MLOps as Enabler of Trustworthy AI

Yann Billeter, Philipp Denzel,
Ricardo Chavarriaga, Oliver Forster,
and Frank-Peter Schilling
Centre for AI (CAI)
Zurich University of Applied Sciences ZHAW
Winterthur, Switzerland
{bily,denp,char,foro,scik}@zhaw.ch

Stefan Brunner, Carmen Frischknecht-Gruber,
Monika Reif, and Joanna Weng
Institute for Applied Mathematics and Physics (IAMP)
Zurich University of Applied Sciences ZHAW
Winterthur, Switzerland
{brrt,frsh,reif,wenj}@zhaw.ch

***Abstract*—As Artificial Intelligence (AI) systems are becoming ever more capable of performing complex tasks, their prevalence in industry, as well as society, is increasing rapidly. Adoption of AI systems requires humans to trust them, leading to the concept of trustworthy AI which covers principles such as fairness, reliability, explainability, or safety. Implementing AI in a trustworthy way is encouraged by newly developed industry norms and standards, and will soon be enforced by legislation such as the EU AI Act (EU AIA). We argue that Machine Learning Operations (MLOps), a paradigm which covers best practices and tools to develop and maintain AI and Machine Learning (ML) systems in production reliably and efficiently, provides a guide to implementing trustworthiness into the AI development and operation lifecycle. In addition, we present an implementation of a framework based on various MLOps tools which enables verification of trustworthiness principles using the example of a computer vision ML model.**

***Index Terms*—AI, MLOps, explainability, trustworthiness**

I. Introduction

Trustworthy AI (TAI) principles are anticipated to shape future AI system regulations [1]. Despite their crucial role, these principles often remain abstract, lacking concrete operational mandates [2]. While emerging frameworks offer high-level ethical guidelines for TAI, providing high-level addressing principles like fairness, autonomy, control, transparency, reliability, security, and privacy (e.g., [3]), there remains a need for technical guidelines and workflows.

In parallel, yet seemingly unrelated, the new and fast-evolving field of Machine Learning Operations (MLOps) [4] takes inspiration from the concept of DevOps (Development and Operations) to establish methods, best practices, and tools to operationalize an ML- or AI-based system (AIS), i.e., to bring it into production. Covering all stages from project setup and requirements to continuous integration/delivery, data management, model development, testing, validation, deployment (including cloud and edge), monitoring, and continual learning, a systems approach to MLOps ensures holistic alignment with specified objectives [5]. We show that MLOps practices, besides streamlining deployment and maintenance of AIS, can be naturally extended to address and assess specific requirements on AI trustworthiness.

The remainder of this paper is structured as follows: section II reviews existing research on integrating TAI into the ML lifecycle. In section III, we provide a comprehensive mapping of TAI principles to both the MLOps lifecycle and the practices underlying it. In section IV, we describe a concrete implementation of TAI principles as part of an MLOps system. Finally, section V discusses gaps and obstacles in the adoption of MLOps for TAI and suggests avenues for future research.

II. Related Work

A. Trustworthy AI in the ML lifecycle

Traditional performance metrics in ML need to be complemented with additional principles as real-world applications rise. Technical concerns like robustness, explainability, transparency, reproducibility, and generalization, along with ethical considerations such as fairness, privacy, and accountability collectively fall under the term “trustworthy AI” [6]. This concept is closely linked to “AI governance”, defined as “the set of rules, regulations, ethical and technical frameworks, and similar mechanisms that guide the development and deployment of artificial intelligence technologies.” [7].

Several studies integrate TAI elements into the ML lifecycle. Laato et al. [8] include AI governance in common system development lifecycle models, resembling MLOps. Ashmore et al. [9] outline assurance desiderata for each ML lifecycle stage and review existing methods. Li et al. [6] propose a systematic approach to incorporate TAI in the ML lifecycle.

Despite recognizing the importance of a holistic approach to trustworthiness across lifecycle stages, the literature lacks integration of trustworthy AI with the growing adoption of MLOps. Although a guide on adopting MLOps in the context of responsible AI exists [10], it does not directly address how MLOps practices map to TAI principles. Additionally, actual real-world evaluations for existing approaches are scarce.

B. Metrics in MLOps

MLOps is driven by automation and metrics. Quality and reliability, for instance, are measured and monitored through metrics like *mean time to restore* and *change failure rate* (percentage of deployments causing failure in production). In machine learning, metrics guide training and testing. In trustworthy AI, metrics serve a dual purpose. Firstly, they enable the construction of effective feedback loops with quantifiable measures for TAI principles in MLOps. Secondly, they

TABLE I
INTERRELATION OF MLOPS STAGES (COLUMNS), MLOPS PRACTICES (ROWS), AND TAI PRINCIPLES, I.E., AUTONOMY AND CONTROL (AC), FAIRNESS (FN), PRIVACY (PR), RELIABILITY (RL), SECURITY (SE), TRANSPARENCY (TR). A“✓” DENOTES MLOPS PRACTICES APPLICABLE TO RESPECTIVE STAGES. TAI PRINCIPLES LISTED AFTER THE “✓” INDICATE A PARTICULAR RELEVANCE OF THE PRACTICE TO THE LISTED PRINCIPLES.

Practice / Stage	Business & Data Understanding	Data Engineering	ML Model Engineering	ML Model Evaluation	Deployment	Monitoring & Maintenance
TAI Principles	FN, PR, RL	FN, PR, TR, RL	AC, FN, PR, TR, RL	FN, RL, TR	AC, PR, RL, SE, TR	AC, FN, RL, SE
Versioning		✓: TR, RL	✓: TR, RL		✓	
Testing		✓: FN, TR, RL		✓: FN, RL, TR		
Automation		✓	✓	✓	✓	
Reproducibility		✓	✓	✓	✓	
Deployment					✓	
Monitoring						✓

may serve in definitions of and compliance to (upcoming) AI regulation. An example in the context of the EU AIA is the Key AI Risk Indicators framework for AI in the financial services industry [11]. Yet, despite their practical relevance and expected importance, the role of metrics in TAI has not been fully addressed in the existing literature.

III. MLOps as Enabler of Trustworthy AI

In this section, we illustrate how MLOps enables TAI by establishing the relationship of trustworthy AI principles with MLOps concepts. For these principles, we rely on the six dimensions of trustworthiness identified by the Fraunhofer Institute [3]. For each principle, we report (a) the key stage(s) of the ML lifecycle where it should be addressed, as well as how it relates to the MLOps practices [12] (versioning, testing, automation, reproducibility, deployment, and monitoring), and (b) which metrics could be tracked to monitor progress of the respective TAI principle. We follow the CRSIP-ML(Q) lifecycle [13] for our MLOps lifecycle model, comprising the stages “Business & Data Understanding”, “Data Engineering”, “ML Model Engineering”, “ML Model Evaluation”, “Deployment”, and “Monitoring & Maintenance”. Table I summarizes the relation of trustworthy AI principles, MLOps lifecycle stages, and MLOps practices.

A. Mapping Trustworthy AI Principles to the MLOps Lifecycle

a) Autonomy and Control (AC): This principle focuses on two key elements: evaluating the appropriate level of autonomy for the artificial intelligence application (e.g., human-in/on/out-of-the-loop) and analysing how well the AI application supports and allows adequate room for the individual’s interaction with it. The key stages in the ML lifecycle for autonomy and control are model engineering, deployment, and monitoring & maintenance. Model engineering considers potential human feedback. Deployment influences human control and model use. Monitoring is vital for human oversight in the maintenance stage. Despite the importance of addressing these concerns, there is limited research on achieving optimal human control in the context of AI systems, and we are unaware of potential metrics for this principle.

b) Fairness (FN): The fairness principle mainly aims to prevent unjust discrimination in AI use, often caused by biased training data or statistical under-representation of certain groups, leading to reduced quality for those groups. Regulatory requirements highlight the importance of addressing fairness in the planning stage. Technical interventions in data engineering, model engineering, and evaluation can address issues like class imbalances and group representation. Constraint enforcement and ongoing monitoring in the model engineering and monitoring stages, respectively, ensures compliance with fairness criteria, with metrics categorized into individual, group, and causality-based fairness [14].

c) Privacy (PR): This principle focuses on safeguarding sensitive data during AIS development and operation, encompassing personal data and business secrets. Similar to fairness, privacy requirements often stem from regulations, necessitating early consideration in the planning stage, such as through a privacy-by-design approach. Technical interventions in data and model engineering involve practices like data minimization, reducing attack surfaces, and implementing differential privacy (DP) [15]. Privacy-related issues, including model extraction and attacks like membership inference and model inversion [16], must be addressed during deployment. Privacy metrics serve a dual purpose: information-theoretic measures quantify system privacy properties, while in the context of DP, metrics like summary statistics assess the inherent privacy-utility tradeoff [17], [18].

d) Reliability (RL): The reliability principle assesses the AIS’ quality, focusing on robustness and output uncertainties. Emphasized during model engineering and evaluation, practices like certified training can enhance robustness [19]. Evaluation includes testing model robustness via adversarial attacks and formal verification [20], with certified accuracy as a popular metric. Reliability extends to deployment, addressing potential user adversarial actions and ensuring fast recovery. Monitoring and maintenance involve close attention to reliability metrics. An emerging approach involves integrating domain knowledge [21], improving reliability and aiding in judging prediction plausibility. Identifying applicable domain knowledge should begin in the planning stage, becoming actionable in the model engineering and evaluation stages.

e) Security (SE): Involving functional security properties and safeguarding against attacks, Security encompasses measures related to embedding the AI component. This includes traditional IT security methods and metrics (e.g. [22]). It directly correlates with deployment and monitoring stages, akin to traditional software solutions. Beyond established practices for securing AIS, it must address the potential of AI to enhance or compromise existing security measures [23].

f) Transparency (TR): Transparency in an AIS involves interpretability, reproducibility, and explainability, assessing user and expert comprehensibility, as well as result reproducibility and explainability. Improving transparency aligns with MLOps practices such as reproducibility and versioning, which are crucial for comprehensive change tracking. The incorporation of explainable algorithms and architectures enhances model transparency. Throughout model evaluation, monitoring, and maintenance, transparency is assessed using explainability methods [24], providing insights for developers and potential users. However, the absence of a standardized transparency measure requires further research, as machine-based metrics may not directly align with human relevance [25].

B. Trustworthy AI as an Iterative Process in MLOps

Due to the evolving nature of AIS and their complex post-deployment, trustworthiness levels may fluctuate dynamically. Continuous risk monitoring is crucial for responsible AI development, aligning with the iterative nature of MLOps driven by versioning, automation, testing, deployment, and monitoring. The core tenet of continuous refinement can be extended by the addition of trustworthiness metrics alongside accuracy and efficiency. This enables the implementation of continuous feedback loops, systematically addressing TAI requirements throughout the entire AI lifecycle.

IV. Case Study

In the preceding section, we presented how TAI principles map onto the MLOps practices and lifecycle stages. This section shows an implementation of MLOps-supported pipeline for the development of Trustworthy AI systems. Using open-source MLOps software components, we automate the assessment of reliability and transparency requirements for computer vision (CV) models across the data engineering, model engineering, and model evaluation stages. We evaluate the system's capability to monitor trustworthiness for various CV models and tasks, employing various reliability and transparency methods [26]. For instance, for reliability, robustness to perturbations is expected, which we evaluate on both the data and model level, e.g. through estimation of noise profiles of the training data, and simulation of domain-specific perturbations. Regarding transparency, one example is the identification of global explanations to model decisions as part of model evaluation. For this, we extract factors of variation and prototypes during the data engineering stage. During model evaluation, the prototypes are used conjointly with explanation methods [24].

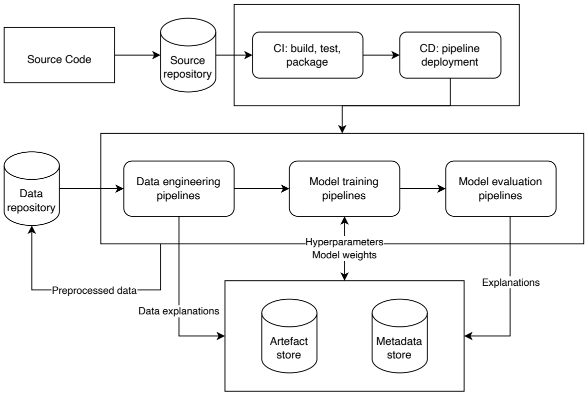

Fig. 1. Overview of the MLOps system architecture.

A. MLOps System Architecture

A spectrum of MLOps tools, encompassing both open-source and proprietary solutions, is available. In alignment with considerations for intellectual property and data protection, we opted for established open-source tools suitable for in-house deployment, except for the code repository, where we utilize git versioning through GitHub[1]. We employ Oxen[2] for data versioning, while MLflow[3] manages experiment and model version tracking. Our system integrates GitHub Actions for CI/CD and relies on Apache Airflow[4] for efficient workload scheduling. A schematic overview is given in fig. 1.

B. Workflow

The system reacts to changes in source code or data, initiating CI/CD processes for code changes and triggering related pipelines. Model code changes lead to the execution of training and explanation pipelines. All resulting artifacts (e.g., model weights, metrics) are stored and versioned in MLFlow for analysis.

Data changes trigger data engineering pipelines, generating training-ready data and artifacts (e.g. prototypes and noisy samples) for model evaluation pipelines. This process includes model retraining and evaluation, in turn generating model explanations and reliability diagrams. The automated reliability and transparency feedback loops prevent inadequate models from reaching production. Monitoring data variation factors and noise profiles aids in understanding cause-effect relations between data changes and model performance, guiding future data collection decisions. Ultimately, consistent adherence to these MLOps feedback loops leads to a continuous and automatic increase in trustworthiness.

V. Discussion and Conclusion

MLOps systems ought to be non-negotiable for organizational productivity when developing or deploying AIS. Beyond productivity, they enable TAI through automated feedback loops and metrics, ensuring continuous assessment of TAI principles

[1] https://github.com/ [2] https://www.oxen.ai/ [3] https://mlflow.org/
[4] https://airflow.apache.org/

across lifecycle stages. Using an appropriately scaling MLOps system, this can be achieved regardless of project scale.

The framework also supports assessing conformity with internal or regulatory criteria. The relationship of MLOps and TAI evolves bidirectionally: Akin to how privacy-by-design was developed from a specific concern, MLOps evolves according to such criteria. Conversely, upcoming standards on TAI are informed by the technological state-of-the-art, which entails MLOps.

The described system for MLOps-guided TAI-compatible ML development is not exhaustive; real-world applications necessitate appropriate organizational structures and practices, and adopting MLOps for TAI presents challenges, warranting further research into specific hindrances:

Firstly, significant gaps exist among different trustworthy AI principles, both in research and practical implementation. Recent attention has focused on robustness and security, while autonomy and control remains largely unexplored in the trustworthiness literature. Secondly, some principles are missing adequate metrics. Transparency and explainability methods mainly rely on human interpretation and thus cannot be monitored automatically as part of an MLOps feedback loop, or are machine-generated yet not guaranteed to be relevant for human comprehension.

Finally, the overall implementation of MLOps is often impeded by a lack of clear standardization among MLOps tools. This compels practitioners to develop custom interfacing code for integrating their data and models, increasing friction and impeding the widespread adoption of MLOps tools.

MLOps has the potential to enable trustworthy AI by creating continuous feedback loops throughout the entire AI lifecycle. To leverage this potential, we believe that future research should address reducing the friction encountered in MLOps adoption and seek to address the gaps in the existing trustworthy AI literature.

Acknowledgment

This work was co-financed by Innosuisse (101.650 IP-ICT).

References

[1] EU Parliament. EU AI Act: first regulation on artificial intelligence. [Online]. Available: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[2] L. Lucaj, P. van der Smagt, and D. Benbouzid, "AI Regulation Is (not) All You Need," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23. ACM, 2023, p. 1267–1279.

[3] M. Poretschkin *et al.*, "Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz," Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Technical Report, 2021. [Online]. Available: https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html

[4] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," *IEEE Access*, vol. 11, pp. 31 866–31 879, 2023.

[5] C. Huyen, *Designing Machine Learning Systems*. USA: O'Reilly Media, 2022.

[6] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy AI: From Principles to Practices," *ACM Comput. Surv.*, p. 1–46, 2023. [Online]. Available: https://doi.org/10.1145/3555803

[7] Carnegie Council for Ethics in International Affairs. AI governance. [Online]. Available: https://www.carnegiecouncil.org/explore-engage/key-terms/ai-governance

[8] S. Laato, T. Birkstedt, M. Mäantymäki, M. Minkkinen, and T. Mikkonen, "AI governance in the system development life cycle: insights on responsible machine learning engineering," in *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, ser. CAIN '22. ACM, May 2022.

[9] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges," *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–39, May 2021.

[10] B. M. A. Matsui and D. H. Goya, "MLOps: a guide to its adoption in the context of responsible AI," in *Proceedings of the 1st Workshop on Software Engineering for Responsible AI*, ser. ICSE '22. ACM, May 2022, p. 45–49.

[11] P. Giudici, M. Centurelli, and S. Turchetta, "Measuring ai safety," *SSRN Electronic Journal*, 2022. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4298352

[12] L. Visengeriyeva, A. Kammer, I. Bär, A. Kniesz, and M. Plöd. (2020) MLOps Principles. [Online]. Available: https://ml-ops.org/content/mlops-principles

[13] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, 2021. [Online]. Available: https://www.mdpi.com/2504-4990/3/2/20

[14] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, no. 1, p. 4209, 2022.

[15] A. E. Ouadrhiri and A. Abdelhadi, "Differential Privacy for Deep and Federated Learning: A Survey," *IEEE Access*, vol. 10, pp. 22 359–22 380, 2022.

[16] S. V. Dibbo, "SoK: Model Inversion Attack Landscape: Taxonomy, Challenges, and Future Roadmap," in *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. IEEE, 2023, pp. 439–456.

[17] M. Bloch, O. Gunlu, A. Yener, F. Oggier, H. V. Poor, L. Sankar, and R. F. Schaefer, "An Overview of Information-Theoretic Security and Privacy: Metrics, Limits and Applications," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 5–22, 2021.

[18] J. P. Near and D. Darais, "Guidelines for Evaluating Differential Privacy Guarantees," National Institute of Standards and Technology, Tech. Rep., 2023. [Online]. Available: https://csrc.nist.gov/pubs/sp/800/226/ipd

[19] N. Jovanovi'c, M. Balunovic, M. Baader, and M. T. Vechev, "On the Paradox of Certified Training," *Trans. Mach. Learn. Res.*, vol. 2022, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:252873705

[20] L. Li, T. Xie, and B. Li, "SoK: Certified Robustness for Deep Neural Networks," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 1289–1310.

[21] N. M. Gürel, X. Qi, L. Rimanic, C. Zhang, and B. Li, "Knowledge enhanced machine learning pipeline against diverse adversarial attacks," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 3976–3987. [Online]. Available: https://proceedings.mlr.press/v139/gurel21a.html

[22] Y. Cheng, J. Deng, J. Li, S. A. DeLoach, A. Singhal, and X. Ou, *Metrics of Security*. Springer International Publishing, 2014, pp. 263–295.

[23] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions," vol. 2, no. 3.

[24] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[25] F. Biessmann and D. Refiano, "Quality Metrics for Transparent Machine Learning With and Without Humans In the Loop Are Not Correlated," in *Workshop on Theoretical Foundations, Criticism, and Application Trends of Explainable AI*, 2021. [Online]. Available: https://arxiv.org/abs/2107.02033

[26] P. Denzel, S. Brunner, P.-P. Luley, C. Frischknecht-Gruber, M. U. Reif, F.-P. Schilling, A. Amini, M. Repetto, A. Iranfar, J. Weng, and R. Chavarriaga, "A framework for assessing and certifying explainability of health-oriented ai systems," nov 2023, explainable AI in Medicine Workshop, Lugano, Switzerland, 2-3 November 2023. [Online]. Available: https://digitalcollection.zhaw.ch/handle/11475/29258