Trustworthy AI—Part III

Riccardo Mariani, Nvidia

Francesca Rossi, T.J. Watson IBM Research Lab

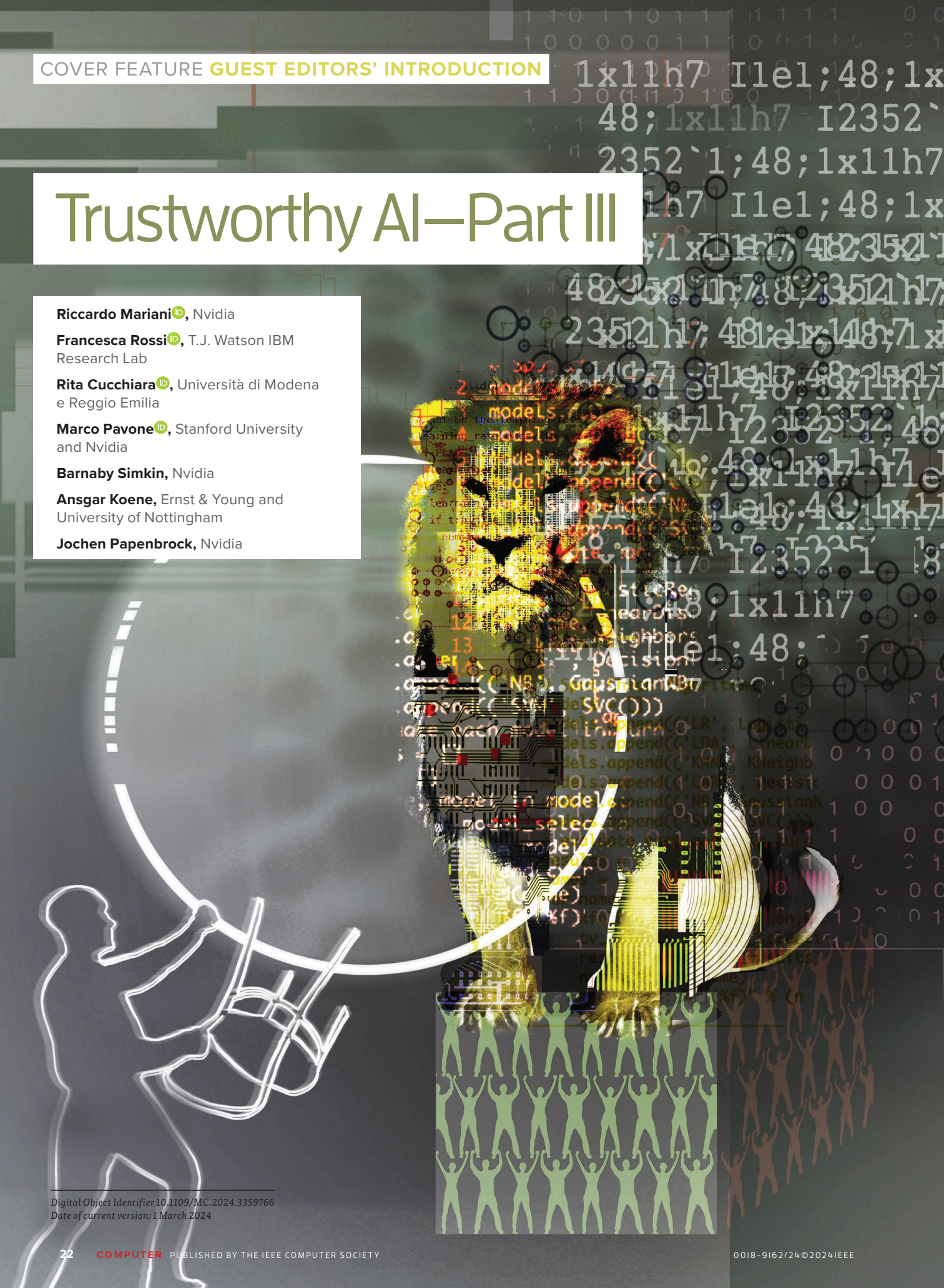
Rita Cucchiara, Università di Modena e Reggio Emilia

Marco Pavone, Stanford University and Nvidia

Barnaby Simkin, Nvidia

Ansgar Koene, Ernst & Young and University of Nottingham

Jochen Papenbrock, Nvidia



Digital Object Identifier 10.1109/MC.2024.3359766
Date of current version: 1 March 2024

The final part of our Trustworthy AI special issue features four articles on AI security, reliability, trust, and AI trustworthiness as a whole.

The first two parts of the Trustworthy AI special issue included contributions centered on trustworthy AI principles such as verifiability, robustness, reliability, explainability, bias, functional safety, fairness, trust, and transparency.

This third and last part of the Trustworthy AI special issue focuses on the following AI principles and related application fields.

The first article[A1] discusses security challenges. The authors look at the risk of malicious manipulation of data to mislead the learning process (poisoning attacks) and related mitigations using basic security principles or by deploying machine learning specific defensive mechanisms.

The second article[A2] discusses reliability and *trust* challenges in a very sensitive application field such as healthcare (managing type 2 diabetes) for elderly patients. The authors propose a trustworthy-AI-based insulin recommender that offers reliable insulin recommendations supported by clinical evidence.

SPECIFICALLY, THE AUTHORS PROPOSE A FRAMEWORK THAT CAN GUIDE NONEXPERTS TO UNLOCK THE FULL POTENTIAL OF USER TRUST IN AI DESIGN.

The following two articles propose frameworks to address AI trustworthiness as a whole.

The third article[A3] discusses trust challenges. The authors propose a framework that can guide nonexperts to unlock the full potential of user trust in AI design.

The final article[A4] discusses how to address the common concern of the lack of precision in the assessment and deployment of AI. The author defines a taxonomy of application types and associated potential harms, related to four main governance dimensions.

APPENDIX: RELATED ARTICLES

A1. A. E. Cinà, K. Grosse, A. Demontis, B. Biggio, F. Roli, and M. Pelillo, "Machine learning security against data poisoning: Are we there yet?" *Computer*, vol. 57, no. 3, pp. 26–34, Mar. 2024, doi: 10.1109/MC.2023.3299572.

A2. T. Padmapritha, K. Bekiroglu, S. Seshadhri, and S. Srinivasan, "Trustworthy AI-based personalized insulin recommender for elderly people who have type-2 diabetes," *Computer*, vol. 57, no. 3, pp. 35–45, Mar. 2024, doi: 10.1109/MC.2024.3352639.

A3. S. Sousa, D. Lamas, J. Cravino, and P. Martins, "Human-centered trustworthy framework: A human–computer interaction perspective," *Computer*, vol. 57, no. 3, pp. 46–58, Mar. 2024, doi: 10.1109/MC.2023.3287563.

A4. J. B. Peckham, "An AI harms and governance framework for trustworthy AI," *Computer*, vol. 57, no. 3, pp. 59–68, Mar. 2024, doi: 10.1109/MC.2024.3354040.

We would like to thank the authors of the four 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.

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 AI-based hardware and software platforms. Contact him at rmariani@nvidia.com.

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

RITA CUCCHIARA is a professor at the Università di Modena e 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 at Nvidia, 0623 Berlin, Germany, where he coordinates Nvidia's overall strategic engagement with regulatory and standards bodies and influences those technical requirements related to AI, automated driving, machine learning, and virtual testing. Contact him at bsimkin@nvidia.com.

ANSGAR KOENE is a global 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, U.K. 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.