Published: 09 April 2024
Accepted: 13 February 2024
Received: 06 February 2024

INTRODUCTION

Introduction to Special Issue on Trustworthy Artificial Intelligence

ROBERTA CALEGARI, University of Bologna, Bologna, BO, Italy

FOSCA GIANNOTTI, School Normal Superior of Pisa, Pisa, RM, Italy

FRANCESCA PRATESI, Italian National Research Council, Rome, RM, Italy

MICHELA MILANO, University of Bologna, Bologna, BO, Italy

Open Access Support provided by:

University of Bologna

School Normal Superior of Pisa

Italian National Research Council

Introduction to Special Issue on Trustworthy Artificial Intelligence

Trustworthy Artificial Intelligence (TAI) systems have become a priority for the European Union and have increased their importance worldwide. The European Commission has consulted a High-Level Expert Group that has delivered a document on Ethics Guidelines for Trustworthy AI to promote Trustworthy AI principles. TAI has three overarching components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations, (2) it should be ethical, ensuring adherence to ethical principles and values, and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm. Each component in itself is necessary but not sufficient for the achievement of TAI. 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. From a practical standpoint, these foundational principles manifest into various TAI dimensions, encompassing robustness, reproducibility, safety, transparency, explainability, diversity, non-discrimination, fairness, auditing, independent oversight, privacy, data governance, sustainability, and accountability.

This special issue was conceived with the purpose of soliciting surveys addressing at least one dimension of TAI, providing a comprehensive and reasoned overview of the current state of the art. Emphasis was placed on the review and comparison of methodologies addressing specific trustworthiness dimensions or exploring the intricate interplay and tensions between different dimensions.

The response to our call for papers was robust, yielding 106 submissions. After rigorous evaluation following the ACM manuscript review guidelines, 29 papers emerged as contributors to this special issue. These papers are thoughtfully grouped into two distinct issues, with the current issue featuring 14 selected papers.

The collection opens with "Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering" by Lu et al., introducing a pattern catalog derived from a Multivocal Literature Review, focusing on operationalizing responsible AI from a system perspective throughout the governance and engineering lifecycle.

Then, a compilation of reviews addressing the topic of trustworthiness within *specific domains* or fields is presented.

Perez-Cerrolaza et al.'s "Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey" concentrates on safety engineering dimensions in industrial and transportation domains, reviewing challenges and methods for developing AI-based safety-critical systems.

ACM Reference Format:
Roberta Calegari, Fosca Giannotti, Francesca Pratesi, and Michela Milano. 2024. Introduction to Special Issue on Trustworthy Artificial Intelligence. *ACM Comput. Surv.* 56, 7, Article 162 (April 2024), 3 pages. https://doi.org/10.1145/3649452

ACM 0360-0300/2024/04-ART162
https://doi.org/10.1145/3649452

Zhang et al.'s "Exploiting Blockchain to Make AI Trustworthy: A Software Development Lifecycle View" explores blockchain-based Trustworthy AI through a software development lifecycle lens, investigating trustworthy issues in data transparency, model transparency, and system deployment/use.

Conlon et al.'s "A Survey of Algorithmic Methods for Competency Self-Assessments in Human-Autonomy Teaming" discusses machine agents' capacity for self-assessment, particularly focusing on methods improving interactions within human-machine teams.

The subsequent articles delve into a more *specific trustworthiness requirement* discussion, including but not limited to fairness, explainability, and robustness.

The article "Fairness in Machine Learning: A Survey" by Caton and Haas endeavours to furnish a comprehensive overview of diverse paradigms and methodologies dedicated to enhancing the fairness of machine learning. The organization of these methodologies is structured within the widely accepted framework of pre-processing, in-processing, and post-processing methods, further subcategorized into 11 distinct method areas.

In a parallel pursuit of addressing fairness concerns, with a specific focus on datasets and an expanded discussion encompassing TAI dimensions such as explainability and reliability, "Trusting My Predictions: On the Value of Instance-Level Analysis" by Lorena et al. conducts a survey of recent advancements in characterizing and understanding the difficulty level of individual instances within a dataset, referred to as instance hardness level.

Delving into dataset issues, particularly in the realm of computer vision, the paper "A Survey of Dataset Refinement for Problems in Computer Vision Datasets" by Wan et al. offers a thorough and structured overview of recent advances in refining problematic computer vision datasets. The addressed issues include class imbalance, noisy labels, dataset bias, and high resource costs, all of which have the potential to impede model performance and diminish trustworthiness. Further examining data-related aspects, the paper "It is All About Data: A Survey on the Effects of Data on Adversarial Robustness" by Xiong et al. reviews a subset of adversarial robustness literature, concentrating on the investigation of properties within training data that influence model robustness under evasion attacks. The paper delineates key properties of data leading to adversarial vulnerability and discusses guidelines and techniques for fortifying adversarial robustness through improvements in data representation, learning procedures, and the estimation of robustness guarantees for specific datasets. Continuing the exploration of robustness, the paper "The Path to Defence: A Roadmap to Characterising Data Poisoning Attacks on Victim Models" by Chaalan et al. adopts an evaluative perspective on robustness, specifically through the lens of Data Poisoning Attacks (DPA). The paper provides a comprehensive summary of the latest research on DPAs and defenses, introducing a DPA characterizing model to investigate the dependency of adversary attacks on the victim model. On the same line, tackling robustness, the paper "Byzantine Machine Learning: A Primer" by Guerraoui et al. delves into the dimension of robustness within Trustworthy AI (TAI), presenting a primer on Byzantine machine learning. The authors systematically organize existing works, allowing readers to discern both the merits and limitations of state-of-the-art solutions and providing a clear trajectory for future advancements in this field.

In the context of data, with a focus on synthetic data generation, the article "Non-Imaging Medical Data Synthesis for Trustworthy AI: A Comprehensive Survey" by Xing et al. conducts a review of state-of-the-art data synthesis algorithms, particularly in the domain of non-imaging medical data, with the overarching aim of facilitating trustworthy AI in this specific domain.

The paper "Explainable Reinforcement Learning: A Survey and Comparative Review" by Milani et al. delves into the intricacies of explainability, aiming to illuminate the decision-making processes inherent in reinforcement learning (XRL). The authors propose a novel taxonomy for organizing the XRL literature, prioritizing its application in the reinforcement learning setting.

Similarly, in the domain of explainability but from a distinct perspective, the paper "A Survey on Graph Counterfactual Explanations: Definitions, Methods, Evaluation, and Research Challenges" by Prado-Romero et al. concentrates on the concept of Counterfactual Explanations for Graph Neural Networks. The paper introduces a taxonomy, a uniform notation, and benchmarking datasets and evaluation metrics to comprehensively explore this aspect.

Concluding the spectrum of discussions, the paper "Secure and Trustworthy Artificial Intelligence-Extended Reality (AI-XR) for Metaverses" by Qayyum et al. endeavors to scrutinize the security, privacy, and trustworthiness facets associated with various AI techniques employed in AI-XR (extended reality) metaverse applications.

Roberta Calegari
Alma Mater Studiorum–Universitá di Bologna, Bologna, Italy

Fosca Giannotti
Scuola Normale Superiore, Pisa, Italy

Francesca Pratesi
National Research Council, Pisa, Italy

Michela Milano
Alma Mater Studiorum–Universitá di Bologna, Bologna, Italy

Guest Editors

Received 6 February 2024; accepted 13 February 2024