AI and Ethics
https://doi.org/10.1007/s43681-023-00397-z

PREFACE

AITA: AI trustworthiness assessment

AAAI spring symposium 2023

Bertrand Braunschweig[1] · Stefan Buijsman[2] · Faïcel Chamroukhi[1] · Fredrik Heintz[3] · Foutse Khomh[4] · Juliette Mattioli[5] · Maximilian Poretschkin[6]

The accelerated developments in the field of Artificial Intelligence (AI) hint at the need for considering "Trust" as a design principle rather than an option. Moreover, the design of AI-based critical systems, such as in avionics, mobility, defense, healthcare, finance, critical infrastructures, etc., requires proving their trustworthiness. Thus, AI-based critical systems must be assessed across many dimensions by different parties (regulators, developers, customers, reinsurance companies, and end-users) for different reasons. We can call it AI validation, monitoring, assessing, or auditing, but the fundamental concept in all cases is to make sure that the AI is performing well within its operational design domain. Such assessment begins from the early stages of development, including the definition of the specification requirements for the system, the analysis, the design, etc. Trust and trustworthiness assessment have to be considered at every phase of the system lifecycle, including sale and deployment, updates, maintenance, or int. It is expected that full trustworthiness in AI systems can only be established if the technical measures to establish trustworthiness are flanked by specifications for the governance and processes of organizations that use and develop AI. Application of Social Sciences and Humanities (SSH) methods and principles to handle human–AI interaction, and aid in the operationalisation of (ethical) values in the design and assessment, with important information provided on their actual impact on trust and trustworthiness is a key issue.

Thus, AI researchers and engineers are confronted with different levels of safety and security, different horizontal and vertical regulations, different (ethical) standards (including fairness, privacy), different homologation/certification processes, and different degrees of liability, which force them to examine a multitude of trade-offs and alternative solutions. In addition, they are struggling with values that need to be translated into concrete standards that can be used in assessment. Collaboration with SSH researchers to specify these standards is a central challenge to make sure that assessments also cover the normative/ethical aspects of trustworthiness.

To judge AI-based systems merely by the accuracy percentage is a highly misleading metric. In addition, the conventional methods for testing and validating software fall short and it is even difficult to measure test coverage in principle. Due to the multi-dimensional nature of trust and trustworthiness, one of the main issues we face is to establish objective attributes, such as accountability, accuracy, controllability, correctness, data quality, reliability, resilience, robustness, safety, security, transparency, explainability, fairness, privacy, etc., map them onto the AI processes and its lifecycle and provide methods and tools to assess them. Thus, this shines a light on quality requirements ("-ilities", or non-functional requirements) which appear particularly challenging in an AI system, although many of them can be considered in any critical system. Furthermore, beyond quality requirements, this can also encompass risk and process considerations. The expected attributes and the expected values for these attributes depend on contextual elements such as the level of criticality of the application, the application domain of the AI-based system, the expected use, the nature of the stakeholders involved, etc. This means that in some contexts, certain attributes will prevail, and other attributes

✉ Juliette Mattioli
juliette.mattioli@thalesgroup.com

1 IRT SystemX, Palaiseau, France

2 TU Delft, Delft, The Netherlands

3 Linköping University, Linköping, Sweden

4 Polytechnique de Montréal and Mila, Montréal, QC, Canada

5 Thales, Palaiseau, France

6 Fraunhofer IAIS, Sankt Augustin, Germany

Published online: 03 January 2024

may be added to the list. Clear specifications of the non-functional requirements will help clarify these conflicts and can also spur innovation that solves some of these conflicts, allowing us to fulfill more of them at the same time.

The goal of this symposium is to establish and grow a community of research and practitioners for AI trustworthiness assessment leveraged by AI sciences, system and software engineering, metrology, and Social Sciences and Humanities (SSH). This symposium aims to explore innovative approaches, metrics, and/or methods proposed by academia or industry, to "assess the trust and trustworthiness" of AI-based critical systems with a particular focus on (but not limited to) the following questions:

- How can we qualify datasets according to the expected trustworthy requirements of the resulting AI-based critical system?
- How to define appropriate quantitative performance indicators and generating test examples to feed into the AI (e.g., corner cases, synthetic data)?
- How can we characterize or evaluate AI systems according to their potential risks and vulnerabilities?
- How can non-functional requirements such as accountability and controllability be evaluated (quantitatively)?
- How could interpretability and explainability algorithms be evaluated from both technical and end-user perspectives?
- How do metrics of capability and generality, and the trade-offs with performance affect trust and/or trustworthiness?
- How can we define suitable processes and governance mechanisms in organizations that develop and deploy AI systems?
- How can we leverage pilot assessments to develop systematic evaluation techniques for AI trustworthiness?

1 AITA 2023 (March 27–29, 2023)

1.1 Trustworthiness measures

"Advances in Automatically Rating the Trustworthiness of Text Processing Services"—*Biplav Srivastava, Kausik Lakkaraju, Mariana Bernagozzi, Marco Valtorta*
"An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering"—*Juliette Mattioli, Henri Sohier, Agnes Delaborde, Kahina Amokrane-Ferka, AfefAwadid, Zakaria Chihani, Souhaiel Khalfaoui, Gabriel Pedroza*
"Assessing Systematic Weaknesses of DNNs using Counterfactuals"—*Sujan Sai Gannamaneni, Michael Mock, Maram Akila*
"Evaluating Trustworthiness of Decision Tree Learning Algorithms based on Equivalence Checking"—*Omer Nguena Timo, Tianqi Xiao, Florent Avellaneda, Yasir Malik, Stefan Bruda*
"Relative Effects of Positive and Negative Explanations on Satisfaction and Performance in Human-Agent Teams"—*Bryan Lavender, Sami Abuhaimed, Sandip Sen*

1.2 Performance indicators

"Neighborhood Sampling Confidence Metric for Object Detection"—*Christophe Gouguenheim, Ahmad Berjaoui*
"On the Evaluation of the Symbolic Knowledge Extracted from Black Boxes"—*Federico Sabbatini, Roberta Calegari*
"Real-time Weather Monitoring and Desnowification through Image Purification"—*Eliott Py, Elies Gherbi, Nelson Fernandez Pinto, Martin Gonzalez, Hatem Hajri*

1.3 Risks and vulnerabilities

"To Be Forgotten or To Be Fair: Unveiling Fairness Implications of Machine Unlearning Methods"—*Dawen Zhang, Shidong Pan, Thong Hoang, Zhenchang Xing, Mark Staples, Xiwei Xu, Lina Yao, Qinghua Lu, Liming Zhu*
"Protecting ownership rights of ML models using watermarking in the light of adversarial attacks"—*Katarzyna Kapusta, Lucas Mattioli, Boussad Addad, Mohammed Lansari*

1.4 Processes and governance

"Risk Assessment Using Ethical Dimensions"—*Alessio Tartaro, Enrico Panai, Mariangela Zoe Cocchiaro*
"Conformity Assessments under the EU AI Act General Approach"—*Eva Thelisson*
"Towards a safe MLOps Process for the Continuous Development and Safety Assurance of ML-based Systems in the Railway Domain"—*Marc Zeller, Thomas Waschulzik, Reiner Schmid, Claus Bahlmann*

1.5 Dataset qualification

"ECS—an Interactive Tool for Data Quality Assurance"—*Christian Sieberichs, Simon Geerkens, Alexander Braun, Thomas Waschulzik*
"QI^2—an Interactive Tool for Data Quality Assurance"—*Simon Geerkens, Christian Sieberichs, Alexander Braun, Thomas Waschulzik*

1.6 Poster session

"Using ScrutinAI for Visual Inspection of DNN Performance in a Medical Use Case"—*Rebekka Go¨rge, Elena Haedecke, Michael Mock*
"Conformal Prediction for Trustworthy Detection of Railway Signals"—*Léo Andéol, Thomas Fel, Florence De Grancey, Luca Mossina*
"Gender mobility in the labor market with skills-based matching models"—*Ajaya Adhikari, Steven Vethman, Daan Vos, Marc Lenz, Ioana Cocu, Ioannis Tolios, Cor J. Veenman*

2 Sponsors

2.1 Confiance.ai

Confiance.ai is the technological pillar of the Grand Défi "Securing, certifying and enhancing the reliability of systems based on artificial intelligence" launched by the Innovation Council. It is the largest technological research programme in the #AIforHumanity plan, which is designed to make France one of the leading countries in artificial intelligence (AI).

2.2 TAILOR network

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

2.3 IVADO

IVADO generates, stimulates, and supports initiatives in artificial intelligence (AI), by bringing together the research community, organizations, and institutions. Canada is among the leaders on the international AI map! In the Tortoise Global AI Index for the first quarter of 2022, the country ranks fourth in the world, and Québec ranks seventh for AI performance. Together with its partners and collaborators, and always at the service of society, IVADO mobilizes, so that knowledge is transformed into future solutions.

2.4 Zertifizierte KI

Together with the German Federal Office for Information Security (BSI) and the German Institute for Standardization (DIN) as well as other research partners, Fraunhofer IAIS is developing test procedures for the certification of artificial intelligence (AI) systems. The aim is to ensure technical reliability and responsible use of the technology. Industrial requirements are taken into account through the active involvement of numerous associated companies and organizations representing various industries, such as telecommunications, banking, insurance, chemicals, and trade.