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

Research Paper

Philip Singer[1], Kathrin Brecker[1], Ali Sunyaev[1]

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
philip.singer@student.kit.edu, {brecker,sunyaev}@kit.edu

Abstract. Developers are confronted with a plethora of guidelines to address trustworthy AI (TAI). However, guidelines remain mainly abstract and implications for the development process are often difficult to address technically in a holistic manner across the AI development process of data preprocessing, model development, model evaluation, and inferencing. We synthesize existing research on TAI guidelines and realization of TAI qualities, such as security or fairness, to derive techniques that TAI development platforms need to prioritize and include to guide developers for addressing TAI. Our study reveals 34 techniques for achieving TAI on development platforms along six main technique categories: trustworthy training data, trustworthy model training, tests for trustworthy model evaluation, monitoring and control for trustworthy inferencing, external and internal transparency, and data protection. This study contributes to a better understanding of realizing TAI so that researchers and practitioners can better address TAI requirements and ensure impact of TAI guidelines.

Keywords: Trustworthy Artificial Intelligence, AIaaS, AI services

1 Introduction

The advancing application of artificial intelligence (AI), even in sensitive areas of peoples' life such as skin cancer diagnoses for patients (Sangers et al., 2021), extreme weather prediction (Qin et al., 2023), or airplane development (Braiek et al., 2023), has raised calls for guidelines to ensure the development of trustworthy AI (TAI) systems (Adadi & Berrada, 2018; Schmager & Sousa, 2021).

Despite, the growing calls for TAI development, we observe that current guidelines (Hagendorff, 2020), principles (Curto & Comim, 2023; Diakopoulos et al., 2023), and best practices (Mazumder et al., 2023) to facilitate TAI development remain abstract and difficult to apply (Hagendorff, 2020; Schmager & Sousa, 2021). Guidelines lack technical details, for example, guidelines may call for ensuring fairness but developers remain uncertain which fairness metrics to choose due to metric diversity (Caton & Haas, 2023). Developers are also confronted with a plethora of algorithms and tools focusing on achieving different TAI qualities such as privacy or fairness. For example robust neural networks (Carlini & Wagner, 2017), the projected gradient descent for

19th International Conference on Wirtschaftsinformatik,
September 2024, Würzburg, Germany

adversarial training algorithm (Madry et al., 2018), or robustness focused data sanitization (Xiong et al., 2022), or tools such as IBM's "Adversarial Robustness Toolbox (ART)" or the open-source library "cleverhans" to address the TAI qualities robustness and security. However, they 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. Reflecting these reasons, guidelines have been rendered as unhelpful to change or improve developers behavior and subsequently achieving TAI in development practice remains challenging (Pant et al., 2024).

Novel cloud-based AI development platforms can be a compelling alternative to foster TAI because these platforms provide developers with best practices and tools to enable and guide the AI development (e.g., to develop, train, deploy, manage, and use AI models) (Lins et al., 2021). Eminent cloud providers, such as Amazon, Google, IBM, Microsoft, and small and medium-sized enterprises have started offering AI developer platform services (AIaaS) with varying capabilities, becoming more and more important for AI developers nowadays (Sundberg & Holmström, 2022; Zapadka et al., 2020). We propose that these AI development platforms should incorporate and prioritize techniques to achieve TAI. These platforms can then provide operationalizable means to developers lacking clear guidance on TAI development (B. Li et al., 2023).

The ever-increasing number of research articles discussing TAI is a valuable starting point to identify suitable techniques that can be integrated into cloud-based AI development platforms. Such work examines, among others, how to use non-trustworthy AI systems in a trustworthy way in different use cases (e.g., D. Wang et al., 2021), what ethical and legal implications arise from AI usage in sensitive domains (e.g., Nguyen et al., 2023), or what aspects and values are part of AI trustworthiness (e.g., Floridi et al., 2018). While current research provides valuable insights for TAI considerations, we observe three key research challenges. First, extant TAI research is spread across various disciplines (e.g., information systems, computer science, or medicine). Second, proposed techniques to achieve TAI qualities are often examined in isolation. Third, AI development platforms require TAI techniques that span the entire AI development lifecycle. Further research is required that synthesizes extant techniques proposed by research in different disciplines, combines identified techniques, and then integrates them into the entire AI lifecycle to support AI developers in every process step. Accordingly, we seek to answer the following research question: ***What are the key techniques for fostering TAI that can be integrated into AI development platforms?***

We conducted a descriptive literature review (Paré et al., 2015) that synthesizes the scattered knowledge on TAI development suggestions that can be operationalized in an AI development platform. Overall, the review revealed six categories of techniques that can be integrated into TAI development platforms: trustworthy training data, trustworthy model training, tests for trustworthy model evaluation, monitoring and control for trustworthy inferencing, external and internal transparency, and data protection. Our 34 techniques for TAI development platforms contribute to research and practice by synthesizing and aggregating the scattered knowledge and providing suggestions for techniques that should be included to set up TAI development platforms.

2 Background

2.1 AI Development Platforms

We follow a broad AI definition and define AI "as the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity" (Rai et al., 2019, p. 3). To foster the development of AI systems, nowadays various AI development platforms have emerged that are software platforms to enable the development and use of AI models by providing access to frameworks, tools, libraries, programming languages, and software development kits (Lins et al., 2021). AI development platforms provide storage and compute resources (e.g., virtual machines, physical servers, or containers), often containing accelerated hardware to manage the compute-intensive training needed for AI development. AI development platforms also provide data storage technologies to manage the large amount of high-quality data required for AI development (Demchenko et al., 2014). Developers can leverage these resources in every step of the AI development process: (1) data preprocessing, (2) model development, (3) model evaluation, and (4) inferencing (Amershi et al., 2019; Schlegel & Sattler, 2023). First, in the data preprocessing phase (1), data scientists prepare the training data and perform tasks such as data collection and selection, data cleaning, data labeling, or feature engineering. Second, during model development (2), AI developers and data scientists work with the training data to perform model training, finetuning and selection. Third, in the model evaluation phase (3) the developers evaluate the model regarding performance and other TAI qualities based on the training data and new data. In the last phase of inferencing (4), the model deployment, outcomes generation and monitoring take place. It is then possible to use the model and generate insights (Amershi et al., 2019; Schlegel & Sattler, 2023). With our study we are looking for the key techniques that can be integrated into AI development platforms to facilitate TAI in each of these development phases.

2.2 Trustworthy AI

AI can be considered trustworthy when it is developed and applied in ways that are compliant with all relevant laws, when it is safe to use, and when general ethical principles are met (Thiebes et al., 2021). AI "must work reliably, in ways that anyone can trust will be for the benefit of humanity and the whole environment" (Floridi, 2019, p. 1). To operationalize TAI, researchers often refer to key qualities of TAI systems, namely privacy, fairness, accountability, robustness, security, transparency, performance (Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019). Privacy covers that data and models must be kept private and protected against unauthorized access (Liu et al., 2023). Fairness describes that AI systems' outcomes are free from bias and discrimination against individuals and groups (Gittens et al., 2022). Accountability describes who is responsible for the AI system use, design, and subsequent consequences and how this entity can justify decisions of the AI system (Wieringa, 2020). Robustness refers to the ability of an AI model to handle unknown scenarios or anomalies well and

avoid adverse effects (e.g., performance reduction) (Fjeld et al., 2020). Security focuses on protection against intentional attacks trying to compromise AI systems' confidentiality, integrity, and availability (Leslie, 2019). Transparency enables understanding of the AI system (e.g. outcomes' explainability) (Barredo Arrieta et al., 2020). Performance describes the AI system's capability to fulfill the given task (e.g., measured as accuracy) (Q. V. Liao & Sundar, 2022).

2.3 Related Research

Research on TAI is steadily evolving and can be categorized in three major research streams: Defining TAI, TAI Requirements, Realizing TAI (Table 1).

Table 1. Related research streams on Realizing TAI

Research stream	Description	Exemplary studies
Defining TAI	Conceptualizing ethical and trust principles for AI	• Floridi et al. (2018) • Fjeld et al. (2020) • Thiebes et al. (2021)
TAI Requirements	Discussing prerequisites and criteria for TAI	• Procter et al. (2023) • Ju et al. (2023) • Otoum & Mouftah (2021)
Realizing TAI	Developing (technical) methods to achieve TAI	• Curto & Comim (2023) • B. Li et al. (2023) • **This study**

First, various articles seek to define TAI by approaching it from an ethical viewpoint or applying human trust perception principles to AI. For example, Floridi et al. (2018) propose beneficence, non-maleficence, autonomy, justice, and explicability as the five foundational principles for TAI derived from traditional bioethics. Fjeld et al. (2020) aims to achieve consensus around privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values as key thematic trends for trustworthy AI. Wieranga (2020) defines algorithmic accountability based on a literature research and includes viewpoints from different disciplines. Thiebes et al. (2021) introduce TAI by adopting the five foundational principles for TAI of Floridi et al. (2018) and propose a data-driven research framework for TAI.

Second, related articles discuss what requirements must be met to develop and use TAI in specific environments. For instance, Procter et al. (2023) examine what AI accountability means in clinical decision making. Ju et al. (2023) discuss high-level characteristics for a governmental chatbot for citizens. Otoum & Mouftah (2021) examine how TAI can be implemented for energy infrastructure management in cities.

Third, there is a large base of articles proposing techniques and algorithms to realize TAI. These articles focus mostly on achieving one TAI quality (e.g., fairness). For example, Curto & Comim (2023) present an end-to-end framework to develop fair AI by continuously including stakeholders' feedback. B. Li et al. (2023) summarize methods

for robustness, generalization, explainability and transparency, reproducibility, fairness, privacy protection, and accountability representing one of the few articles covering multiple TAI qualities together.

While related research contributes valuable insights to the understanding of TAI, we still lack knowledge on how to attain all (or a set of) TAI qualities in combination and how to combine proposed techniques to realize TAI efficiently. Therefore, we require further knowledge which techniques can be integrated into an AI development platform to foster TAI, which techniques can be effectively combined to achieve TAI qualities and in which step of the AI lifecycle these techniques need to be applied. In this study we review existing research to combine concepts, methods, and components to derive techniques for a TAI development platform supporting the entire lifecycle.

3 Research Approach

3.1 Literature Search

We conducted a literature review (Paré et al., 2015) to synthesize the state of current research on requirements and techniques for a TAI development platform, while applying guidelines for literature reviews (vom Brocke et al., 2015). The search string was constructed to reveal articles dealing with TAI development in general and platforms, resulting in the following search string: *AB (Trustworth* AND ((Artificial Intelligence) OR AI OR ML OR (Machine Learning)) AND (Development OR Platform))*

The search was conducted on five scientific databases, selected for their access to a wide range of peer-reviewed articles in the various disciplines: the ACM Digital Library, EBSCOhost, AIS eLibrary, IEEEXplore, and ProQuest. We applied our search to the abstracts to reveal a broad set of (potentially) relevant articles across disciplines. The initial search yielded 350 potentially relevant articles, as of October 16, 2023.

We conducted a relevancy check in two stages. In the first step, the 350 papers were assessed for fit based on the title and abstract, resulting in the exclusion of 269 articles. We applied the following exclusion criteria: categorized as off-topic and not dealing with TAI (90), not in English (2), duplicates (36), not a research article (5), other TAI related research not providing techniques (51), and focusing on very specific use cases (e.g., Trustworthy AI-based health chatbot; H. Wang et al., 2022) (85). In the second step, the 81 potentially relevant articles were analyzed in their entirety. 39 relevant articles were remaining for analysis after excluding 42 articles mainly dealing with non-technical aspects of TAI.

3.2 Literature Analysis

Thematic analysis was applied to the final set of 39 articles. To identify the relevant techniques for TAI development platforms, we performed six steps: ‘familiarizing with the data’, ‘generating initial codes’, ‘searching for themes’, ‘reviewing themes’, ‘naming and defining themes’ and ‘producing the report’ (Braun & Clarke, 2006).

During *familiarizing with the data*, we noticed that some articles take a rather ethical or sociotechnical perspective, while others describe very specific technical implementations towards one or multiple qualities of TAI. Few articles are also reviewing open-source tools and implemented frameworks to build TAI systems (e.g.; Liu et al., 2023; Marzouk et al., 2023).

In the *generating initial codes* phase, papers were read and text passages containing techniques or specific methods to fulfill TAI qualities for the TAI development platform were marked. For example, the text passage "Homomorphic Encryption (HE) enables computation functions on the data without accessing the plaintext by allowing mathematical operations to be performed on ciphertext without decryption. […]" (Liu et al., 2023, p. 32) was coded as "Homomorphic encryption". In total 214 initial codes were assigned (e.g., "Access control", "Data augmentation").

In *searching for themes*, the 214 initial codes were further combined when they were part of the same technique or describing a similar technique. For example, the codes "Anomaly Detection" and "Data Sanitization tests and removes abnormal samples" were combined to the theme of techniques "Training Data Monitoring". Searching for similar techniques resulted in 34 themes of techniques (e.g., debiasing, training data monitoring).

In the phase *reviewing themes*, the themes relating to similar techniques were revisited and adapted. The criteria of internal homogeneity and external heterogeneity were considered for the revision to form themes (Patton, 2002). The review was guided by the question "What is the goal of this theme of techniques within the TAI development platform?". Two themes of techniques were merged if their goals were the same. For example, "Debiasing" and "Training Data Monitoring" were merged to the higher-level category of themes "Trustworthy Training Data". Finally, we aggregated 34 technique themes into six higher-level categories of techniques (Table 2).

For the *naming and defining themes* phase the final six higher-level categories of techniques were subsumed as: "Trustworthy Training Data", "Tests", "Trustworthy Model Training", "Internal and External Transparency", "Inferencing Monitoring and Control", and "Data Protection".

Afterwards, the categories of techniques were compared to the four main phases of the AI development process: data preprocessing, model development, model evaluation, and inferencing. Contents of the respective technique category were compared to the tasks within the corresponding AI development phases and sorted accordingly. The technique category "C1: Trustworthy Training Data" comes into play mainly in the data preprocessing phase, the technique category "C2: Trustworthy Model Training" is mainly applicable in the model development phase. "Tests" are most relevant in the model evaluation phase and were thus named to "C3: Trustworthy Model Evaluation" and "Inferencing Monitoring and Control" is mainly relevant in the inferencing phase and was thus named "Trustworthy Inferencing". The technique categories "C5: Internal and External Transparency" and "C6: Data Protection" are considered cross-phase because they are relevant in all AI development phases. Table 2 shows the final technique categories including the mapped TAI lifecycle phase, exemplary techniques and addressed TAI qualities.

4 Techniques for Trustworthy AI Development Platforms

Table 2. Technique Categories from the Literature Review

Technique Category Description	**AI Dev. Lifecycle Phase**	**Exemplary Techniques**	**TAI Quality Addressed**
C1: Trustworthy Training Data Techniques for monitoring and preprocessing the training data.	Data Preprocessing (1)	Issue Detection, Debiasing, Data augmentation, Preserving Privacy	Privacy Fairness Security Robustness Performance
C2: Trustworthy Model Training Techniques to build and train robust, fair, and privacy-preserving models.	Model Development (2)	Robust Training, Model Debiasing, Differential Privacy	Privacy Fairness Robustness
C3: Trustworthy Model Evaluation Techniques to evaluate model's fairness, performance, and robustness; and ensure explainability.	Model Evaluation (3)	Fairness Evaluation, Robustness Evaluation, Ensuring Explainability	Fairness Accountability Robustness Transparency Performance
C4: Trustworthy Inferencing Techniques to monitor and actively control inferencing. Applicable in inferencing phase.	Inferencing (4)	Input Monitoring, Inferencing control, Output Monitoring	Robustness Security Transparency
C5: Internal and External Transparency Techniques to enable transparency of AI development decisions and process, incl. internal / external communication.	Applicable in all lifecycle phases	Documentation, Collaboration and Communication, Process control	Accountability Security Transparency
C6: Data Protection Techniques to transmit, store and process sensitive data securely.	Applicable in all lifecycle phases	Access Control, Homomorphic Encryption	Privacy Security

4.1 Trustworthy Training Data

The core objective of integrating TAI techniques for trustworthy training data during the data preprocessing phase is to contribute to privacy, fairness, security, robustness, and performance of these training data.

Detection: Involves identifying data issues and potential attacks. Techniques include distribution-based methods and classifier-based approaches (Liu et al., 2023). For example, maximum mean discrepancy tests (Gretton et al., 2012) and pre-trained classifiers (Gong et al., 2017) can be used to detect attacks. Further, „risk difference“ can be used to predict sensitive variables to assess dataset fairness (D. Xu et al., 2019).

Cleaning: Involves removing outliers, adversarial samples, and noise, and correcting data. Techniques include denoisers (F. Liao et al., 2018), GANs (Goodfellow et al., 2014), and data compression (X. Wang et al., 2019). For example, GANs can be used to generate clean data by predicting or generating original labels or features to replace corrupted ones (Xiong et al., 2022).

Preserving Privacy of AI users and data subjects. Methods include perturbation-based mechanisms like differential privacy (Díaz-Rodríguez et al., 2023) and methods for data anonymization and pseudonymization (Mourby et al., 2018). To anonymize data different cryptographic algorithms need to be part of the platform (Choudhury et al., 2020).

Data Augmentation: Involves filling missing values, creating new samples, or replacing privacy-sensitive variables. Techniques include simulation-based methods and using GANs or auto-encoders (Xing et al., 2023). Also distribution-based methods (Muralidhar et al., 1999) and linear regressions can impute data points (Burridge, 2003).

Debiasing: Involves reducing bias in the dataset by altering it, which can lead to decreased performance (X. Gu et al., 2022). Methods include resampling, reweighting, and representation transformation (Marzouk et al., 2023). A popular example is the Synthetic Minority Over-sampling Technique (SMOTE) to resample by creating new samples and balance the dataset (Chawla et al., 2002).

4.2 Trustworthy Model Training

During model development and training, TAI techniques support the development of robust, fair, and privacy-preserving models.

Robust Training: Ensuring models are less susceptible to anomalies or manipulation (Das et al., 2023). Example techniques are adversarial regularization, where models or learning algorithms are adopted to improve robustness (S. Gu & Rigazio, 2014), and using robust statistical methods, also known as robust learning (G. Xu et al., 2017). For instance, projected gradient descent for adversarial training (PGD-AT) is an adapted learning algorithm to learn adversarial robustness accuracy at the same time (Madry et al., 2018).

Model Debiasing: Techniques to mitigate biases in models via pre-processing (see previous section) in-processing (adapting the model itself), and post-processing (adapting the model outputs) methods (Gittens et al., 2022). For example, adversarial debiasing is an in-processing technique and aims to minimize the disclosure of sensitive data by the model while the model is trained (Zhang et al., 2018).

Preserving Privacy is also relevant during model development and training. This includes applying differential privacy to optimization algorithms, like DP-SGD (Abadi et al., 2016) and teacher-student architectures to protect the actual model against spying (Papernot et al., 2018). In federated learning environments, adapted methods ensure

data privacy by keeping unprotected data on client-side and exchanging only (noised) model parameters during training (Cao et al., 2020).

4.3 Trustworthy Model Evaluation

Once the model is built, the TAI development platform offers techniques for model evaluation that assess fairness, accountability, robustness, and performance, while contributing to transparency by explaining the model.

Fairness Evaluation: Various metrics such as demographic parity (Dwork et al., 2011), predictive equality (Hardt et al., 2016), and equalized odds (Agarwal et al., 2018) are employed to assess model fairness. A TAI development platform should provide processing methods and frameworks for fairness evaluations, along with tools like Fairlearn (Weerts et al., 2023) and FairTest (Tramer et al., 2017).

Performance: Techniques such as sandbox and simulation environments such as the simulator "Gazebo" (Koenig & Howard, 2004) are used for comprehensive and effort effective performance evaluation (Schaich Borg, 2021). Those reusable environments support reproducibility and accountability through standard test cases (Shneiderman, 2020).

Robustness Evaluation: Metrics like Security Evaluation Curve (Biggio et al., 2014), Loss Sensitivity (Arpit et al., 2017), and Empirical Robustness (Moosavi-Dezfooli et al., 2016) are utilized to quantify model robustness. Benchmarking techniques are employed to evaluate models against adversarial attacks (Croce & Hein, 2020). Toolkits such as DeepRobust are utilized for robustness assessment (DSE-MSU, 2019/2024).

Explainability including global and local methods like including surrogate models (Wu et al., 2018), feature importance techniques (Zien et al., 2009), and attention-based methods (Atkinson et al., 2020), are used to increase model transparency. A TAI development platform offers interpretable models (*InterpretML*, 2023) and combines visual and textual explanations to enhance understanding in users' language (Shneiderman, 2020; Zeiler & Fergus, 2013).

4.4 Trustworthy Inferencing

The last phase is the inferencing phase. In this phase the developed and successfully tested model is deployed and monitored during inferencing. External transparency and user communication are especially relevant in this phase as external stakeholders play a key role in meeting trust requirements.

Input Monitoring: Involves checking incoming data for attacks and anomalies. Techniques include misuse detection to identify harmful data or odd usage patterns (Javadi et al., 2021). Methods such as subnetworks (Metzen et al., 2017), feature squeezing (W. Xu et al., 2017), and ensemble defender modules are used for attack detection (Darvish Rouani et al., 2019). Additionally, AI-based quality checks ensure high-quality input data (X. Wang et al., 2019). For example, similarity-based data quality scores (Carrara et al., 2019) and real-time artifact detection (de Fauw et al., 2018) on images are employed.

Input Transformation: After monitoring and detecting data issues, input transformation techniques are applied to mitigate attacks and rectify issues. Techniques include compression to reduce the impact of adversarial attacks (X. Wang et al., 2019) and perturbation rectifying networks to remove artificial perturbations from data (Akhtar et al., 2018). Conditional GANs can also be used for data rectification (G. Li et al., 2020).

Output Monitoring regarding model performance issues before sending model output back to the client. Methods include performance monitoring (Das et al., 2023), concept drift detection (Klinkenberg & Joachims, 2000), and ethical validation to ensure fairness (Lu et al., 2023). Tools like "Evidently AI" can be used for monitoring model performance and usage (Murindanyi et al., 2023).

Inferencing Control: Focuses on human intervention during inferencing to increase accountability, robustness, and security. For example, fail-safe mechanisms (B. Li et al., 2023), and emergency stops (Lu et al., 2023) are implemented to ensure safe usage and enable immediate deactivation of the AI component in risky situations.

4.5 Transparency and Data Protection Techniques Facilitating TAI

Next to the phase-specific techniques, the TAI development platform provides cross phase techniques that contribute to privacy, accountability, transparency, and security.

Data Protection for confidentiality, integrity, and availability throughout the whole lifecycle. A special concern is external data protection including methods to protect data as close to the data source as possible during data acquisition (e.g., TLS as cryptographical transmission protocol; Park et al., 2022). Access control mechanisms ensure authorized access (Alexander et al., 2023). Additionally, encryption methods like homomorphic encryption are utilized to keep data secured during preprocessing and training (Díaz-Rodríguez et al., 2023). Trusted Execution Environments (TEE) provide isolated execution environments on hardware levels, enabling privacy-preserving computing (Hoekstra et al., 2013).

Documentation encompasses phase-dedicated and overarching system information. Comprehensive AI system information contains data factsheets including data provenance from the preprocessing phase (Toreini et al., 2020), model cards from the model development and evaluation phases (Mitchell et al., 2019) and service fact sheets for the inferencing phase (Arnold et al., 2019). Further, co-versioning of data and models contribute to accountability with tools like Data Version Control (Perez-Cerrolaza et al., 2023).

Communication and Collaboration. Effective communication channels within development teams and with external stakeholders contribute to transparency. For instance, standardized feedback mechanisms like active learning improve model performance (Q. V. Liao & Sundar, 2022). Sharing incidents openly and in a standardized way contributes to accountability and reduces reputational costs for the organization in the long-run (Das et al., 2023). Communicating data quality issues or uncertainty enhances system robustness as it enables humans to intervene and take control to avoid harm and increase performance (Dietterich, 2017).

Control. (Organizational) data and AI governance is a vast field of policies, processes, roles and tools to control the application and development of datasets and AI

models (Butcher & Beridze, 2019; Dama International, 2017). Data management plans enforce governance frameworks that guide developers and governance team members in handling data securely (Perez-Cerrolaza et al., 2023). MLOps frameworks, like MLFlow (Zaharia et al., 2018), extend DevOps practices to AI development and integrate trust enhancing methods (Borg, 2022; B. Li et al., 2023).

5 Discussion

Principal findings. Our literature review revealed 34 techniques categorized in six themes that can be integrated into AI development platforms to operationalize TAI. We also mapped the techniques with the AI development process phases (data preprocessing, model development, model evaluation, and inferencing). Our findings show that TAI needs to be pursued and prioritized in every development phase and how a TAI development platform can support this process end to end. The derived TAI development platforms' techniques contribute to several TAI qualities simultaneously, showing how TAI development can be enabled holistically and by default.

Implications for Research. Existing literature has discussed TAI guidelines and qualities that should be addressed for TAI systems (Fjeld et al., 2020; Floridi et al., 2018; Thiebes et al., 2021). For example, that AI systems should ensure fairness (e.g., no discrimination) (Das et al., 2023; Díaz-Rodríguez et al., 2023). However, research mainly studies TAI qualities and techniques in isolation (e.g.; Carlini & Wagner, 2017; Madry et al., 2018), ultimately lacking knowledge to understand how theoretical guidelines can be put into practice to help achieving TAI systems that ensure TAI qualities. Our aggregated six categories along 34 techniques provide a synthesized overview based on the literature that shows how various TAI qualities can be addressed by these techniques in parallel, such as GANs for general purpose cleaning, data augmentation, and data rectification (Goodfellow et al., 2014; G. Li et al., 2020; Xing et al., 2023). Thereby we provide insights into TAI quality combination and respective techniques for developing TAI systems in line with guidelines. This paves the way for future research to further investigate the consequences of combining TAI qualities and techniques (e.g., synergies or adverse effects; Petkovic, 2023; Steimers & Schneider, 2022). Future research can build on our findings by testing and complementing the proposed techniques for addressing TAI qualities holistically.

We advance research by providing starting points for realizing TAI to guide developers on AI development platforms. Current research provides only limited orientation for developers to realize TAI along the development process (e.g. Cao et al., 2020; Toreini et al., 2020). We show how TAI qualities can be addressed on AI development platforms along the development process (data preprocessing, model development, model evaluation, inferencing). We mapped the proposed TAI techniques to development lifecycle's key phases to contribute to a better understanding of how each technique contributes to one or more TAI qualities along the lifecycle. Future research can test and complement suggestions and evaluate suitable combinations of identified TAI techniques that can be integrated into AI platforms.

Implications for Practice. The results provide a starting point for AI developers and platform providers to construct TAI development platforms by providing concrete techniques to include at every stage of the AI development process. The 34 techniques derived show which algorithms and tools developers can choose from on a TAI development platform (as a one stop shop) to fulfill TAI qualities.

Our results indicate what techniques to include for the construction of TAI development platforms, thereby accelerating the process to develop TAI. We show how organizations could harness development platforms and integrate extant techniques to provide TAI development guidance for developers. Thereby, organizations can benefit from a ready to use development platform without having to deal with operationalizing TAI guidelines individually.

Limitations and future research. This literature review does not come without limitations paving the way for future research. First, there is an inherent bias for the selection of the search string and the identification of relevant literature which we attempted to mitigate by describing our choices as transparent as possible. Second, TAI development platform techniques can significantly contribute to TAI development, but there are other aspects that need to be considered for TAI development as well that are not reflected in this study. For example, the construction of TAI systems also requires a safety-aware organizational culture in the planning and requirements engineering phase. Third, existing research evaluates individual TAI techniques, but so far there has been no evaluation of the components in combination as proposed in our study. Therefore, future research should further investigate how different TAI qualities can potentially interfere with each other when introduced combined as suggested. The understanding of ethical concepts such as TAI might also further evolve over time so that the derived techniques might need to be adapted or updated as theoretical concepts change.

6 Conclusion

Policymakers and society are increasingly demanding AI systems that prioritize TAI qualities (privacy, fairness, accountability, robustness, security, transparency, performance as qualities). However, the demand can only be fulfilled if theoretical and abstract guidelines can be applied by developers in their actual development work. To understand how TAI can be realized, techniques relevant to be included and prioritized for a platform guiding the TAI development process need to be identified. We conducted a literature review and identified 34 relevant TAI techniques linked to 6 technique categories: trustworthy training data, trustworthy model development, tests for trustworthy model evaluation, monitoring and control for trustworthy inferencing, internal and external transparency, and data protection. Our results indicate that TAI needs to be pursued throughout the entire AI development process (data preprocessing, model development, model evaluation, inferencing). The combination of TAI qualities into one platform that guides all phases of AI development to ensure TAI will be especially relevant for future TAI advancements.

References

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

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

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A Reductions Approach to Fair Classification. *International Conference on Machine Learning*, 60–69.

Akhtar, N., Liu, J., & Mian, A. (2018). Defense Against Universal Adversarial Perturbations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2018.00357

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

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

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, *63*(4/5), 6:1-6:13. https://doi.org/10.1147/JRD.2019.2942288

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., & Lacoste-Julien, S. (2017). A Closer Look at Memorization in Deep Networks. *International Conference on Machine Learning*. https://doi.org/10.5555/3305381.3305406

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence*, *289*, N.PAG-N.PAG. https://doi.org/10.1016/j.artint.2020.103387

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

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

Borg, M. (2022). Agility in Software 2.0 – Notebook Interfaces and MLOps with Buttresses and Rebars. In A. Przybyłek, A. Jarzębowicz, I. Luković, & Y. Y. Ng (Eds.), *Lean and agile software development* (pp. 3–16). Springer International Publishing and Springer. https://doi.org/10.1007/978-3-030-94238-0_ 1

Braiek, H. B., Reid, T., & Khomh, F. (2023). Physics-Guided Adversarial Machine Learning for Aircraft Systems Simulation. *IEEE Transactions on Reliability*, *72*(3), 1161–1175. https://doi.org/10.1109/TR.2022.3196272

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing*, *13*(4), 321–327. https://doi.org/10.1023/A:1025658621216

Butcher, J., & Beridze, I. (2019). What is the State of Artificial Intelligence Governance Globally? *The RUSI Journal*, *164*(5–6), 88–96. https://doi.org/10.1080/03071847.2019.1694260

Cao, H., Liu, S., Zhao, R., & Xiong, X. (2020). IFed: A novel federated learning framework for local differential privacy in Power Internet of Things. *International Journal of Distributed Sensor Networks*, *16*(5), 155014772091969. https://doi.org/10.1177/1550147720919698

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

Carrara, F., Falchi, F., Caldelli, R., Amato, G., & Becarelli, R. (2019). Adversarial image detection in deep neural networks. *Multimedia Tools and Applications*, *78*(3), 2815–2835. https://doi.org/10.1007/s11042-018-5853-4

Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*. https://doi.org/10.1145/3616865

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

Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., & Das, A. (2020). A Syntactic Approach for Privacy-Preserving Federated Learning. In *ECAI 2020* (pp. 1762–1769). IOS Press. https://doi.org/10.3233/FAIA200290

Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *International Conference on Machine Learning*. http://proceedings.mlr.press/v119/croce20b/croce20b.pdf

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

Dama International. (2017). *DAMA-DMBOK: data management body of knowledge*. Technics Publications, LLC.

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

Das, S. D., Bala, P. K., & Mishra, A. N. (2023). Towards Defining a Trustworthy Artificial Intelligence System Development Maturity Model. *Journal of Computer Information Systems*, 1–22. https://doi.org/10.1080/08874417.2023.2251443

de Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, *24*(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

Demchenko, Y., de Laat, C., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. In W. W. Smari (Ed.), *2014 International Conference on Collaboration Technologies and Systems (CTS 2014)* (pp. 104–112). IEEE. https://doi.org/10.1109/CTS.2014.6867550

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

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

Dietterich, T. G. (2017). Steps Toward Robust Artificial Intelligence. *AI Magazine*, *38*(3), 3–24. https://doi.org/10.1609/aimag.v38i3.2756

DSE-MSU. (2024). *DSE-MSU/DeepRobust* [Python]. https://github.com/DSE-MSU/DeepRobust (Original work published 2019)

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

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

Floridi, L. (2019). Establishing the Rules for Building Trustworthy AI. *Nature Machine Intelligence*, *1*, 1–2. https://doi.org/10.1038/s42256-019-0055-y

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

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

Gong, Z., Wang, W., & Ku, W.-S. (2017). Adversarial and Clean Data Are Not Twins. *International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*. https://doi.org/10.1145/3593078.3593935

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *Communications of the ACM*. https://doi.org/10.1145/3422622

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, *13*, 723–773.

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

Gu, X., Li, J., Raymond Choo, K.-K., Zhang, T., Wei Ren, W., & Tianqing, Z. (2022). Privacy, accuracy, and model fairness trade-offs in federated learning. *Computers & Security*, *122*, 102907. https://doi.org/10.1016/j.cose.2022.102907

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

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

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

InterpretML. (2023). http://interpretml.github.io/

Javadi, S. A., Norval, C., Cloete, R., & Singh, J. (2021). Monitoring AI Services for Misuse. In M. Fourcade (Ed.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 597–607). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462566

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Ju, J., Meng, Q., Sun, F., Liu, L., & Singh, S. (2023). Citizen preferences and government chatbot social characteristics: Evidence from a discrete choice experiment. *Government Information Quarterly*, *40*(3), N.PAG-N.PAG. https://doi.org/10.1016/j.giq.2022.101785

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

Koenig, N., & Howard, A. (2004). Design and use paradigms for gazebo, an open-source multi-robot simulator. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2149–2154. https://doi.org/10.1109/IROS.2004.1389727

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

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

Li, G., Ota, K., Dong, M., Wu, J., & Li, J. (2020). DeSVig: Decentralized Swift Vigilance Against Adversarial Attacks in Industrial Artificial Intelligence Systems. *IEEE Transactions on Industrial Informatics*, *16*(5), 3267–3277. https://doi.org/10.1109/TII.2019.2951766

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2018.00191

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

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

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

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

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

Marzouk, M., Zitoun, C., Belghith, O., & Skhiri, S. (2023). The Building Blocks of a Responsible AI Practice: An Outlook on the Current Landscape. *IEEE Intelligent Systems*, 1–10. https://doi.org/10.1109/MIS.2023.3320438

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

Metzen, J. H., Genewein, T., Fischer, V., & Bischoff, B. (2017). On Detecting Adversarial Perturbations. *International Conference on Learning Representations*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.282

Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., Smith, H., Aidinlis, S., & Kaye, J. (2018). Are `pseudonymised' data always personal data?

Implications of the GDPR for administrative data research in the UK. *Computer Law & Security Review*, *34*(2), 222–233. https://doi.org/10.1016/j.clsr.2018.01.002

Muralidhar, K., Parsa, R., & Sarathy, R. (1999). A General Additive Data Perturbation Method for Database Security. *Management Science*, *45*(10), 1399–1415. https://doi.org/10.1287/mnsc.45.10.1399

Murindanyi, S., Nagwovuma, M., Nansamba, B., & Marvin, G. (2023). Explainable Ensemble Learning and Trustworthy Open AI for Customer Engagement Prediction in Retail Banking. *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, 198–206. https://doi.org/10.1145/3607947.3607983

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical Principles for Artificial Intelligence in Education. *Education and Information Technologies*, *28*(4), 4221–4241. https://doi.org/10.1007/s10639-022-11316-w

Otoum, S., & Mouftah, H. T. (2021). Enabling Trustworthiness in Sustainable Energy Infrastructure Through Blockchain and AI-Assisted Solutions. *IEEE Wireless Communications*, *28*(6), 19–25. https://doi.org/10.1109/MWC.018.2100194

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in the Age of AI: An Analysis of AI Practitioners' Awareness and Challenges. *ACM Transactions on Software Engineering and Methodology*, *33*(3), 80:1-80:35. https://doi.org/10.1145/3635715

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., & Erlingsson, Ú. (2018). Scalable Private Learning with PATE. *International Conference on Learning Representations*.

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

Park, S., Kim, S., & Lim, Y. (2022). Fairness Audit of Machine Learning Models with Confidential Computing. *Proceedings of the ACM Web Conference 2022*, 3488–3499. https://doi.org/10.1145/3485447.3512244

Patton, M. Q. (2002). Two Decades of Developments in Qualitative Inquiry: A Personal, Experiential Perspective. *Qualitative Social Work*, *1*(3), 261–283. https://doi.org/10.1177/1473325002001003636

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

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

Procter, R., Tolmie, P., & Rouncefield, M. (2023). Holding AI to Account: Challenges for the Delivery of Trustworthy AI in Healthcare. *ACM Trans. Comput.-Hum. Interact.*, *30*(2). https://doi.org/10.1145/3577009

Qin, Y., Su, C., Chu, D., Zhang, J., & Song, J. (2023). A Review of Application of Machine Learning in Storm Surge Problems. *Journal of Marine Science and Engineering*, *11*(9), 1729. https://doi.org/10.3390/jmse11091729

Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: Toward human-AI hybrids. *MIS Quarterly*, *43*(1). https://wrap.warwick.ac.uk/113653/

Sangers, T. E., Wakkee, M., Kramer–Noels, E. C., Nijsten, T., & Lugtenberg, M. (2021). Views on mobile health apps for skin cancer screening in the general population: An in–depth qualitative exploration of perceived barriers and facilitators*. *British Journal of Dermatology*, *185*(5), 961–969. https://doi.org/10.1111/bjd.20441

Schaich Borg, J. (2021). Four investment areas for ethical AI: Transdisciplinary opportunities to close the publication-to-practice gap. *Big Data & Society*, *8*(2), 205395172110401. https://doi.org/10.1177/20539517211040197

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

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

Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Trans. Interact. Intell. Syst.*, *10*(4). https://doi.org/10.1145/3419764

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

Sundberg, L., & Holmström, J. (2022). Towards 'Lightweight' Artificial Intelligence: A Typology of AI Service Platforms. *AMCIS 2022 Proceedings*. https://aisel.aisnet.org/amcis2022/sig_odis/sig_odis/13

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

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

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A., & Lin, H. (2017). FairTest: Discovering Unwarranted Associations in Data-Driven Applications. *2nd IEEE European Symposium on Security and Privacy*, 401–416. https://doi.org/10.1109/EuroSP.2017.29

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, *37*(1). https://doi.org/10.17705/1CAIS.03709

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

Wang, H., Gupta, S., Singhal, A., Muttreja, P., Singh, S., Sharma, P., & Piterova, A. (2022). An Artificial Intelligence Chatbot for Young Peopleâ\text\euro^{TM}s Sexual and Reproductive Health in India (SnehAI): Instrumental Case Study. *Journal of Medical Internet Research*, *24*(1), N.PAG-N.PAG. https://doi.org/10.2196/29969

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

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023). Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, *24*(257). http://jmlr.org/papers/v24/23-0389.html

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

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

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

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

Xu, D., Yuan, S., Zhang, L., & Wu, X. (2019). FairGAN +: Achieving Fair Data Generation and Classification through Generative Adversarial Nets. In C. Baru (Ed.), *2019 IEEE International Conference on Big Data* (pp. 1401–1406). IEEE. https://doi.org/10.1109/BigData47090.2019.9006322

Xu, G., Cao, Z., Hu, B.-G., & Príncipe, J. (2017). Robust support vector machines based on the rescaled hinge loss function. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2016.09.045

Xu, W., Evans, D., & Qi, Y. (2017). Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2018.23198

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., Xie, F., & Zumar, C. (2018). Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Engineering Bulletin*. https://doi.org/10.1145/3399579.3399867

Zapadka, P., Hanelt, A., Firk, S., & Jana, O. (2020). *Leveraging "AI-as-a-Service"–antecedents and consequences of using artificial intelligence boundary resources*. https://scholar.archive.org/work/t746u3nb4zay5mutuwzts3x7ci/access/wayback/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1209&context=icis2020

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *European Conference on Computer Vision*, *8689*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_ 53

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.

Zien, A., Krämer, N., Sonnenburg, S., & Rätsch, G. (2009). The Feature Importance Ranking Measure. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine learning and knowledge discovery in databases* (Vol. 5782, pp. 694–709). Springer. https://doi.org/10.1007/978-3-642-04174-7_ 45