From Guidelines to Practice: Integrating Techniques in Development Platforms to Achieve Trustworthy AI

Philip Singer, Kathrin Brecker, Ali Sunyaev

philip.singer@student.kit.edu, {brecker, sunyaev}@kit.edu

Research Motivation

- Growing calls for guidelines to ensure the development of trustworthy AI (TAI) systems [1-2]
- Current guidelines [3], principles [4-5], and best practices [6] remain abstract and difficult to apply [2-3]
- Tools are only of limited use for developers as they co-exist in isolation, focus on only one or few TAI qualities, and lack alignment to guide the entire AI development lifecycle [4, 7-10]
- Cloud-based AI development platforms can foster TAI because these platforms provide developers with best practices and tools to enable and guide the AI development [11]
- Extant TAI research is spread across various disciplines (e.g., information systems, computer science, or medicine [12])

Research Question:
What are the key techniques for fostering TAI that can be integrated into AI development platforms?

(Descriptive Literature Review [13])

	Technique Category Description	AI Dev. Lifecycle Phase [14-15]	Exemplary Techniques	TAI Quality Addressed
1	**Trustworthy Training Data** Techniques for monitoring and preprocessing the training data.	Data Preprocessing (1)	Issue Detection [16], Debiasing [17], Data augmentation [18], Preserving Privacy [19]	Privacy [16], Fairness [35], Security [36], Robustness [37], Performance [38]
2	**Trustworthy Model Training** Techniques to build and train robust, fair, and privacy-preserving models.	Model Development (2)	Robust Training [20], Model Debiasing [21], Differential Privacy [22]	Privacy [16], Fairness [35], Robustness [37]
3	**Trustworthy Model Evaluation** Techniques to evaluate model's fairness, performance, and robustness; and ensure explainability.	Model Evaluation (3)	Fairness Evaluation [23], Robustness Evaluation [24], Ensuring Explainability [25]	Fairness [35], Accountability [40], Robustness [37], Performance [38], Transparency [39]
4	**Trustworthy Inferencing** Techniques to monitor and actively control inferencing.	Inferencing (4)	Input Monitoring [26], Input transformation [27], Inferencing control [28], Output Monitoring [29]	Robustness [37], Security [36], Transparency [39]
5	**Internal and External Transparency** Techniques to enable transparency of AI development decisions and process, incl. internal / external communication.	Applicable in all lifecycle phases	Documentation [30], Collaboration and Communication [31], Process control [32]	Accountability [40], Security [36], Transparency [39]
6	**Data Protection** Techniques to transmit, store and process sensitive data securely.	Applicable in all lifecycle phases	Access Control [33], Homomorphic Encryption [19], Trusted Execution Environment [34]	Privacy [16], Security [36]

Implications for Research

- Synthesized overview how to address various TAI qualities by these techniques in parallel
- Paves the way for future research to further investigate the consequences of combining TAI qualities and techniques (e.g., synergies or adverse effects [41-42])

Implications for Practice

- Starting point for AI developers and platform providers to construct TAI development platforms by providing concrete techniques
- Organizations can harness extant techniques to provide TAI development guidance for developers

Karlsruhe Institute of Technology

Critical Information Infrastructures Research Group

Institute of Applied Informatics and Formal Description Methods
Prof. Dr. Ali Sunyaev

From Guidelines to Practice: Integrating Techniques in Development Platforms to Achieve Trustworthy AI

Philip Singer, Kathrin Brecker, Ali Sunyaev

philip.singer@student.kit.edu, {brecker, sunyaev}@kit.edu

References

[1] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Schmager, S., & Sousa, S. (2021). A Toolkit to Enable the Design of Trustworthy AI. 536–555. https://doi.org/10.1007/978-3-030-90963-5_41

[3] Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

[4] Curto, G., & Comim, F. (2023). Fairness: From the ethical principle to the practice of Machine Learning development as an ongoing agreement with stakeholders. Social Science Research Network. https://doi.org/10.2139/ssrn.4397259

[5] Diakopoulos, N., Friedler, Arenas, Barocas, Hay, Howe, Jagadish, Unsworth, Sahuguet, Venkatasubramanian, Wilson, Yu, & Zevenbergen. (2023). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms: FAT ML. https://www.fatml.org/resources/principles-for-accountable-algorithms

[6] Mazumder, S., Dhar, S., & Asthana, A. (2023). A Framework for Trustworthy AI in Credit Risk Management: Perspectives and Practices. Computer, 56(5), 28–40. https://doi.org/10.1109/MC.2023.3236564

[7] Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. 2017 IEEE Symposium on Security and Privacy (SP), 39–57. https://doi.org/10.1109/SP.2017.49

[8] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. International Conference on Learning Representations

[9] Xiong, P., Buffett, S., Iqbal, S., Lamontagne, P., Mamun, M., & Molyneaux, H. (2022). Towards a robust and trustworthy machine learning system development: An engineering perspective. Journal of Information Security & Applications, 65, N.PAG-N.PAG. https://doi.org/10.1016/j.jisa.2022.103121

[10] Li, B., Qi, P., Liu, B., Di Shuai, Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. ACM Computing Surveys, 55(9), 1–46. https://doi.org/10.1145/3555803

[11] Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2021). Artificial Intelligence as a Service. Business & Information Systems Engineering, 63(4), 441–456. https://doi.org/10.1007/s12599-021-00708-w

[12] Wang, D., Wang, L., Zhang, Z., Wang, D., Zhu, H., Gao, Y., Fan, X., & Tian, F. (2021). “Brilliant AI Doctor” in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3411764.3445432

[13] Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. Information & Management, 52(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

[14] Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

[15] Schlegel, M., & Sattler, K.-U. (2023). Management of Machine Learning Lifecycle Artifacts. ACM SIGMOD Record, 51(4), 18–35. https://doi.org/10.1145/3582302.3582306

[16] Liu, H., WANG, Y., FAN, W., Liu, X., Li, Y., JAIN, S., LIU, Y., JAIN, A., & Tang, J. (2023). Trustworthy AI: A Computational Perspective. ACM Transactions on Intelligent Systems & Technology, 14(1), 1–59. https://doi.org/10.1145/3546872

[17] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[18] Xing, X., Wu, H., Wang, L., Stenson, I., Yong, M., Del Ser, J., Walsh, S., & Yang, G. (2023). Non-Imaging Medical Data Synthesis for Trustworthy AI: A Comprehensive Survey. ACM Computing Surveys. https://doi.org/10.1145/3614425

[19] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion, 99, N.PAG-N.PAG. https://doi.org/10.1016/j.inffus.2023.101896

[20] Gu, S., & Rigazio, L. (2014). Towards Deep Neural Network Architectures Robust to Adversarial Examples. International Conference on Learning Representations.

[21] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. AIES ’18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 350–340.

[22] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. Proceedings of The, 308–318. https://doi.org/10.1145/2976749.2978318

[23] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. Information Technology Convergence and Services. https://doi.org/10.1145/2090236.2090255

[24] Biggio, B., Fumera, G., & Roli, F. (2014). Security Evaluation of Pattern Classifiers under Attack. IEEE Transactions on Knowledge and Data Engineering, 26(4), 984–996. https://doi.org/10.1109/TKDE.2013.57

[25] Wu, H., Wang, C., Yin, J., Lu, K., & Zhu, L. (2018). Sharing Deep Neural Network Models with Interpretation. Proceedings of the 2018 World Wide Web Conference, 177–186. https://doi.org/10.1145/3178876.3185995

[26] Darvish Rouani, B., Samragh, M., Javidi, T., & Koushanfar, F. (2019). Safe Machine Learning and Defeating Adversarial Attacks. IEEE Security & Privacy, 17(2), 31–38. https://doi.org/10.1109/MSEC.2018.2888779

[27] Wang, X., Li, J., Kuang, X., Tan, Y., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. Journal of Parallel and Distributed Computing, 130, 12–23. https://doi.org/10.1016/j.jpdc.2019.03.003

[28] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2023). Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. ACM Computing Surveys. https://doi.org/10.1145/3626234

Karlsruhe Institute of Technology

CRITICAL INFORMATION INFRASTRUCTURES RESEARCHGROUP

Institute of Applied Informatics and Formal Description Methods
Prof. Dr. Ali Sunyaev

From Guidelines to Practice: Integrating Techniques in Development Platforms to Achieve Trustworthy AI

Philip Singer, Kathrin Brecker, Ali Sunyaev

philip.singer@student.kit.edu, {brecker, sunyaev}@kit.edu

References

[29] Klinkenberg, R., & Joachims, T. (2000). Detecting Concept Drift with Support Vector Machines. Proceedings of the Seventeenth International Conference on Machine Learning, 487–494. https://doi.org/10.5555/645529.657791

[30] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 272–283. https://doi.org/10.1145/3351095.3372834

[31] Liao, Q. V., & Sundar, S. S. (2022). Designing for Responsible Trust in AI Systems: A Communication Perspective. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1257–1268. https://doi.org/10.1145/3531146.3533182

[32] Perez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., Cerquides, J., Cazorla, F. J., Englund, C., Tauber, M., Nikolakopoulos, G., & Flores, J. L. (2023). Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. ACM Computing Surveys. https://doi.org/10.1145/3626314

[33] Alexander, C. S., Yarborough, M., & Smith, A. (2023). Who is responsible for `responsible AI'?: Navigating challenges to build trust in AI agriculture and food system technology. Precision Agriculture, 1–40. https://doi.org/10.1007/s11119-023-10063-3

[34] Hoekstra, M., Lal, R., Pappachan, P., Phegade, V., & Cuvillo, J. (2013). Using innovative instructions to create trustworthy software solutions. HASP '13: Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy. https://doi.org/10.1145/2487726.2488370

[35] Gittens, A., Yener, B., & Yung, M. (2022). An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML. IEEE Access, 10, 120850–120865. https://doi.org/10.1109/ACCESS.2022.3218715

[36] Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. The Alan Turing Institute. https://doi.org/10.2139/ssrn.3403301

[37] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet& Society. https://dash.harvard.edu/handle/1/42160420

[38] Liao, Q. V., & Sundar, S. S. (2022). Designing for Responsible Trust in AI Systems: A Communication Perspective. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1257–1268. https://doi.org/10.1145/3531146.3533182

[39] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[40] Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18. https://doi.org/10.1145/3351095.3372833

[41] Petkovic, D. (2023). It is Not "Accuracy vs. Explainability"—We Need Both for Trustworthy AI Systems. IEEE Transactions on Technology and Society, 4(1), 46–53. https://doi.org/10.1109/TTS.2023.3239921

[42] Steimers, A., & Schneider, M. (2022). Sources of Risk of AI Systems. International Journal of Environmental Research and Public Health, 19(6), 3641. https://doi.org/10.3390/ijerph19063641

Karlsruhe Institute of Technology

CRITICAL INFORMATION INFRASTRUCTURES RESEARCHGROUP

Institute of Applied Informatics and Formal Description Methods
Prof. Dr. Ali Sunyaev