Trustworthy AI—Part II

Riccardo Mariani, Nvidia
Francesca Rossi, T.J. Watson IBM Research Lab
Rita Cucchiara, Università di Modena e Reggio Emilia
Marco Pavone, Stanford University and Nvidia
Barnaby Simkin, Nvidia
Ansgar Koene, University of Nottingham
Jochen Papenbrock, Nvidia

Digital Object Identifier 10.1109/MC.2023.3253980
Date of current version: 3 May 2023

With the widespread use of artificial intelligence (AI) systems, trustworthiness is becoming relevant for several application fields. This introduction provides a summary of the articles contributing to the second part of this special issue on AI trustworthiness.

The first part of the Trustworthy AI special issue (the February 2023 issue) included contributions on trustworthy artificial intelligence (AI) principles such as verifiability, robustness, reliability, explainability, bias, and transparency. In this issue of *Computer*, contributions are focusing on the following AI principles and related application fields.

The "Documenting High-Risk AI: A European Regulatory Perspective"[A1] article discusses *transparency* obligations introduced in the AI Act, the recently proposed European regulatory framework for AI.[1] Specifically, the authors look at requirements for providers of high-risk AI systems in terms of the provision of information to users and technical documentation.

The "A Framework for Trustworthy AI in Credit Risk Management: Perspectives and Practices"[A2] article addresses *trustworthiness* for the

APPENDIX: RELATED ARTICLES

A1. I. Hupont, M. Micheli, B. Delipetrev, E. Gómez, and J. S. Garrido, "Documenting high-risk AI: A European regulatory perspective," *Computer*, vol. 56, no. 5, pp. 18–27, May 2023, doi: 10.1109/MC.2023.3235712.

A2. S. Mazumder, S. Dhar, and A. Asthana, "A framework for trustworthy AI in credit risk management: Perspectives and practices," *Computer*, vol. 56, no. 5, pp. 28–40, May 2023, doi: 10.1109/MC.2023.3236564.

A3. A. Brando, I. Serra, E. Mezzetti, F. J. Cazorla, J. Perez-Cerrolaza, and J. Abella, "On neural networks redundancy and diversity for their use in safety-critical systems," *Computer*, vol. 56, no. 5, pp. 41–50, May 2023, doi: 10.1109/MC.2023.3236523.

A4. S. Tariq, S. Jeon, and S. S. Woo, "Evaluating trustworthiness and racial bias in face recognition APIs using deepfakes," *Computer*, vol. 56, no. 5, pp. 51–61, May 2023, doi: 10.1109/MC.2023.3234978.

A5. J. R. Tong and T. X. Lee, "Trustworthy AI that engages humans as partners in teaching and learning," *Computer*, vol. 56, no. 5, pp. 62–73, May 2023, doi: 10.1109/MC.2023.3234517.

A6. E. D. Degefe, Y. D. Prabowo, K. Savani, and A. Sheetal, "Functional analogies increase trust in black-box AI systems among lay consumers: The case of GeNose C-19," *Computer*, vol. 56, no. 5, pp. 74–83, May 2023, doi: 10.1109/MC.2023.3235880.

A7. L. Migliorelli, S. Tiribelli, A. Cacciatore, B. Giovanola, E. Frontoni, and S. Moccia, "Accountable deep-learning-based vision systems for preterm infant monitoring," *Computer*, vol. 56, no. 5, pp. 84–93, May 2023, doi: 10.1109/MC.2023.3235987.

ABOUT THE AUTHORS

RICCARDO MARIANI is the vice president of industry safety at Nvidia, 57036 Porto Azzurro, Italy. He is responsible for developing cohesive safety strategies and cross-segment safety processes, architecture, and products that can be leveraged across Nvidia's artificial intelligence-based hardware and software platforms. Contact him at rmariani@nvidia.com.

FRANCESCA ROSSI is an IBM Fellow and the IBM AI Ethics Global Leader. She is based at the T.J. Watson IBM Research Lab, Yorktown Heights, NY 10598 USA. Her research interests focus on artificial intelligence (AI), with a special focus on constraint reasoning, preferences, multiagent systems, computational social choice, neuro-symbolic AI, cognitive architectures, and value alignment. Currently, she is the president of AAAI. Contact her at francesca.rossi2@ibm.com.

RITA CUCCHIARA is a professor at the University of Modena and Reggio Emilia, 41121 Modena, Italy, where she is the director of the Artificial Intelligence Research and Innovation Center and director of the European Labs of Learning and Intelligent Systems Unit. Contact her at rita.cucchiara@unimore.it.

MARCO PAVONE is an associate professor at the Department of Aeronautics and Astronautics at Stanford University, Stanford, CA 94305 USA, and the director of autonomous vehicle research at Nvidia. Contact him at pavone@stanford.edu.

BARNABY SIMKIN is a guest editor of this issue and is affiliated with Nvidia, 0623 Berlin, Germany, where he coordinates Nvidia's overall strategic engagement with regulatory and standards bodies and influences those technical requirements related to artificial intelligence, automated driving, machine learning, and virtual testing. Contact him at bsimkin@nvidia.com.

ANSGAR KOENE is a global artificial intelligence (AI) ethics and regulatory leader at Ernst & Young, 1000 Brussels, Belgium, where he supports the AI Lab's policy activities on trusted AI. He is also a senior research fellow at the Horizon Digital Economy Research Institute at the University of Nottingham, Nottingham, UK. Contact him at ansgar.koene@nottingham.ac.uk.

JOCHEN PAPENBROCK is the head of financial technology in the Europe, Middle East, and Africa region at Nvidia, 60305 Frankfurt, Germany. Contact him at jpapenbrock@nvidia.com.

specific case of credit risk management, in particular for three key tenets (usable, reliable, and transparent), and proposes a holistic approach that covers relevant concerns in practical implementations. The "On Neural Networks Redundancy and Diversity for Their Use in Safety-Critical Systems"[A3] article addresses *functional safety*, and specifically, how neural network-based safety functions can leverage redundancy and diversity to satisfy the requirements of existing safety standards such as IEC 61508[2] or ISO/IEC TR 5469.[3]

The "Evaluating Trustworthiness and Racial Bias in Face Recognition APIs Using Deepfakes"[A4] article addresses *fairness, trust,* and *racial bias* for the specific case of facial recognition. In fact, racial bias in web-based face recognition services can lead to inaccurate results, causing severe technical and social issues and widespread distrust in AI-based systems. The authors use deepfake generation methods to introduce small imperceptible changes to the real images to shift the racial class of predictions with critical findings.

The "Trustworthy AI That Engages Humans as Partners in Teaching and Learning"[A5] article focuses on educational AI applications for which human-in-the-loop approaches that partner with learners and educators in an AI-involved teaching process present an opportunity to build trust in educational AI applications by allowing them to participate actively in the process of educational decision making. The "Functional Analogies Increase Trust in Black-Box AI Systems Among Lay Consumers: The Case of GeNose C-19"[A6] article presents a case study of a black-box AI-based Covid-19 detection product, GeNose C-19, developed by the Indonesian government. The authors found that explaining how GeNose works using functional analogies increases both Indonesian and American lay

consumers' trust in GeNose. The "Accountable Deep-Learning-Based Vision Systems for Preterm Infant Monitoring"[A7] article discusses preterm infants' movement monitoring in neonatal intensive care units (NICUs) and proposes an ethical framework that highlights possible ethical risks (in particular, bias) in the design and use of deep learning-based vision systems for monitoring infants' movements in NICUs.

We would like to thank the authors of the seven articles in this issue for sharing their knowledge and experiences on how to improve the trustworthiness of AI systems. We also thank all the reviewers for helping us evaluate the articles and selecting those of high quality to be included in this theme issue.

REFERENCES

1. "Proposal for a regulation laying down harmonised rules on artificial intelligence," European Commission, Brussels, Belgium, 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence
2. *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems – Parts 1 to 7*, IEC 61508:2010, 2010.
3. *Artificial Intelligence — Functional Safety and AI Systems*, ISO/IEC CD TR 5469. [Online]. Available: https://www.iso.org/standard/81283.html

Digital Object Identifier 10.1109/MC.2023.3263955