwashing"), are currently not taken into account by the legal framework of the EC for lack of a probability-based assessment method. Thus, although the AI Act is a first necessary

step in President von der Leyen's commitment to "legal certainty on AI and digital technology", it is currently severely challenged by the ***ethics-versus-law dilemma*** (Figure 1).

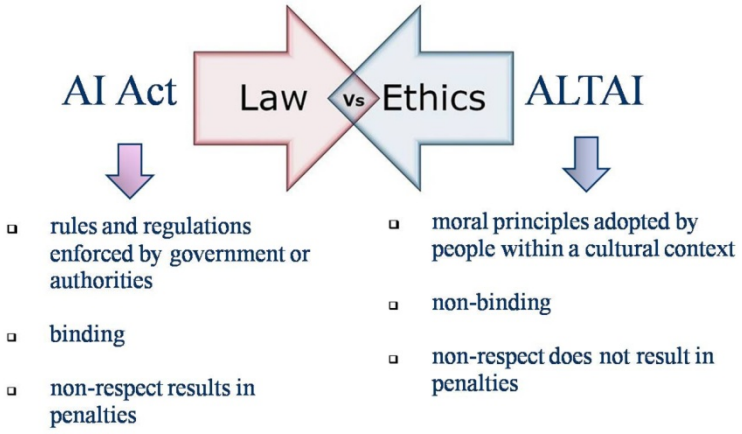

Figure 1: The Assessment List for Trustworthy AI (ALTAI) is based on universal human rights and Western-culture-centered moral principles while the European AI Act aims for legal certainty imposed by new EU rules and legislation for digital technologies and AI.

Under the light of this dilemma, it appears that a massive collective mobilization in science and society will be necessary to ensure that all AI is thoroughly screened for compatibility with fundamental human rights and trustworthiness criteria as stipulated in the ALTAI before it causes irreversible damage to citizens and society. The contribution by Roberto Zicari specifically addresses how trustworthy AI can be assessed in practice through international collaboration between research labs and stakeholders from government, industry, and the public. Insisting on the fact that ALTAI only provides a rather static checklist (that one may go through as when checking a product before labeling it "fit" or "unfit" for consumption), but does not validate clear claims, or take into account changes in such claims as they may occur in time as knowledge increases. In addition, the fact that the criteria in the ALTAI and the AI Act are not contextualized for a specific domain represents a limitation that needs to be overcome. To this end, the Z-inspection© initiative and the Trustworthy AI labs worldwide [3] mobilize international domain experts to evaluate AI systems in a participatory process that is to help all stakeholders assess the specific risks pertaining to a specific system. In several case studies, Z-inspection© has permitted to determine the trustworthiness of AI for cardiovascular risk detection, AI-based skin lesion classification for the early detection of skin cancer and precancerous lesion, and AI-based determination of the degree of compromised lung function in Covid19 patients. In a pilot project spanning over the years 2022-2023, the trustworthiness of automated tracking of natural landscapes through analysis of satellite imagery by AI has helped determine that the AI system under scrutiny, which is an important environmental monitoring tool for the Dutch government, passed the FRIA as well as all the self-assessment steps for the ALTAI criteria. In addition, since Z-inspection© is a collaborative and holistic approach rather than a mere checklist-based approach, bringing in different stakeholders from science, government and the public at different stages of the whole life-cycle of the AI system (design, development, testing/simulations, deployment, post-deployment monitoring) has also permitted to identify tensions relating to the AI systems. Such may exist between "winning" and "losing" aspects of the system for different stakeholders, between "short-term" and "long-term" effects of the system or goals, or between "local" and "global" consequences and effects engendered. Multi-domain stakeholder and expert interactions help identify such tensions and propose solutions beyond the limitation(s) of the static checklists provided in the FRIA and the ALTAI. The overarching goal of the Z-inspection© process under the premise of the seven criteria of the ALTAI (Figure

2) is to achieve a ***consensus-based mapping*** of the advantages and the drawbacks of an AI system, and to assess its trustworthiness under the light of best-case and worst-case scenarios and potentially arising tensions, for which a solution is proposed.

Figure 2: The Z-inspection© process is tailored for a multi-stakeholder consensus-based-mapping of domain-specific advantages and drawbacks of AI systems with respect to their trustworthiness placed under the premise of the seven criteria (stated in the text in the circles here) of the ALTAI. Source*: Ethics Guidelines for Trustworthy AI*. The High-Level Expert Group on Artificial Intelligence (HLEGAI) of the European Commission, April 08, 2019.

Specific issues relating to individual data protection laws in the case of AI-assisted or fully automated decision making in clinical trials and clinical research are then stressed in the contribution by Paola Aurucci and Giovanni Sartor. Although the power of AI as part of the care process [4] offers unprecedented opportunities for tackling the full complexity of multifactor disease processes, it poses concrete challenges to the General European Data Protection Regulations (GDPR). For the effective monitoring of clinical trials, patient and process safety, and post-market safety by automated processes and AI, the sharing of ***health data*** is essential. This entails communicating genetic, biometric and other individual data across experts, clinics, and studies, yet, these data are legally protected and belong in principle to the patient and should not be shared. The presentation points towards new solutions for sharing protected health data by a computational process that leads to ***pseudonymization***. It consists of replacing directly identifying data with non-identifying data. Unlike anonymization, pseudonymization is a reversible process and the data retain a personal character. As stipulated by the European Commission in July 2022, pseudonymized data remain subject to the GDPR. Complex legal issues pertaining to how pseudonymized data are used by AI to generate knowledge on the one hand and to assist medical decision making on the other are up to present not resolved. Each of these processes represents specific problems with regard to trustworthiness, relating to the way in which the data are to be structured and/or the way in which this affect an AI-assisted medical decision. Cloud-based analytics platforms for medical data already exist to enable data sharing across clinical cohorts, in Europe and beyond, challenging the ALTAI, the AI Act and the GDPR at various levels that will have to be clarified.

Thus, although the adoption of the AI Act by the European Parliament in June 2023 should pave the way for inter-institutional negotiations and test case studies world-wide, it still remains to be seen how secondary legislation, the GDPR and the General Product Safety Regulations (GPSR) being part of such, will ensure legal certainty in the first place. This specific problem is closely related to the problem of how to define and identify different levels of risk

associated with Artificial Intelligence. The contribution by Geneviève Fieux-Castagnet and Gérald Santucci argues for novel, trans-disciplinary and collaborative methods of risk assessment such as Z-inspection©, which will allow getting to the heart of this complex matter. Under the light of the points stressed in the previous contributions and the recognized need for large datasets, feed-back mechanisms, trans-disciplinary expertise, multi-stakeholder involvement and long-term evaluation procedures, the three-phase-dynamics of the Z-inspection© process are brought forward (Figure 3).

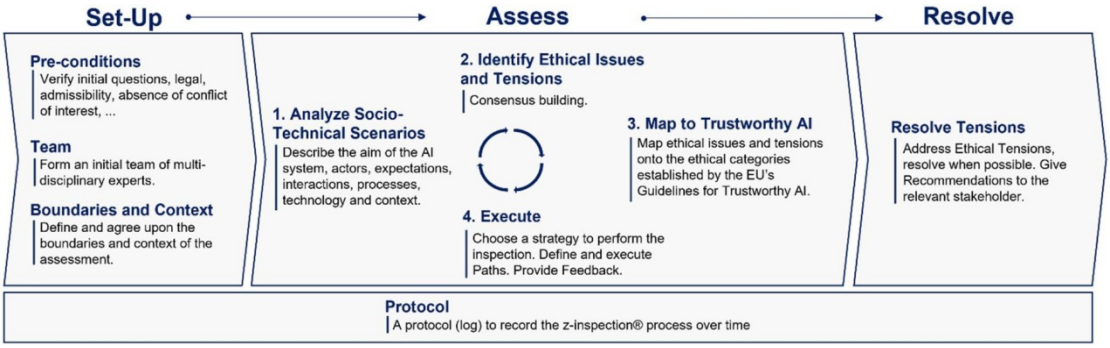

Figure 3: The three phases of the Z-inspection© process to ensure life-cycle specific long-term assessment of AI trustworthiness under the premise of the ALTAI. Each use case will be mapped and have a protocol (log) of the collaborative evaluation process as a function of domain-specific problems.

The presentation stresses the criticality of this moment in time, where international efforts towards collaboration between stakeholders and domain experts are needed to ensure that AI will be developed and deployed to the benefit, not the detriment of citizens and society [5]. Gérald Santucci insists that at this moment in time, where computing power will increase exponentially over the next decade and bring about a critical acceleration of change in technology, we need to ask profound moral questions relative to who we are and where we want to go with AI.

Conclusions

The ALTAI and the European AI Act represent an important step forward towards a European regulatory framework for AI and digital technologies. However, only ensuring full consistency of both with secondary EU legislation will provide legal certainty for AI in the digital society, President von der Leyen's ultimate goal. This remains a major challenge. Ultimately, this problem extends beyond Europe with the steady growth of global markets and world-wide use of AI in almost every domain of science and society. The contributions from this tutorial, briefly summarized here, make a clear case for the need of international collaboration between stakeholders and domain experts to master questions relative to whether or not we can trust a specific AI system in a specific use-case scenario. Z-inspection© and similar processes, including platforms that involve citizen stakeholders in this assessment [6], will need all the support they can get.

References

[1] The European AI Act https://artificialintelligenceact.eu/the-act/

[2] Boulanin V, Verbruggen M. Mapping the Development of Autonomy in Weapon Systems. *The Stockholm Peace Research Institute*, 2017. https://www.sipri.org/publications/2017/other-publications/mapping-development-autonomy-weapon-systems

[3] The Z-inspection© process and labs https://z-inspection.org/affiliated-labs/

[4] Hines PA, Herold R, Pinheiro L, Frias Z, Arlett P. Artificial intelligence in European medicines regulation. *Nat Rev Drug Discov,* 2023; 22(2):81-82. https://pubmed.ncbi.nlm.nih.gov/36411368/

[5] The "Oppenheimer moment" of AI https://www.nbcnews.com/news/us-news/movie-director-christopher-nolan-warns-ais-oppenheimer-moment-rcna95612

[6] The BIAS project https://www.biasproject.eu/