AITA: AI trustworthiness assessment

Juliette Mattioli, Bertrand Braunschweig

DOI: 10.1002/aaai.12096

SYMPOSIUM REPORT

AITA: AI trustworthiness assessment

Juliette Mattioli | **Bertrand Braunschweig**

Correspondence
Juliette Mattioli.
Email: juliette.mattioli@thalesgroup.com

Juliette Mattioli is associated with Thales, France; and Bertrand Braunschweig is associated with Confiance.ai, France.

Abstract

We report about the first ever symposium on the assessment of AI trustworthiness, leading to the birth of a new research community on the matter.

The accelerated developments in the field of AI hint at the need for considering trustworthiness as a design principle in particular for AI-based critical systems. Thus, AI trustworthiness characterization is multi-dimensional and multi-criteria as assessed by different parties (regulators, developers, customers, (re-)insurance companies, end-users). Moreover, to improve trustworthiness in AI-based systems, assessment through process and product measures are needed covering algorithms and social-technical systems, taking into account context, usage, different levels of safety, security, reliability, robustness, explainability and transparency, horizontal and vertical regulations, (ethical) standards—including fairness, privacy, homologation/conformity assessment processes, and different degrees of accountability and liability.

This first AAAI symposium on AI Trustworthiness Assessment (AITA) was triggered by a group of initiatives on responsible and trustworthy AI that came together with the belief that in order to increase the trustworthiness of AI-based critical systems, one needs to be able to measure it: namely the French Confiance.ai industrial and academic program developing an engineering environment for trustworthy application in critical systems; the German Zertifizierte KI program developing the technological basis for certification of AI business applications; the Canadian IVADO institute, currently launching an initiative on similar topics; the Australian Responsible AI Network (RAIN) and operationalizing responsible AI initiative at CSIRO, and the TAILOR European network which puts trustworthiness at the core of its research activities.

During this 2.5 day workshop, 35 researchers and engineers from four continents attended to 18 enlightening presentations; exciting keynotes from Freddy Lecue (J.P. Morgan—USA) on explainable AI in finance, Stefan Wrobel (Fraunhofer IAIS—GE) on reliable AI algorithms and AI certified systems, Christophe Labreuche (Thales—FR) on multi-criteria decision aiding to support trustworthiness assessment, Elham Tabassi (NIST—USA) on the AI risk management framework and Maximilian Poretschkin (Fraunhofer IAIS—GE) on the Zertifizierte KI initiative; Liming Zhu (CSIRO's Data61 – AU) also gave a presentation on Australia's approach to responsible AI engineering.

Stefan Wrobel gave the conclusion during the SSS Plenary in front of participants from all parallel symposia: three domains of actions must be considered: algorithms, norms and regulations, systems, and context.

Some of the main findings emerging from the set of presentations and lively discussions that took place in the meeting room and in the nice facilities provided by the AAAI Spring Symposium (SSS) organization are as follows:

- Assessment is a critical enabling factor for improvement: "if you want to improve trustworthiness, you have to measure it".
- A holistic and systemic multidimensional view is needed since applications are context-dependent and addressing a variety of users and stakeholders;
- The assessment of trustworthiness is a combination of process-based and product-based characteristics. None of these two dimensions alone is sufficient;

- Checklists and qualitative approaches are only the beginning: even if more difficult, quantitative assessment of factors such as robustness, bias, safety, data quality and so on is indispensable;
- It is also critical to pay attention to the cost-benefit ratio of this quantitative assessment.

Conventional methods for testing and validating algorithms fall short due to the multi-dimensional nature of trustworthiness (accountability, accuracy, controllability, correctness, data quality, reliability, resilience, robustness, safety, security, transparency, explainability, fairness, privacy etc.). AI-based system design shines a light on quality requirements ("-ilities", or non-functional requirements) which appear particularly challenging. Beyond quality requirements, this can also encompass social-technical system 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, and so on. This means that in some contexts, certain attributes will prevail, and other attributes may be added to the list. Last but not least, full trustworthiness in AI systems can only be established if all technical activities to establish trustworthiness are flanked by regulations, norms and standards to support the governance, processes of organizations that use, develop and deploy AI.

This is the start of new research community that will develop over the years thanks to this first event supported by AAAI.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

How to cite this article: Mattioli, Juliette, and Bertrand Braunschweig. 2023. "AITA: AI trustworthiness assessment." *AI Magazine* 1–2. https://doi.org/10.1002/aaai.12096

AUTHOR BIOGRAPHIES

Juliette Mattioli, Juliette Mattioli is considered a reference inartificial intelligence not only within Thales but also in France. In 2017, shewas one of the five representatives of France at the G7 Innovators Conference,contributing to the issue of AI, member of the #FranceIA mission. Since 2019,she is President of the "Data Sciences & Artificial Intelligence" Hub of the Systematic Paris-Region competitiveness cluster. Recognizedfor her expertise in Hybrid AI, her excellent knowledge of industrial AIissues, she contributes in the field of algorithmic engineering with aparticular focus on trusted AI to accelerate the industrial deployment of AI-based solutions in critical systems. Juliette Mattioli is alsoco-author of a book with Michel Schmitt on mathematical morphology, haspublished numerous scientific papers and filed seven patents. She has also lednumerous R&D projects for Thales programs and European projects (FP6, FP7,H2020) and is now strongly involved in the French National Grant Challenge "Confiance.ai" dedicated to Trustworthy AI Engineering toward certification.

Bertrand Braunschweig, After a career as a researcher and project managerin simulation and AI in the field of energy, Bertrand Braunschweig headed the ICTdepartment of the ANR, two Inria research centers (Rennes then Saclay), produced the Inria white paper on AI and coordinated the research component of the national artificial intelligence program. He is now an independent consultant and provides scientific support to various organizations including as scientific coordinator of the Confiance.ai programme operated by the IRT System X.