Artificial Intelligence Ethics and Trust: From Principles to Practice

Fang Chen and Jianlong Zhou, *University of Technology Sydney, Sydney, NSW, 2007, Australia*
Andreas Holzinger, *University of Natural Resources and Life Sciences Vienna, A-1190, Vienna, Austria*
Kenneth R. Fleischmann, *The University of Texas at Austin, Austin, TX, 78712, USA*
Simone Stumpf, *University of Glasgow, Glasgow, G12 8QQ, U.K.*

Despite the proliferation of ethical frameworks of artificial intelligence (AI) from different organizations such as government agencies, large corporations, and academic institutions, it is still a challenge to implement and operationalize ethical and legal frameworks for AI in practice due to its complexities. The implementation and operationalization involve different aspects in original theoretical and practical research on designing, developing, presenting, testing, and evaluating approaches, which are supported by advanced AI techniques and interdisciplinary research, in particular, social science, law, and cognitive science. This editorial provides an overview of the field of operationalization of AI ethics and trust, and highlights a few key topics covered in this special issue, i.e., the current landscape of AI ethics implementation, trust and trustworthiness in AI, ethical framework for trust calibration, approaches to build morality in AI, implementation of AI ethics with a pattern-oriented engineering approach, and inclusive user studies.

Artificial intelligence (AI), driven by successes in machine learning, now permeates virtually all areas of our daily lives to make or at least influence decisions.[1] In areas that impact human life (agriculture, climate, forestry, health, and so on), ethical and legal aspects such as transparency, fairness, and trust in such decisions are receiving increasing attention.[2] As a result, hundreds of ethical frameworks have now been published by organizations such as government agencies, large corporations, and academic institutions.[3] Adopting these principles is widely seen as one of the best ways to ensure that AI does not cause unintended harm and is used safely and responsibly. However, due to the complexity of AI, it remains a challenge to implement and operationalize ethical and legal frameworks for AI in practice.[4,5]

The implementation and operationalization of ethical and legal frameworks for AI in practice involve different aspects such as original theoretical and practical research on designing, developing, presenting, testing, and evaluating approaches for AI framework implementations supporting trust in AI, including cutting-edge theories, foundations, actionable tools, and impactful case studies of AI ethical framework implementations, supported by advanced AI techniques and interdisciplinary research, in particular, social science, law, and cognitive science.

Motivated by the challenges and demand for the implementation and operationalization of ethical principles, we organized this special issue in *IEEE Intelligent Systems* on this topic in September 2022. The aim is to foster interdisciplinary and transdisciplinary approaches and stimulate cross-domain integration of diverse disciplines for making AI ethical principles operable in applications.

This special issue covers selected recent research results along this line. Fourteen articles were submitted in response to our call for articles, and five articles were competitively selected for this special issue after a thorough review process by international experts. Here, we highlight key topics touched on by this special issue. More specifically, the accepted articles in this special issues cover six areas on the implementation of AI ethical principles:

1541-1672 © 2023 IEEE
Digital Object Identifier 10.1109/MIS.2023.3324470
Date of current version 13 November 2023.

1) The current landscape provides an outlook of current status of the research effort in implementation of ethical principles into practice through a survey study.
2) Trust and trustworthiness conducts a literature review on trust and trustworthiness in AI and demonstrates the need to consider the wider context of AI system development and use to better understand trust in AI.
3) Ethical framework for trust calibration presents an ethical framework for trust calibration in AI aiming at AI reliability and AI safety in the AI ethics context.
4) Approaches to build morality in AI discusses top-down and bottom-up approaches to build morality in AI.
5) Implementation with pattern-oriented engineering approach presents an implementation approach for responsible AI (RAI) with a case study.
6) Inclusive user studies outlines principles for more inclusive explainable AI (XAI) user studies.

In the remainder of this article, we briefly discuss the articles' contributions from each of these perspectives.

THE CURRENT LANDSCAPE

An examination of actionable methods for the operationalization of ethics is helpful to bridge the gap between AI ethical principles and practice. The first article, by Marzouk et al.,[A1] conducts an extensive survey to outline the current landscape of the operationalization of ethics. This survey article introduces a set of criteria to answer two research questions: 1) "What": What defines an RAI framework? and 2) "How": How do these frameworks operationalize ethics? To frame the findings of the survey study, this article introduces an all-encompassing metamodel that aims to establish the building blocks for the development of an RAI practice. This metamodel positions the current landscape and suggests means to adopt and scale AI safely, transparently, and responsibly.

TRUST AND TRUSTWORTHINESS

Users' attitudes toward AI systems play an important role in real-world applications. Whether people trust AI or whether an AI system is trustworthy affect the final adoption of AI for decision making. The article by Duenser and Douglas[A2] conducts a literature review on trust in AI and AI trustworthiness. The concepts of trust and trustworthiness are examined and the differences between them identified. The article argues that trust in AI involves not only reliance on the system itself but also trust in the developers of the AI system. AI ethics principles such as explainability and transparency are often assumed to promote user trust, but empirical evidence of how such features actually affect how users perceive the system's trustworthiness is not as abundant or not that clear. This leads to a discussion of trust in AI as sociotechnical systems, and the need to consider the wider context of AI system development and use to better understand trust in AI. The article argues that more empirical evidence is needed to better understand how and if certain AI ethics principles, combinations of principles, and potential trade-offs or conflicts between principles impact people's trust.

ETHICAL FRAMEWORK FOR TRUST CALIBRATION

Up to now, numerous ethical frameworks containing fundamental ethical principles have been developed by researchers, organizations, or governments, however, they have lacked specific guidance on their practical implementation and operationalization. Although trust is one of key ethical concerns with AI, the article by Schmid and Wiesche[A3] presents an ethical framework by integrating the concept of trust calibration into the context of AI ethics for practical implementation of AI ethics. The connection is established as the ethical principle of "safety" is closely linked to the important dimension of "reliability" for trust calibration. Moreover, both characteristics, AI safety and AI reliability, are of striking importance for the verification and validation as part of the AI lifecycle. The effectiveness of the presented framework is evaluated based on 17 interviews within an international automotive supplier to demonstrate its effectiveness.

BUILDING MORALITY IN AI

Ethical AI, and more generally, building morality in AI, are important for the responsible use of AI. How might we begin to go about doing so? The article by Seeamber and Badea[A4] highlights the different aspects that should be considered when building moral AI, such as those of technical approaches, culture, emotion, sentience, and governance. The article also discusses the top-down and bottom-up approaches to designing moral AI and the role of emotion and sentience in morality. It presents a hybrid approach using a combination of top-down and bottom-up and a further combination of different moral paradigms to get around the different obstacles and limitations. The article also

APPENDIX: RELATED ARTICLES

A1. M. Marzouk, C. Zitoun, O. Belghith, and S. Skhiri, "The building blocks of a responsible artificial intelligence practice: An outlook on the current landscape," *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 9–18, Nov./Dec. 2023, doi: 10.1109/MIS.2023.3320438.

A2. A. Duenser and D. M. Douglas, "Whom to trust, how and why: Untangling artificial intelligence ethics principles, trustworthiness and trust," *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 19–26, Nov./Dec. 2023, doi: 10.1109/MIS.2023.3322586.

A3. A. Schmid and M. Wiesche, "The importance of an ethical framework for trust calibration in AI," *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 27–34, Nov./Dec. 2023, doi: 10.1109/MIS.2023.3320443.

A4. R. Seeamber and C. Badea, "If our aim is to build morality into an artificial agent, how might we begin to go about doing so?" *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 35–41, Nov./Dec. 2023, doi: 10.1109/MIS.2023.3320875.

A5. Q. Lu, Y. Luo, L. Zhu, M. Tang, X. Xu, and J. Whittle, "Developing responsible chatbots for financial services: A pattern-oriented responsible artificial intelligence engineering approach," *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 42–51, Nov./Dec. 2023, doi: 10.1109/MIS.2023.3320437.

A6. U. Peters and M. Carman, "Unjustified sample sizes and generalizations in explainable artificial intelligence research: Principles for more inclusive user studies," *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 52–60, Nov./Dec. 2023, doi: 10.1109/MIS.2023.3320433.

discusses the relevant practical issues and techniques for implementing this approach.

PATTERN-ORIENTED RAI ENGINEERING APPROACH

To implement RAI, many governments and organizations have released AI ethics principles. However, those principles often remain too abstract for practitioners to apply directly. Moreover, the existing RAI solutions primarily target the model level, which is a narrow perspective and requires broader system-level thinking. RAI issues can arise at any stage of the AI system engineering lifecycle, from the initial planning stage to the final monitoring stage. To tackle these challenges, the article by Lu et al.[A5] builds an RAI pattern catalog for various stakeholders across different stages of an AI system's lifecycle. The usefulness of the pattern catalog is showcased in a real-world scenario using a case study on the development process of a financial service chatbot. In summary, the article proposes a pattern-oriented RAI engineering approach to address the end-to-end system-level challenges of RAI.

INCLUSIVE USER STUDIES

User studies are one approach in our toolset to test the effectiveness of implementations of AI ethical principles. Although explainability is one of the mandatory AI ethical principles, XAI models are frequently tested for their adequacy in user studies. As different people may have different explanatory needs, it is important that participant samples in user studies are large enough to represent the target population to enable generalizations. However, it is unclear to what extent XAI researchers reflect on and justify their sample sizes or avoid broad generalizations across people. By analyzing XAI user studies (N=220) published between 2012 and 2022, the article by Peters and Carman[A6] outlines principles for more inclusive XAI user studies. Those principles are 1) the principle of sample-size justification: small-scale user studies may be acceptable at the beginning due to funding or feasibility concerns, however, to promote best scientific practice, AI journals should require user studies to include (where feasible) power analyses or other sample-size justifications. 2) The principle of reporting relevant background: XAI user studies need to consider, report, and justify the inclusion and exclusion of relevant control variables such as technical background because it may affect XAI user perception. 3) The principle of generalization checks: user studies should provide generality-constraint statements, i.e., statements that articulate generalizability limits and justify the scope of result-related claims to avoid overgeneralizations.

ACKNOWLEDGMENT

We would like to thank Prof. San Murugesan, editor in chief of *IEEE Intelligent Systems*, and Prof. Longbing Cao, former editor in chief of *IEEE Intelligent Systems*, for making this article possible. We would also like to express our sincere appreciation to the administrative team of the magazine for their support in managing this issue. We thank all the authors for their submissions. Special thanks are also given to the reviewers for their diligent work with constructive and timely feedback in evaluating these submissions.

REFERENCES

1. A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, and H. Müller, "AI for life: Trends in artificial intelligence for biotechnology," *New Biotechnol.*, vol. 74, no. 1, pp. 16–24, 2023, doi: 10.1016/j.nbt.2023.02.001.
2. A. Holzinger et al., "Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence," *Inf. Fusion*, vol. 79, no. 3, pp. 263–278, 2022, doi: 10.1016/j.inffus.2021.10.007.
3. J. Zhou, F. Chen, A. Berry, M. Reed, S. Zhang, and S. Savage, "A survey on ethical principles of AI and implementations," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, 2020, pp. 3010–3017, doi: 10.1109/SSCI47803.2020.9308437.
4. J. Zhou and F. Chen, "AI ethics: From principles to practice," *AI Soc.*, pp. 1–11, Nov. 2022, doi: 10.1007/s00146-022-01602-z.
5. K. Stoeger, D. Schneeberger, and A. Holzinger, "Medical artificial intelligence: The European legal perspective," *Commun. ACM*, vol. 64, no. 11, pp. 34–36, 2021, doi: 10.1145/3458652.

FANG CHEN is a distinguished professor of artificial intelligence at the Data Science Institute, University of Technology Sydney, Sydney, NSW, 2007, Australia. Contact her at fang.chen@uts.edu.au.

JIANLONG ZHOU is an associate professor of computer science at the Data Science Institute, University of Technology Sydney, Sydney, NSW, 2007, Australia. Contact him at jianlong.zhou@uts.edu.au.

ANDREAS HOLZINGER is head of the Human-Centered AI Lab at University of Natural Resources and Life Sciences, Vienna, Vienna, A-1190, Austria. Contact him at andreas.holzinger@human-centered.ai.

KENNETH R. FLEISCHMANN is a professor in the School of Information, The University of Texas at Austin, Austin, TX, 78712, USA. Contact him at kfleisch@ischool.utexas.edu.

SIMONE STUMPF is a reader in responsible and interactive artificial intelligence with the School of Computing Science, University of Glasgow, Glasgow, G12 8QQ, U.K. Contact her at simone.stumpf@glasgow.ac.uk.