

# Software Engineering for Responsible AI



Qinghua Lu<sup>ID</sup>, Liming Zhu<sup>ID</sup>, and Jon Whittle<sup>ID</sup>,  
CSIRO's Data61

James Bret Michael<sup>ID</sup>, Naval Postgraduate School

Digital Object Identifier 10.1109/MC.2023.3242055  
Date of current version: 5 April 2023

*The unique characteristics of artificial intelligence (AI) systems pose new challenges to traditional software engineering approaches. Thus, new software engineering approaches are required to develop AI systems in a responsible manner.*

**A**rtificial intelligence (AI) continues to demonstrate its positive impact on society and achieve widespread adoption in data-intensive industries. The value of the global AI market was US\$93.5 billion in 2021 and is predicted to have an annual growth rate of 38.1% from 2022 to 2030 (<https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market>). To fully realize these benefits, it is vital to ensure the AI systems are responsibly developed and trusted by citizens and communities who will rely on them. This has prompted significant global effort to pursue and realize responsible AI in the world.

organizations, and enterprises. A principle-based approach provides general guidelines that can be applied in a flexible manner to different technologies and contexts, allowing for adaptation to changing circumstances. Those high-level ethical principles are an important starting point, but they alone do not guarantee the trustworthiness of AI systems.

For example, operationalizing the human-centered value principle is a complex and challenging task, requiring consideration in design, deployment, and monitoring throughout the entire lifecycle of AI systems. Furthermore, significant efforts have been put into algorithm-level solutions that primarily focus on a subset

models. There is a lack of linkage to the software development processes. In addition, the unique characteristics of AI systems bring new challenges to traditional software engineering approaches.<sup>1</sup> These new challenges cannot be tackled using only extensions of existing methods. New software engineering approaches are required to develop AI systems in a responsible manner.<sup>2</sup> Therefore, in this special issue, we focus on system-level methods that can be used to operationalize responsible AI. Seven articles have been accepted for this special issue after review and revision processing.

Maalej et al.<sup>A1</sup> focus on the requirements engineering (RE) for responsible AI. Specifically, the authors discuss six areas that need particular attention by researchers and practitioners, including acceptable levels of quality requirements, data- and user-centered prototyping for AI, expanding RE to focus on data, embedding responsible AI terminology into engineering workflows, tradeoffs, and requirements as a foundation for AI quality and testing.

Li et al.<sup>A2</sup> summarize 17 quality attributes of trustworthy AI through a literature review. The authors establish a framework for trustworthy AI systems by dividing the identified attributes into five categories. Additionally, the authors identify the major research gaps and outline a research agenda in this area.

**RESPONSIBLE AI IS THE DEVELOPMENT AND USE OF AI SYSTEMS THAT BENEFIT INDIVIDUALS, GROUPS, AND THE WIDER SOCIETY WHILE MINIMIZING THE RISK OF NEGATIVE CONSEQUENCES.**

Responsible AI is the development and use of AI systems that benefit individuals, groups, and the wider society while minimizing the risk of negative consequences. To ensure responsible AI, many ethical principles have been released by governments, research

of ethical principles that are mathematics amenable (such as privacy and fairness). However, responsible AI issues can arise at any stage of the development lifecycle and span multiple AI and non-AI components of systems, beyond just AI algorithms and

Bao et al.<sup>A3</sup> have proposed a four-tier software architecture for industry drawing digitization based on responsible AI principles. The authors present the overall architecture design with a focus on federated learning training and digitization tasks. The proposed solution is evaluated in terms of multiple criteria, including accuracy, style transferability, performance, availability, usability, and interactivity.

Li et al.<sup>A4</sup> propose a novel explainable AI method to ensure consistency and reduce time consumption. Specifically, the proposed solution consists of three phases: prediction difference computation, feature contribution value computation, and feature importance order conversion. The authors define the measure

of consistency via explanation summary distances. The proposed method is evaluated via multiple case studies.

Raja and Zhou<sup>A5</sup> conduct a survey to analyze accountability in AI. The authors discuss the concept of accountability and its necessity in AI systems. They further examine the approaches to ensure accountability and the factors that may affect this principle, and they outline the three different levels of accountability and address accountability in different sectors. The authors provide insightful discussions and point out the challenges of AI accountability.

Badran et al.<sup>A6</sup> evaluate three extant AI fairness preprocessing algorithms, including Reweighting, Learning Fair Representations, and Optimized

Preprocessing. Since the algorithms have different priorities, which lead to variations in fairness and accuracy tradeoffs, the authors explore the feasibility of ensembling the algorithm results. In this article, the authors share their experience and provide practitioners with actionable recommendations.

Zhang et al.<sup>A7</sup> focus on the ethical and legal considerations in cyberphysical-social systems (CPSSs). The authors propose a data-driven system-level design framework consisting of two main modules: a design module and an analytic module. Specifically, the design module can continuously optimize a CPSS based on the feedback from the analytic module, while the analytic module analyzes the data from the design module using multimodal data analysis methods.

## APPENDIX: RELATED ARTICLES

- A1. W. Maalej, Y. D. Pham, and L. Chazette, "Tailoring requirements engineering for responsible AI," *Computer*, vol. 56, no. 4, pp. 18–27, Apr. 2023, doi: 10.1109/MC.2023.3243182.
- A2. G. Li, B. Liu, and H. Zhang, "Quality attributes of trustworthy artificial intelligence in normative documents and secondary studies: A preliminary review," *Computer*, vol. 56, no. 4, pp. 28–37, Apr. 2023, doi: 10.1109/MC.2023.3240730.
- A3. Z. Bao et al., "Software architecture for responsible artificial intelligence systems: Practice in the digitization of industrial drawings," *Computer*, vol. 56, no. 4, pp. 38–49, Apr. 2023, doi: 10.1109/MC.2023.3240416.
- A4. D. Li, Y. Liu, J. Huang, and Z. Wang, "A trustworthy view on explainable artificial intelligence method evaluation," *Computer*, vol. 56, no. 4, pp. 50–60, Apr. 2023, doi: 10.1109/MC.2022.3233806.
- A5. A. K. Raja and J. Zhou, "AI accountability: Approaches, affecting factors, and challenges," *Computer*, vol. 56, no. 4, pp. 61–70, Apr. 2023, doi: 10.1109/MC.2023.3238390.
- A6. K. Badran et al., "Can ensembling preprocessing algorithms lead to better machine learning fairness?" *Computer*, vol. 56, no. 4, pp. 71–79, Apr. 2023, doi: 10.1109/MC.2022.3220707.
- A7. S. Zhang, L. T. Yang, Y. Zhang, X. Zhou, and Z. Cui, "Data-driven system-level design framework for responsible cyber-physical-social systems," *Computer*, vol. 56, no. 4, pp. 80–91, Apr. 2023, doi: 10.1109/MC.2023.3243645.




### ABOUT THE AUTHORS

**QINGHUA LU** is a principle research scientist and the team leader of software engineering for the artificial intelligence research team at CSIRO's Data61, Sydney 2015, Australia. She is a Senior Member of IEEE. Contact her at qinghua.lu@data61.csiro.au.

**LIMING ZHU** is a research director at CSIRO's Data61, Sydney 2015, Australia, and a conjoint full professor at the University of New South Wales. He is a Senior Member of IEEE. Contact him at liming.zhu@data61.csiro.au.

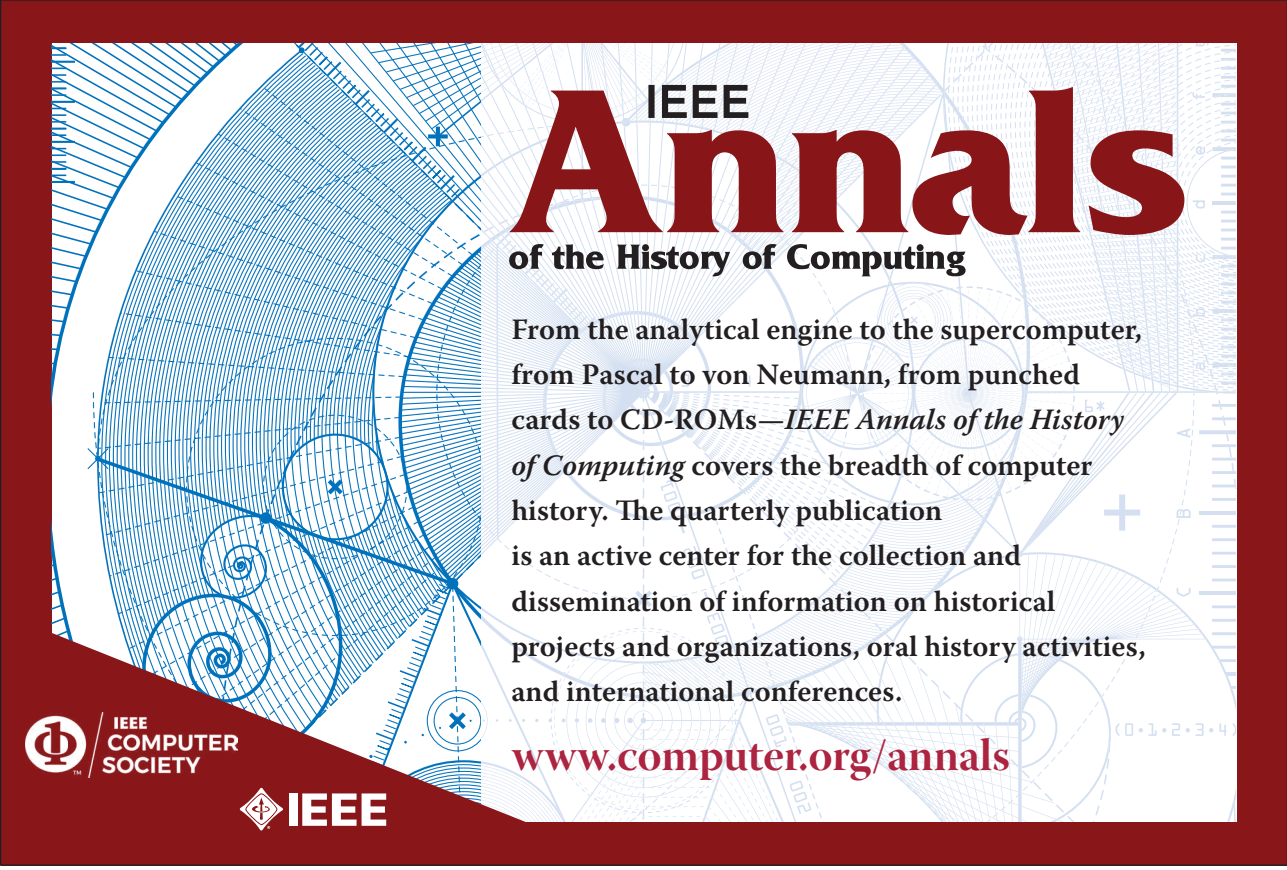
**JON WHITTLE** is director of CSIRO's Data61, Melbourne 3168, Australia, the digital and data science arm of Australia's national science agency. Contact him at jon.whittle@data61.csiro.au.

**JAMES BRETT MICHAEL** is a professor of computer science and electrical engineering at the Naval Postgraduate School, Monterey, CA 93943 USA. He is a Fellow of IEEE. Contact him at bmichael@nps.edu.

The guest editors would like to thank all of the authors who submitted their work to this special issue. We also express our gratitude to the reviewers for their great efforts. Finally, we sincerely appreciate the support from the editor in chief, Dr. Jeff Voas. We hope researchers will find the articles in this special issue enjoyable. 

### REFERENCES

1. M. Shaw and L. Zhu, "Can software engineering harness the benefits of advanced AI?" *IEEE Softw.*, vol. 39, no. 6, pp. 99–104, Nov./Dec. 2022, doi: 10.1109/MS.2022.3203200.
2. Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, "Towards a roadmap on software engineering for responsible AI," in *Proc. 1st Int. Conf. AI Eng., Softw. Eng. AI*, May 2022, pp. 101–112, doi: 10.1145/3522664.3528607.





The graphic features a dark red background with a complex, light blue geometric pattern of overlapping circles and lines, resembling a technical drawing or a stylized globe. The text is in white and red.

**IEEE**  
**Annals**  
of the History of Computing

From the analytical engine to the supercomputer, from Pascal to von Neumann, from punched cards to CD-ROMs—*IEEE Annals of the History of Computing* covers the breadth of computer history. The quarterly publication is an active center for the collection and dissemination of information on historical projects and organizations, oral history activities, and international conferences.

[www.computer.org/annals](http://www.computer.org/annals)

 IEEE COMPUTER SOCIETY

 **IEEE**