Introduction to Special Issue on Trustworthy Artificial Intelligence (Part II)

Trustworthy Artificial Intelligence (**TAI**) systems have become a priority for the European Union and have increased worldwide importance. 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 elements work in harmony and overlap in their operation. If, in practice, tensions arise between these components, society should endeavor 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 to solicit 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 reviewing and comparing methodologies addressing specific trustworthiness dimensions or exploring the intricate interplay and tensions between different dimensions.

The response to our call for articles was robust, yielding 106 submissions. After rigorous evaluation following the ACM manuscript review guidelines, 28 articles emerged as contributors to this special issue. These articles are thoughtfully grouped into two distinct issues, with the current issue featuring 14 selected articles.

The collection begins with the article IMPACTS Homeostasis Trust Management System: Optimizing Trust in Human-AI Teams by Hou et al. This article provides a comprehensive review of the current state-of-the-art in trust research as it pertains to Human-AI Teaming. It summarizes the development of a Trust Management System, setting the groundwork for the broader understanding of trust in AI. Following this, the collection presents a series of reviews that address trustworthiness within specific requirements.

The first set of articles focuses on the requirement of *fairness*. In this spectrum, we find A Systematic Review of Fairness, Accountability, Transparency, and Ethics in Information Retrieval by Bernard and Balog. This review explores issues of fairness, accountability, transparency, and ethics in information retrieval systems, pointing out the lack of standard definitions and discussing

ACM Reference Format:
Roberta Calegari, Fosca Giannotti, Michela Milano, and Francesca Pratesi. 2025. Introduction to Special Issue on Trustworthy Artificial Intelligence (Part II). *ACM Comput. Surv.* 57, 6, Article 134 (February 2025), 3 pages. https://doi.org/10.1145/3711127

ACM 0360-0300/2025/02-ART134
https://doi.org/10.1145/3711127

unresolved challenges. Next, Preserving the Fairness Guarantees of Classifiers in Changing Environments: A Survey by Barrainkua et al. investigates fair classification in environments where data distributions shift. It offers a taxonomy of adaptive and robust approaches to maintain fairness despite changes in training and testing data and explores benchmarking methods, related fields, and future research directions. Two additional contributions address fairness in vision and language. Fairness in Deep Learning: A Survey on Vision and Language Research by Parraga et al. reviews debiasing methods for fairness-aware neural networks in vision and language tasks, highlighting their tendency to model biases and proposing a novel taxonomy to organize the existing literature. In Gender Bias in Natural Language Processing and Computer Vision: A Comparative Survey by Bartl et al., an interdisciplinary approach to detecting and mitigating gender bias in NLP, computer vision, and visual-linguistic AI models is examined. This survey emphasizes methodological adaptations across disciplines and the increasing integration of social science frameworks to align AI bias analytics with human values and address non-binary gender categories.

The collection then shifts to the requirement of *robustness*. The first article in this section, The Triangular Tradeoff between Robustness, Accuracy and Fairness in Deep Neural Networks: A Survey by Li and Li, delves into the intersection of fairness and robustness. It examines the tradeoffs among accuracy, robustness, and fairness in deep learning, exploring their origins, evolution, and influencing factors, while highlighting the challenges in achieving true intelligence in AI systems across these dimensions. Specific studies on robustness include A.I. Robustness: A Human-Centered Perspective on Technological Challenges and Opportunities by Tocchetti et al., which reviews recent progress in AI robustness. This article introduces taxonomies for methods addressing robustness in the machine learning pipeline, specific model architectures, and evaluation methodologies while discussing research gaps and opportunities and emphasizing the crucial role of human evaluation. Another specific contribution is Adversarial Robustness of Neural Networks From the Perspective of Lipschitz Calculus: A Survey by Zühlke and Kudenko, which explores adversarial robustness of neural networks through Lipschitz calculus. It covers estimation methods, regularization techniques, robustness guarantees, and their connection to generalization. Finally, Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance by Wei and Liu reviews techniques for ensuring trustworthy distributed AI. This article addresses robustness while also extending the discussion to privacy and governance issues, providing a taxonomy of countermeasures and broadening the scope to include other requirements.

The final requirement addressed in this part of the special issue is *explainability*. A article bridging robustness and explainability is Toward **Trustworthy Artificial Intelligence (TAI)** in the Context of Explainability and Robustness by Chander et al., which discusses the need for trustworthy AI by examining current technologies for explainability and robustness, addressing risks and uncertainties in AI development, and proposing future research directions to enhance AI's reliability, security, and alignment with human values. Specific contributions on explainability include Benchmarking Instance-Centric Counterfactual Algorithms for XAI: From White Box to Black Box by Moreira et al., which evaluates the impact of machine learning models on counterfactual explanation generation across different model types. The study shows minimal impact of model type, the inefficacy of proximity-based algorithms, the necessity of plausibility for meaningful evaluations, and the importance of counterfactual inspection to identify biases. Another article on the same topic, Causality for Trustworthy Artificial Intelligence: Status, Challenges and Perspectives by Rawal et al., provides a comprehensive survey of causal inference methods and frameworks, highlighting recent advancements in machine learning and AI, and detailing a taxonomy of approaches, evaluation techniques, and open challenges in understanding cause-and-effect relationships. Additionally, Toward a Privacy-Preserving Face Recognition System: A Survey of Leakages and Solutions by Laishram et al., presents a detailed review of privacy-preserving face

recognition methods using a six-level framework, addressing privacy concerns in face recognition technology and summarizing current research trends and challenges in mitigating privacy leakage. Finally, the collection concludes with a domain-specific article in healthcare, Explainable AI for Medical Data: Current Methods, Limitations, and Future Directions by Hossain et al. This article examines state-of-the-art explainable AI techniques designed to improve the transparency of deep neural networks in healthcare, discussing their strengths, limitations, evaluation metrics, and future research opportunities to enhance AI acceptance among medical professionals.

The guest editors extend their heartfelt thanks to all the authors who contributed articles to this special section. We are also profoundly appreciative of the referees for their thorough reviews and constructive feedback, which were delivered within tight deadlines. We hope that this collection of work will inspire further innovative research in the field of trustworthy AI.

Roberta Calegari
University of Bologna, Bologna, Italy
email: roberta.calegari@unibo.it

Fosca Giannotti
CNR Italian National Research Council, Pisa, Italy
email: fosca.giannotti@isti.cnr.it

Michela Milano
DISI, University of Bologna, Bologna, Italy
email: michela.milano@unibo.it

Francesca Pratesi
Istituto di Scienza e Tecnologie, Istituto di Scienza e Tecnologie, Pisa, Italy
email: francesca.pratesi@isti.cnr.it

Guest Editors

Received 10 December 2024; revised 10 December 2024; accepted 19 December 2024