

Trustworthy AI—Part I

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

To improve the trustworthiness of artificial intelligence systems, organizations should address the risks relevant to the use case whilst taking into account the stakeholders that interact directly or indirectly with the system.

As the influence of artificial intelligence (AI) grows across the world, industry, academia, and governments are all exploring the best ways to encourage the development and use of AI that is human centered and trustworthy.^{3,5,6,7,8,10,11,12,13} The quest for trustworthy AI is a fundamental issue as it is becoming increasingly difficult to determine whether an AI system protects the individual rights and democratic values of an ever-widening group of stakeholders.

To improve the trustworthiness of AI systems, organizations should address the risks relevant to the use case whilst taking into account the stakeholders that interact directly or indirectly with the system. The applications using machine learning in particular carry a diverse and sometimes unique set of risks that should be mitigated before a system may be trusted.

To help identify the broad range of risks, it is common to define a set of trustworthy AI principles (“Principles”) that are derived from human rights, ethical norms, and legal properties. Each Principle is necessary but not sufficient for achieving trustworthiness. Thus far, there is a broad international consensus around a core set of Principles, these include the following:

- › **Transparency:** the property of an AI system or organization that ensures appropriate information is communicated to relevant stakeholders.
- › **Functionality and performance:** the property of an AI system that supports actions or processes to meet specified objectives.

- › **Explainability:** the property of an AI system that expresses important factors influencing outcomes in a way that humans can understand.
- › **Verifiability:** the property of an AI system that confirms it was built correctly and fulfills specified requirements.
- › **Security:** the property of an AI system related to achieving and maintaining integrity, authenticity, and reliability of a system.
- › **Privacy:** the property of an AI system related to achieving and maintaining confidentiality of data and attributes of a system.
- › **Autonomy and control:** the property of an AI system that allows actions or processes to take place automatically, without the need for natural persons to be directly involved in their execution.
- › **Safety:** the property of an AI system that enables a freedom-from-safety risk, which is not tolerable.
- › **Robustness and reliability:** the property of an AI system related to demonstrating the ability or inability of the system to have comparable performance on atypical data as opposed to the data expected in typical operations.
- › **Nondiscrimination, bias, and fairness:** the property of an AI system that restricts the systematic difference in treatment of certain objects, people, or groups in comparison to others.
- › **Sustainability:** the property of an AI system that avoids the depletion of natural resources to maintain an ecological balance.

Organizations have some flexibility to determine the applicability of each Principle and refine them to each individual use case. However, implementation of the Principles should remain an agile and iterative process throughout the AI lifecycle. An organization may apply a risk-based approach to identify possible impacts to the organization, intended users, and society, and to mitigate the risks appropriately. Organizations should develop methods to ensure the relevant Principles are translated into appropriate technical-organizational measures and mapped to a portion of the AI lifecycle. This could be achieved through a combination of

- › rules (behaviors or states) that the system should always follow
- › restrictions on behaviours or states that the system should never transgress
- › organizational process with specific measurable outcomes.

Organizations should evaluate the effectiveness of the measure to mitigate risk and justify changes to the implementation processes on an ongoing basis.

Trustworthy AI requires a strong long-term commitment to be successful. Even though there is no universally correct solution to this challenge, it is critical that industry, academia, and governments share best practices, learn together, and grow as a community. Although AI regulations are relatively premature, the standards landscape is becoming more robust, with IEEE^{4,9}, ISO/IEC, and CEN-CLC^{1,2,14} developing an array of standards and

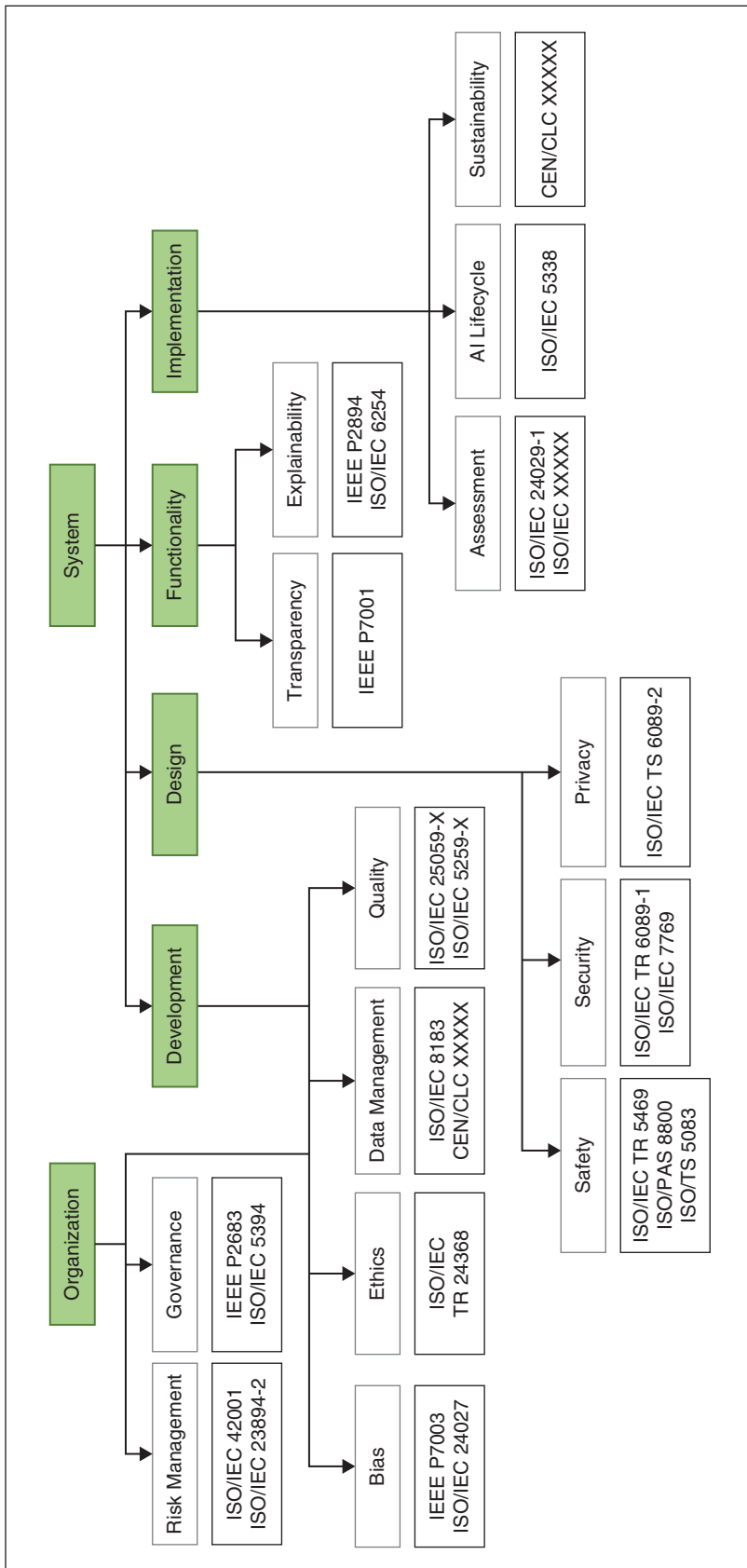


FIGURE 1. A nonexhaustive list of standards covering the AI lifecycle. TR: technical report.

technical reports that support the industry with guidance on how best to develop trustworthy AI (see Figure 1).

In this issue of *Computer*, we have invited stakeholders who contributed articles toward exploring innovative and practical methods for implementing trustworthy AI.


The first article^{A1} addresses trustworthiness as a whole and proposes a novel conceptual model for trustworthy AI based on an adaptation of the well-known ability-benevolence-integrity model of trust to trustworthiness. The next article^{A2} addresses trustworthiness for the specific case of autonomous driving and identifies the maturity level of the different requirements.

Our next article^{A3} focuses on verifiability in the specific case of autonomous systems and describes research carried out by a consortium to address that Principle. Robustness is addressed in the fourth article,^{A4} and an AI model-inspection framework is proposed for detecting and mitigating robustness risks.

A survey of the main reliability assessment methodologies, focusing mainly on fault-injection techniques, is presented in our next article.^{A5} The sixth article^{A6} addresses explainability based on the Taguchi method and evaluates the diversity and average number of evaluations of a solution.

The next article^{A7} addresses bias and provides motivation, theory, code, and examples on how to perform a disciplined discovery of systematic deviations in data and models at the subset level. Our final article^{A8} addresses transparency in three levels: algorithmic, interaction, and social.

A second special issue is scheduled for May 2023, with other selected contributions on trustworthy AI.

We would like to thank the authors of the eight 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. 

ACKNOWLEDGMENT

Riccardo Mariani is the corresponding author.

REFERENCES

1. *Information Technology — Artificial Intelligence — Artificial Intelligence*
 2. *Information Technology — Artificial Intelligence — Overview of Trustworthiness in Artificial Intelligence*, ISO/IEC TR 24028, International Organization for Standardization, Geneva, Switzerland, 2020. [Online]. Available: <https://www.iso.org/standard/77608.html>
 3. “German standardization roadmap for AI,” DIN DKE, Frankfurt am Main, Germany, Nov. 2020. [Online]. Available: <https://www.dke.de/resource/blob/2017010/99bc6d952073ca88f52c0ae4a8c351a8/nr-ki-english—download-data.pdf>
 4. “Ethically designed AI version 2,” IEEE, New York, NY, USA, 2017. [Online]. Available: <https://ethicsinaction.ieee.org/wp-content/uploads/eadle.pdf>
 5. “Empowering AI leadership - An oversight toolkit for boards of directors,” World Economic Forum, Geneva, Switzerland, 2020. [Online]. Available: https://wef-ai.s3.amazonaws.com/WEF_Empowering-AI-Leadership_Oversight-Toolkit.pdf
1. *Information Technology — Artificial Intelligence — Artificial Intelligence*
 2. *Information Technology — Artificial Intelligence — Overview of Trustworthiness in Artificial Intelligence*, ISO/IEC TR 24028, International Organization for Standardization, Geneva, Switzerland, 2020. [Online]. Available: <https://www.iso.org/standard/77608.html>
 3. “German standardization roadmap for AI,” DIN DKE, Frankfurt am Main, Germany, Nov. 2020. [Online]. Available: <https://www.dke.de/resource/blob/2017010/99bc6d952073ca88f52c0ae4a8c351a8/nr-ki-english—download-data.pdf>
 4. “Ethically designed AI version 2,” IEEE, New York, NY, USA, 2017. [Online]. Available: <https://ethicsinaction.ieee.org/wp-content/uploads/eadle.pdf>
 5. “Empowering AI leadership - An oversight toolkit for boards of directors,” World Economic Forum, Geneva, Switzerland, 2020. [Online]. Available: https://wef-ai.s3.amazonaws.com/WEF_Empowering-AI-Leadership_Oversight-Toolkit.pdf

APPENDIX: RELATED ARTICLES

- A1. A. M. Singh and M. P. Singh, “Wasabi: A conceptual model for trustworthy artificial intelligence,” *Computer*, vol. 56, no. 2, pp. 20–28, Feb. 2023, doi: 10.1109/MC.2022.3212022.
- A2. D. Fernández-Llorca and E. Gómez, “Trustworthy artificial intelligence requirements in the autonomous driving domain,” *Computer*, vol. 56, no. 2, pp. 29–39, Feb. 2023, doi: 10.1109/MC.2022.3212091.
- A3. M. R. Mousavi et al., “Trustworthy autonomous systems through verifiability,” *Computer*, vol. 56, no. 2, pp. 40–47, Feb. 2023, doi: 10.1109/MC.2022.3192206.
- A4. P.-Y. Chen and P. Das, “AI maintenance: A robustness perspective,” *Computer*, vol. 56, no. 2, pp. 48–56, Feb. 2023, doi: 10.1109/MC.2022.3218005.
- A5. A. Ruospo, E. Sanchez, L. M. Luza, L. Dilillo, M. Traiola, and A. Bosio, “A survey on deep learning resilience assessment methodologies,” *Computer*, vol. 56, no. 2, pp. 57–66, Feb. 2023, doi: 10.1109/MC.2022.3217841.
- A6. A. Martínez-Vargas, J. A. Gómez-Avilés, M. A. Cosío-León, and Á. G. Andrade, “Explaining the walking through of a team of algorithms,” *Computer*, vol. 56, no. 2, pp. 67–81, Feb. 2023, doi: 10.1109/MC.2022.3212998.
- A7. S. Speakman, G. A. Tadesse, C. Cintas, W. Ogallo, T. Akumu, and A. Oshingbesan, “Detecting systematic deviations in data and models,” *Computer*, vol. 56, no. 2, pp. 82–92, Feb. 2023, doi: 10.1109/MC.2022.3213209.
- A8. K. Haresamudram, S. Larsson, and F. Heintz, “Three levels of AI transparency,” *Computer*, vol. 56, no. 2, pp. 93–100, Feb. 2023, doi: 10.1109/MC.2022.3213181.

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, with 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 director of the AI 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 in the Department of Aeronautics and Astronautics at Stanford University, Stanford, CA 94305 USA, and 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 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. Contact him at ansgar.koene@nottingham.ac.uk.

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

systems." SSRN. Accessed: 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3914105

6. "Building trust in AI and ML principles." Workday. Accessed: 2021. [Online]. Available: <https://www.workday.com/content/dam/web/en-us/documents/whitepapers/building-trust-in-ai-ml-principles-practice-policy.pdf>
7. "Trustworthy AI implementation framework for AI systems." SSRN. Accessed: 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091
8. "Ethical guidelines for trustworthy AI: High level expert group on AI," European Commission, Brussels, Belgium, 2019. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
9. *IEEE Standard Model Process for Addressing Ethical Concerns during System Design*, IEEE Standard 7000-2021, 2021. [Online]. Available: <https://standards.ieee.org/ieee/7000/6781/>
10. "capAI - A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act." SSRN. Accessed: 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091
11. "NOREA guiding principles trustworthy AI investigations," NOREA, Amsterdam, The Netherlands, 2021. [Online]. Available: <https://www.norea.nl/uploads/bfile/a344c98a-e334-4cf8-87c4-1b45da3d9bc1>
12. R. V. Zicari et al., "How to assess trustworthy AI in practice," DeepAI, San Francisco, CA, USA, 2022. [Online]. Available: <https://deepai.org/publication/how-to-assess-trustworthy-ai-in-practice>
13. "Tools for trustworthy AI," OECD, Paris, France, 2021. [Online]. Available: <https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm>
14. *Artificial Intelligence — Functional Safety and AI Systems*, ISO/IEC TR 5469, International Organization for Standardization, Geneva, Switzerland, unpublished. [Online]. Available: <https://www.iso.org/standard/81283.html>