PhD-FSTM-2024-012
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 04/03/2024 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN INFORMATIQUE

by

Mary Katherine ROSZEL
Born on 24 October 1993 in Florida, United States of America

TOWARDS TRUSTWORTHY ARTIFICIAL INTELLIGENCE IN PRIVACY-PRESERVING COLLABORATIVE MACHINE LEARNING

Dissertation defence committee

Dr. Radu STATE, dissertation supervisor
Professor, Université du Luxembourg

Dr. Gilbert FRIDGEN, Chairman
Professor, Université du Luxembourg

Dr. Vijay GURBANI, Vice Chairman
Research Associate Professor, Illinois Institute of Technology / Chief Data Scientist, Vail Systems, Inc.

Dr. Andrey MARTOVOY
Senior Advisor - Innovation & Digital, Association des Banques et Banquiers, Luxembourg (ABBL)

Dr. Jean HILGER
Head of Finnovation Hub, Université du Luxembourg

To my beloved husband

Acknowledgments

I would like to express my sincere gratitude to my supervisor Prof. Dr. Radu State for the opportunity to pursue my PhD and for the guidance and support during my studies. I appreciate his encouragement, guidance, and advice, especially during times of wavering motivation.

I would like to extend my appreciation to my CET members: Dr. Jean Hilger and Dr. Vijay Gurbani for providing their insights and constructive feedback. Further, I would like to express my appreciation to Prof. Dr. Gilbert Fridgen and Dr. Andrey Martovoy for agreeing to join my defense jury and taking the time to review my dissertation.

My sincerest thanks are extended to Association des Banques et Banquiers, Luxembourg (ABBL) and its Fondation ABBL pour l'éducation financière for financially supporting my PhD and providing continuous input and facilitation of the project. Special thanks to Dr. Andrey Martovoy for his coordination and support throughout the duration of the project.

I would like to thank my colleagues in SEDAN lab for providing endless entertainment throughout the years. I would like to specifically extend my thanks to Dr. Beltran Fiz and Dr. Robert Norvill for all of their assistance in our research collaborations.

Lastly, I want to thank my husband, Dr. Sean Rivera, for his never-ending support, countless hours spent discussing research, late nights working on papers together, and always believing in me. I could not have succeeded without him.

Abstract

Artificial Intelligence (AI) systems are proliferating in our society due to their capacity to simulate human intelligence, behaviors, and processes. The increased utilization of AI systems in society, especially in high-risk settings such as autonomous systems and healthcare, has been accompanied by an increased concern about the impact of AI systems on society. In recent years, vulnerabilities to algorithmic bias, adversarial attacks, and data breaches have resulted in the critical assessment of how AI systems can be designed to be inherently trustworthy.

This dissertation presents the key concepts of trustworthiness in AI systems, with a focus on identifying the challenges associated with designing, developing, and deploying collaborative AI. Towards this purpose, key elements of trustworthy AI are identified, culminating in a set of concise guidelines that developers can leverage in the development of trustworthy AI. Further, this dissertation explores how techniques initially created solely for privacy, specifically federated learning, can be leveraged to build trust in machine-learning environments.

Federated learning is assessed for its implications on trustworthy principles, with a particular focus on how privacy is established to enable collaboration between participants without the sharing of private data. The security of federated learning is further assessed by demonstrating the impact of targeted model poisoning attacks and an assessment of Byzantine-tolerant defense mechanisms to prevent and defend against such attacks. Further, the potential for federated learning to be leveraged for compliance with regulatory requirements is assessed.

Index

List of Figures

List of Tables

List of Abbreviations

AP Accuracy Parity

AML Anti-Money Laundering

AI Artificial Intelligence

AIA Artificial Intelligence Act

CDD Customer Due Diligence

EO Executive Order

FP / FPR False Positive / False Positive Rate

FN / FNR False Negative / False Negative Rate

FedAvg Federated Averaging

FL Federated Learning

GDPR General Data Protection Regulation

HLEG High-Level Expert Group

iid Independent and Identically Distributed

IoT Internet of Things

KYC Know Your Customer

KYM Know Your Model

ML Machine Learning

NLP Natural Language Processing

RFA Robust Federated Averaging

Chapter 1

Introduction

Artificial Intelligence (AI) refers to the development of machines, especially computer systems, that simulate human intelligence, behaviors, and processes. These encompass learning, reasoning, problem-solving, perception, language understanding, and self-correction, among others. In recent years, the use of AI technologies has proliferated in our society due to advancements in fields as diverse as healthcare, finance, transportation, and entertainment. AI systems are being utilized in nearly every sector, with notable applications in autonomous systems [392], education [68], manufacturing [198], and healthcare [231]. Advancements in big data, computational power, and algorithm sophistication continue to propel AI's capabilities. However, with this promise also come challenges with the trustworthiness of AI-powered systems, making the study and application of trustworthy AI critical.

The trustworthiness of an AI system implies that the development of such a system considers the greater ethical, technical, and practical impacts on humanity. In recent years, growing interest and advancements in the field of AI have drawn attention to the con-

cerns of the large-scale impacts of such AI systems [9]. In particular, recent research has raised concern about issues with lack of explainability [18], bias [234], loss of privacy [157], and safety [349]. Several regulating and policy-making agencies have suggested that the development of Trustworthy AI systems is vital to ensure that AI systems do not cause unintended harm [79, 261, 354].

Consider, for example, the use of AI in medical diagnoses and decisions where an AI assists medical personnel in making decisions based on diagnostic scans. In theory, these systems should provide faster and more accurate diagnoses, allowing hospitals to treat additional patients, and reduce human error. However, while these systems perform well in controlled environments, in real-world scenarios they often perform poorly and increase the time required to make diagnoses [30]. In this example, the issue of trust in such a system comes into consideration on multiple levels: the system must have the trust of the medical personnel utilizing it, the patient, and its greater community. How do we trust that the decisions these systems make are accurate, and who takes responsibility if a patient is misdiagnosed, or if time is wasted on attempting to use a faulty system?

The concept of trustworthiness should be at the forefront of consideration when we think about AI development and deployment into society. In general, the technical aspects of trustworthy AI may focus on the concepts of fairness and non-discrimination, privacy, safety and security, and transparency and explainability, whereas ethical components may focus on human control and the promotion of human values. These aspects govern how AI makes decisions about vulnerable populations, ensures user privacy, maintains security against data leaks and malicious attacks, and explains the decisions it makes. To encourage large-scale AI adoption and increase trust, the burden is on the creators to address trust in their deployed systems. However, even though there is a plethora of research, literature, and policy texts on AI, there is little consistency in the definition of Trustworthy AI and which elements are required to develop a trustworthy system.

One emerging technology with notable applications in increasing trustworthiness in AI systems is Federating Learning (FL). FL is a popular distributed machine learning (ML) setting where multiple clients can collaboratively train an ML model without sharing private data [233]. Typically orchestrated by a central server, FL follows a multi-round, multi-agent-based strategy. In each round, the server distributes a current global ML model to a random subset of participants, who then separately leverage private data to locally update the model. The updated models are sent back to the server, which aggregates the updates into a new global model. Due to its strength in allowing many participants to collaborate, FL has gained popularity, with applications in mobile devices [206], speech and image recognition [222], finance [215], and medicine [200].

In this dissertation, the concept and technicalities of trustworthiness in AI will be explored. The various texts about Trustworthy AI are explored, revealing commonality between trustworthy concepts around six main principles: accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security. These principles appear in the majority of trustworthy texts, with an additional emphasis on improving transparency in each principle. Guidelines towards developing and deploying trustworthy AI are defined, with connections to the main concepts in Trustworthy AI literature. Further, several applications that leverage FL to increase trust in AI are proposed. FL has a multitude of benefits that, if implemented properly, have significant applications in increasing trust in AI. In particular, data privacy, security, and robustness are at the forefront of the benefits of FL systems, ensuring that private data remains under user and organizational control. FL can assist organizations comply with stringent data protection regulations, such as the GDPR and AI Act in the European Union. By emphasizing local computation and maintaining user control, FL can lead to more transparent, private, secure, and accountable AI systems.

Specifically, the following research questions are explored:

1. What are the main principles associated with Trustworthy AI, and how can AI systems be designed, developed, and deployed to be inherently trustworthy?
2. Which design principles are essential to ensure that AI systems can be audited for trustworthiness?
3. How can privacy-preserving methods such as FL be leveraged to increase trust while facilitating secure and efficient collaboration amongst different individuals, institutions, or other players?
4. What are the security implications of FL, and how can FL systems be fortified against adversarial attacks, especially when processing sensitive data?
5. What is the impact of attacker-resistant aggregation mechanisms on the performance of FL models?
6. How can FL aid in ensuring AI regulatory compliance?

1.1 Dissertation Structure

In this dissertation, methods to increase trust in AI are discussed, primarily focusing on leveraging FL for secure distributed ML collaboration among multiple parties. A novel approach is introduced to increase trust in AI implementation by providing a set of clear guidelines for creators to leverage in the development of AI systems. Several applications for the use of FL are discussed, focused on encouraging collaboration while ensuring data privacy, model performance, and secure computation, and the use of FL to ease AI regulation is discussed. Finally, considerations on the security and safety of federated ML are explored with an analysis of model poisoning in several scenarios.

Chapter 2 provides in-depth background information on the field of trustworthy AI. The various international and national approaches to trustworthy AI are discussed, including the ethical considerations proposed by leading authorities.

Chapter 3 introduces the Know Your Model (KYM) concept. This concept is influenced by the idea that all models have a unique identity and that model characteristics can be leveraged to know and trust models. To ”know” a model implies collecting, recording, and storing detailed records of the processes undergone during the development of a model, subsequently establishing *model identity*. Twenty key guidelines are proposed to establish a model's identity, particularly around 4 core principles: **efficacy**, **reliability**, **safety**, and **responsibility**.

Chapter 4 introduces FL as a privacy-preserving collaborative learning method. The types of architectures and aggregation mechanisms are discussed. Several use cases are explored, including healthcare, finance, and manufacturing. The implications of FL on increasing trust in AI are outlined. Lastly, a use case applying FL to anti-money laundering is explored.

Chapter 5 analyzes the security and privacy considerations of FL. In this chapter, a thorough analysis of the behavior of byzantine aggregation mechanisms against model poisoning in an FL setting is explored. In particular, the performance of popular defenses such as Krum, Multi-Krum [45], Norm-Difference Clipping [332], and Robust Federated Averaging (RFA) [273] are discussed. Model poisoning is conducted to explore the impact of adversarial attacks on each aggregation mechanism. The impact of each defense mechanism on the performance of the FL model is measured.

Chapter 6 explores applying FL to fulfill the requirements of the AI Act proposed by the European Commission in 2023. In this chapter, an FL regulatory sandbox is proposed to foster an environment for developer/regulator collaboration in a privacy-preserving manner.

1.2 Contributions

The following published contributions are included in this dissertation. Contributions are listed in chronological order.

- Roszel, M., Norvill, R., & State, R. An Analysis of Byzantine-Tolerant Aggregation Mechanisms on Model Poisoning in Federated Learning. In *International Conference on Modeling Decisions for Artificial Intelligence*, MDAI 2022, Sant Cugat del Vallès, Spain, 2022, pp. 143-155. Cham: Springer International Publishing. [297] (*Included in Chapter 5*)
- Roszel, M., Fiz, B., Norvill, R., Hilger, J., & State, R. Know Your Model (KYM): Increasing Trust in AI and Machine Learning. In *Deployable AI (DAI) Workshop of the 37th AAAI Conference on Artificial Intelligence*, AAAI 2023, Washington DC, USA, 2023. [299] (*Included in Chapter 3*)
- Roszel, M., Fiz, B., & State, R. FLAIRS: Federated Learning AI Regulatory Sandbox. In *Machine Learning and Knowledge Discovery in Databases: Workshop on ML, Law, and Society: European Conference*, ECML PKDD 2023, Turin, Italy, September 18–22, 2023. Springer Nature Publishing. [298] (*Included in Chapter 6*)

Chapter 2

Trustworthy Artificial Intelligence

AI is a field of study focused on the development of machine intelligence, particularly the development of machines and systems that simulate human intelligence, behaviors, and processes, such as problem-solving, decision-making, and learning [301]. While the definition of AI is widely debated, the field encompasses systems with a variety of functions, including Expert Systems, Machine Learning, Robotics, Natural Language Processing, Computer Vision, and Speech Recognition [78]. The use of AI systems has transformed society, with wide applications and utilization in a multitude of domains, including autonomous systems [392], e-commerce [28], education[68], finance [62], healthcare [158], power electronics [397], medicine [231], smart manufacturing [198], and supply chain efficiency [343].

The increased prevalence of AI systems in society has been accompanied by an increased concern about the impact of AI systems on society. When developing AI systems we often consider its accuracy in decision making, but accuracy alone is not enough in high-stake scenarios (such as judicial decisions, and fraud detection) where an incorrect decision may

have undesirable consequences [394]. In recent years, there has been an increasing concern that AI systems are vulnerable to algorithmic bias and discrimination, such as bias due to gender, race, religion, age, nationality, or socio-economic status [6, 257]. There are several notable examples of serious repercussions of AI resulting in biased results, such as the COMPAS algorithm predicting criminal recidivism with a bias against African-American offenders [236], and the Amazon recruitment tool recommending candidates in a gender-biased manner [183]. Further, there are concerns about personal privacy, accountability, security, and safety of AI systems, and even concerns about the long-term impact of AI on job availability and human control over AI [89].

The increased concerns about the large-scale impact of AI deployment on society have sparked an interest in developing Trustworthy AI. The domain of Trustworthy AI, similar to the domains of *benevolent AI*, *responsible AI*, and *ethical AI*, focuses on the development of AI that can be trusted; systems in which the benefits are maximized and the risks and dangers are minimized [341]. The large-scale adoption of AI systems greatly depends on developing trust in not only their performance but also their greater purpose and transparency. Developing trust in AI is a dynamic process requiring continuous trust development and maintenance throughout all stages of development and deployment, this process being crucial for the greater adoption of AI systems [318].

Trustworthy AI has gained the attention of policy-makers, governments, regulatory bodies, and scientific communities. Several regulating and policy-making agencies have suggested that the development of Trustworthy AI systems is vital to ensure that AI systems do not cause unintended harm to humanity [79, 142, 261, 354]. The International Organization for Standardization (ISO), has developed standardization guidance to establish trust in AI systems using the concepts of availability, resiliency, reliability, accuracy, safety, security, and privacy [156]. Further, many policy-makers have proposed requirements for the development of AI systems that will lead to Trustworthy AI, such as the High-Level Expert

Group reports, and the Artificial Intelligence Act in the European Union [142, 338], and the *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* in the United States of America [43].

These frameworks define various key aspects that would lead to a ”trustworthy” AI system. While these frameworks will be detailed further later in this chapter, the majority of publications, regulations, and policies calling for Trustworthy AI include the following concepts [113, 175, 197]:

- Accountability: Accountability refers to the obligation to explain and justify the actions and decisions of an AI system to the users to which the system interacts or to a relevant authority (such as a regulatory body and/or policy-makers) [256].
- Explainability & Interpretability: These concepts refer to the need to explain, interpret, and understand the operations and outcomes of an AI system [210].
- Fairness & Non-Discrimination: Fairness refers to the fairness and lack of discrimination in the outcomes of an AI system, particularly toward the absence of bias toward any specific group or individual based on characteristics irrelevant to the decision-making process [236, 311].
- Privacy: Privacy primarily refers to the protection of personal data, particularly from the unauthorized or unlawful gathering and use of data. Privacy typically includes calls for users' ability to consent and control the use of their private data, users' rights to restriction, rectification, and erasure, and privacy by design [113].
- Robustness & Reliability: Robustness calls for AI systems to be technically robust to errors, incorrect inputs, or unseen data [197], and reliability refers to consistency in behavior and results of an AI system [156].
- Safety & Security: Safety of an AI system refers to the safe design and function of an

AI system, ensuring that an AI system does not harm living beings or the environment by design or misuse [113]. Security refers to the ability of an AI system to remain secure against external threats, such as cyber-attacks, data breaches, and malicious actors [174].

The development of a truly 'trustworthy' AI requires careful consideration of the interplay between these concepts and the requirements that need to be fulfilled to satisfy each during development.

In the following sections, the relevant publications, regulations, and policies on Trustworthy AI will be discussed to explore Trustworthy AI as defined by the various groups. Then, the commonality between these frameworks will be further detailed to understand their importance in Trustworthy AI development.

2.1 Ethical and Regulatory Guidelines for Trustworthy AI

In recent years, many policy-makers, government agencies, academic institutions, and private organizations have published guidelines, principles, and frameworks for the development of Trustworthy AI [53, 114, 115, 175, 197, 245, 272, 346]. However, currently, there is a lack of agreement on the exact requirements that should be focused on in the development of Trustworthy AI, with each report producing a different set of guidelines. As stated in the previous section, these publications share a common set of concepts, each mentioning in some way the concepts of accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security.

Among the publications on Trustworthy AI, three are particularly relevant in the context of this dissertation for defining Trustworthy AI and guiding regulation toward the development of Trustworthy AI systems: *Ethics Guidelines for Trustworthy AI: European High Level*

Expert Group on AI [142], the *Artificial Intelligence Act* [142, 338], and the *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* [43]. The concepts previously defined are reflected in each of these publications with subtle differences in definition and application. Additionally, each publication includes unique concepts to the interests or priorities of the group of interest. Each of these publications will be discussed with a focus on the differences in how trust is defined in each.

2.1.1 Ethics Guidelines for Trustworthy AI: European High Level Expert Group on AI

In 2019, the European High Level Expert Group (HLEG) [142] made significant progress in defining Trustworthy AI by defining a set of guidelines and requirements for trustworthy AI development. The established guidelines are driven by three components that make an AI trustworthy: law, ethics, and robustness. The text proposed that a Trustworthy AI should be lawful (abide by relevant laws and regulations), ethical (adhere to ethical values and principles), and robust (perform in a safe and reliable manner). This proposal was based on fundamental rights, defining four ethical principles to which trustworthy AI should adhere to: *respect for human autonomy*, *prevention of harm*, *fairness*, and *explicability*. The principle of *respect for human autonomy* ensured the fundamental rights of the freedom and autonomy of human beings, encouraging Trustworthy AI systems to be developed with a human-centric design that allows for human choice and primarily supports humans in pursuit of life and work. *Prevention of harm* is the concept that AI systems should not do undue harm to humans, protecting the mental and physical health of humans as well as human dignity. The principle of *fairness* refers to ensuring that AI systems are free from bias and discrimination, as well as calling for human choice, social fairness, and equal opportunity. Lastly, the principle of *explicability* calls for transparency, traceability, and/or auditability in the communications on the capabilities, purposes, and decision-making of an AI system.

Within these guidelines, the group identified seven requirements: *Human agency and oversight*, *technical robustness and safety*, *privacy and data governance*, *transparency*, *diversity, non-discrimination and fairness*, *society and environmental well-being*, and *accountability*. These requirements are briefly described below:

- *Human Agency and Oversight*: The AI system should have a human-centric approach, supporting human autonomy, decision-making, and fundamental rights. This requirement proposes human-in-the-loop, human-in-the-loop, and human-in-command approaches to ensure human agency and oversight.
- *Technical Robustness and Safety*: AI systems should be reliable, secure, safe, and reproducible. Results of AI systems should be reliable and reproducible, and the AI system should be robust to malicious actors, with recovery plans in case of attack or failure. There is a strong emphasis on preventing *unintended harm* by ensuring proper preparation, remediation, and action in case of issues with robustness, accuracy, and/or safety.
- *Privacy and Data Governance*: Privacy must be guaranteed via proper data governance procedures throughout an AI systems lifecycle. Private data must be kept protected, secured, with proper access protocols in place. Further, quality of data should be ensured by tests for biases, inaccuracies, errors and mistakes throughout.
- *Transparency*: The AI system, models, and data should be transparent. This requirement calls for clear documentation on the processes taken during AI development, including data gathering and labelling. Further, this requirement calls for explainability in AI systems, their decision-making processes, their capabilities, and their limitations.
- *Diversity, Non-Discrimination and Fairness*: Inclusion and diversity should be considered throughout the AI systems lifecycle, avoiding unfair biases and discrimination

for sensitive groups. Accessibility and feedback from stakeholders should be prioritized.

- *Societal and Environmental Well-Being*: AI systems should benefit society, including benefiting human beings and being environmentally friendly and sustainable. It is important to consider the societal and social impacts of an AI system throughout its lifecycle, now and in the future.
- *Accountability*: AI systems should be held responsible and accountable for outcomes during the entire lifecycle, and should be able to report, respond, and remedy actions and decisions in each stage. This requirement calls for *auditability* to assess algorithms, data, and design processes.

The HLEG expanded upon this report by presenting the "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment" [143], which establishes particular questions that AI developers and creators can ask to assess their attention to each of the key requirements.

2.1.2 Artificial Intelligence Act

The Artificial Intelligence Act (AIA) is a proposed regulation by the European Union. It was proposed on April 21st, 2021 and a provisional agreement was reached between lawmakers on December 9th, 2023. At the time of writing this dissertation, the final agreement has not been published; therefore, an analysis of only the *proposed* AIA is provided here. For this section, the AIA proposal published on April 21st, 2021 is primarily referenced, taking into account the amendments to the act published on June 14th, 2023 [338, 340].

The AIA is a proposal towards establishing "harmonized" AI rules, taking a risk-based approach by defining risk levels and regulatory processes. The primary objectives of the AIA are to provide a set of rules for AI within the European Union and its markets to create law-

ful and safe AI that are aligned with existing legal structures. Like the HLEG report, there is a focus on protecting the fundamental rights of persons, as well as respecting Union values and encouraging effective enforcement of existing Union laws and safety requirements. Further, the act aims toward fostering a market for ”lawful, safe and trustworthy AI applications”, providing an avenue toward assessing the risk-levels and safety of AI systems without incurring large costs and constraining the ability of creators to place AI systems on the market. Regarding the elements of Trustworthy AI, as both the AIA and the HLEG publication were written and published by the European Union, the AIA references the principles laid out by the HLEG for the development of ethical and trustworthy AI.

There are two major elements proposed by the AIA to be discussed: classifying AI systems by risk-level and establishing rules and regulations for such systems based on their risk level. The act defines several different risk levels (Title II): unacceptable-risk, high-risk, low- or minimal-risk.

Unacceptable risk systems are detailed in Title II, specifying unacceptable risk systems as those systems that ”whose use is considered unacceptable as contravening Union values”. This includes systems such as social scoring systems, AI systems that manipulate individuals through subliminal messages, systems that exploit vulnerable and protected individuals, and other systems that may violate fundamental rights. AI systems that are classified in the unacceptable risk category are prohibited from use within the European Union as well as from export to third-party countries.

High-risk systems (Title III) are those that pose a high risk to health, safety, or fundamental rights. A high-risk system is classified as an AI system that (i) is indented to be used as a safety component and (ii) the safety component is subject to a third-party conformity assessment. Any system that fulfills both conditions as a safety system classifies it as a high-risk system. In addition, the act defined several specific types of AI systems that may

be classified as high-risk, including [339]:

- systems utilizing biometric data for identification and/or classification of persons
- systems involved with the management of critical infrastructure, such as those involved with safety for transportation and utilities
- systems involved with the assessment or access for educational or vocational training
- systems involved with employment (recruitment, promotion, termination)
- systems that determine access to essential private and public services, such as assessing eligibility by governments for public assistance benefits, credit worthiness assessments, and the dispatching of emergency services
- systems involved with law enforcement
- systems involved with asylum, migration, and/or border control
- election systems, and systems involved with democratic or judicial processes

Low-risk and minimal-risk systems include any other systems that are not considered as high-risk or higher. This risk classification level include a wide variety of AI applications. The requirements for these systems are minimal, the Union calling for minimal transparency requirements (Title IV). In particular, these transparency requirements are targeted to systems that interact and engage with humans in such a way that it can detect emotions, determine social group membership based on biometric data, or generate or manipulate certain content. The transparency requirements in this case require that an individual be informed and given the choice on whether to proceed.

The AIA also defines a set of regulatory principles that applies both to providers and users of AI systems within the European Union. Namely, it applies to 1) providers of AI systems placing such systems onto the market anywhere in the Union, 2) users of AI systems

physically located within the Union, and 3) providers and users located in a third country, provided that the output of the AI system is used in the Union.

Unacceptable-risk systems are considered too risky, and prohibited. Limited-risk and minimal-risk systems do not have mandatory regulatory requirements. Instead, a *Code of Conduct* has been developed to encourage the voluntary application of the rules that apply to high-risk systems (Title IX).

The majority of requirements are specifically targeting high-risk systems, and these requirements are laid out in Title II, Chapter 2 of the AIA. High-risk systems must comply with a list of requirements, including (but not limited to):

- *Article 9: Risk management system*: The establishment of a risk-management system for the entire lifecycle of the AI system. This risk-management system should include: (i) identification of any known and foreseeable risks that the AI system might be associated with; (ii) evaluation of any risks that may occur due to the intended use of the AI system, as well as an estimation of risk during misuse of the AI system; (iii) evaluation of risks foreseen after entering the market; (iv) adoption of a risk management system that complies with the requirements of the article, including elimination/reduction of foreseen risks during design, communication of risk with users, implementation of mitigation measures, among others.
- *Article 10: Data and data governance*: For systems that use data processing techniques that segment data into training, testing, and validation sets, data and data governance techniques shall be implemented. The data and data governance practices include assessment of design choice, data collection and preparation tasks, assessments of relevant a priori assumptions, availability, quantity, and suitability of data, assessment of possible biases, and identification of any data ”gaps and shortcomings” and establishing mitigation strategy for them.

- *Article 11: Technical documentation*: The establishment of thorough technical documentation. This technical documentation, further detailed in Annex IV [339], include general descriptions of the AI systems purpose and specifications (software, firmware, hardware), description of the development processes and design specifications, description of the functionality and monitoring of the AI system (including a description of expected accuracy, foreseen risks and unintended outcomes, and human oversight measures).
- *Article 12: Record-keeping*: The design and development of AI systems with the ability to automatically collect logs during the operation of the system for the automated record-keeping of events. These logs will ensure traceability and monitoring of events throughout the systems lifecycle.
- *Article 13: Transparency and provision of information to users*: The output of an AI system should be transparent and interpretable to the users of the system. This article defines transparency requirements for the design and development of an AI system such as the need to include instructions for users with information on the intended use, level of accuracy, robustness and cybersecurity to which the system has been validated, risks to safety during use/misuse, input data, human oversight measures, and other relevant qualifications.
- *Article 14: Human oversight*: The design and development of an AI system should be with human oversight in mind. The AI system shall enable a human overseeing the functioning of an AI system to: (i) understand the capabilities and limitations of the system and monitor its operation for ”anomalies, dysfunctions, and unexpected performance”; (ii) remain aware of automation bias; (iii) interpret the output; (iv) bypass in some way the output of the system; (v) intervene to stop or interrupt the operation of the AI system.
- *Article 15: Accuracy, robustness and cybersecurity*: Achieve an ”appropriate” level of

accuracy, robustness, and cybersecurity. Accuracy (and other relevant metrics) shall be reported in the instructions to users (Article 13). Robustness shall be achieved with back-up or fail-safe processes or procedures. The system shall also be considered robust to security threats, with appropriate cybersecurity mechanisms given the intended purpose of the AI system, and resiliency to malicious actors.

The act continues by defining the obligations of both users and providers of high-risk AI systems (AIA Chapter 3), requirements for notifying bodies (AIA Chapter 4), specific details on standardization, conformity, certification, and registration. The regulatory requirements of the AIA for high-risk systems will be further discussed in Chapter 6.

Throughout the AIA, there is a strong focus on the protection of the fundamental rights of persons, with several mentions of the respect for the privacy, non-discrimination, transparency, reliability, safety, security, and the promotion of the fundamental rights of peoples.

2.1.3 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

The Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence was issued by President Joseph Biden on October 30th, 2023 [43]. The EO establishes standards for AI safety and security toward protecting the privacy of individuals, civil rights, consumers, and workers. Further, there is a focus on promoting innovation and competition, as well as working with other nations to further Trustworthy AI efforts.

The EO defines eight principles for the development and advancement of AI (Section 2):

- *Artificial Intelligence must be safe and secure*: This principle requires evaluations of AI systems to understand and mitigate the risks that AI systems post. It also calls

for addressing security risks, including cybersecurity threats, and performing testing and evaluation on the AI system for any risks due to misuse. Further, this principle mentions that output should be labelled as such that users can identify if the content is generated using AI.

- *Promoting responsible innovation, competition, and collaboration*: This principle calls for investments in "AI-related education, training, development, research, and capacity" and IP protection efforts. The government pledges to promote opportunities for marketplace development in AI to drive innovation.
- *Commitment to supporting American workers*: Encouraging job creation, training, and education to increase access to job opportunities created by AI development.
- *Advancing equity and civil rights*: The use of AI to disadvantage sensitive groups is not tolerated. AI systems shall comply with all Federal laws to avoid discrimination and bias with technical evaluations and oversight to advance "civil rights, civil liberties, equity, and justice for all".
- *User and consumer protections*: Users and purchases of AI systems are still protected by existing consumer protection laws and principles. In particular, existing laws should be leveraged to protect consumers against "fraud, unintended bias, discrimination, infringements on privacy, and other harms".
- *Protection of privacy and civil liberties*: Personal data should be protected, and the collection, use, and retention of data should be lawful, secure, and confidential. Privacy tools and technologies shall be used to protect privacy and mitigate any risks to privacy.
- *Responsible use of AI by the Federal Government*: This principle focuses on the responsible use of AI by the government. It largely calls for recruitment of sufficiently trained staff to ensure responsible AI adoption at the governmental level.

- *Global societal, economic, and technological progress*: This principle calls for the United States' involvement in promoting safe, secure, and trustworthy AI around the world, via engagement with international groups to promote responsible AI.

Section 4 outlines the actions the Federal government will undertake to foster safe and secure AI technology. The EO established that following publication of the EO, the following actions shall be taken: development of further guidelines and standards by leading governmental and academic authorities (Section 4.1), defining of requirements for AI systems targeting national defense and critical infrastructure (Section 4.2 and 4.3), identification and reduction of risks from AI associated with chemical, biological, radiological and nuclear threats (Section 4.4), identification and reduction of risk of synthetic content produced by governmental agencies (Section 4.5), seeking input on widely available model weights for dual-use foundation models (Section 4.6), promotion of the safe release of Federal data for AI training (Section 4.7).

Section 5 focuses on promoting innovation and competition, Section 6 on promoting workers, Section 7 on advancing civil rights and equity, Section 8 on protecting "consumers, patients, passengers, and students", Section 9 on protecting the privacy of individuals and private data, Section 10 on AI utilization in the Federal government, and Section 11 on encouraging collaboration with other nations on the development of safe, secure, and trustworthy AI.

At the time of writing this dissertation, no further actions have been taken on enforcing the requirements established in the EO.

2.1.4 Trustworthy Principles in Ethical and Regulatory Guidelines

As is evident, many of these principles and requirements overlap among the publications listed in this section. For example, each publication makes mention of the need for fairness

and non-discrimination, privacy, reliability, robustness, safety, security, and the promotion of societal well-being. In each, the concept of accountability is approached, with both the AIA and EO approaching accountability with a regulatory and standardization, and the HLEG suggesting auditability. However, the European publications had a significant focus on explainability, interpretability, and transparency that was not reflected as strongly in the EO. Mentions on this concept in the EO are limited to ensuring that users understand when content is generated by AI, rather than the need to explain and interpret outputs.

With these works, it is clear that there are commonalities in the defining of Trustworthy AI, including both technical and societal/ethical qualities that AI systems can exemplify to be considered Trustworthy AI. In the next section, these concepts will be further explored to understand their greater role in trustworthy AI design and development.

2.2 Trustworthy AI Concepts

Building upon the culmination of works defining Trustworthy AI, in this dissertation the following concepts are held paramount in developing a Trustworthy AI system: accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliablity, and safety & security. While no consensus has been found on the formal definition of trustworthy AI, focus has been placed on these key principles. Increasing trust in AI requires that providers, creators, designers, and developers closely analyze how they address these key principles during the development of their systems. In this section, these key principles will be further explored to uncover the challenges they pose in the development of trustworthy AI systems, both from an ethical and technical perspective.

2.2.1 Accountability

With AI becoming increasingly prevalent in society, there is increasing concern about who will be accountable for the decisions and impact of AI technologies. The principle of accountability calls for an obligation to explain, justify, and in cases of failure, mediate the

actions and decisions of AI systems, as well as calling attention to the need for auditing, regulatory requirements, and responsibility [174]. Fjeld et al. [113] found that 97% of the relevant documents published prior to 2020 on principled AI in all sectors (private, civil society, government, inter-governmental, multi-stakeholders) mentioned the concept of accountability: 69% mentioned accountability in terms of identifying a party that is responsible for the technology and any harms it may cause, particularly identifying responsibility in terms of the AI systems creators, developers, and/or providers; 53% recommended regulatory systems and impact assessments, and 47% mentioned a need for auditability.

Assigning responsibility for the outcomes of AI systems could potentially prevent significant failures and harm to society, particularly for AI systems involved in high-stakes scenarios. Consider, for example, the case of the Boeing 737 MAX crashes in 2018 and 2019 where significant errors with the maneuvering characteristics augmentation system (MCAS) caused two fatal accidents claiming the lives of 346 people [159]. While this case is rather extreme, many other AI failures have captured international attention, such as a robotic arm in a car factory malfunctioning and killing a man, a self-driving car causing a fatal accident, and cases of malfunctioning AI systems resulting in bias, racism, and malicious behavior [365]. Proper regulation, audit structures, governance, and otherwise system and outcome verification could have prevented these failures. However, there is no consensus on *who* is responsible for failures. Who is responsible, ethically and/or lawfully, for these failures: creators who develop the AI systems, data collectors of training and validation data, stakeholders and providers who deploy the systems, or even governments that do or do not regulate how AI systems interact with their citizens?

The need for accountability is paramount for building trust. Accountability requires both *answerability* [256] and *explanation* [92]. Accountability requires that something or someone is *answerable* to another, such as to a higher power, authority, or entity [256]. Explanations increase trust in a multitude of ways: explanations can reveal information about

decisions made during the design, development, and deployment of an AI system without revealing the precise technical details about the AI system's decision-making process, explanations can be used to determine whether proper procedures were followed, and they can be used to prevent and/or correct failures or errors [92].

Therefore, considering these two aspects, accountability is an important aspect of developing trust in AI. Establishing accountability mechanisms during the design, development, and planning of an AI system is equally as important as maintaining accountability throughout deployment and, in cases of error, after problems have occurred.

Proposed Solutions

Researchers have proposed a number of solutions to increase accountability in AI systems. Accountability in AI can be proactive or reactive. Proactive accountability occurs *before* the development and deployment of algorithms, focusing on the design, development, and subsequent planning for events and failures; reactive accountability occurs *after* the deployment of an AI system, or after a failure/errors, and concerns primarily the reporting and enforcement of mitigation and sanctions.

Proactive accountability focuses on developing AI systems to be accountable from the offset, including the planning for adverse behaviors. Research in this area focuses primarily on the design specifications and implementing mechanisms during the development of the AI system. Impact assessment methods are proposed to report potential impacts to stakeholders and users, such as the Human Rights Impact Assessment (HRIA) [177], Privacy Impact Assessment (PIA) [293], Ethical Impact Assessment (EIA) [360], and Surveillance Impact Assessment (SIA) [361]. Kaminski and Malgieri [162] propose impact assessments via a multi-layer transparency process based on the Data Protection Impact Assessments in the European Union (EU)'s General Data Protection Regulation (GDPR) [353]. However, Metcalf et al. [240] argue that these impact assessment frameworks are disconnected from

assessing the actual harms of an AI system, and argue for adaptations to such assessments to increase accountability and foster public trust.

Further, researchers have proposed mechanisms for increased accountability during development, such as developing AI systems with significant human oversight in the AI systems lifecycle from the offset, enhancing human agency without removing responsibility [350]. Many researchers propose human-in-the-loop frameworks that prioritize human interpretability and control over outputs [182] and emphasize the need for documentation [186].

Reactive accountability focuses on the developed/deployed AI system and reacting to the outcomes and/or failures of AI systems. Risk management and assessment mechanisms are proposed to assess AI systems for risk upon deployment [230, 262, 338], as well as for identifying, explaining, and mitigating failures [259]. Researchers in this area have proposed audit architectures that collect information from AI system outputs to monitor, repair, and potentially redirect AI systems [242, 270, 306]. Regulatory and policy frameworks typically fall under this type of accountability, applying recommendations for transparency, compliance, certification, and risk assessments to existing AI systems [52, 57, 103].

Other researchers focus on providing end-to-end accountability frameworks. Raji et al. [282] propose a process to increase governance utilizing audits, outlining an internal audit framework to be utilized during AI development. They define five stages: scoping, mapping, artifact collection, testing, and reflection, wherein providers of AI systems increase accountability by the design of their systems aligned with trustworthy principles, collect documentation throughout development including interviews with stakeholders, audit checklists, model cards, data sheets, and technical documentation, and develop remediation plans, among other qualifications. Broeders et al. [52] propose a framework for the use of Big Data, including considerations for the use of Big Data from the design to reg-

ulation of a system. Toward establishing accountability in ML, Kim and Doshi-Velez [181] analyze proposed techniques to increase accountability in ML, outlining five categories of approaches to increase accountable ML: transparency, interpretability, post-hoc inspection of model outputs, pre-market and post-market performance evaluation, and design properties.

Technical Challenges

There is no clear solution to increase accountability in AI. While many solutions have been proposed, there is little consensus on how exactly accountability can be developed, such as if accountability can be reached after the development of an AI system via audits or assessments, or if accountability must be built into each stage of the AI systems lifecycle. The concept of accountability also often overlaps with other aspects of trustworthy AI, particularly transparency, explainability, fairness, and human oversight, indicating that accountability requires an interplay of several domains. Therefore, it is difficult to create just one method of increasing accountability; rather, focus should be placed on a holistic approach to improving trustworthy AI, and accountability will follow.

2.2.2 Explainability & Interpretability

AI decision-making has replaced human decision-making in many aspects of day-to-day life. In order to trust the decisions that AI systems make, users must understand the outcomes of the system. However, as these systems are increasingly complex and difficult to understand, explaining and interpreting their processes and outcomes is a significant challenge.

Increasing trust in AI requires increased explainability and interpretability of AI systems. However, in AI research and publications these terms are intertwined. While uses of the terms often overlap, there are subtle differences in how each term is defined and utilized [208]. Explainability is primarily concerned with ensuring that the operations and out-

comes of an AI model are understandable to a human, such that a human can understand the decision-making process [243], whereas interpretability concerns the intuition behind how inputs are mapped to outputs of a system [4]. Researchers often use these terms synonymously, often with calls for increased transparency in outcomes and decisions, data use and governance, development, and deployment processes via transparency by design [109, 362].

Both explainability and interpretability play a vital role in increasing trust in AI systems and outcomes. Indeed, DARPA highlights the importance of explainability in trust, claiming that producing explainable models and enabling effective explanation techniques improves trust in AI systems [131]. Further, many researchers argue that building understandable systems leads to increased user trust, such as Mercado et al. [239] who found that increased transparency resulted in increased trust and perceived usability of an automated AI system. Similarly, Shin [316] found that if users understand the decisions and limitations of an AI system, a higher level of trust and acceptance was observed. Establishing explainability and interpretability mechanisms during the design, development, and deployment of an AI system is important in fostering trust from users.

Proposed Solutions

This area of research is quite active, with many researchers proposing various methods for increasing transparency and explainability in AI systems. Arrieta et al. [18] identified nine primary goals in explainable and interpretable AI research: *trustworthiness*, *causality*, *transferability*, *informativeness*, *confidence*, *fairness*, *accessibility*, *interactivity*, and *privacy awareness*. The authors reveal that the literature makes a clear distinction between AI systems that are interpretable by design, and those that require external tools or methods to be applied to explain them after the fact, classifying two types of explainability as (1) transparency in modeling and (2) post-hoc explainability.

Transparency in modeling, also called intrinsic interpretability, refers to the concept that the models themselves should be interpretable via algorithmic transparency, decomposability, and simulatability [18]. Typically, this type of explainability includes simplistic models such as linear/logistic regression, decision trees, K-Nearest Neighbors, rule-based systems, general additive models, and Bayesian methods [244]. Each of these exhibits some level of transparency in the modeling process itself, such as the ability of the model to be easily thought about by a human, the ability to explain every input, parameter, or calculation of a model, and/or the ability of a human to understand the process by which the system produced an output from a given input via mathematical analysis and methods. For example, some researchers have proposed techniques to foster explainability in their modeling processes [64, 195].

Post-hoc explainability occurs when the model itself is not interpretable by design, but rather methods can be undertaken to enhance interpretability via explanations [18]. The goal of these explanations is to increase the level of interpretability, whether via explaining the whole logic, processes, and decision-making of a model (Global Interpretability), or explaining individual decisions made by the system (Local Interpretability) [4]. These types of explanations include text explanations, visual explanations, local explanations, explanations by example, explanations by simplification, and feature relevance explanations [18]. For example, Local Interpretable Model-Agnostic Explanations (LIME) and its variations were proposed to provide simplications and local explanations for predictions of ML models [289, 290]. Further, feature relevance explanations methods to explain feature influence, relevance, and importance have been proposed, such as SHapley Additive exPlanations (SHAP) [220], Quantitative Input Influence (QII) [85], Automatic STRucture IDentification method (ASTRID) [139], and others [184, 294, 330]. Visualization techniques have also been proposed in a variety of ways to explain the data, guide feature selection, and assess the performance of algorithms [7].

The goal of explainability and interpretability is to increase understandability, foster trust, and increase transparency in AI systems. There are a multitude of methods and solutions focused on making AI systems explainable by design or via providing post-hoc explanations. For a more comprehensive review of explainability and interpretability methods, please reference Arrieta et al. [18], Linardatos et al. [208], and Adadi and Berrada [4].

Technical Challenges

As modeling becomes more complex, understanding becomes more difficult and opaque. Providing an interpretation of how an AI model works becomes a significant issue, as does providing a metric for measuring a model's explainability. In addition, the type of explanation needed depends on the user and model type, and therefore so does the metric needed to measure explainability, further complicating the issue [146]. Although this is a very active research area, it is not clear how these methods are being used in deployment. Bhatt et al. [41] find that the methods are being used primarily by ML engineers during development and debugging, not by the users and stakeholders themselves, revealing a gap between theory and practice. It is clear that an explanation of some sort is required, but there is little consensus on the audience, depth, and degree of specification of explanations for AI systems.

2.2.3 Fairness & Non-Discrimination

With the proliferation of AI systems in society, there is increased attention on the fairness of such systems, especially in high-stakes scenarios. The fairness and non-discrimination principle calls for the consideration, detection, and prevention of discrimination and bias in the development of AI systems. As these systems are often used in sensitive or high-stakes areas, it is vital that decisions are made without discriminatory or biased influence toward particular demographic groups or populations. In recent years, biased and discriminatory practices in AI have been identified in nearly every type of system, including advertisements, chatbots, employment decisions, legal decisions, facial and voice recognition, and

search engines [234]. For example, several notable examples of biased criminal detection and criminal recidivism systems have been found, where racial bias has been revealed in the predictions [11, 234]. Gender bias is also prevalent, with notable AI systems producing bias in predictions, particularly due to biased data sources [296]. For example, gender bias in AI systems has been demonstrated in medical settings where inconsistency in diagnosis accuracy was observed for males and females based on x-ray images [190].

Fairness of an AI system implies that the outcomes are *fair* if they do not cause disparate harm against any particular subgroup [26]. Fairness in AI has a clear relationship with trust. Fairness, non-discrimination, and the mitigation of bias are mentioned in the majority of Trustworthy AI literature, with claims that fairness in the development, deployment, and outcomes of AI systems is vital to increasing user trust [113]. However, there is no consensus on the formal or mathematical definition of fairness, as ”fairness” in an AI system is largely context-dependent and varies from application to application. Fairness literature defines *individual* fairness and *group* fairness, where individual fairness seeks equality among similar types of individuals, whereas group fairness seeks equality across groups [228]. Regardless of the type of fairness, this principle calls for methods that both mitigate discrimination and measure bias impacting the AI system. One common focus in the literature is that fairness in AI is strongly related to measuring and mitigating bias.

To mitigate biases, it is important to understand them and their causes. Several types of bias are prevalent in AI literature. Kaur et al. [175] broadly define three types of bias: data bias, model bias, and evaluation bias.

Data bias concerns the inputs to the AI system. This type of bias refers to the systematic skew of data that leads to unfair or discriminatory outcomes when processed by AI systems. Data bias can be a result of data collection, processing, or how it is leveraged during algorithmic development. Indeed, this is a significant focus in fairness research, where it

has been demonstrated that inherent biases in how humans structure their datasets can lead to bias in their AI systems [191]. Data bias generally occurs because population dynamics of a dataset do not align with reality, such as an imbalance of representation among classes (particularly protected classes such as age, gender, race, socioeconomic status, etc.), unlearned data cases, or manipulation of the data to skew the distribution [296].

Model bias occurs at the algorithmic level when the algorithm itself introduces the bias due to errors or improper development practices resulting in a model that does not reflect the full logic of the prediction. For example, this may occur when nonsensical features are being used during prediction or features are being given more weight than others without valid logic. Obermeyer et al. [260] found that a widely used health-tech algorithm incorrectly predicted patient risk levels in a racially biased way because it was using financial information rather than health information. In another case, it was discovered that the way that textual data was processed in a popular ML method was leading to gender bias [46].

Evaluation bias occurs when models are evaluated incorrectly, such as when an unsuitable evaluation metric is used or when an inappropriate and disproportionate benchmark is used [234]. For example, this might occur when biased datasets are used to evaluate the performance of an algorithm [56], or during the use of the AI system user behavior results in a feedback loop that biases outcomes [193].

Therefore, to develop a fair AI system, it is crucial to ensure that attempts are made to mitigate bias at the data level, model level, and during evaluation. Establishing fairness and non-discrimination mechanisms during the design and deployment of AI systems can help mitigate bias and reduce discriminatory practices, thereby increasing trust in these systems.

Proposed Solutions

Research in this area provides many solutions for auditing, and improving bias and fairness in AI systems. Approaches to increase fairness in AI can be broadly classified as *pre-processing*, *in-processing*, and *post-processing* methods [234].

Pre-processing methods attempt to transform or process the data to remove any bias or discrimination [82]. Zhang et al. [388] demonstrated that causal models and graphs can be leveraged to identify meaningful partitions for proving discrimination in datasets and propose a method to remove discriminatory data points from datasets. Messaging is also shown to have strengths in removing bias in datasets by changing class labels to produce a more balanced dataset [163]. Building upon that work, Kamiran and Calders [164] introduces a preferential sampling to remove bias without relabeling [165]. Brunet et al. [54] propose a method to remove bias on word embedding via approximation of differential bias. Calmon et al. [61] propose a probabilistic framework for data transformation that controls for discrimination, limits distortion in samples, and preserves utility, demonstrating reduced discrimination in criminal recidivism applications. Sharma et al. [314] perform data augmentation to improve fairness in data using an ”ideal world” metric to sample datasets for bias and simulate equality by augmenting additional data points. Other approaches include removal of disparate impact [108], discrimination prevention [223, 300], and input feature modification [303].

In-processing methods directly act upon the algorithms to prevent and mitigate bias during the development process [82]. For example, augmented cost functions with augmented ’fairness’ regularizer can reduce discrimination by penalizing the algorithm depending on how it learns protected vs. non-protected classes [167, 378]. Berk et al. [37] propose a weighted regularizer that computes the accuracy-fairness tradeoff of an algorithm. Kamiran et al. [166] alter Decision Tree architecture to minimize the impact of discriminatory data

and improve performance. Similarly, Calders and Verwer [60] propose three methods to reduce bias in the naive Bayes classifier via modifications to the algorithm. Other approaches include fairness via decision boundaries [376], application of Simpson's paradox[192], and adversarial learning [382].

Post-processing methods are concerned with mitigating bias using the outputs of the AI system. These methods typically include alteration of outputs to produce an unbiased result or post-hoc testing of the trained algorithm using data unseen during training. For example, in the earlier model of gender bias, Bolukbasi et al. [46] remove bias by modifying the embedding from the learned model. Hardt et al. [137] demonstrate a method to optimally adjust any learned predictor to remove bias and discriminatory behavior while preserving the privacy of the system. Other approaches include applying decision thresholds to reduce bias based on known groups [80, 238].

Another approach to address fairness is via fairness toolkits. For example, IBM developed the AI Fairness 360 toolkit, a set of tools that can be used in industry to evaluate algorithmic fairness [33]. The toolkit includes fairness metrics for both datasets and models, algorithms to mitigate bias, as well as a web interface to educate users on fairness principles. Another toolkit was developed by Saleiro et al. [302] for the auditing of fairness and bias in systems. These types of toolkits are useful for evaluating bias in their systems and reducing discrimination.

It is clear that many methods have been proposed to improve fairness in AI systems. This section could not provide an exhaustive analysis of all methods, and for a more thorough survey fairness methods, please reference the work by Mehrabi et al. [234].

Technical Challenges

Although research into fairness in AI is very active, there are many challenges to its development. A significant challenge in this domain is in clearly defining fairness, and what

constitutes a ”fair” system. Mehrabi et al. [235] argue that while much attention has been placed on *equality*, not enough attention has been placed on *equity*. They propose that fairness should be motivated by equity, considering the existing and historical biases that are present today, and making decisions based on those biases. However, the majority of definitions do not account for this, revealing a disconnect in how different groups view fairness. In further defining the definition of fairness, existing methods can be applied to reduce unfairness in data, modeling, and evaluation processes.

2.2.4 Privacy

Privacy is a significant concern in all systems where the use of personal data has significant social and economic impact [157]. As the performance of AI systems largely depends on the data that it is trained on, the availability and usage of that data is very important. In particular, privacy principles in Trustworthy AI focus on private and personal data, where unlawful gathering, misuse, and/or loss of data can lead to harmful consequences. Concerns over privacy in AI systems are particularly prevalent with the high volume of data used for sensitive decisions, such as in advertisement, surveillance, health-care decisions, and money lending [113]. For example, consider the case of the Equifax data breach in 2017: 145 million US consumers had sensitive information leaked, exposing them to identity theft and potential financial and legal repercussions [36]. Further, threats to privacy are prevalent, such as adversarial attacks to expose data, expose and poison models, or otherwise manipulate the results of the AI system [386]. Increasing trust in AI systems is closely connected with protecting the privacy of the system and users.

Privacy is not only an issue from a systems' and users' perspective, but also at regulatory and governmental levels as well. For example, in the European Union, the GDPR highly regulates the use of personal and private data in AI, with large financial penalties levied on businesses who do not comply [353]. In particular, the clause that provides users with the ”right to be forgotten”, where users have the right to request that their personal data

be deleted, is particularly important in AI privacy. Researchers have called attention to the issues with applying this clause to AI, pointing out that the complete deletion of private data in AI may be impossible as AI systems do not ”forget” data in the same way as humans [352].

To develop a trustworthy AI system, privacy protection is vital. The use of personal data in AI systems is affected during all phases of AI development, including data collection, model development, and deployment as well as the storage and utilization of data. To secure against issues at each stage, privacy protections must be in place, including user consent, privacy by design, control over the use of data, and the ability to restrict processing, among others [113].

Proposed Solutions

Research in increasing privacy in AI systems primarily targets ensuring that data is kept secure and private. Methods to ensure privacy target the data and system development with techniques such as de-identification, privacy-preserving modeling, risk identification, and customized data use and management.

De-identification techniques aim to remove identifying information from datasets to prevent data from being linked with specific individuals [121]. This includes anonymized and pseudo-anonymized data practices, where data is processed to remove or obscure any personal information [155]. For example, k-anonymity algorithms have been utilized to de-identify data by ensuring that each data record is similar to a set threshold of other records, ensuring that individual data cannot be easily identified [29, 100]. Improvements upon these methods include L-Diversity [227] and t-closeness [201]. However, these methods have several weaknesses, including including a loss of data utility and relevancy [375], and potential re-identification [101].

Other data protection methods include censoring, suppression, encryption, exclusion, and

obfuscation [189]. For example, encryption techniques aim to protect data confidentiality and integrity through encryption, such as utilizing homomorphic encryption for the development of secure and robust systems [153]. Another example is obfuscation, where data is modified with methods such as data masking, tokenization, or scrambling, such as Ardagna et al. [14] who utilize obfuscation techniques on location data taken by sensing technologies to protect the private location of individuals.

Privacy-preserving technologies are another approach to enable AI to use data without compromising individual privacy. For example, differential privacy techniques have been leveraged in a variety of domains and use cases, becoming a popular choice for data privacy [99]. Another approach is Federated Learning (FL), which has specific applications in protecting user data via decentralized model sharing [202]. Other privacy-preserving approaches in recent years include Multi-Party Computation [94], Zero Knowledge Proof [111], Partial sharing [211, 317], and data augmentation [118].

Technical Challenges

Protecting the privacy of personal data in AI systems is paramount in building trust. However, the protection of data comes with a multitude of challenges. While de-identification methods are common, there are risks associated with the data remaining useful and relevant for decision-making systems [375], and significant risks of re-identification, where previously personal information and links are revealed about de-identification data [101]. Additionally, while many privacy-preserving techniques have been proposed, there is variability in the effectiveness and cost of each, requiring a balance between performance and overhead [67].

Further, there are several risks to data privacy posed by adversarial attacks, including attacks on the data and models [263]. While the security of AI systems will be further discussed in 2.2.6, there are significant implications that attackers may be able to access per-

sonal data, threatening the privacy of AI systems. It is vital that these challenges are addressed in developing a trustworthy AI system.

2.2.5 Robustness & Reliability

A majority of Trustworthy AI texts place a significant focus on the performance of AI systems. The principles of both robustness and reliability concern the performance of AI systems. Specifically, robustness calls for AI systems to be technically robust to errors, incorrect inputs, or unseen data [197], and reliability refers to consistency in behavior and results of an AI system [156].

Both of these concepts are paramount for building trust. The lack of either of these may result in unintended behavior by the system. The definitions of these two terms often overlap, but here they are defined as two separate concepts that both contribute to the trustworthiness of an AI system. The two concepts are interlinked, in part, due to the impact that errors with robustness might have on the reliability of the system.

For example, consider the example of autonomous driving. Autonomous driving systems must be trained on diverse datasets, including data from different scenarios and environments, such as ensuring the inclusion of training data taken in all weather conditions and at all hours of the day. While autonomous driving works well in sunny and clear conditions, quite a lot of research has been conducted on improving performance in conditions that are dark or with inclement weather [337, 389]. This research has both an impact on the robustness of autonomous driving systems to perform equally on all inputs, and the reliability of autonomous driving under all environmental conditions. This example demonstrates the link between these two concepts: robustness to errors, inputs, and unseen data guarantee reliability in the behavior of AI systems.

Robustness and reliability can apply to the data, algorithms, and overarching developed

and deployed systems. Data-level robustness focuses on the data that is used to train AI systems. For example, the above example demonstrates "Robustness to distributional shift" [9]. AI systems are reliant on the data used to train them, and if that data is very different from unseen data it may be difficult for the AI system to generalize new inputs. AI systems must be robust to distributional shifts, or it may result in undesired behavior in the AI system.

The robustness of algorithms highly focuses on robustness to adversarial attacks [197]. Robustness to attacks refers primarily to ensuring that AI systems are defended against attacks and that malicious actors or inputs cannot alter the behavior of a system. Adversarial attacks and robustness to malicious actors will be further explored in Section 2.2.6.

At the system level, robustness primarily concerns robustness to errors of execution or robustness to illegal inputs. For example, Li et al. [197] refers to system-level robustness as robustness against illegal inputs such as high-resolution images in an image recognition system causing the system to crash. Some researchers expand reliability at the systems level to include reliability to perform at a certain accuracy or with high performance on another metric [23].

Robustness and reliability also depend on the verifiability, replicability, and reproducibility of AI systems. Results of an AI system must be *reproducible* to ensure that outputs can be repeated, verified, and trusted [19]. Reproducibility concerns the methods, results, and inferences of an AI system [130]. For example, if another cannot repeat the results of one researcher, how can those results be trusted for accuracy?

Ensuring the robustness and reliability of AI systems is paramount to increasing trust in AI. Without either, users cannot have confidence in the results and outputs of such systems.

Proposed Solutions

This section has already discussed a multitude of methods that have an impact on the robustness and reliability of AI systems. Solutions for robustness and reliability concern enhancing all principles of Trustworthy AI. For example, a biased system is not robust to diverse inputs, and therefore unreliable in unseen cases. Additionally, many approaches focus on robustness to adversarial attacks, discussed further in Section 2.2.6. As solutions for robustness and reliability largely overlap with other principles of Trustworthy AI, this section focuses primarily on solutions that directly target ensuring robustness through performance and maintenance.

Testing and monitoring the performance of AI systems can be conducted at each stage of the AI lifecycle to confirm that the expected behavior of an AI system aligns with its real behavior [197]. Many researchers propose testing and monitoring of AI systems throughout development and deployment, such as testing for correctness, relevance, security, privacy, efficiency, fairness, and interpretability [387]. Simulation techniques allow developers to verify that systems are performing as expected in real-world scenarios [96], allowing robust design and deployment. Performance benchmarking can be leveraged to assess performance on standardized datasets to test AI systems and algorithms for robustness and reliability in its outputs [401].

Additionally, the performance of AI systems concerns its *generalizability*. The concept of generalization refers to the ability of an algorithm to learn and properly predict patterns of unseen data [126]. Li et al. [197] argue that generalization is closely related to AI trustworthiness, expanding that the problem of ”Robustness against distributional shifts” is a problem of generalization. Ensuring and evaluating the generalization of AI systems may therefore improve robustness and reliability. Solutions to increase generalization include benchmarking [401], measures of generalization error [31], and targeting robustness and

generalization during algorithmic modeling [32, 384, 402].

Other approaches for robustness and reliability focus on increasing the transparency of AI system development and deployment processes, including the provenance of data and algorithms. As the outcomes of AI systems depend directly on training data use (and misuse), data transparency, including transparency in data collection, utilization, and storage, is an area of significant concern in trustworthy AI.

AIOps [84], and MLOps [150, 229] aim toward streamlined, efficient, and effective AI system design, development, and deployment. These methods develop workflows for building AI, providing the backbone for the development of trustworthy AI [197]. Research has focused both on end-to-end tracking of provenance information and on the evaluation of models for performance and trust. Several algorithm-provenance solutions have been proposed. Schelter et al. [312] propose a system for the extraction and storage of meta-data and provenance information commonly observed in the modeling lifecycle. Hummer et al. [150] propose ModelOps, a cloud-based framework for end-to-end AI pipeline management, including support for addressing several trustworthy principles, such as reliability, traceability, quality control, and reproducibility. Further, several tools for complete asset tracking of AI pipelines have also been developed, focusing on tracking modeling inputs, results, and production processes [124, 152, 377].

Building upon this, data provenance (or data lineage) methods aim to improve replication, tracing, quality assessment in data use, and data transformation processes [141]. Several researchers have proposed data provenance and lineage solutions for the tracking of data and data transformations during the AI lifecycle [326, 327, 395].

While these solutions assist with internal data provenance, several researchers have also advocated for private, secure, and standardized methods for data tracking. Gebru et al.

[123] proposed *datasheets for datasets*, a standardization method for the documentation of datasets. These datasheets include information on "operating characteristics, test results, recommended uses, ... motivation, composition, collection process, [and] recommended uses", offering a detailed questionnaire for dataset creators to provide [123]. Similarly, Bender and Friedman [35] propose *data statements* for dataset characterization in natural language processing, also considering the generalization of experiments and composition of datasets concerning bias. Further, Holland et al. [147] propose a standardized diagnostic method for an overview of the core components of a dataset with the *dataset nutrition label*.

Many researchers argue that AI documentation is a step toward robustness and reliability via an increase in transparency. For example, a recent trend is the use of *FactSheets*. Arnold et al. [17] proposes FactSheets to communicate "purpose, performance, safety, security, and provenance information" from the creator to the user of an AI service. Sokol and Flach [323] extended this with a taxonomy for characterizing and assessing explainability in AI with *Explainability FactSheets*. However, Hind et al. [144] found that developers found these FactSheets challenging and time-consuming to complete, noting issues with developer recall about modeling details, data transformation documentation, privacy, and ownership concerns. Considering legality and regulations, Yanisky-Ravid and Hallisey [372] propose the *AI Data Transparency Model*, encouraging data audits by both stakeholders and third parties to assess data use and storage, to encourage replicability and compliance.

Technical Challenges

Achieving robustness and reliability in AI systems is complex and challenging. One significant challenge is the interplay between these concepts and others in Trustworthy AI, indicating that a robust and reliable system requires attention to other principles. Additionally, AI systems are typically trained on historical data, which may not fully represent future conditions, leading to issues with generalization and performance consistency when faced

with unseen scenarios [402]. The threat of adversarial attacks, where data, algorithm, or system inputs can be perturbed to lead to incorrect outputs, further complicates robustness (discussed further in Section 2.2.6). Further, while several approaches have been proposed for end-to-end provenance and tracking, AI systems are still prone to errors and robustness shortcomings. Balancing the trade-offs between performance and provenance, along with ongoing validation across diverse real-world conditions, makes building truly reliable and robust AI systems a difficult task.

2.2.6 Safety & Security

In recent years, advancements in the field of AI have drawn attention to concerns about the large-scale impacts of such AI systems [9], urging awareness of the potential harm that these systems may cause. With the increasing utilization of AI systems, particularly in high-stakes areas such as autonomous vehicles, healthcare services, and surveillance, we must consider their trustworthiness.

Safety and security are both vital to consider when developing trustworthy AI. In recent years, damages caused by autonomous vehicles, manipulation of public-facing AI systems, and software problems have harmed public perceptions of the safety and security of AI systems in society [9]. This principle covers assessing the safety of AI systems, how secure an AI system is, and ensuring the robustness of an AI system from adversarial attacks.

AI safety is both a technical and ethical concern, where potentially negative impacts on society could occur due to unintended accidents or failures [349]. Amodei et al. [9] define AI safety problems based on where they occur in the AI lifecycle, with safety threats including negative side effects, reward hacking, non-scalable oversight, unsafe exploration, and distributional shift. ”Robustness to distributional shift” appears again here, as systems may be impacted by this phenomenon and result in harm to their environments. For example, using the automotive car example from the previous section, systems trained on only sunny

data may crash in rainy weather, potentially harming their occupants or pedestrians.

Often, the principles of safety and security are interconnected, where issues in one domain are likely to have an impact on the other [212]. Security flaws can contribute significantly to safety failures, where attacks by malicious actors can misclassify inputs to worsen or manipulate performance or gain information about the model and data it was trained on [151].

Security in Trustworthy AI can apply to the data, algorithms, and overarching developed and deployed systems. At the data level, unauthorized access to data via attacks or other data breaches threatens user privacy, potentially exposing the private information of users and putting the provider at risk of violating compliance with privacy laws [197]. At the algorithm level, adversarial attacks threaten the performance and security of AI systems. Threats at the algorithm level might expose the data used to train algorithms [214] or somehow alter the performance [5]. For example, poisoning attacks inject perturbed samples into the training data to impact the behavior of the model in a way that benefits the attacker [283]. At the systems level, the hardware, software, and networks must be secured against vulnerabilities that could compromise the AI system [148]. Protecting and defending against these security threats is vital to ensure a safe and trustworthy AI system.

Proposed Solutions

Several solutions have been proposed to improve the safety and security of AI systems. To achieve safety, one may consider the four principles of safety: Inherently safe design, Safety reserves, Safe fail, and Procedural safeguards [244]. Ensuring safety involves comprehensive risk assessments, adherence to ethical guidelines, and the implementation of fail-safes and redundancy mechanisms to mitigate potential harm. Safety constraints, such as fail-safe mechanisms, cause the system to fall into a ”safe” state upon error using active or passive controls to prevent harm from the system [196]. Other approaches include risk

assessment to measure potential points of failure of AI systems and any resulting safety concerns [52, 57, 103].

In regards to security, while no solution is complete, systems can be developed to detect and address adversarial issues and attacks [253]. In particular, anomaly monitoring and detection can be leveraged to identify adversarial inputs and defend against attacks that can affect the robustness and safety of systems [65, 266]. Further, adversarial robustness attempts to mitigate malicious behavior at the algorithm level by introducing perturbed data to the model training stage [127] or introducing a regularization term to penalize malicious inputs [129]. In addition, Raghunathan et al. [280] proposes certifications of robustness against adversarial attacks.

While every solution for ensuring the safety and security of AI systems cannot be mentioned here, comprehensive reviews on the various areas of AI security can be found in the literature [5, 263].

Technical Challenges

Defending against security risks at the data, algorithms, and systems levels is vital to secure the AI system and ensure trustworthiness. A multitude of attacks can impact the behavior of an AI system, potentially threatening user privacy, system robustness, and potentially the safety of its users. However, truly securing against adversarial attacks is a challenge, particularly with the various types of attacks and goals of attackers, and the ever-changing landscape of adversarial threats [263].

2.3 The Role of Transparency

All of the above aspects are important in developing a trustworthy AI system. For example, Shin [315] found that fairness, accountability, and explainability influence how users of an AI system perceive an AI system. AI systems that are perceived as fair, accountable,

transparent, and explainable are seen as more trustworthy.

However, developing a trustworthy AI system that addresses all of these guidelines is a difficult task. The development of such systems leads to problems with scalability and feasibility: with the development of complex models, how do companies and individuals ensure that each is accounted for, and at what cost? For example, consider the privacy requirements of the GDPR, such as the ”right to explanation”: addressing this regulatory requirement increases technical challenges required in the development of an AI system [353, 355], and is often infeasible for smaller scale operations. Even if this requirement is satisfied, proper care must be given to the other elements of trust as well, further complicating the development of a trustworthy AI system. Developing such systems is costly and time-consuming, a barrier to entry to trustworthy AI development.

That said, all aspects of trustworthy AI can be improved in some way by increasing transparency; transparency in design, transparency in purposes, transparency in fairness, transparency in security, and so on. Fundamentally, transparency is the key to trustworthy AI. At its core, transparency involves clear communication about how AI systems are designed, function, and their decision-making processes. This communication is crucial for allowing assessments of fairness and bias, understanding the risks to privacy, security, and safety due to potential errors or unexpected behavior, enabling users to understand the outcomes of systems, and ensuring that AI systems perform robustly throughout their lifecycle. Transparency aids in regulatory and compliance, enabling regulatory and governmental bodies to assess AI systems for compliance with local, national, and international rules. It also can serve as an indicator as to what other aspects are lacking, guiding improvements in other areas.

Transparency is a significant theme in Trustworthy AI literature and is mentioned in each of the texts mentioned at the beginning of this chapter. Increasing transparency is the first,

final, and most vital step in building trust in AI systems.

2.3.1 Summary

Academia and industry alike have called for practices to encourage transparency and increase trust in AI development and deployment. Various ways that trustworthiness can be established have been proposed, with the majority of calls focusing on the concepts described in this chapter. Trustworthy AI seeks to ensure that AI systems are developed and deployed ethically, transparently, and with attention to the greater principles of trustworthiness: accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security.

However, the literature on Trustworthy AI is diverse and inconsistent, with the various frameworks, guidelines, and principles proposed by different organizations, researchers, and policy-makers using different definitions and targeting varying principles of trustworthiness. For example, the HLEG proposal discussed in Section 2.1.1 places a strong emphasis on the ethical considerations of AI deployment in society, but falls short in guiding technical implementations. The AIA (Section 2.1.2) provides more of a targeted approach for technical implementation regarding accountability, bias, privacy, robustness, and safety, but falls short when it comes to explainability and interpretability. Likewise, the EO discussed in Section 2.1.3 targets accountability, privacy, safety, and security, but makes little reference to the other principles and lacks any implementation guidance. These inconsistencies and varying emphasis on different principles can lead to challenges in implementing AI systems.

Moreover, an important distinction should be made between the trustworthiness of *AI systems* vs. the trustworthiness of *AI models*. An AI model can be simply defined as a specific algorithm or set of computational processes that are typically designed to perform a specific task, such as learning, decision-making, classification, or problem-solving. For

example, Sarker [310] argue that AI models can be broken down into ten categories: ML; neural networks and deep learning; data mining, knowledge discovery, and advanced analytics; rule-based modeling and decision-making; fuzzy logic-based approaches; knowledge representation, uncertainty reasoning, and expert system modeling; case-based reasoning; text mining and natural language processing; visual analytics, computer vision, and pattern recognition; hybridization, searching, and optimization. AI models are a component of AI systems, playing an important role in the application of an AI model toward real-life problems in various domains.

An AI system, by that logic, encompasses a broader scope. AI systems include not only the AI models but also the infrastructure, interfaces, and mechanisms that enable models to function in a real-world environment. For example, in health tech, an AI model might be used to predict abnormalities in a patient scan to predict illness, but an AI system would take a broader approach and use additional patient information (potentially including the predicted result of the scan) to provide a treatment plan [241].

Increasing trust in AI models and AI systems involves different approaches and considerations. While all components of Trustworthy AI apply to both, the primary focus of trust may be different for each. For AI models, trust primarily depends on accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security. The challenge primarily lies in making model and model outcomes transparent, understandable, and fair, particularly as perceived by a human user. Whereas, for AI systems, the expectations may expand to the infrastructure supporting them and become more complex. Trust not only encompasses the principles above, but also includes the overall safety, security, and accountability of the system. For example, Hengstler et al. [140] found that trust in an AI system depends on operational safety, data security, contextualization of the purposes of an AI system, and clear communication about stakeholders and developmental processes. Further, they found that there is a connection between a

user's perceived opinion of the AI system and their values, indicating that the principle of promotion of human values is also a stronger focus for AI systems.

Therefore, it is important to consider trust from both perspectives. Increasing trust in both AI models and AI systems is crucial for the widespread acceptance and effective use of AI in various domains. Current solutions focus primarily on one stage of the AI lifecycle, or only a handful of trustworthy principles, neglecting to give proper attention to the "whole picture" required in developing a trustworthy system. In the next chapters, methods to increase trust will be discussed.

Chapter 3

Establishing Requirements for Trustworthy AI

The previous chapter established that the trustworthiness of an AI system implies that the development of such a system addresses issues with accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security. These concepts should be at the forefront of consideration when we think about AI development and deployment into society. However, developing a completely trustworthy AI system is a difficult task. The summary of Trustworthy AI literature in the previous chapter reveals inconsistencies in which requirements are truly necessary to build trust in AI systems. While some texts focus on only one principle of trustworthiness, some call for attention to a combination of principles. Likewise, some texts place great emphasis on technical aspects such as security and privacy, while others emphasize the ethical and societal impact of AI systems. To design, develop, and deploy a Trustworthy AI system, there must be formalized methods for tracking and reporting how developers address issues of trust.

While some methods have approached this goal, there are no formal methods that address all principles of trustworthiness. For example, FactSheets and its variations take a step toward increasing transparency via documentation to approach many trustworthy principles, but they lack direct technical implementations for issues such as fairness and non-discrimination [17]. Likewise, MLOps methods go far toward technical measurements of processes, provenance, robustness, and reliability, but they do not specifically focus on addressing issues with fairness, explainability, or accountability [150]. Without a targeted method of addressing each principle of Trustworthy AI, the wide-scale development of trustworthy AI systems is greatly hindered.

Providers, creators, developers, and all other stakeholders involved in the development and deployment of an AI system must pay careful attention to these aspects, as they govern how users will trust the system outcomes and results. Errors in implementation, such as improper attention to fairness, data storage, security mechanisms, and explainability, may impact how an AI system makes decisions about vulnerable populations, protects user privacy, and maintains security against malicious actors, as well as impacting the way a user understands and trusts the decisions it makes. To encourage large-scale AI adoption and increase trust, the burden is on the creators to address these principles in their deployed systems.

Further, in the previous chapter, a clear distinction between AI models and AI systems was established, clarifying the unique approaches to trustworthiness that must be taken with each. While the ultimate goal is to develop trust in AI systems, it is vital to first develop trust specifically in AI models. AI models play a critical role as the core of AI systems, with the capacity for learning, decision-making, classification, problem-solving, and other functionalities. However, the efficacy and acceptance of these models largely hinge on the degree of trust users place in their outcomes.

Toward developing trust in AI, this chapter introduces the concept of Know Your Model (KYM), the idea that all models have a unique identity and that model characteristics can be leveraged to know and trust models. To ”know” a model implies collecting, recording, and storing detailed records of the processes undergone during the development of a model, subsequently establishing *model identity*.

To know a model, this chapter proposes 20 key guidelines that creators can address to establish a model's identity, particularly around 4 core principles: **efficacy**, **reliability**, **safety**, and **responsibility**. These guidelines provide a general framework that applies to any AI implementation, rather than prescribing a particular implementation. The proposed guidelines are concise suggestions of important aspects that creators should be able to address about their AI systems regarding processes, methodology, and trust. These guidelines can be leveraged by creators to increase transparency and trustworthiness in their AI development processes.

The goal of KYM is not to provide a definitive solution for developing trust in AI, but rather to encourage the need for transparency in AI toward establishing model identity and trustworthiness in AI. The guidelines bridge the gap between current Trustworthy AI texts by providing simplistic guidelines that target all principles of trustworthiness. The guidelines suggest key areas for increased attention in development, considering technical, ethical, and legal aspects in addition to trust. Therefore, the primary aim of this chapter is to outline a method to establish model identity with a general framework that all creators can apply to AI system development. The information required to fulfill the guidelines will vary by the complexity of each system, with more complex systems requiring greater attention to nuances in their use of data and modeling processes. This attention to detail will benefit creators by ensuring that the appropriate information is collected during each stage of AI development and easing the burden of proof for the effectiveness and trustworthiness of their systems.

3.1 Related Work

The fields of Know Your Customer (KYC) and Know Your Data (KYD) have paved the way for the collection of relevant information toward the goal of transparency, auditing, verification, and compliance. KYM is tangentially related to both of these fields, collecting relevant information on AI models and systems rather than on customers and data.

KYC refers to the requirement for financial institutions to ”monitor, audit, collect, and analyze” relevant information about their customers before engaging in business with them [44]. KYC policies are utilized to comply with a variety of financial regulations and laws, governing illegal behavior by customers such as money laundering, identity theft, fraud, and terrorist financing [44, 116]. This collection of customer data increases the legibility for a financial institution about a customer before and during their business with them, allowing them to assess whether the customer is engaging in legal behavior, and enabling them to comply with legal and regulatory requirements. Know Your Transaction (KYT) builds upon this to evaluate transactions for fraudulent behavior [180].

The field of KYD has recently been established to encourage data-driven assessments and regulation. The processes involved with data collection, management, use, and storage are very complex and nuanced, often complicated by regulatory or legal requirements [16]. KYD focuses on understanding the datasets utilized in AI development, assessing data quality and issues such as bias and explainability [295]. For example, Hawken and Munck [138] identified issues with common corruption benchmark datasets, revealing that data needs to be assessed for quality and validity. Toolkits are available for KYD, including Google’s Know Your Data tool [128].

Various other fields apply this same concept. For example, in the medical sciences, ”Know your target, know your molecule” [55], and ”Know your dose” [265] also provide a similar

basis for knowledge sharing targeting specific problems in science, biology, and medicine. Know Your Employee (KYE) encourages pre-employment screening for businesses to hire appropriate employees and to encourage anti-bribery and anti-corruption [104]. Know Your Vendor (KYV) advocates for laws and regulations on third-party vendors and suppliers to prevent fraud and ensure ethical business practices [87].

The concept of ”knowing” something in this context typically involves assessing the target with specific inquiries or questions, such as assessing bias in datasets in KYD or gathering customer data in KYC. For example, Bunnage et al. [55] establish a list of questions to understand drug candidates, in an attempt to understand drug discovery targets and enable targeted research.

This same concept is leveraged in KYM. KYM applies the concept of knowledge gathering and sharing to AI systems and models to assess, audit, and manage their efficacy, reliability, safety, and responsibility.

3.2 Know Your Model (KYM)

Developing trust in AI models is paramount. For a user to trust the outcomes of an AI model, increasing trust primarily depends on accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security. Models are the backbone of AI systems, and building trust in these models is the first step toward building a truly trustworthy AI system.

To build a trustworthy AI system, the trustworthy principles cannot be considered in silo. All principles of Trustworthy AI must be accounted for at the forefront of AI development. It is to this end that the KYM framework is proposed. The goal of the KYM framework is to increase trust in AI development by outlining a method in which providers, creators, developers, and other stakeholders can establish trust in their AI models and systems.

The KYM framework is built upon the idea that all models have a unique identity. This unique identity is established by model characteristics, such as model application, model type, features of the models, robustness, and reliability of model outcomes, mechanisms used to protect the privacy and other safety of users, and the individuals who are involved in the development and deployment of an AI system. These model characteristics can be leveraged to know and trust models. To this end, to ”know” a model implies collecting, recording, and storing detailed records of the processes undergone during the development of a model, subsequently establishing *model identity*. Here, model identity refers to the minimum information to distinguish one model from another, or establish a model's uniqueness. KYM strives for all models to have a unique model identity, allowing model characteristics to be leveraged to know and trust models.

As established in the previous chapter, transparency is key to trust, with research showing that increased transparency in many aspects of an AI system also increases trust in the AI system in turn. Therefore, to encourage large-scale AI adoption, transparency is vital to increase the trust that users have in the AI systems' processes and outcomes. While a holistic solution to create a perfect trustworthy AI system is not currently available, by increasing transparency in how developers address each Trustworthy AI principle, users will have the information needed to assess a model's trustworthiness.

The need for transparency highlights the necessity of the KYM framework. KYM provides a framework for providers, creators, developers, and other key stakeholders to increase transparency and trust in their AI models and systems. To establish trust, a provider has the responsibility to clearly establish its identity. This framework leverages four principles to guide increased transparency in AI models and systems: **efficacy**, **reliability**, **safety**, and **responsibility**. These four key concepts aim to encompass the requirements of Trustworthy AI in a simple, concise, and widely applicable framework.

While the framework of KYM addresses a multitude of trustworthy principles, it is the creators' responsibility to decide which guidelines are most important to address for their system. Of course, focus must be placed on increasing transparency in the use of data, the development of models, and how issues of AI trustworthiness are addressed. However, at the same time, it is important to recognize that this transparency is different among the various types of AI systems. As definitively proving that a model is trustworthy is quite difficult, a developer should keep and maintain thorough records on the application of KYM on their system. Developers should record techniques and tools they used to address issues in each area, and justifications for why a check was not completed or required. Due to the rapidly evolving nature of AI development and AI research, KYM suggests that developers remain vigilant in addressing issues of trust with regular checks and updates, particularly in respect to fairness, privacy, security, safety, user understanding, and reliability.

3.2.1 Key Principles of KYM

The four key principles of KYM summarize Trustworthy AI literature in an approachable and applicable manner. These principles distill the principles of trustworthy AI down into four targeted concepts that can be applied to the design, development, and deployment of AI models toward trustworthy AI systems.

Efficacy

Efficacy in KYM ensures that models produce an honest, fair, understandable, and desirable result. With the increase in the use of AI in everyday applications, it is vital to ensure that the outcomes of models are appropriate for their intended purpose, that the model performs well, and that outcomes are fair and beneficial to society. As systems can have unintended outcomes, it should be verified that models perform in the way that the developer intended.

In KYM, the principle of efficacy calls for:

- Transparency in the purpose, intentions, and outcomes of models, including intended purpose, target groups, and expected outputs.
- Efforts toward improving human understanding of processes, operations, and outcomes of the AI pipeline.
- Careful attention to fairness and non-discrimination in data and modeling to reduce bias and discrimination in outcomes.

This first principle establishes a strong basis for establishing trust in the intents and expectations of AI systems. A strong focus is placed on transparency in the inputs and outcomes of the system and model(s), including the expected applications, the use and treatment of data, and expected generalizability. As model outcomes are heavily influenced by model inputs, this principle also targets how data is utilized toward the model or system's intended purpose. This may include transparency in the treatment of data, feature extraction, training and testing, and prediction outcomes.

Towards addressing fairness and explainability in modeling and outcomes, Efficacy proposes transparency in how bias, discrimination, and understandability were addressed. Careful attention to fairness and non-discrimination is encouraged to reduce bias and discrimination in outcomes. Further, efforts to increase the explainability and understandability of the results and outcomes are encouraged to increase user trust.

Reliability

Reliability in KYM ensures that models are reliable in their outputs and developmental processes. Here, it is important to consider the processes that are used in development: Are the methods and processes leveraged during design and development appropriate and robust? Are the outcomes and processes verifiable, reproducible, and reliable? Would another method produce more reliable results? Are the appropriate regulatory and legal processes followed?

The principle of reliability calls for:

- Transparency in developmental processes, including the use and transformation processes of data, feature extraction, training, testing, and prediction outcomes.
- Reliability in outcomes and developmental processes, including the appropriate use of methods, availability, and consistency.
- Replicability or verifiability of outcomes and processes.
- Attention to data quality to avoid bad, inadequate, or inappropriate data collection, utilization, or transformation processes.
- Data and model provenance.
- Adherence to ethical, legal, and regulatory environments and requirements.

Of critical importance in this concept is replicability: developers should be able to reproduce the outcomes of their models and trace the model back to its origin. This includes ensuring proper provenance with records of data used, data transformations undergone, modeling processes (development environment, model type, hyperparameter tuning, etc.), and inference verification. Users should be able to verify the developmental products of models. Reliability advocates for clear documentation of data and model provenance, modeling processes and methods, and initial and ongoing performance so that a thorough auditing process is possible.

Attention should be given to data and modeling quality. The collection, preparation, and treatment of data are vital to consider when considering model identity. Data has a profound impact on the modeling process. The type and quality of data used for the development of an AI system have direct consequences on the quality of the models and inferences. For instance, data of poor quality or poorly leveraged data can lead to unreliable and incor-

rect models [305]. Therefore, careful attention to data processes is required. This principle requires reliability in the data use and data transformation in modeling systems.

Reliability does not specifically outline all of the specific requirements with data treatment. Instead, focus is placed on the major issues directly required to establish model identity. Developers should pay close attention to their use and treatment of data and be aware of any regulatory or legal requirements to determine the level at which they record their data. Modeling requirements are perhaps the most detailed records that developers should keep. Developers should ensure that their models are reliable and replicable. KYM advocates that developers keep clear records of their model development so that a clear auditing process can be completed.

This concept is especially relevant in regulatory environments, where developers may be required to verify the exact processes undergone during model development and reproduce relevant results.

Safety

The large-scale adoption of AI requires that users are confident that AI systems are safe to use and do not pose undue harm to the user or society as a whole. The need for safety is considered with great importance in KYM. Here, the concept of Safety includes assessing the safety, security, and privacy of AI systems from unintended accidents, breaches, and threats to user privacy.

This principle calls for:

- Building AI systems with careful attention to safety, including safe design, contingencies in case of error or failure, and audits or standards to assess initial and continuous system safety.
- System and model stability, including attention to failures and their causes, mainte-

nance to address and fix failures upon occurrence, and reducing failure rates [309].

- Robustness to threats to security, including robustness to attacks from adversaries or malicious actors and continual attention to state-of-the-art security techniques.
- Careful attention to user privacy, including (personal) data collection, utilization, and storage. This also includes any legal or regulatory requirements for securing user information.

This principle considers safety with a holistic approach: safety of the system from failure, security against breaches and errors that would cause harm to individual privacy, robustness against security threats, and otherwise safety against the AI system causing harm to humanity. Namely, this principle calls for risk-assessment measures of the AI system's impact on users and society and safety, security, and privacy failure points. This encourages developers and creators to conduct critical assessments of their systems to identify, assess, and mitigate failure points. By conducting these risk assessments, AI safety risks should be reduced.

One critical component of Safety in KYM is user privacy. Privacy is a significant issue in Trustworthy AI, especially due to the increased threat of personal data breaches in modern times. Protecting the private data utilized for the AI system is vital to ensure user trust. Therefore, this principle places a strong emphasis on the storage, use, and protection of personal data to ensure trust in the AI system.

Lastly, as security threats contribute to safety threats, securing AI systems from malicious actors is important [151]. Safety encourages measures of security robustness, such as compliance with security requirements and/or securing AI systems from malicious threats and attackers.

Responsibility

The principle of Responsibility targets the impact and benefit of AI systems on society, including the impact professional entities involved in the design, development, and deployment of an AI system, the societal and social impact of AI systems, the role of explainability and human control, and the regulatory and legal landscape relevant for the AI system. Responsibility in KYM bridges the gap between technical implementations and the legal and ethical implications of AI systems. It specifically calls attention to social and ethical responsibility and accountability. In addition to technical information about AI systems, creators must pay close attention to societal, social, and developer roles in the overarching impact of their systems.

The principle of responsibility calls for:

- Transparency about the developer or creator identity, including transparency about stakeholders and entities involved in the design and deployment of AI systems.
- Careful attention to the level at which human control is required and provided, including clarity on the implementation of human control in a system, opportunities for human intervention and review, and safeguards in the absence of human control.
- Consideration of the societal impact, purpose, and value of AI systems, and methods to maximize their benefit to society.

The principle of responsibility in KYM is perhaps the most abstract. With the large variation in the applications of AI systems, responsibility will have a different meaning for each creator. Rather than providing concrete guidelines in this area, KYM encourages creators to be transparent about the impacts and purposes of their systems, who was involved in their creation, and the level at which human control is required and provided.

An important aspect of responsibility is developer identity. In some cases, it is important to be transparent about what entity or entities were involved in the development of a model. It may be important to consider the experience and credentials of developers and the investments and interests of developers in model development.

Further, the aspect of human-AI interactions targets the explainability of an AI system. It highlights the need for transparency between the AI system and its users. Interactions between AI and user should be transparent, and it should be clear where the AI system makes decisions alone, requires human review, and/or allows for human intervention. This may be particularly relevant in today's landscape, where many regulators are calling for there to be transparency between the AI system and the user in that the AI system cannot attempt to conceal its nature and pretend to be a human [142].

This principle encourages transparency in the impact, benefit, and value of an AI system. This may align with accountability principles of explaining the purpose of an AI system, but it may also align with justifying the value of an AI system to society. For example, the EO calls for AI development to "positively augment human work", calling for responsibility in the development of AI to encourage human work rather than remove humans from work entirely [354]. This principle supports this notion by encouraging developers to justify the value and benefit of their systems toward users and/or society.

3.3 Key Guidelines of KYM

This section outlines the 20 guidelines of KYM, aligned with the four core principles efficacy, reliability, safety, and responsibility. These guidelines summarize the information developers are encouraged to record to establish model identity. The keywords "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [49].

Each guideline is followed by an example application. A summary of the KYM guidelines and additional, shortened explanations can be found in Table 3.1, Table 3.2, Table 3.3, and Table 3.4.

3.3.1 Efficacy

E1: Creators MUST describe the intended purpose, use, target user, and outputs of the system.

Creators MUST record information on the intentions of their AI systems. d and time-frame-of-validity of the system. Where possible, clear communication of the target user of the system or its outcomes should be provided. Further, Creators MUST be transparent about the expected outputs of the system, as the expected and observed outputs may differ.

In cases where intentions and outcomes are misaligned, the extent of the misalignment and any positive and/or negative impacts MUST be known and recorded. As it is vital to understand the impact of an unexpected output, creators have the responsibility to identify the causes of any misalignment in intents and outcomes.

Consider the following example:

> *This AI system is an Agricultural Drone Monitoring System. This system is designed to optimize agricultural practices by providing high-resolution drone imagery for crop monitoring, health assessment, climate assessment, and yield prediction. It is specifically intended for farmers, agricultural consultants, and businesses aiming to improve farming techniques and crop assessment and yield. The system utilizes aerial drones equipped with sensors and cameras to collect data on crop conditions, moisture levels, pest presence, and overall crop health. The expected outputs from this system include detailed maps on crop growth patterns. Based on these maps, the system pro-*

Table 3.1: The Know Your Model (KYM) Guidelines for Efficacy (E#). An example application of each guideline is provided. These examples are simplified. Real-world applications may require longer and more technical justifications.

Know Your Model Guidelines		Example Application of Guideline
E1	**Creators MUST describe the intended purpose, use, target user, and output of the system.**	[Navigation] "The system is a navigation system that users can use to map the most efficient path from one location to another. The system outputs the shortest path as defined by the estimated travel time from one input to another, utilizing available geographical information at the time of request. The targeted users are individuals who utilize iOS devices for navigation. "
E2	**Creators MUST record (statistical) metrics about training and test datasets.**	[Logistics AI] "The system was trained on a combination of our weekly, quarterly, and annual volume information. This data shows an average purchase of 10,000 units. (sd = 1,000), with higher throughput events with an average of 15,000 units (sd=2,500) occurring around holidays. It was confirmed that the training and test datasets exhibit identical distributions."
E3	**Creators SHOULD describe the expected performance on unseen data**	[Medical AI] "Data from low-quality or outdated equipment will result in poor performance. Shadowing or blurring in images may negatively affect model performance."
E4	**Creators MUST record *methods* taken to reduce bias, discrimination, and fairness issues in data and modeling outcomes, and SHOULD record specific *metrics* on bias, discrimination, and fairness.**	[Criminal Sentencing AI] "In order to ensure fairness in sentences, all potentially identifying sensitive demographic information has been removed from the dataset. Additionally, the system was evaluated by experts in justice and equality in order to mitigate potential problems with bias. Bias remediation was performed using [state-of-the-art tool]. A bias was identified and mitigated with a re-weighing method."
E5	**Creators SHOULD aim for increased understandability.**	[Medical AI] "This model is designed to be used by trained doctors. The system provides diagnostic information and justifications that explain the features utilized for each decision with importance ranking. Furthermore, the system provides documentation to provide additional information."

vides recommendations for targeted interventions such as irrigation, pesticide application, and fertilizer distribution.

This example illustrates the level of detail required for this principle. The creators demon-

strate transparency in the purpose, target users, and expected outputs of their systems.

E2: Creators MUST record (statistical) metrics about training and test datasets.

Creators must be transparent about the data utilized during the development of an AI system. Creators MUST record relevant metrics about the training (and test, where relevant) datasets. For example, this might include the volume of data utilized, dataset split utilized, class distributions, data quality metrics, data sources, feature distribution, feature correlation, temporal metrics, and geographical metrics. If applicable, this SHOULD include metrics on demographic representation and bias indicators.

The information required to fulfill this requirement will vary depending on the system. In cases of private data and strict regulations governing the reporting on private data, reporting on dataset metrics might be limited to less specific factors such as the volume of data utilized, quality metrics, and bias indicators.

Consider the following example:

> *Regarding a Medical Image Diagnosis System designed to identify and classify tumors from medical imaging scans, detailed statistical metrics about both the training and test datasets used during the system's development were recorded. These metrics included the number of images, patient demographics, distribution of tumor types, distribution of tumor location, and geographical distributions of the location the patient received care. All images were assessed for quality by setting inclusion criteria of a specific resolution threshold. The dataset features a representative sample of each tumor type and location. An imbalance is noted in the distribution of the treatment center in which a patient received care, revealing that more data is available for care centers in city centers compared to rural treatment facilities.*

This documentation provides insights into the system's potential limitations and biases. It aids in identifying where the dataset may need to be expanded to improve the system's performance and reliability. It can also reveal the system's relevance to different problems, in case users would like to apply the system to use cases outside of the established purpose.

E3: Creators SHOULD describe the expected performance on unseen data.

Once deployed, AI systems may experience data that is vastly different than the data used to train/test the system. Creators SHOULD describe the expected performance on unseen data, such as data from different distributions or data collected in different conditions.

For example, consider a speech recognition system developed to perform medical dictation:

> *The system has an expected accuracy rate of 95% on standard medical dictations in clear, noise-free environments. However, a slight decrease in performance to an accuracy of 90% is expected in real-world clinical settings where there is a lot of background noise, such as emergency rooms or busy clinics. This variance is caused by ambient sounds not fully represented in the dataset.*

This explanation reveals that the developers have noted a difference in model performance in rare conditions. This attention to detail and anticipation of variations in performance guides users in their use of the system.

Of course, it is not always feasible to anticipate performance on unseen data. Utilizing robust development processes is therefore encouraged to avoid issues with generalizability.

E4: Creators MUST record *methods* taken to reduce bias, discrimination, and fairness issues in data and modeling outcomes, and SHOULD record specific *metrics* on bias, discrimination, and fairness.

Creators SHOULD record any methods taken to address bias, discrimination, and fairness issues in data or modeling outcomes. This may include data treatment techniques and remediation, model checks and remediation, and outcome verification.

Even in cases where careful attention is paid to reducing bias in input data, algorithms may still exhibit biased behaviors. Developers are encouraged to pursue methods to measure fairness in their outcomes, using state-of-the-art methods and tools.

For example, consider a criminal sentencing AI that suggests sentencing ranges utilizing historical data, legal precedents, and case-specific details. The developers clearly describe identified bias and attempts made to mitigate it.

> *During development, the data was carefully reviewed and sanitized to remove any variables that could directly or indirectly lead to biased outcomes, such as race, gender, zip code, or socioeconomic factors. The data represented a balanced group of cases and demographic groups. The team utilized measures of group fairness to assess the sentencing recommendations across the various demographic groups. A minor bias was noted in drug-related offenses related to certain socio-economic groups and was mitigated utilizing [state-of-the-art] techniques. Before deployment, the decision-making process was audited for bias by a panel of external experts who analyzed model decisions across various demographic groups.*

As this principle is primarily targeting bias, discrimination, and fairness in personal data, it may not be relevant for systems that do not utilize private data.

E5: Creators SHOULD aim for increased understandability.

Developers SHOULD attempt to increase understanding of all stages of AI development to different users and groups. Detail efforts are taken to improve explainability & interpretability, and human-AI interactions (review, validation, etc.) of the developmental processes and outcomes of AI systems. This may include providing explanations on model decisions, clarity in model processes and techniques utilized, and interpreting model development and functionality in language appropriate to the target user.

For example, an AI system trained to provide educational content recommendations to educators based on individual student needs, learning styles, and performance levels might provide the following explanation:

> *The system's recommendations are easily understood by both educators and students due to the implementation of transparent explanations of the decision-making process. Each recommendation is accompanied by a clear explanation of which features of a student's profile contributed to the decision-making process. Further, the system has a user-friendly interface that highlights the student's strengths, areas of improvement, and how the recommended content aligns with their learning goals. This interface can be updated by students and educators to adapt the recommendations.*

Where possible, developers can greatly increase trust by improving explainability, interpretability, and understandability. By enabling users to understand the decision-making process and its influences, trust can be developed.

3.3.2 Reliability

RL1: Creators MUST record the processes followed in the development of the AI system.

Document and justify the requirements, design specifications, data collection, preparation, and storage processes, implemented algorithms and techniques, verification and testing methods, output generation, ethical/legal/regulatory compliance, and deployment monitoring of an AI system. Documentation MUST be thorough and include all information needed to identify and justify utilized methods, identify storage locations, and replicate outcomes. The extent of the required information will vary greatly depending on the system type.

For example, considering an AI personal finance advisor:

> *The AI system was designed with a user-friendly interface, secure data integration with banking institutions, and real-time financial analytics. Data collection included gathering financial transactions from accredited partner institutions, market trends, and user feedback. All data was anonymized and stored in compliance with financial regulations and data protection laws. Reinforcement and natural language processing algorithms were leveraged for the decision-making process based on individual user profiles and preferences. Verification and testing were conducted through simulated financial scenarios and A/B testing. The system is continually monitored for performance, user satisfaction, and compliance with financial regulations.*

This documentation should provide a summary and justification of all processes followed, whereas the following rules will provide a more in-depth explanation of data and model provenance.

This guideline ensures transparency in the development process, allowing all stakeholders, users, regulators, and other relevant entities to understand how the system was designed, built, and deployed. The decisions made at each stage should be explained and justified. This documentation increases accountability, enables reproducibility, encourages quality assurance, allows for assessment and management of risks, and allows brief assessments of ethical and legal compliance.

RL2: Creators MUST ensure adequate provenance for data.

Creators MUST maintain clear records of data collection, utilization, and transformation processes. Records MUST be adequate, clear, and complete enough to determine the origin of the data, assess data quality, and understand any transformations that occurred. Records may include but are not limited to, data collection process and techniques, the identity of data owner or licensing entity, dataset creation time, type and amount of data utilized, dataset utilization in development, and data updating practices.

Consider an example of a traffic management system that utilizes AI to optimize traffic flow, reduce traffic congestion, and improve road safety:

> *This system leverages data from a variety of sources, including traffic cameras, sensor data from roads, GPS data from vehicles and smartphones, and historical traffic patterns. Data management includes records of the origin of each dataset, including locations where data was collected, device metadata, time stamps, and data prepossessing steps. Metadata describing the data provenance was stored alongside the datasets in a structured format, including information about the data's source, collection time, and any processing it underwent. Historical data was collected for traffic patterns between 1990 - 2020, and real-time data was collected at 30-second intervals. Data storage included secure cloud storage and on-premises data centers. All data was*

Table 3.2: The Know Your Model (KYM) Guidelines for Reliability (RL#). An example application of each guideline is provided. These examples are simplified. Real-world applications may require longer and more technical justifications.

Know Your Model Guidelines		Example Application of Guideline
RL1	**Creators MUST record the processes followed in the development of the AI system.**	[E-Commerce AI] "This model leverages neural network technology, building on research previously published in the domain. Model training and testing were tracked locally and will be stored for three years following the end-of-life of the product. Data is collected and stored in accordance with international regulation."
RL2	**Creators MUST ensure adequate provenance for data.**	[Social Media AI] "Textual data was parsed from three social media websites between the dates of January and May 2020, and stored on a private server. Data were not checked for quality. Datasets are documented internally. The system maintains an index of all data as well as a log of all changes to the data set. Unigram transformation and punctuation removal were utilized."
RL3	**Creators MUST ensure adequate provenance for end-to-end model development.**	[Advertising AI] "Complete records of metadata from model training, testing, and prediction were taken utilizing an end-to-end asset tracking tool."
RL4	**Creators MUST record evaluation and performance metrics.**	[Classification AI] "Models were trained using a 70/30 test/train split, 10-fold cross-validation, and evaluated using prediction accuracy and AUC. The chosen model has an 80.2% accuracy rate, with a sensitivity/specificity rate of 74.5%/61.8% respectively."
RL5	**Creators SHOULD track model update performance and information ingestion.**	[Social Media AI] "We capture user data upon each deployment and retrain the model with the captured data. Model performance is analyzed with each update and must remain within $\pm 15\%$."
RL6	**Creators SHOULD record metrics on outcome replicability.**	[Robotics AI] "In order to reproduce the system results, a docker file has been provided. By leveraging this dataset and docker file, the system will produce the same results. This docker file was created using the following dataset and model settings."

encrypted both at rest and in transit.

RL3: Creators MUST ensure adequate provenance for end-to-end model development.

Developers MUST maintain clear records of developmental processes undergone in AI design, development, and deployment. These records MUST be complete enough to be able to replicate model results and outcomes. Records may include data (and/or metadata) on feature extraction, training and testing, and prediction outcomes, date and time of modeling stages, development environment (development language, packages used, etc.), model version, time of the last update, changes in performance between updates, algorithms, and techniques used, training conditions (i.e. hyperparameters), use of the dataset in each stage, testing performance & results, etc.

Expanding upon the AI system for educational content described in E5:

> *Each stage of model development, including initial training, testing, and model refinements, was documented utilizing end-to-end ModelOps tools. The development environment utilized Python 3.10 and scikit-learn packages for machine learning and Pandas for data processing. Package requirements and version control were maintained with each iteration of the model. Detailed logs of each update were collected, including changes to model performance and recommendation effectiveness.*

Technical documentation to fulfill this requirement MUST be thorough. Leveraging asset management tools such as MLOps and ModelOps is highly encouraged.

RL4: Creators MUST record evaluation and performance metrics

Developers MUST record detailed records of the evaluation and performance processes used. Creators MUST maintain a record of the metrics and techniques that were used to measure the performance of their systems, such as accuracy, precision/recall, error rates,

F1 scores, AUC, etc. It is suggested that significant technical data is recorded. Metrics for both intermediary and final models are encouraged.

For example, consider the Medical Image Diagnosis System from guideline E2:

> *Standard evaluation metrics were utilized during training of the AI model, including accuracy, precision, recall, and F1 score. Area Under the Receiver Operating Characteristic Curve (AUROC) was utilized to evaluate the model's ability to distinguish between different tumor types. To benchmark performance, a well-known and widely utilized benchmarking dataset was utilized to assess performance on unseen data. The system achieved an accuracy of 94%, with a precision of 92% and a recall of 93%, indicating a high level of reliability in identifying tumors. The F1 score was recorded at 92.5%.*

These records SHOULD be specific enough to assess the performance of the model and if the appropriate metrics are being utilized.

The example above illustrates transparency in recording evaluation metrics. This allows for auditing of the performance of the AI system and may provide information on the effectiveness of the performance assessments. For example, if the benchmark utilized to assess the performance was found externally to be biased, it may reveal bias in the AI system and warrant further assessments.

RL5: Creators SHOULD track model update performance and information ingestion.

Developers SHOULD clearly track model updates and how new data is used and affects performance. If new data is ingested after deployment, developers SHOULD record the origin of the new data, how it is integrated into the system, and if there are any bounds for performance changes.

For example, an AI system trained on social media data to monitor sentiment and trends:

> *A monitoring system captures key performance indicators after each model update. Accuracy, precision, recall, and F1 score of sentiment classification are assessed with each model iteration. The system continuously monitors various social media platforms and feeds, collecting new data at regular intervals. The monitoring system tracks the rate of messages processed per minute and measures latency in processing times. Model performance is analyzed with each update and reaches a minimum F1 score of 85%.*

RL6: Creators SHOULD record metrics on outcome replicability.

Developers SHOULD measure the replicability of outcomes of their AI systems, utilizing state-of-the-art metrics.

Consider an AI news recommendation system:

> *A reproducibility measurement process was developed. The [state-of-the-art] metric was utilized to assess how well the recommendations were replicated across multiple user interactions. The system aimed for a replicability rate of at least 90%.*

3.3.3 Safety

S1: Creators MUST assess safety to users and society.

The development of systems MUST consider safety at the forefront. Developers MUST pay careful attention to safe design, failure contingencies, and safety standards. Consideration MUST be given to how the AI system impacts its surroundings, individuals, and society as a whole, and whether its use or deployment poses any safety risks. In the case that there are safety concerns, creators MUST be transparent about any safety concerns or issues the AI system may have.

Consider an AI-powered autonomous vehicle was designed to autonomously navigate traffic:

> *A comprehensive risk assessment was conducted, including an evaluation of potential safety risks such as collisions, pedestrians, and adverse weather conditions. Simulations and real-world testing were conducted to identify behavior in these various conditions. Mitigation strategies were identified to address safety concerns, and human-override features were prioritized in the design of the AI system. The system adheres to all international and national safety standards applicable to motor vehicles. Continuous monitoring and feedback mechanisms were implemented to adapt the system to unknown scenarios.*

Assessments of safety are vital to ensure user and stakeholder trust.

S2: Creators MUST assess potential security, safety, and privacy failure points.

Assessments of potential security, safety, and privacy failure points present in models (and solutions if available) MUST be undertaken.

For example, consider an AI chatbot system:

> *The system was assessed for potential failure points, and a number of areas of concern were identified. The system was found to be vulnerable to potential data breaches. Robust security measures were implemented to protect user data, including encryption and vulnerability assessments. Safety risks were identified associated with the chatbot providing improper feedback and/or incorrect information. These risks were mitigated through continuous monitoring, frequent updates, and warnings on the system about inaccuracies.*

These assessments MUST assess potential failure points concerning data breaches, system

Table 3.3: The Know Your Model (KYM) Guidelines for Safety (S#). An example application of each guideline is provided. These examples are simplified. Real-world applications may require longer and more technical justifications.

Know Your Model Guidelines		Example Application of Guideline
S1	**Creators MUST assess safety to users and society.**	[Robotics AI] "In the event of detected compromise, the system can be placed into a fail-safe state by the activation of a hardware cutoff or a software shutdown. To comply with safety standards, this system has several human-tracking safety features that override the AI in situations where humans can potentially be harmed."
S2	**Creators MUST assess potential security, safety, and privacy failure points.**	[Finance AI] "The system was designed with the following threat model in mind. The system is an online banking platform with the potential for both denial-of-service and database attacks. Additionally, the model is trained on user data that has been anonymized, however, attacks do exist that could de-anonymize users. Finally, the model itself is vulnerable to data poisoning or similar attacks. "
S3	**Creators SHOULD record metrics for security robustness.**	[E-Commerce AI] "Our system is regularly tested to comply with PCI DSS standards. We have also received ISO/IEC 27001:2013 certification for our handling of critical data."
S4	**Creators MUST ensure user privacy and appropriate treatment and use of private data.**	[E-Commerce AI] "Only data that is relevant to the product is collected, with the consent of the individual. Private data is stored on an encrypted server."
S5	**Creators SHOULD ensure secure data utilization and storage.**	[Personal Services AI] "Data is stored on an encrypted disk, where access is granted by keys. All data changes are signed by key, for easy traceability."

and/or component failure, and potential unauthorized access to the system.

S3: Creators SHOULD record metrics for security robustness.

Creators SHOULD record metrics taken for improving the robustness of their systems from adversarial attacks and malicious actors (i.e. checks undergone for adversarial concerns). The previous requirement (S2) requires the identification of security issues, and in cases where security risks are identified, creators SHOULD record the metrics taken to mitigate these security risks.

Continuing the previous example of the AI chatbot:

> *The security assessment revealed the potential for security issues associated with unauthorized access. A potential SQL injection vulnerability was identified. To measure the impact of this security threat, a simulation was conducted to explore detection time, response time, data compromise, and effectiveness of the mitigation strategies. To mitigate this type of attack, user inputs were sanitized to ensure malicious code was not executed.*

Due to the rapidly evolving nature of AI security, developers SHOULD continuously engage in improving security robustness by utilizing state-of-the-art techniques.

S4: Creators MUST ensure user privacy and appropriate treatment and use of private data.

Developers MUST be acutely aware of the treatment of user data and the role of user data in their systems development and outcomes.

Regarding an AI personal finance assistant:

> *Private user data, including financial transactions, account details, and communications, were encrypted both in transit and at rest. Explicit user consent was required before the collection and processing of user data. Only data relevant to the core functioning of the system was collected. Personal financial data was pseudo-anonymized whenever possible.*

Private data is not utilized in every system, and therefore this requirement will vary depending upon the data relevant to the AI system. For private data, creators MUST consider regulatory requirements for storage, deletion, and use of data, including requirements for consent.

S5: Creators SHOULD ensure secure data utilization and storage.

Creators SHOULD ensure that all data is used and stored securely.

For example, consider a smart home automation system for the control of in-home IoT devices:

> *User data, including voice commands, were encrypted end-to-end. Data was stored on an encrypted database. Robust access controls were implemented to restrict access to unauthorized authorities. Data retention policies were adopted to remove data after a specific time period to reduce the risk of data exposure. Logging and monitoring policies were implemented to track unauthorized access attempts and real-time alerts were configured to identify administrators of any potential threats.*

3.3.4 Responsibility

RP1: Creators SHOULD disclose or record all entities involved in system development.

Creators SHOULD record the identities (or affiliations), qualifications, and diversity of all entities involved (including stakeholders, businesses, domain experts, individuals, teams, etc.) in the design, development, and deployment of the AI system. This may include the experience and credentials of developers, team diversity, and the investments and interests of developers (and other stakeholders) in model development.

Consider an AI translation program:

> *The core development team is comprised of engineers, data scientists, and linguists. Each individual's role in design and development was clearly documented. A third-party company was engaged for data collection and security*

Table 3.4: The Know Your Model (KYM) Guidelines for Responsibility (RS#). An example application of each guideline is provided. These examples are simplified. Real-world applications may require longer and more technical justifications.

Know Your Model Guidelines		Example Application of Guideline
RP1	**Creators SHOULD disclose or record all entities involved in system development.**	[Human Resource AI] "Our team is composed of machine learning engineers, statisticians, and social scientists, all graduates of accredited universities. We consulted with an AI domain expert during development."
RP2	**Creators SHOULD detail the implementation of human-AI interactions.**	[Medical AI] "The system uses patient characteristics and health information to formulate diagnoses. The decisions must be confirmed by a human before a diagnosis can be made."
RP3	**Creators SHOULD describe the impact, value, and benefit of the system.**	[Chatbot] "The system allows for rapid interactions with customers. This increases availability, provides immediate assistance to customers, and reduces the need for customer service staff. The system is only used for our business and does not have any larger foreseen societal impacts."
RP4	**Creators MUST comply with legal and regulatory requirements.**	[Finance AI] "Our system complies with GDPR regulations on the use of private data, and internal regulations on the use of private data and clarity in decisions."

robustness assessments. Further, translators of rare languages were consulted during the evaluation of the system.

The disclosure of the entities involved in the design, development, deployment, and management of an AI system enforces accountability and transparency. Transparency in this regard may allow users and regulators to assess robustness, honesty, and compliance with relevant regulations.

RP2: Creators SHOULD detail the implementation of human-AI interactions.

Creators SHOULD understand the implementation of human-AI interactions in the system. This may include areas where human review is allowed and/or required, opportunities for human intervention, and human role in AI decisions.

Consider an AI personal assistant:

> *The system involves user interactions via web interfaces and mobile apps. The AI system interacts directly with the user on all platforms. The system takes user input in the format of text and audio and responds to the user based on sentiment analysis, voice recognition, and context prediction. User feedback is collected to assess the system's effectiveness and user satisfaction. All communication occurs between the AI system and the human user. Customer service agents do not participate in the communication with the human user through the interface and must be contacted separately.*

The implementation of human-AI interactions SHOULD consider both elements of explainability and safety of an AI system. Transparency on where a user is interacting with AI is encouraged.

RP3: Creators SHOULD describe the impact, value, and benefit of the system.

Creators SHOULD justify the impacts, values, and benefits that the AI system has to society. This may also include any potential detriments to society (and justifications for why the AI system maintains value).

An example of an e-commerce recommendation system for online shoppers:

> *The system analyzes user behavior, purchase history, and user preferences to provide tailored product recommendations. Users experience high satisfaction with the relevance of the products recommended to them compared to other recommendation systems, resulting in a 25% increase in user satisfaction and a 30% increase in user retention rates. The system improved user engagement and sales, resulting in a 10% increase in sales over a three-month period.*

The value and benefit of the system should be clearly communicated to assess the impact of the AI system. In the example above, this is clearly demonstrated with concrete impacts on customer satisfaction and sales.

RP4: Creators MUST comply with legal and regulatory requirements.

With the rising legal and regulatory requirements for AI development, careful attention MUST be given to national, international, and vocational requirements for AI design, development, and deployment.

Consider a financial fraud detection system:

> *The system was developed in compliance with financial regulations regarding requirements for data security, transaction monitoring, and reporting of suspicious activities. Additionally, the system adhered to GDPR data privacy regulations on the use of private data and data handling processes.*

These legal and regulatory requirements will vary greatly depending on the system. It is advised for developers to be keenly aware of the relevant laws and regulations during the design of the AI system.

3.4 Discussion and Future Work

With the proliferation of AI in greater society, members of academia and industry alike should strive for the development of robust, trustworthy systems. The complexity and wide array of applications in AI systems complicate the process of creating trust, placing the burden on each creator to establish a method for building and maintaining trust. Further, the complex landscape of Trustworthy AI literature, guidelines, regulations, and recommendations makes it difficult to define and apply trustworthy principles in regard to AI development.

The Know Your Model (KYM) framework guides the development of trustworthy AI via a simplistic, straightforward approach. The KYM guidelines aim to provide a comprehensive framework for creators to leverage to address both provenance and principles of trust in the design, development, and deployment of their AI systems. A set of 20 guidelines was established, centered around four key principles of trustworthiness: efficacy, reliability, safety, and responsibility. These four principles highlight the critical areas that need to be addressed by creators and developers in all stages of AI development. These guidelines can be leveraged to establish model identity and increase transparency and trust in AI.

This chapter has explored the 20 proposed guidelines. Each of the 20 guidelines has a direct connection to the principles of Trustworthy AI explored in Chapter 2. Further, example applications of guidelines were provided for various AI systems that have relevance today. These example applications provide only brief summaries of the type of information suggested to fulfill each guideline. In practice, explanations, justifications, and documentation should be more extensive and detailed.

Although previous efforts have been made to increase transparency and trust in AI, the focus has been placed primarily on provenance rather than trust. In those methods that do address trust, attention is only given to one or two principles, neglecting the importance of others. The KYM framework aims to merge provenance methods with a focus on trust, providing a complete framework for creators to assess their current and future AI processes. Further, this framework considers the importance of technical, ethical, and legal responsibility, providing guidelines that bridge the gap between research and industry.

As definitively proving that a system or model is trustworthy is quite difficult, it is suggested that developers maintain thorough records on methods taken to address trust concerns. KYM assists developers in improving transparency toward increasing trust. By increasing transparency, developers ensure clarity on how key issues are addressed, and users have the

information needed to assess trust where necessary. Developers should record techniques and tools they used to address issues in each area, and justifications for why a check was not completed or required. These should be frequently assessed and updated for every new model or model update. Increasing transparency in this way is the next step in increasing overall trust in AI. However, it is important to note that KYM is not an exhaustive list of requirements. Developers should continuously ensure that they engage in transparent practices in tracking their modeling efforts.

Further, it should be noted that while KYM provides a set of guidelines for creators to leverage, it does not provide a method for *sharing* this information among users, or outside of an organization. This distinction should be made by the creator depending on the unique factors of their systems. In general, thorough records and complete transparency within the development team and internal users (where applicable) are advised. In many cases, it is unlikely that full transparency in external sharing of data and model provenance is feasible for a multitude of creators due to proprietary or privacy concerns. The guidelines discussed in this chapter provide the ability for creators to guide internal records on these topics but do not provide an avenue for securely sharing this information. Additional attention will be needed in situations where this level of transparency with external stakeholders is required.

Additional attention is warranted on developing a formalized system for KYM. As the state of AI research is rapidly evolving, it would be beneficial to develop a formalized system that includes up-to-date methods to analyze the guidelines. Further, it would be of extreme value if these guidelines could be streamlined into an automated system for record-keeping for creators to leverage. Future work may include clear avenues for the sharing of information with external users, such as customers or the general public.

Chapter 4

Collaborative Learning: Leveraging Federated Learning to Increase Trust

Developing Trustworthy AI systems is a difficult task, requiring an interplay between a multitude of concepts and principles during all stages. While many solutions have been proposed, there are no complete solutions that address all principles of Trustworthy AI. In recent years, FL has emerged as a compelling solution, addressing several key principles of trustworthiness such as fairness & non-discrimination, privacy, robustness, and security. FL represents a groundbreaking step forward toward collaborative learning, fundamentally altering how data is utilized and processed while allowing multiple actors to build a robust and secure model.

In traditional architectures, data, and modeling are centralized, often leading to issues with robustness, transparency, and fairness. AI systems rely heavily on the data they are trained with, and issues with data quality and volume often result in fairness issues [191], privacy [36], and algorithmic robustness [88]. As algorithms become more complex, the amount

of data required to train them significantly increases [88]. Traditional approaches require providers and developers to train algorithms on their own data. This can potentially lead to issues when there is a scarcity or lack of diversity in the data that is available to them, resulting in algorithms that are poorly trained and are not able to generalize to reality. While large organizations or state actors may be able to collect sufficient data, it is a significant challenge for many smaller organizations or researchers to obtain the appropriate volume and quality of data to train robust algorithms.

For example, AIs utilized in healthcare are often observed to be biased due to training on data predominantly from a specific demographic group, data collected in specific clinical conditions, or data collected with misspecified outcomes [66]. The algorithms that underpin these models do not perform on unseen data, producing results that potentially amplify existing medical biases [260]. The collection of robust healthcare datasets is often hindered by requirements to protect patient privacy [250], and therefore healthcare organizations are often limited to training algorithms on the data available to them, potentially resulting in poorly trained and non-generalizable algorithms. While a large hospital may be able to collect sufficient data, how would small clinics get sufficient data to train robust algorithms when they are likely to see only a small subset or a specific demographic of patients?

FL addresses these issues by enabling multiple clients or participants to collaboratively train a model while keeping their data local and secure. FL is an ML setting where the training of a model is distributed across multiple clients to create a collaborative model that can learn from a larger subset of data than is available locally. FL maintains data privacy by design, as it trains algorithms across multiple decentralized clients, devices, or servers without exchanging the data itself [233].

The growth in popularity of FL is primarily due to the ability to benefit from private data without having access to it, the power of multiple clients collaborating to update the model

without sharing private data between themselves, and the vast number of clients that can participate, up to thousands or even millions for large-scale systems [161, 185, 233, 370]. However, privacy is not the only benefit of FL from a trustworthy perspective. FL can also improve fairness in AI due to its ability to learn from a wide array of data sources, providing opportunities for the AI model to learn from more diverse and heterogeneous datasets [24, 344]. Additionally, FL can enable transparency in AI systems operations and governance, allowing for better understanding and accountability of AI systems [321]. As such, FL stands as a robust approach in the pursuit of Trustworthy AI, fostering confidence in AI systems among users and stakeholders.

In this Chapter, FL is described with its relevant implications on trust. Some relevant use cases of FL are explored, and challenges are identified. The strengths of FL are explored briefly with a case study on transaction monitoring for applications in Anti-Money Laundering. This case study reveals the benefits of FL in encouraging collaborative learning without the sharing of data, ensuring privacy without performance loss. In Chapters 5, and 6, FL is further explored to assess its impact on trust in AI.

4.1 Federated Learning

McMahan et al. [233] proposed FL in 2016 to address the privacy concerns with data acquisition in decentralized devices performing collaborative training. Simply, FL is a distributed ML setting where multiple clients can collaboratively train a model without sharing private data [233]. The decentralized approach ensures that sensitive user data remains private and secure by enabling collaborative model training without the need for data sharing. Traditional ML methods require centralizing the data, which can be challenging when dealing with large-scale or distributed data. In contrast, FL enables the training of models across multiple devices, such as smartphones or Internet of Things (IoT) devices, while keeping the data localized. This allows organizations or platforms to leverage the collective knowl-

edge from a vast number of devices without compromising data privacy. Due to its strength in allowing many participants to collaborate, FL has gained popularity, with applications in mobile devices [206], speech and image recognition [222], finance [215], and medicine [200].

Typically orchestrated by a central server, FL follows a multi-round, multi-agent-based strategy. In each round, the server distributes a current global ML model to a random subset of participants, who then separately leverage private data to locally update the model. Each participant separately and concurrently sends the difference between the current global model and their updated model back to the server, potentially in the form of masked gradients or weights with encryption [13, 383], differential privacy [317], or secret sharing techniques [47]. The updated models are sent back to the server, which aggregates the updates into a new global model. This type of FL is referred to as the client-server, or centralized architecture [281].

However, decentralized architectures exist as well, removing the need for a central server and replacing it with peer-to-peer architecture [281]. In this setting, the central server is replaced with specific aggregation mechanisms that entire peer-to-peer communication and privacy. Participants have local models, and improvements are made with communication with their neighbors [348]. Decentralized architectures are out of scope for this dissertation; for further information about these methods please reference [34].

In both cases, aggregation is performed by an aggregation algorithm, a critical component that combines the learning updates from multiple devices to create a global model. These aggregation mechanisms are the core of FL research. Various aggregation mechanisms have been proposed, with each having a unique ability to handle various challenges in the FL environment. These challenges include (among others): data heterogeneity, where data across different clients may not be independent and identically distributed (*iid*); system het-

erogeneity, considering client-level capacity for communication, computation, and storage; communication efficiency, particularly targeting update size and security; and robustness against attacks such as model poisoning and backdoor attacks [281].

Perhaps the most common and straightforward aggregation mechanism is Federated Averaging (FedAvg) [233]. In FedAvg, selected clients or participants in the federated network train a local model on their own data using the procedure described above. Orchestrated by a central server, selected participants receive a pre-trained or basic generic global model. Participants separately train this global model locally using their own data for multiple epochs on a stochastic gradient descent (SGD) optimization algorithm and send the updated model to the central server. FedAvg aggregates these models to update the global model using a simple weighted average of the received updates. This process is repeated iteratively until the server outputs a final global model.

Other types of aggregation mechanisms include methods such as FedBoost [134], Scaffold [170], FedProx [203], FedMA [358], Adaptive Federated Optimization [285], and Secure Aggregation [47], among countless others. Each aggregation mechanism differs in the approach for combining model updates, and each has strengths and weaknesses. For example, Adaptive Federated Optimization is the only one that has an adaptive learning rate, but it exhibits issues with client drift [281].

However, the type of aggregation mechanism is also informed by the type of FL utilized. There are three main types of FL systems: horizontal FL, vertical FL, and transfer FL [370]. The type utilized is characterized by the type of data partitioning and communication architectures used. Specifically, data partitioning concerns the *sample space* and *feature space*, where the sample space includes all dataset instances (i.e. samples), and the feature space includes all dataset attributes (i.e. features) [370].

Horizontal FL is utilized in scenarios where datasets share feature spaces but differ in the sample space. For example, different hospitals might collect records on the same type of data (feature space), but collect that data for different sets of patients (sample space). Horizontal FL enables model updates using the same model architecture due to the overlaps in feature space. Using their local data, participants train the local model using the same architecture and send the updates back for aggregation into the global model. The approach proposed by McMahan et al. [233] utilizes this type of learning: utilizing Android phone updates, individual Android phones update model parameters locally and upload them, where all model updates are aggregated into a global model to improve the model for all users. Other approaches utilizing horizontal FL include approaches utilizing Federated Optimization [185], Deep Gradient Compression [207], and Stochastic Gradient Descent [324], among others [120, 317].

Vertical FL, on the other hand, applies to scenarios where datasets share sample spaces but differ in feature space [370]. For example, different types of financial institutions such as investment firms and retail banks, may have the same customers (sample space), but gather different data on those customers (features space). Vertical FL aggregates the different features and calculates the model parameters in a privacy-preserving manner. One such method is Secureboost, a privacy-preserving tree-boosting algorithm [75]. Proposed methods for vertical FL include secure linear regression [95, 122, 171, 307], ridge-regression [125], and privacy-preserving logistic regression [255].

Transfer FL applies in settings where datasets differ in both sample and feature space, but common representations of their overlapping feature and/or sample spaces can be leveraged to learn [213]. This is an extension to existing FL systems to solve problems that exceed the bounds of the other two types of systems. An example of this type of FL might be when different types of companies (such as an e-commerce company and a financial institution) in different locations or markets (such as different countries) want to collaborate. Due to

differences in the types of business conducted and geographical location, there is likely very little overlap in feature and sample spaces. This type of FL might be leveraged as a solution to build a common system. Methods of transfer FL include Fedhealth, a transfer FL method for wearable medical devices [73], and others [160, 178, 313, 368, 393].

4.2 Federated Learning Use Cases

FL has a wide array of practice use cases across various domains, with applications in IoT devices, speech and image recognition, and predictive analysis in domains such as healthcare, finance, manufacturing, transportation, infrastructure, e-commerce, and other industries [200].

Speech, text, and image recognition tasks are incredibly prevalent in FL research. In particular, several approaches have been proposed for text prediction using mobile device keyboards to predict user input [136, 194, 329]. Language modeling, text classification, speech recognition, sequence tagging, and recommendation systems are other popular natural language processing (NLP) tasks observed in FL research [209]. Image recognition tasks are also prevalent, with researchers training FL systems on a variety of computer vision tasks [179]. For example, Yang et al. [368] propose FedSteg for secure image steganalysis that improves upon existing steganalysis methods.

Applications of FL apply these methods to solve problems in specific domains. In healthcare, FL is enabling collaboration between healthcare institutions to develop diagnostic tools, personalized healthcare, and monitoring tools. For example, Chen et al. [73] proposed FedHealth, a transfer FL solution for integrating wearable medical device data to build personalized models of activity. Learning digital medical information from electronic health records (EHR) is another significant application, enabling collaborative analysis between healthcare institutions that protect patient data [51, 83, 219, 347]. Several researchers have also applied FL to medical imaging, with applications in MRI imaging of the brain [133, 319],

x-ray and medical imaging radiology scans [107, 249, 287], and pathology [218]. Applications in electroencephalography (EEG) signal classification problems also show significant promise [120, 160]. Other applications in healthcare include remote health monitoring, disease diagnosis and detection, and other health monitoring [252, 363].

In the finance sector, FL has been leveraged to enhance how financial institutions handle data and collaborate for problems such as fraud detection, risk management, and personalized financial services. Cheng et al. [76] utilize FL for loan risk assessment, Kawa et al. [176] for assessment of credit risk, and Yang et al. [371] and [58] apply it to credit card fraud. Long et al. [215] propose leveraging FL for open banking to develop superior AI models for financial services. Other researchers apply it to financial text classification [27], predicting financial distress [154], credit scoring [398], and financial crime detection [333, 334].

Further, there have been several notable contributions to the manufacturing and supply chain. Kevin et al. [178] propose a transfer FL method for cross-domain prediction in smart manufacturing and production processes. Similarly, Zhang and Li [393] proposes a fault-diagnosis method.

In the transportation domain, FL has been applied for the development of route planning, traffic management, and autonomous vehicle guidance, among others [336]. Applications in this domain include vehicular communications [304], energy demand prediction for efficient use of charging stations by electric vehicles [308], traffic management [232], and other intelligent transport systems [391]. Likewise, FL has been applied to infrastructure, where FL has been applied to the development of smart grids [209, 209, 331], smart utility meters [359], and smart cities [400].

The use cases in this section are perhaps the most prevalent in FL research, but the relevance of FL goes beyond these use cases. Additional use-cases include e-commerce [199],

recommendation systems [356, 369], personalized systems [335], and information retrieval [275], among others. For a more thorough review of applications in FL, please reference [200].

4.3 Case Study: Anti-Money Laundering

In the previous section, a multitude of FL applications were explored. Building upon this, this section aims to provide a more focused exploration of a specific use case, Anti-money laundering (AML). Quite a few applications mentioned concern the finance sector, where FL has applications in risk assessment, fraud prediction, and the improvement of financial services. This section aims to explore how FL can improve performance in AML applications.

Money laundering is the illegal process of concealing the origin of money by converting it to an official source. Money laundering involves three steps: placement (introducing the illegal funds into the financial system), layering (concealing the source of the money through a series of transactions and other tricks), and integration (the money is reintroduced into the economy and is used for legitimate purposes) [81].

Preventing money laundering is a critical concern in the financial sector. AML refers to the laws, regulations, and procedures intended to prevent money laundering [116]. AML efforts include a wide range of measures and activities taken by financial institutions and regulatory bodies to detect and prevent money laundering, such as customer due diligence (CDD), KYC, monitoring and reporting of suspicious transactions, and compliance with regulatory requirements [110]. AML frameworks are designed to combat money laundering by requiring financial institutions to report activities that might be associated with criminal activity. There are a multitude of techniques already utilized for AML, including transaction monitoring systems, risk assessment methods, and algorithmic detection using AI. [8]. Further, CDD and KYC procedures require financial institutions to collect, monitor, and

analyze relevant information about their customers to prevent identity theft, fraud, money laundering, and terrorist financing [44].

In AML, a successful detection system can identify trends and/or patterns indicative of suspicious behavior and levy an alert that signals potentially fraudulent behavior. However, several challenges in the AML domain make it difficult to develop a sufficient system, such as quality and volume of data, class imbalance, concept drift, class overlap, and class mislabelling [74].

This section explores the use of FL in creating a robust AML detection system. FL shows great promise in the field of AML due to the ability to learn from data from heterogeneous sources. Often, financial institutions are hindered by a lack of quality or plentiful data on fraudulent transactions. Financial institutions process vast amounts of transactions per day, and only a small proportion of those are fraudulent [74]. Further, current approaches often result in low detection rates and high false positive (FP) alerts, as fraudulent cases are not always caught, and some valid transactions are incorrectly labeled as fraudulent [264]. By enabling multiple institutions to collaborate to create a global model utilizing multiple resources, detection rates may improve, and FPs may become less frequent.

The experiments were conducted to understand three primary research questions:

1. Can FL be leveraged to encourage collaboration between financial institutions without a decrease in model performance?
2. How can FL be leveraged to specifically target a reduction of false-positive alerts?
3. How can participant contribution be assessed?

Related Work

FL applications in AML have shown significant strengths in improving model performance and improving fraud detection. Suzumura et al. [334] demonstrate the strengths of federated graph learning on detecting money laundering using the UK FCA TechSprint dataset, resulting in a 20% performance increase over local models. Similarly, Du et al. [93] propose GraphSniffer, a graph-based learner that they demonstrate on malicious Bitcoin transactions. Kanamori et al. [168] propose DeepProtect to detect financial fraud in both transactions and bank accounts using a real dataset from five banks in Japan. Myalil et al. [247] perform federated learning on fraud detection where active and malicious adversaries are involved and propose a method to remove malicious participants. Several approaches have been proposed for credit card fraud detection [367, 371, 399], credit risk assessment [176], and credit scoring [398].

Experimental Setup

Due to the privacy associated with financial data, there are few publicly available datasets for the study of AML. To solve this problem, several researchers have proposed simulators for generating transaction data based on real datasets [217]. Here, AML detection is explored utilizing the PaySim dataset [216]. This dataset is a simulated dataset composed of mobile money transactions generated based on real transactions from a multinational company providing mobile financial services to over 14 countries. A portion of the PaySim dataset with 6.3 million transactions was used.

Metrics	**Value**
Total Transactions	6,362,620
Legitimate Transactions	6,354,407
Fraudulent Transactions	8,213
Average Legitimate Transactions	$178,197
Average Fraudulent Transactions	$1,467,967

Table 4.1: A summary of the PaySim dataset, regarding the total transactions, legitimate transactions, fraudulent transactions, and average transaction amount for each.

Transaction Type	Count	Average Amount
CASH_IN	1,399,284	$168,920
CASH_OUT	2,237,500	$176,274
DEBIT	41,432	$5,484
PAYMENT	2,151,495	$13,058
TRANSFER	532,909	$910,647

Table 4.2: A summary of the *type* variable in the PaySim dataset. A total of five payment types are recorded with variable frequencies.

The dataset contains 11 variables including features measuring transaction type, transaction amount, time passed, and several features representing the origin and destination account information and balances. Transactions are labeled with '0' if they are legitimate, and '1' if they are fraudulent. Further, the authors added an additional feature that measures if the transaction is flagged by a simple rule-based system. Categorical variables representing client identification codes were removed from the dataset, leaving seven variables. The models were trained with an 80%/20% train/test ratio. A summary of the dataset can be found in Table 4.1, and the number of transactions for each transaction type can be found in Table 4.2.

One significant challenge in this domain is the low proportion of positive samples in datasets. In the domain, fraudulent transactions make up less than 1% of transactions [38]. This is reflected in this dataset with 0.13% of all transactions being fraudulent. Due to this imbalance, detection systems are a challenge to train. If the system is trained poorly, it could completely miss all positive cases, alerting in very few cases, or it could result in too many alerts, adding more work to financial institutions to check each alert for legitimacy. Neither case is ideal, so a balance must be struck.

The experiments were conducted using Flower, a comprehensive framework designed to facilitate the development and testing of federated learning algorithms [39]. A centralized architecture was employed, utilizing the FedAvg aggregation mechanism in conjunction

with an XGBoost model for enhanced performance and efficiency [70]. All FL experiments were conducted for 10 rounds, with 5 clients per round. For comparison purpose, a global model trained on all data, and local models trained only on a proportion of the dataset.

Area Under the Precision-Recall Curve (AUCPR) was utilized as the primary evaluation metric to assess model performance, due to its strengths with imbalanced datasets [48]. The following metrics were also utilized to analyze the model results in this section:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$
$$\text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{TP}{TP+FN}$$
$$\text{F1 Score} = \frac{2*Precision*Recall}{Precision+Recall}$$
$$\text{F0.5 Score} = (1+0.5)^2 * \frac{Precision*Recall}{(0.5^2*Precision)+Recall}$$
$$\text{False Positive Rate (FPR)} = \frac{FP}{TN+FP}$$
$$\text{False Negative Rate (FNR)} = \frac{FN}{TP+FN}$$

where TP is the number of True Positives, TN the number of True Negatives, FP the number of False Positives, and FN the number of False Negatives.

Experiment 1

The first experiment aims to understand how FL can be leveraged to encourage collaboration between financial institutions. The goal is to demonstrate the strength of FL in enabling collaboration between financial institutions, without negatively impacting performance. As mentioned, financial institutions are often hindered in their AML detection efforts by a lack of data, particularly correctly labeled fraudulent transactions. FL can greatly improve the detection of fraudulent transactions by enabling collaboration between financial institutions, allowing each to benefit from a larger volume of data without directly sharing data.

In this experiment, the dataset is split between 5 participants (i.e. *financial institutions*). To simulate equality between participants, the dataset is split equally among classes to ensure that each has an equal proportion of each type of transaction. To analyze the federation, the following are considered: (1) global model: a model trained on the entire dataset; (2) local models: each participant trains a local model utilizing only their data; (3) FL model: the results of the FL model trained iteratively with all participants. The average result of five runs for each model is taken. The results are displayed in Table 4.3.

	Accuracy	F1	F0.5	FPR	FNR	Precision	Recall
Global	0.9995	0.83	0.78	0.0004	0.06	0.75	0.94
Local 1	0.9994	0.80	0.75	0.0005	0.09	0.72	0.91
Local 2	0.9994	0.80	0.75	0.0005	0.09	0.71	0.91
Local 3	0.9994	0.80	0.75	0.0005	0.09	0.72	0.91
Local 4	0.9994	0.80	0.75	0.0004	0.09	0.72	0.91
Local 5	0.9994	0.81	0.76	0.0004	0.1	0.73	0.90
FL	0.9997	0.88	0.93	>0.0001	0.19	0.87	0.81

Table 4.3: Experiment 1: Results include model performance for a global model (a model trained on the entire dataset), local models (models trained on only data available for each participant), and an FL model trained on all participant data. The results demonstrate the superior performance of the FL model.

The global model, trained with the entire dataset, performs as expected. The classifier performs with a high accuracy of 99.95%. However, due to the large misbalance in the dataset, using accuracy is not the best metric. In this case, perhaps the best metric is the F1 score. Here, the global model has a high score as compared to the local models, but there are clear issues with model precision.

All five local models are trained on only a portion of the dataset. In this case, each participant had access to 1/5 of the dataset with a class distribution similar to the original dataset. This simulates individual financial institutions with limited datasets. The results of the local models indicate good performance but with metrics lower than the global model. This is to be expected as the local models are trained on less data, with each financial institution

having access to less data than the global model. Precision is once again a concern with these models.

The FL model is trained iteratively for 10 rounds utilizing all five clients. This model performs well, demonstrating its strength in this domain. The FL model has the highest accuracy, and F1 score, with the F1 score improving by 6% over the global model. Further, the False Positive Rate (FPR) has decreased to below 0.0001, indicating improvements in reducing FP alerts. This demonstrates the strength of the FL model in prioritizing a reduction of FP alerts, while still performing well. A noted improvement in precision is observed, along with a decrease in recall. These differences indicate strength in the FL model in more balanced prediction compared to the global and local models.

However, a potential drawback is an increased False Negative Rate (FNR). The FNR measures the proportion of actual positive fraudulent cases that are incorrectly classified (identified as negative). In the FL model, an increased FNR is observed as compared to the local and global models, indicating that the FL model may misclassify truly fraudulent data points.

This experiment demonstrates that FL can be leveraged to improve model performance. Both the global and federated models notably outperformed the local models as expected, and the federated model outperforms the global model. While there was a performance trade-off between precision and recall, the federated model had a superior F1 score. Further, the federated model preserved client privacy lowering the barriers to entry for different competitors.

Second-Layer System

Having demonstrated that an FL system can improve upon existing AML by way of improving access to data, the next set of experiments were designed to explore alternative methods of AML detection. In particular, the second and third experiments examine FP alerts. One

of the challenges in AML is that the class imbalance results in models that generate a large amount of FP alerts, causing financial institutions to dedicate resources to check each alert for validity. To combat this, Experiment 2 and Experiment 3 explores utilizing FL as a second layer to existing detection systems. By taking only those transactions flagged as fraud by existing detection systems, FL is leveraged to measure the improvement in identifying positive alerts and reducing FP alerts.

Two ways to assess a reduction of FP alerts are considered. First, in Experiment 2, optimized local detection systems were developed utilizing five commonly used ML methods. The positive predictions of these models were entered into the federation to train an FL model to refine the prediction and reduce the FPR. Next, in Experiment 3, local detection systems were purposely perturbed to prioritize a high recall value. This ensures that the local models miss very few positive predictions, but results in high FPRs.

Experiment 2: Optimized Local Detection Systems

For this experiment, five optimized local detection systems were simulated. To simulate the detection systems, the dataset was split into five distinct datasets, representing separate financial institutions. Customers do not overlap between datasets. Fraud detection systems were trained using five highly utilized supervised algorithms in the domain, Decision Trees [325], Random Forest Decision Trees [50], XGBoost [70], Multilayer Perceptron (MLP) Classifier [276], and k-Nearest Neighbors (kNN) [271]. These algorithms have been applied previously for transaction fraud detection [172, 292]. All local models were trained utilizing scikit-learn [269].

This scenario analyzes optimized detection systems. These models are optimized with balanced prediction rates for each class (fraud and non-fraud). The results of individual detection systems can be found in Table 4.4. All models perform similarly, with an average accuracy of 0.9994, precision of 0.75, and recall of 0.81. These models perform well on the

dataset for both classes. As the goal of this experiment is to utilize FL as a second-layer system to reduce FPs, only the data points predicted as positive (true positives and FPs) were selected for inclusion in the federation. The goal of the experiment is to retrain the FL model on these samples to reduce the FPR.

As in the first experiment, the results of the FL model are compared to a global model and local models. In this case, the global model is trained on the entire subset of data collected by selecting only positive predictions from the five detection systems. The local models are trained using each financial institution's data, simulating a scenario wherein each institution decides to train its own second-layer system utilizing its own data. A FL model was trained utilizing the predictions from all five local models. The results of this experiment are displayed in Table 4.5. Once again, the results of the FL model improve upon the local models, providing a robust model for reducing FP counts. If both models are utilized together, the participants would benefit from a 75% reduction in FPs compared to if only the first detection system were used.

One significant point of note about this experiment is that the characteristics of the dataset changed in the second model. In the local fraud detection models, each financial institution had access to one-fifth of the data with class partitioning equal to the original dataset. However, as only the positive predictions were provided to the federation, this resulted in a sample with approximately 25% negative samples (valid transactions) and 75% positive

Model	Accuracy	F1	F0.5	FPR	FNR	Precision	Recall
DT	0.9994	0.8	0.77	0.0004	0.13	0.75	0.87
RF	0.9995	0.82	0.78	0.0004	0.12	0.76	0.88
XG	0.9993	0.78	0.72	0.0005	0.09	0.68	0.91
MLP	0.9994	0.75	0.79	0.0002	0.31	0.82	0.69
kNN	0.9993	0.74	0.74	0.0003	0.28	0.75	0.72
Average	0.9994	0.78	0.76	0.0004	0.19	0.75	0.81

Table 4.4: Experiment 2: Results of individual detection systems on the PaySim dataset.

Model	Accuracy	F1	F0.5	FPR	FNR	Precision	Recall
Global	0.9083	0.94	0.94	0.1885	0.06	0.94	0.94
Local 1	0.8561	0.91	0.89	0.3861	0.06	0.88	0.94
Local 2	0.8394	0.9	0.88	0.4027	0.08	0.87	0.92
Local 3	0.8411	0.9	0.88	0.4314	0.07	0.86	0.93
Local 4	0.8478	0.9	0.89	0.3292	0.09	0.89	0.91
Local 5	0.8072	0.89	0.85	0.5893	0.06	0.83	0.94
FL	0.8828	0.92	0.93	0.2135	0.09	0.93	0.91

Table 4.5: Experiment 2: Results of second-layer FL system.

samples (fraudulent transactions). This changed the classification problem, resulting in a model with a high FPR and lower accuracy. It is critical to note that while the global model slightly outperforms the federated model, it does so at the cost of sharing all of the data between each of the institutions something with both privacy and business concerns.

Further, this approach does have drawbacks. For example, while each fraud detection model is optimally trained, the average FNR is 0.19. This high FNR appears in the FL scenario for Experiment 1 as well. In either case, this is not ideal as fraudulent transactions will be missed.

Experiment 3: Local Detection Systems with High Recall

The previous experiments revealed concerns with the high FNR. However, the imbalance in classes in AML datasets, and models that are developed with a low FNR often results in a large number of FPs. This increases the burden on financial institutions to check each alert for truly fraudulent transactions. To explore this problem, another set of experiments are conducted to reduce both the FNR and the FPR.

In this experiment, the same logic as the previous experiment applies. This experiment also simulates utilizing FL as a second-layer detection system, but instead of optimizing the fraud detection models using traditional evaluation metrics, the models are trained with high recall for fraudulent transactions. This results in the local detection methods (Table

4.6) predicting positive cases with high accuracy, regardless of how many FPs it results in. Compared to the previous experiments, the local models predict FPs at 10 times the rate as previous.

The FL system was trained utilizing FPs and TPs making up 75% and 25% of samples respectively. Model training for this experiment (Table 4.7) once again improves upon local detection models if each financial institution were to develop a second layer system locally.

To truly measure the benefit of the second layer systems explored in both Experiment 2 and Experiment 3, the results of local detection systems and the FL model were combined. As the goal of these experiments was to reduce the FPs predicted by the local detection systems, the results of both experiments were recalculated to account for the reduction of FPs predicted utilizing both the local system and the second layer FL system. The final metrics after the second-layer system is applied to reduce FPs can be found in Table 4.8.

In particular, the second-layer system increased all performance metrics in both experiments. However, the first setting maintains an increased FNR, due to the original performance of the local detection models missing some positive predictions. However, the results of the second setting with high recall reveal both a low FPR and a low FNR. This im-

Model	Accuracy	F1	F0.5	FPR	FNR	Precision	Recall
1	0.9961	0.4	0.29	0.0038	0.01	0.25	0.99
2	0.9961	0.4	0.29	0.0039	0.01	0.25	0.99
3	0.996	0.4	0.29	0.004	0.01	0.25	0.99
4	0.9961	0.4	0.29	0.0038	0.01	0.25	0.99
5	0.9963	0.4	0.29	0.0036	0.02	0.25	0.98
Average	0.9962	0.4	0.29	0.0038	0.01	0.25	0.99

Table 4.6: Experiment 3: Results of individual detection systems trained to prioritize high positive prediction accuracy.

Model	Accuracy	F1	F0.5	FPR	FNR	Precision	Recall
Global	0.891	0.81	0.76	0.1209	0.07	0.72	0.92
Local 1	0.771	0.66	0.58	0.2694	0.11	0.52	0.89
Local 2	0.7563	0.65	0.56	0.291	0.11	0.51	0.89
Local 3	0.754	0.65	0.56	0.2899	0.12	0.51	0.88
Local 4	0.7583	0.65	0.56	0.2863	0.11	0.51	0.89
Local 5	0.785	0.68	0.59	0.2499	0.11	0.55	0.89
FL	0.9246	0.84	0.88	0.0266	0.22	0.91	0.78

Table 4.7: Experiment 3: Results of Second Layer FL System.

Model	Accuracy	F1	F0.5	FPR	FNR	Precision	Recall
First Layer - Experiment 2 (Average)	0.9994	0.78	0.76	0.0004	0.19	0.75	0.81
First Layer - Experiment 3 (Average)	0.9962	0.4	0.29	0.0038	0.01	0.25	0.99
Second Layer - Experiment 2	0.9996	0.85	0.87	0.0001	0.2	0.89	0.8
Second Layer - Experiment 3	0.9999	0.96	0.94	0.0001	0.01	0.93	0.99

Table 4.8: Results Experiment 2 & 3 After Utilizing Both Local Detection Systems and FL Model. The first layer model results from Table 4.6 and Table 4.4 are repeated here for comparison.

proves upon the results of the prediction of fraud greatly, producing a prediction accuracy of 99.99% and F1 of 0.96. Additionally, the model results are better than those in Experiment 1, where the entire dataset was provided to the federation.

Experiment 4

In the previous experiments, it was demonstrated that FL can be leveraged to improve model performance. It was assumed that each institution had access to an equal amount of data, and benefited from the federation equally via an improvement in model outcomes. However, it is not always the case that participants will have the same amount of data, or will benefit from the federation in the same way.

This experiment aims to understand how participants contribute to the federation, and how variable contributions relate to the benefit of being a part of the federation. To simulate unequal contributions, the dataset is split into decreasing proportions of 50%, 25%, 12.5%, 6.25%, and 3.125%. For example, the first participant had access to 50% of the data, the

second participant to 25% of the data, and so on. The FL model is trained with 5 clients for 10 rounds.

The results of this experiment can be found in Table 4.9. From the local models, it is clear that the performance decreases as the proportion of data that the client has access to decreases. Once again, as the majority of the samples are negative, accuracy does not provide a useful metric, and rather the F1 score is a better metric for model performance. Clients 1 and 2 have the highest performance, with a model that performs with an accuracy of 99.94% with an F1 score of 0.96. These results are comparable to the global model, indicating that model performance can be achieved with a relatively low proportion of the dataset. Comparatively, Client 5, with access to only 3.125% of the data, results in a model with F1 of 0.76, a notable decrease.

Similar to the previous experiments, the FL model outperforms all others with an accuracy of 99.97% and F1 score of 0.86. Once again, there is a trade-off between precision and recall, with the FL model resulting in a very high precision of 0.98 and a recall of 0.77, significantly reducing FPs as a consequence of a slight increase in FNs.

Regarding the benefit of client participation, the results demonstrate that in all cases, participants see an increase in model performance. However, the benefit for clients with a larger proportion of the data is markedly less than for clients with a smaller proportion of the data. For example, Client 1 benefits from the federation with a 0.03% increase in accuracy and a 6.17% increase in F1 score. However, Client 5 benefits with a 0.04% increase in accuracy and a 13.16% increase in F1 score.

Summary

In this section, the use of FL in combating money laundering was explored. It was demonstrated that FL can train a highly accurate and precise model, on par with a model trained locally with all data. FL outperformed local models trained on only data available to each

Model	Proportion of Data	Accuracy	F1	F0.5	FPR	FNR	Precision	Recall
Global	100%	0.9995	0.83	0.78	0.0004	0.06	0.74	0.94
Local 1	50%	0.9994	0.81	0.76	0.0005	0.07	0.73	0.93
Local 2	25%	0.9994	0.81	0.76	0.0004	0.09	0.73	0.91
Local 3	12.5%	0.9994	0.79	0.74	0.0005	0.1	0.71	0.9
Local 4	6.25%	0.9993	0.77	0.74	0.0004	0.16	0.72	0.84
Local 5	3.125%	0.9993	0.76	0.69	0.0006	0.1	0.66	0.9
FL	100%	0.9997	0.86	0.93	0.0001	0.23	0.98	0.77

Table 4.9: Results Experiment 4: Measuring the impact of unequal client participation

participant.

Four experiments were conducted to assess the benefit of utilizing FL in AML settings. In Experiment 1, FL was demonstrated to increase the performance of the model by 6%, demonstrating the strength of utilizing FL for AML problems. Experiments 2 and 3 explored applying FL to only positive predictions, attempting to reduce the FPR. The results of these two experiments indicate that it is possible to decrease FPRs using a second-layer FL approach, resulting in a final model that predicts all cases with 99.99% accuracy and an F1 score of 0.96. Experiment 4 aims to assess client participation, revealing that clients with variable access to data benefit from participation differently. In this case, it was demonstrated that all clients benefited from participation, but that participants with a larger proportion of the data benefited less than those with access to a very small proportion of the data.

In all cases, a notable trade-off between precision and recall was observed between the local models and FL models. This is likely due to the implementation of the FedAvg and XGBoost aggregation mechanism architecture in Flower in conjunction with the class imbalance in the dataset. FL is known to be sensitive to class imbalance, and the inherent averaging process during aggregation can result in a bias toward the majority class [97]. Developers who implement federated algorithms should consider this carefully when implementing FL networks.

Further, it should be noted that only iid data was explored. In the field of AML, the nature and structure of financial transactions and related data tend to exhibit patterns and characteristics that are inherently more uniform, aligning more closely with the assumptions of iid data. This uniformity is caused by regulated financial behaviors, standardized transaction protocols, and compliance frameworks. Therefore, non-iid data is not typical in the AML domain, and was thus not explored here.

Overall, this use case shows the strengths of utilizing FL in AML. The ability to collaborate between financial institutions provides opportunities for better-performing models. The improved FL model has the potential to reduce the amount of FP alerts, thus reducing the burden on financial institutions to check each for legitimacy. By utilizing the FL system as a second-layer detection targeting the reduction of FPs, the performance on AML tasks is improved greatly, showing great potential for leveraging FL for this domain.

Further, the assessment of client participation reveals that even when clients have access to the majority of the data, they still benefit from participation in the FL network. However, the benefit of participation is variable and therefore there may be questions about whether participation in the network is worth it for a small gain in performance. While not discussed in this dissertation, incentivization strategies can be utilized to encourage client participation [379]. Further, while participation while be beneficial to performance, the issue of client participation is more nuanced and may be impacted by other features such as time and cost sensitivity [42].

4.4 Challenges in Federated Learning

While it offers significant advantages in terms of privacy and data decentralization, FL also faces several challenges that have been shown to impact its effectiveness and widespread adoption.

Several of these challenges occur during communication between clients and the server. In particular, communication overhead is a significant issue in FL systems due to both the need to account for the size of the updates and the differences between the system capacities of each client. Researchers have addressed concerns over overhead due to communication costs with several strategies including local updating, compression schemes, decentralized training, and importance-based updating [281]. For example, local updating schemes pass communication costs onto the client, such as using mini-batch optimization [291] or distributed local-updating methods [390]. Compression methods reduce model update size sent to the server each round, utilizing such methods as subsampling, probabilistic quantization, sparsification, dropout, and lossy compression [59, 135, 173]. Other examples include Communication-Mitigated Federated Learning (CMFL), which prevents unnecessary updates from being sent to the aggregator by enabling clients to measure their contributions to each update and disregarding irrelevant updates [224].

An additional challenge is in systems heterogeneity, where the capabilities between clients differ in factors such as network connectivity, memory, CPU, or power level [366]. A popular approach to mitigate this issue is to select participants based on system resources available at each round. For example, Nishio and Yonetani [254] propose FedCS, a method to increase aggregation efficiency and reduce training time by aggregating only clients with sufficient resources. Other methods include Federated Dropout [59], priority-aware aggregation [10], dynamic compression of straggling clients [364], server-side algorithms that ensure efficient aggregation [12], and fault tolerance [322].

Data heterogeneity between clients is also a challenge. In many federated settings, the data collected from participants is typically non-identically distributed [281]. In addition, there is variation in the amount of data clients have and/or contribute, with some participants potentially having a lot of data and some having very little. Data heterogeneity and volume can cause issues with model performance, generalizability, and participation. Arivazhagan

et al. [15] address issues with statistical heterogeneity by adding a personalization layer to the global model, reducing performance issues, and increasing generalizability. Further, participation can be an issue when participants are reluctant to participate in a network if there is not a significant benefit. To this end, incentive mechanisms have been proposed to incentivize clients to participate [381].

Privacy and security of FL networks are also of significant concern. As participants update the model during each round, they transmit relevant information about their contribution to the server. These updates require careful attention to ensure that the information remains truly local and that no information is leaked or otherwise available to others. A significant advantage of FL is the privacy protections afforded by ensuring that model training occurs without sharing data and that model updates are secure. However, researchers have demonstrated that sensitive information can be exposed during communication rounds [403]. To address privacy concerns, methods integrating differential privacy, homomorphic encryption, and Secure Aggregation have been proposed. Differential privacy methods add noise to the data or model updates, making it difficult to expose participant information [102]. Homomorphic encryption secures model updates via a cryptographic method that allows computations to be carried out without decryption, generating encrypted results that can then be securely decrypted by the server without revealing participant information [21]. Secure Aggregation is a specific aggregation method where participant updates are aggregated in a combined way, ensuring that the server can only see combined updates and not individual contributions [47]. All of these methods aim to enhance the privacy of FL systems and prevent the exposure of participant information.

However, even with these methods in place, researchers have revealed that FL has vulnerabilities during the model update process, and is not always sufficiently secure from malicious actors. For example, researchers have demonstrated weaknesses to data poisoning [91, 258, 342], model poisoning[22, 40, 106], inference attacks such as membership and

training input inference [145, 237, 277, 396, 403], and other types of attacks [225, 248]. For example, Zhu et al. [404] demonstrates the vulnerability of FL toward free-rider attacks, where adversaries submit fake updates to participate in the network without providing real information, showing that their free-rider attack where an attacker can stealthily construct counterfeit updates and evade existing defense mechanisms. To prevent these types of attacks, many researchers have proposed Byzantine-tolerant aggregation mechanisms [45, 274, 332]. These types of attacks and aggregation mechanisms will be further explored in Chapter 5.

4.5 Implications on Trust

While the growth in popularity of FL is primarily due to the privacy affordances it provides, FL offers significant benefits toward building trust in AI. Leveraging FL can increase trust via increased data and participant privacy, improving fairness and reducing bias by gathering data from diverse and heterogeneous resources, enabling transparency in the robustness and accountability of systems, and fostering secure systems.

In this chapter, the privacy benefits of FL have been highlighted. This approach enables collaborative model training while keeping sensitive data localized and secure, mitigating risks posed by data leaks and centralized data storage breaches [263]. By allowing data to remain locally and only sharing model updates, FL provides a robust privacy-preserving framework, greatly increasing privacy in AI systems [202]. Of course, trust can only be maintained if the proper privacy-preserving mechanisms are in place [374].

Improving fairness is also made possible by FL, as the AI models can be trained by more diverse and heterogeneous datasets. For example, FL may be able to reduce bias with heterogeneous and diverse datasets from clients in different areas, locations, or domains [344]. Indeed, Zhang et al. [385] propose FairFL, a framework for reducing discriminatory bias utilizing FL. Likewise, Ezzeldin et al. [105] propose Fairfed to reduce bias and enhance group

fairness and Djebrouni et al. [90] propose ASTRAL for fair weight aggregation, While this same heterogeneity may introduce bias, several methods have been developed to improve fairness in FL systems with very skewed data distributions [3].

FL also provides opportunities to increase transparency and accountability. Toward increasing accountability in FL, IBM Research introduced the Accountable FL FactSheet framework [25]. This framework addresses the accountability of the parties involved in federated networks, providing opportunities to clearly demonstrate accountability in a federated setting and enable auditing, transparency, and fact-checking. Further, FL networks can be audited and kept accountable by several mechanisms, to ensure transparency and accountability of both participants and the overarching global model [20, 86, 246]. Further, while FL is not explainable by design, several methods have been proposed to increase the explainability of FL predictions [69, 112, 288, 367].

Notably, several methods have been proposed for the inherent trustworthy design of FL systems. Cao et al. [63] propose FLTrust, a method for FL trust bootstrapping by assigning trust scores to model updates based on their contributions. Rehman et al. [286] propose TrustFed, a method to detect malicious actors and attacks, enable fair training settings, and monitor participant behavior. Bao et al. [24] propose FLChain, a FL marketplace for incentivized and trusted networks and learners. Papadopoulos et al. [267] propose Verified Credentials as a method to increase trust in the users of an FL system, ensuring that only reputable clients may participate. Further, several researchers also propose reputation mechanisms to encourage trust and honesty from participants by providing them with reputation scores [132, 169]. Blockchain has also been proposed as a mechanism to increase trust in FL [24, 251, 286] All of these methods have enhanced the trustworthiness that FL provides.

It is clear that there is a relationship between FL and trust. In this dissertation, this relation-

ship will be further explored to identify the strengths and weaknesses of FL in developing Trustworthy AI. The use case with AML has already demonstrated that FL has a direct impact on privacy and robustness & reliability. In the next chapter, methods for improving safety & security in FL will be demonstrated. Following that, in Chapter 6, leveraging FL for regulatory requirements is connected to the remaining trustworthy principles.

Chapter 5

Defending Federated Learning

The previous chapter demonstrated the potential of FL to increase trust in AI. Namely, the benefits of FL on privacy and robustness & reliability were highlighted. The benefits of FL toward privacy are significant, enabling participants to collaborate to train a shared ML model without sharing data.

This chapter aims to understand the potential for FL regarding the safety and security of AI systems from malicious attackers. It has been shown that FL mitigates risks of data leaks and breaches by keeping sensitive data localized and secure [263]. This greatly enhances the privacy in AI systems; however, the proper security measures must be in place to ensure these privacy benefits [374]. While FL has the potential to increase the security of AI systems, and thus increase trust, current research has revealed several flaws in the security of FL settings. The crux of FL lies in the fact that no single entity owns or verifies the training data that participants utilize to train model updates. In theory, this should prevent many types of malicious attacks and secure the federated system from malicious actors. However, many scholars have shown that FL is still vulnerable to adversarial attacks

[22, 40, 72, 119, 161, 357]. As FL allows an attacker to have access to the modeling process, attackers can leverage model poisoning in a federated environment to significantly impact the performance of a global model. One way this can be done is through the insertion of backdoors during the learning process, where the goal is to corrupt the global model to lead to a misclassification of a *specific* task, rather than affecting the performance of the entire model. Model poisoning greatly outperforms traditional data poisoning and is of great concern among FL researchers [161].

Along with this increase in concern about model poisoning has been an increase in research on methods to defend and harden FL systems against adversarial attacks through alternative aggregation mechanisms. Such aggregation methods are typically Byzantine-tolerant, ensuring convergence even in the presence of Byzantine participants, and acting as a defense mechanism against adversarial attacks. However, many of these mechanisms can be circumvented by sophisticated attacks [22, 106, 119, 332, 357]. As such, creating robust FL against model poisoning attacks is an open problem. While the majority of works focus on how attackers can circumvent specific defenses, there are no current works that address the performance of such defenses on model poisoning in general.

In this chapter, an analysis of the behavior of byzantine aggregation mechanisms against model poisoning in a FL setting is provided. In particular, the performance of popular defenses such as Krum, Multi-Krum [45], Norm-Difference Clipping [332], and Robust Federated Averaging (RFA) [274] are analyzed, and model poisoning is conducted within FL environments under various adversarial settings. These defenses are chosen due to their applicability and strength in defending FL systems. This is demonstrated using two concrete learning tasks commonly used in the domain: image classification on the CIFAR-10 dataset and digit classification on the EMNIST dataset, replicating the learning environments in [357] with a basic model poisoning attack scenario.

This chapter aims to understand how FL can be secured against adversarial attacks. Security against such attacks is a significant element of trust, as data breaches and manipulation of AI systems can result in privacy concerns via personal data loss, and potential safety concerns if the system performs unexpectedly. While inherently FL has security benefits, assessing the security of FL in this way enables a critical evaluation of its trustworthiness concerning several key principles.

5.1 Background

5.1.1 Byzantine-Tolerant Aggregation

The basic aggregation mechanisms utilized in FL were discussed in Chapter 4. These aggregation mechanisms, while robust for performance and convergence, are vulnerable to adversarial attacks such as data and modeling poisoning [22]. The most basic aggregation mechanisms work through averaging local model parameters but rely on the assumption that all participants are honest. An attacker can take advantage of simple aggregation mechanisms to compromise worker devices [45, 373], or model updates [22, 40, 106], compromising the global model for all participants.

Recent work has focused on the development of Byzantine-Tolerant aggregation mechanisms, where the goal is to ensure convergence in the presence of Byzantine participants [45, 72, 332, 373]. These mechanisms are specifically designed to ensure the robustness and reliability of the FL model in the presence of malicious participants. These malicious updates can degrade or manipulate the performance of the global model. Instead of utilizing averaging for model aggregation, these approaches use alternative aggregation approaches such as geometric median or trimmed mean that are less sensitive to extreme or perturbed values. These alternative aggregation mechanisms make them less vulnerable to malicious updates, allowing the FL network to prevent the malicious actor from degrading or manipulating the model.

However, many of these byzantine-robust methods assume that the attacker intends to prevent *convergence* of the model, which is not the case in a backdoor attack scenario. In a backdoor attack scenario, the adversary's goal is to manipulate the model such that the model performs its given task normally while behaving maliciously on a specific, attacker-chosen task [205]. Byzantine-tolerant aggregation mechanisms, while secure against attacks that prevent or slow convergence, have been demonstrated to be weak to backdoor attacks in FL [22].

In this chapter, several byzantine-tolerant aggregation mechanisms are explored for their robustness against adversary backdoor attacks. Namely, Krum, Multi-Krum, Norm-Difference Clipping, and RFA are explored for their ability to defend against adversarial attacks in various model poisoning attacks.

Krum & Multi-Krum

Krum and Multi-Krum are alternative byzantine aggregation methods that intend to tolerate Byzantine participants in a distributed setting by selecting fewer models for aggregation, attempting to exclude malicious participants [45]. In Krum, only one of the participants' local models is chosen to be used as the global model. It is designed to tolerate c compromised participants out of n. For each round, the pairwise distances between all local models submitted are computed. Then, the server sums up the $n - c - 2$ closest distances, and the model with the lowest sum is chosen as the global model for the next round. This process continues for each round, selecting one local model that is geometrically closest to all other local updates as the update for the global model. Multi-Krum is a variation of Krum where instead of one model being chosen, the top $m = n - c$ models are chosen to be averaged into a new global model. This extension to Krum allows for a selection of multiple updates, instead of limiting the updates to just one. This is particularly advantageous in environments with a larger subset of participants, leading to improved performance.

Norm-Difference Clipping

This method relies on the theory that malicious models are likely to produce large norms and that a simple clipping defense could thwart attackers [332]. Norm-difference clipping works by examining the norm-difference of local models submitted to the server, as compared to the current global model, and clipping model updates that have a norm difference larger than threshold M. This method modifies participant updates before aggregation to ensure that no single update disproportionately affects the aggregated result. In this way, the contribution of any model with a large norm difference is small and therefore poisoned models are in theory less influential to the global model.

RFA

Robust Federated Averaging (RFA) replaces the typical approach of aggregating utilizing the arithmetic mean with a modified method to compute a weighted geometric median using the *smoothed Weiszfeld's algorithm* [274]. This aggregation mechanism leverages the geometric median to aggregate model updates, minimizing the influence of malicious or outlying participant updates. By utilizing this mechanism, the aggregated model update is less vulnerable to skew by malicious contributions, enhancing the reliability and security of the global model. The authors demonstrate that this method is robust from federations with up to half of the participants being corrupted.

5.1.2 Related Work

Traditional poisoning attacks focus on altering model behavior at test time through poisoning of the data used to train models [149]. Such attacks include *data poisoning*, where a user's training data is compromised to change the model behavior on a specific task [71], or through the insertion of a backdoor directly into the model to compromise it [98]. However, it has been demonstrated that these attacks are not effective in FL, where defense and privacy-preservation methods prevent compromise and attacker models are aggregated among thousands of participants, limiting the impact that a single attacker can have on the

global model [22].

Nonetheless, the presence of these protective mechanisms does not preclude FL from attacks. Many recent works have discovered methods to insert backdoors in FL using *model poisoning*. In FL, this is conducted with the aim of causing the global model to misclassify a set of chosen inputs while maintaining high accuracy in the original classification tasks. The first of such works demonstrated that model poisoning was effective in a FL system, utilizing a novel method to allow the attacker to send back *any* model they want to be aggregated into the global model, known as *model replacement* [22]. Similarly, Bhagoji et al. proposed a modification that leveraged boosting to increase the learning rate of the backdoor inputs [40]. Further, Wang et al. proposed a method of inserting edge-case backdoors, further demonstrating that the FL settings are vulnerable to both model poisoning and model replacement attacks [357].

5.2 Threat Model

For these experiments, a number of assumptions are made.

It is assumed that the attacker has control over the local training process and system of one random participant, including training data, hyperparameters, and training process. Attackers are assumed to be singular entities and it is assumed that none are working toward a common goal with other participants. In this setting, only the attacker is behaving maliciously and all other participants are behaving honestly and correctly. In all experiments, the scenario is limited to having no more than one attacker per round. The attacker does not have access to the training data of other participants, nor does it know their identities. The attacker does not have control of the server and does not control the defense mechanism utilized to aggregate local models into a new global model each round.

5.2.1 Backdoor Attacks

For consistency with previous work in the domain, this threat model is inspired by existing literature [22, 332, 357]. Here, only backdoor attacks are considered, where an attacker aims to manipulate the performance of a model on a particular subtask (hereby called the 'attack task') while maintaining high accuracy on the model's intended tasks (hereby called the 'main tasks'). The main goal of the attacker is to manipulate the FL system to produce a global model that performs with high accuracy on the model's intended tasks as well as an attacker-chosen subtask. For example, a given model's intended task may be to correctly classify pictures of animals or numbers. In these scenarios, the attack task may be to classify pictures of cats as birds, or the number '6' as the number '2', without impacting the model's performance on its original tasks. By maintaining high accuracy on the model's main tasks, it is more likely that the attack task will go unnoticed.

5.2.2 Model Replacement

Attacks *with* and *without* model replacement are considered. In scenarios *without* model replacement, the attacker trains the current global model with their data to achieve high accuracy in both the main tasks and the chosen attack task. The poisoned model is submitted to the server and aggregated into the global model, according to the associated aggregation method.

Alternatively, in model replacement scenarios, the attacker aims to replace the global model with any model of their choice. Model replacement occurs in conjunction with the backdoor attack. Generally, this can be achieved through a weight re-scaling method, where the attacker re-scales the weights of the global model to resubmit as an adversarial model along with their goals. For all experiments, the weights are scaled using the constrain-and-scale technique developed by Bagdasaryan et al. [22] This approach typically requires that the attacker has knowledge about the current global model and the federated environment, and

requires model convergence.

5.3 Experiments

5.3.1 Experimental Setup

The simulated FL environment is modeled after [233]. The setup consists of K clients, each with access to data. This data is not shared with the server S. For each FL round t, the server randomly selects a subset of clients k and provides the current global model to each. Participants conduct local model training separately and compute a model update. Each participant sends back updated model weights to the server for aggregation.

Experiments are conducted on the effectiveness and limits of Byzantine-tolerant aggregation mechanisms in preventing attacks by adversaries in a federated environment. Four different aggregation mechanisms are considered (Krum, Multi-Krum, RFA, and Norm-Difference Clipping) and compared to a setting where no defense and the standard aggregation method is used, Federated Averaging [204].

Several key things are explored, including (1) the impact of the frequency of adversarial attacks, and In particular, fixed-attack frequencies were explored by altering the attack rate. The following settings were explored: one attack per round (i.e. an attack every round), one attack every 5 rounds, and one attack every 10 rounds. In all settings, only one random

Experiment	Scenario 1 CIFAR-10	Scenario 2 EMNIST
Model	VGG-9	LeNet
Data Points	50,000	341, 873
Classes	10	10
Clients K	3,383	200
Clients per Round k	10	10
Epochs E	2	5
Learning Rate	0.2	0.1

Table 5.1: Parameters for experimental set up including datasets and model type used.

client is the attacker.

As a baseline, a comparison to a setting with no adversaries is provided. In these settings, hereby called a 'no attack' scenario, there is no attack. However, the attack tasks are still measured to provide an analysis of the behavior of the attack task. This serves to ensure that the attack task does not naturally increase in accuracy through normal model training. For comparison purposes, the experiments utilize the same data and experimental setups utilized by previous work in the domain [22, 332, 357], and the values of all hyperparameters can be found in Table 5.1. For all experiments, the subset of clients k is set to 10 (i.e. 10 clients participate per round), and the number of federated rounds t is set to 500. All experiments are implemented in PyTorch [268]. Experiments were run on a server with two NVidia Tesla K80 GPUs and 132 GB of RAM.

5.3.2 Datasets & Learning Models

As the goal of this chapter is limited to analyzing the defense characteristics of aggregation mechanisms and not to introducing novel datasets or poisoning attacks, only poisoned datasets used previously in the literature were used [40, 332, 357].

Experiment Scenario 1 focuses on image classification using the CIFAR-10 dataset [187]. The experimental setup in [357] is replicated, where photos of Southwest Airlines planes are collected and poisoned to be labeled 'truck'. In total, there are 784 and 196 examples in the training and test sets. The VGG-9 model [320] is initialized with 77.53% accuracy. The model is initialized with a learning rate of 0.2 for two epochs.

Experiment Scenario 2 focuses on digit classification. In this experiment, the datasets include EMNIST [77] and ARDIS [188] datasets. The EMNIST dataset is an extended version of the MNIST handwritten character digit dataset and the ARDIS dataset includes 15,000 handwritten Swedish church records from the nineteenth and twentieth centuries.

For the non-malicious participants, there are 660 images used for training. For malicious participants, 66 images of the number '7' are labeled '1' and mixed with 100 randomly sampled images from the EMNIST dataset. For evaluation, 1000 images from the ARDIS dataset are used. The LeNet-5 architecture for image classification is utilized as in the PyTorch MNIST example [278]. The model is initialized with a model with 88% accuracy, with a learning rate of 0.1 for five epochs.

5.3.3 Experimental Fairness

In general, Byzantine-tolerant aggregation methods focus primarily upon ensuring the convergence of the model in the presence of adversaries. However, this does not directly imply that the aggregation method will be fair. Indeed, some aggregation methods have been found to negatively impact the main performance of the model [45, 357].

In this context, the algorithm is considered 'fair' if the success of the main task is left unhindered while the defense is deployed, and 'unfair' if the defense has a significant negative impact on the success of the algorithm's main tasks, regardless of whether or not the defense was successful at mitigating a potential attack. Further, a 'fair' model should accurately classify all tasks consistently, without misclassifying one or more tasks (i.e. if the algorithm classifies 1 task incorrectly consistently, it is not a fair algorithm).

To measure the impact of this fairness concern, **Accuracy Parity** (AP) ratio as formulated in [357] is used. This ratio measures the fairness of the model on each task. As formulated, AP ratio is calculated as $APratio = \frac{p_{min}}{p_{max}}$. A classifier satisfies AP if $p_i = p_j$ for all pairs i, j where p_i is the accuracy of class i. This metric would equal 1 if perfect parity exists (i.e. all classes are measured correctly), and 0 only if one or more classes are completely misclassified.

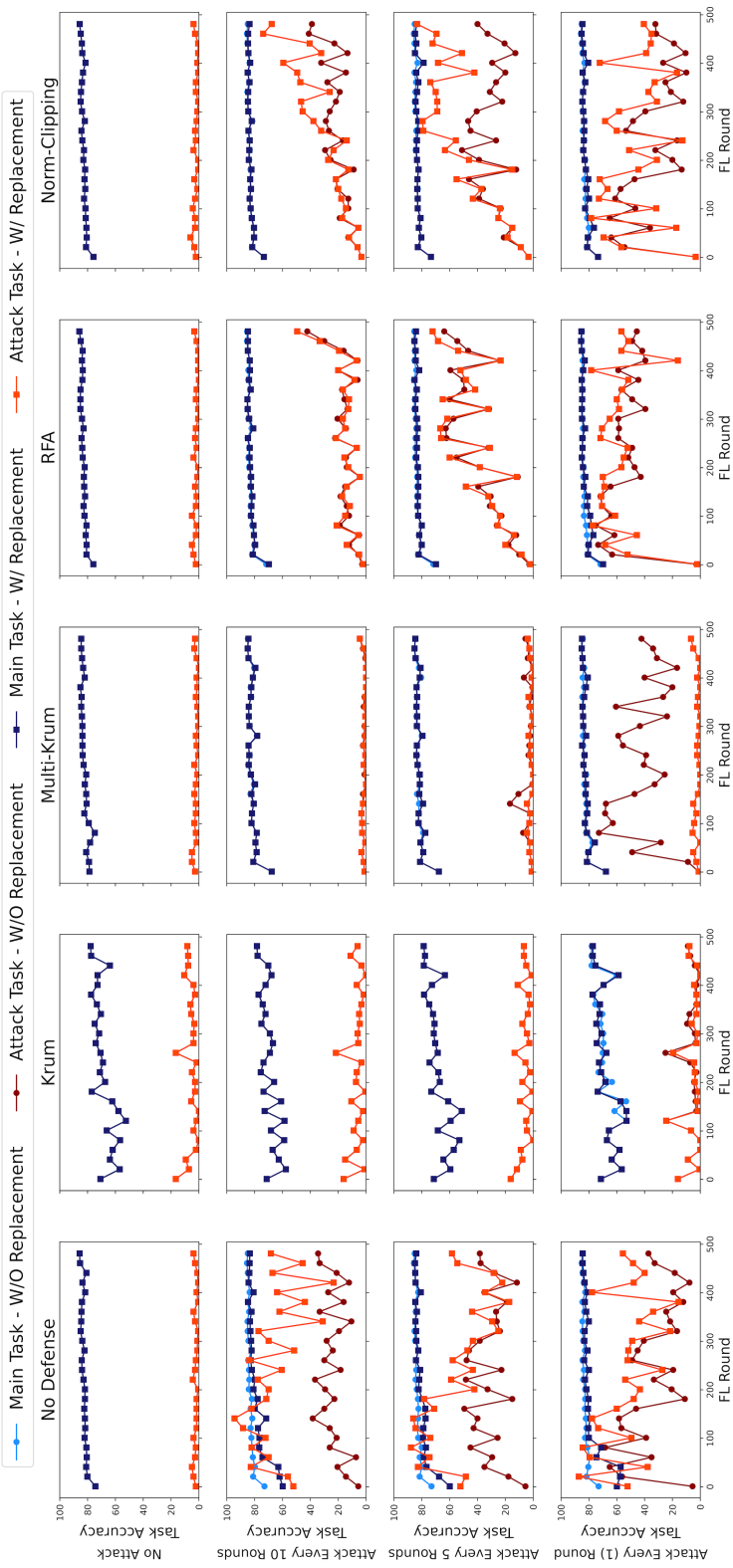

Figure 5.1: Experiment Scenario 1: Accuracy of model performance on the main and attack tasks under the four attack settings and five defense scenarios.

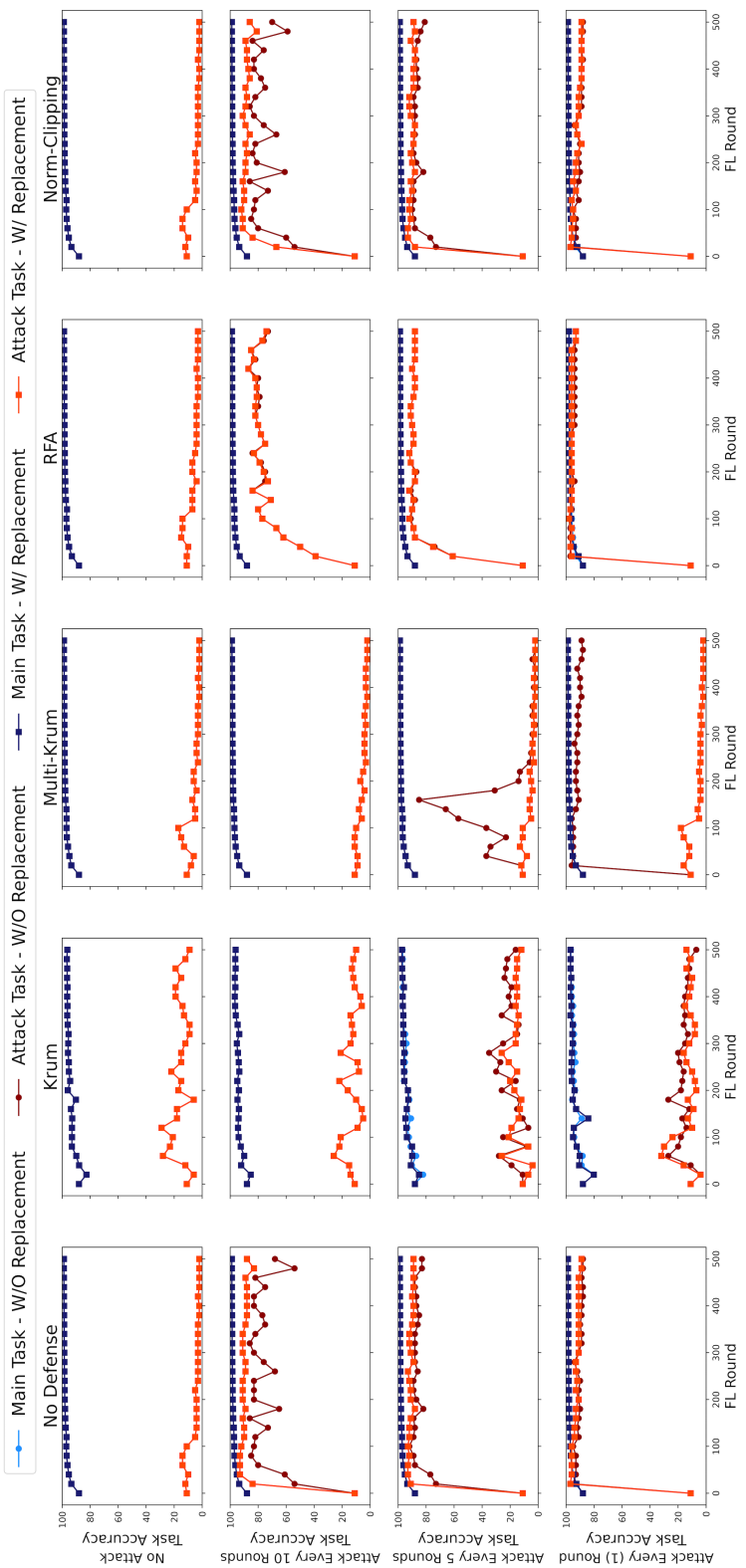

Figure 5.2: Experiment Scenario 2: Accuracy of model performance on the main and attack tasks under the four attack settings and five defense scenarios.

5.4 Experimental Results

The results of Experiment Scenario 1 and 2 are displayed in Figure 5.1 and 5.2. The accuracy rates of the attack task can be found in Tables 5.2 and 5.3.

In all cases, the main model is unaffected by the backdoor poisoning method, and an increase in the accuracy of the backdoor task is noted. This indicates that the poisoning method was successful in poisoning only a specific subtask and maintaining high accuracy on the model's intended tasks. For both datasets, there is no accuracy growth observed for the attack task in the 'no attack' scenario, indicating that the rise of the attack task is in fact due to the backdoor attack. Further, as expected, the frequency of the attack highly impacts its success, with more frequent attacks resulting in higher attack task accuracy. In regards to model replacement, in all cases, there is a clear delimitation between the effectiveness of defenses with and without model replacement. The effectiveness of mitigation of model replacement by each defense method is detailed in the following sections.

		No Defense	Krum	Multi-Krum	RFA	Norm-Difference
No Attack	minimum	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
	maximum	11.1 (11.1)	43.9 (43.9)	15.6 (15.6)	10.6 (10.6)	10.0 (10)
	mean	2.2 (2.2)	6.4 (6.4)	2.1 (2.1)	2.1 (2.1)	2.0 (2)
Attack Every 10 Rounds	minimum	0.0 (2.2)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
	maximum	44.4 (94.4)	45.6 (45.6)	17.8 (11.7)	42.2 (49.4)	43.9 (73.9)
	mean	11.4 (33.3)	6.2 (6.2)	1.8 (1.6)	8.2 (8.5)	10.0 (13.4)
Attack Every 5 Rounds	minimum	0.0 (1.1)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
	maximum	58.9 (92.8)	55.0 (55)	39.4 (12.8)	70.6 (75.6)	61.1 (83.3)
	mean	20.7 (39.7)	6.0 (6)	3.0 (2.2)	23.7 (25.7)	18.7 (28.2)
Attack Every (1) Round	minimum	1.1 (2.2)	0.0 (0)	0.0 (0)	2.8 (1.1)	1.7 (1.1)
	maximum	74.4 (93.3)	53.9 (53.9)	75.0 (14.4)	77.8 (85)	71.7 (86.7)
	mean	34.8 (53.6)	6.8 (5.6)	41.4 (2.1)	52.2 (57)	35.0 (47.9)

Table 5.2: Experiment Scenario 1: Without Model Replacement (With Model Replacement). Attack task accuracy percentages for all scenarios.

		No Defense	Krum	Multi-Krum	RFA	Norm-Difference
No Attack	minimum	2.0 (2)	3.0 (3)	1.0 (1)	2.0 (2)	2.0 (2)
	maximum	17.0 (17)	42.0 (42)	23.0 (23)	20.0 (20)	17.0 (17)
	mean	5.1 (5.1)	14.8 (14.8)	5.4 (5.4)	6.4 (6.4)	5.1 (5.1)
Attack Every 10 Rounds	minimum	11.0 (11)	2.0 (2)	2.0 (2)	11.0 (11)	11.0 (11)
	maximum	93.0 (100)	40.0 (40)	23.0 (23)	93.0 (93)	93.0 (99)
	mean	81.9 (90.8)	12.8 (12.8)	5.4 (5.4)	79.0 (79.5)	81.6 (89.2)
Attack Every 5 Rounds	minimum	11.0 (11)	3.0 (3)	2.0 (2)	11.0 (11)	11.0 (11)
	maximum	95.0 (100)	54.0 (50)	96.0 (18)	95.0 (96)	95.0 (99)
	mean	87.5 (91.8)	20.6 (15.7)	20.3 (5.4)	87.9 (88.2)	87.6 (90.8)
Attack Every (1) Round	minimum	11.0 (11)	4.0 (4)	9.0 (2)	11.0 (11)	11.0 (11)
	maximum	97.0 (100)	55.0 (62)	97.0 (19)	98.0 (98)	97.0 (99)
	mean	90.8 (92.6)	16.6 (13.8)	91.3 (5.6)	94.9 (95.6)	90.9 (92.3)

Table 5.3: Experiment Scenario 2: Without Model Replacement (With Model Replacement). Attack task accuracy percentages for all scenarios.

5.4.1 Aggregation Mechanisms & Defenses

No Defense

For comparison purposes, a 'no defense' setting was utilized. In scenarios where the standard aggregation method is used (Federated Averaging), it is considered an undefended federation due to the inability to prevent poisoning attacks. This is considered a scenario where there is 'no defense' for an adversarial attack, as this aggregation method simply averages the contributions of all participants, including the malicious participant.

In both cases, where there is no defense and no attack, the attack task maintains low accuracy while the main task maintains a stable, high accuracy rate. This is expected behavior and indicates that the model is not poisoned at the start and that it improves over time through iterations. However, in each case where an attack is observed (every 10 rounds, 5 rounds, and each round), an increase in the success of the attack task is observed, with model replacement typically resulting in higher success rates. An increase in the success of the attack task is observed in both scenarios, with more frequent attacks typically resulting in higher attack task accuracy rates.

Overall, these results indicate that 1) the attack task is successful in both cases, and 2) the

experimental setup is robust enough to measure the success of the defenses in a basic model poisoning scenario. Where there is no attack, the consistent observation of low attack task accuracy indicates that this setup is robust enough to measure the impact of an attack on both main and attack task accuracy. Further, this provides a benchmark of comparison for the effectiveness of the byzantine defenses on decreasing attack success.

Krum

Overall, the Krum defense is successful in defending against the attack task in every case. Even in the most aggressive case, it protects against the attack task, maintaining a low accuracy (below 40% in all cases), with less than 20% accuracy observed after 500 rounds. This is observed in both cases with and without model replacement. However, the Krum method negatively impacts the performance of the model even where there is no attack. A notable decrease in the performance of the main tasks is observed in all cases. This is likely due to the protocol choosing only one local model to use as the global model, decreasing the information gained in each round. This issue will be discussed further in Section 5.4.2.

Multi-Krum

As an extension to Krum, Multi-Krum produces similar results. In all cases, Multi-Krum successfully defends against model replacement scenarios, where the attack task accuracy is kept below 20% throughout all 500 rounds. In scenarios without model replacement, Multi-Krum fails in three cases.

In the first scenario, Multi-Krum can defend against attacks successfully up to an attack every round. When an attack is observed every round without model replacement, the accuracy of the attack task oscillates throughout the 500 rounds, with a minimum accuracy of 0% and maximum accuracy of 75%. The defense is overall not effective, as a steady increase is observed in the attack task accuracy to 47.2% at 500 rounds, with a mean accuracy across all rounds of 41.36%. A similar trend is observed in scenario 2 in regards to protecting from attacks with low frequency and failing to defend where an attack is conducted each round.

However, with a frequency of five attacks per round, the accuracy of the attack task rapidly increases to nearly 40% by round 200, where it appears Multi-Krum detected and eliminated the attack within the provided 500 rounds.

RFA

RFA is not successful in completely mitigating attacks in any case but does indicate some effectiveness in protecting against model replacement attacks. In all experiments without model replacement, RFA does not succeed in decreasing the effectiveness of the attack, often actually increasing the overall accuracy of the attack task. It appears that this defense is particularly weak to aggressive (frequent) attacks, where the success of the attack task increases even more aggressively than observed in the no-defense scenario.

However, RFA does show moderate success in the case of model replacement. At first glance, the success of RFA appears consistent between replacement and non-replacement scenarios, as nearly equal attack task accuracy levels in all sets of experiments were observed. However, as model replacement is generally deemed more aggressive, this equality indicates that RFA is more robust against replacement attacks. Indeed, RFA greatly decreases the success of the attack in model replacement scenarios that without a defense were observed to excel immediately. For example, in scenario 1 the mean attack success decreased from 33.3% and 39.7% with no defense to 8.5% and 25.6% with RFA, for attacks every 10 and 5 rounds respectively. From these experiments, it appears that RFA does not aid in scenarios without replacement in decreasing attack task accuracy, often increasing it in more aggressive scenarios. Further, while RFA is not as successful as Krum and Multi-Krum, there is no impact of the method on the overall success of the main tasks.

Norm-Difference Clipping

The norm-difference clipping defense produces similar results as RFA. The most notable difference between the two is that norm-difference clipping does not exhibit the same behavior of increasing the effectiveness of the attack in any case.

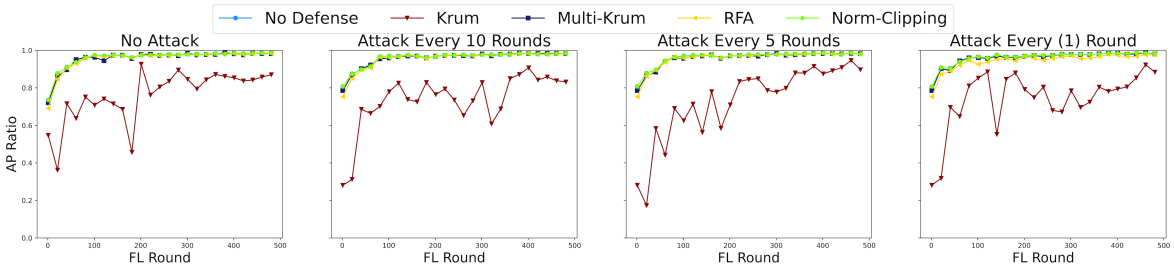

Figure 5.3: AP Ratio under each aggregation mechanism in Scenario 2.

Considering scenarios without model replacement, in all cases utilizing the norm-difference clipping defense, observed attack task accuracy rates are nearly identical to those observed in the no-defense scenario (Tables 5.2 and 5.3). However, this defense was successful in decreasing the success of attacks aided with model replacement, with varying degrees of success.

5.4.2 Defense Fairness

The scenarios focus on backdoor attacks, where the goal of the attacker is to increase the accuracy of a poisoned sub-task while maintaining high accuracy on the models' main tasks. In all cases, this assumption is confirmed. However, there are some fairness concerns with respect to defenses impacting the accuracy of the main tasks, even when there is no attack. Utilizing AP Ratio, the fairness of the algorithm in correctly classifying the intended input can be observed (Figure 5.3). Here, only the results of Scenario 2 are shown, however, these results are similar to those observed in Scenario 1.

In all cases, the AP ratios of no defense, Multi-Krum, RFA, and Norm-Difference Clipping are observed as higher than the AP ratios observed under Krum. This indicates a problem with the fairness of Krum. This can also be readily observed in the accuracy values of the main tasks. In the case of the first scenario, a significant decrease in the accuracy of the main tasks is observed. While the model should have an accuracy over 80%, the main

tasks are observed to have an accuracy below 70%, with an average accuracy of 69% in all cases. This effect is less pronounced but still present in the second scenario, where the main model accuracy drops to 93% on average. While the Krum defense is the most effective at mitigating attacks, it is the only defense that produces this behavior.

In an environment where the accuracy of the model is already low, such as scenario 1, this decrease in the success of the model on its intended tasks could have significant detrimental impacts on the model's performance and cause it to be completely ineffective for legitimate use. Although it is a variation of Krum, Multi-Krum does not exhibit this flaw. This is likely because choosing more than one model allows for richness in the global model throughout rounds. Users should consider this fairness concern while utilizing Krum in regard to their specific use case.

5.5 Discussion & Conclusion

It has been established here, and in the literature, that model poisoning is a significant concern in a FL environment. An attacker can manipulate the global model to produce high accuracy on hidden tasks while maintaining appropriate behavior on main tasks, potentially exposing federated participants to manipulated models that produce an undesired result.

However, defending against these attacks is a difficult task. It has been demonstrated that current Byzantine defenses, such as Krum, Multi-Krum, RFA, and Norm-Difference Clipping, have inconsistent effectiveness in defending against backdoor attacks. The results of the experiments indicate that Krum is the most effective at mitigating attacks, followed by Multi-Krum, RFA, and Norm-Difference Clipping. All defenses perform better than a no-defense scenario, indicating success in protecting against backdoor attacks.

However, while Krum has the most success mitigating against malicious attackers in a back-

door attack scenario, it has a negative impact on the main model, calling into question the fairness of such a defense. Multi-Krum does not share this same fairness concern, indicating that it may be a better defense for all cases except those with the most aggressive attacks, where Multi-Krum may not damage the main model but could fail to prevent attacks.

Backdoor attacks and Byzantine-tolerant aggregation mechanisms in FL have significant implications on trust. While these aggregation mechanisms aim to protect FL from adversaries, backdoor attacks represent a sophisticated threat against the robustness of the global model. This chapter has highlighted that the simplistic aggregation mechanisms such as FedAvg cannot adequately prevent model poisoning attacks without affecting the main performance of the model. However, some aggregation mechanisms can counterbalance this security threat, up to a certain threshold of malicious participants. While these mechanisms show promise in securing a FL system, additional work is needed to harden FL against attacks, especially where the global model is used in sensitive areas, such as health care. Alternative aggregation methods may provide more thorough protection without hindering the success of the model.

This chapter has explored the concept of security in FL systems, revealing some strengths and flaws in current aggregation mechanisms. The adversarial setting explored in this chapter demonstrates FL's weakness to model poisoning backdoor attacks. However, it is important to consider that this type of attack is only one of many types of attacks that can be utilized against FL. Here, it has been shown that FL can be protected against model poisoning utilizing the Krum defense methods. In other scenarios, solutions have been proposed to mitigate data poisoning [91], sybil attacks [119], and other targeted and untargeted attacks [226]. Overall, FL has the potential to enhance security, as long as the proper security techniques and aggregation mechanisms are leveraged.

Chapter 6

AI Regulation: Leveraging Federated Learning for the Artificial Intelligence Act

6.1 Introduction

In Chapter 2, the AIA was introduced as a recently proposed AI law and regulatory framework that was by the European Commission [338]. The AIA defines a regulatory environment that assigns AI applications to various risk categories, outright banning high-risk applications such as AI-based social scoring by public authorities, and providing specific legal requirements and rules for the development, marketing, and use of AI applications and systems. This law is one of the first of potentially many regulatory frameworks that attempt to regulate AI, fueled by the need for a uniform legal framework to encourage safe and ethical AI systems, particularly encouraging the protection of health, safety, and the fundamental rights of individuals and nations.

This proposed law has significant implications for the development and deployment of AI. It is a broadly defined set of regulatory rules that apply to all providers of AI systems in service within the European Union and users of AI systems located within the EU. AI systems will be categorized into one of four categories: unacceptable risk, high-risk, limited risk, and minimal or no risk. Those applications categorized as an *unacceptable risk* will be prohibited from deployment within the EU, while high-risk applications will be required to comply with a set of strict regulatory rules. Breaches of these regulations can incur fines of up to EUR 40 million or 7% of the global annual turnover of the violating party.

Further, the AIA leverages other publications on Trustworthy AI to set standards for the development of trustworthy systems. The guidelines and rules outlined in the AIA take a step toward increasing trust by placing emphasis on the concepts of transparency, privacy, safety & security, accountability, fairness, and robustness & reliability. In the AIA, they propose regulatory sandboxes for the design, development, and safe deployment of AI systems, providing for an enclosed environment for the development of trustworthy and compliant AI.

In this chapter, it is ideated that FL can act as an adequate regulatory sandbox environment at the national and international levels. The benefits of utilizing FL as the basis for a sandbox are proposed to increase trust, ease the regulatory burden on developers and providers, and provide an avenue for connecting developers with the regulatory authorities securely and privately.

6.2 Background

6.2.1 AI Regulatory Sandboxes

To foster innovation and ease the burden on developers, the AIA proposes that Union authorities implement *AI regulatory sandboxes* (Title V). An AI regulatory sandbox is a con-

trolled environment for developers and providers to develop, test, and validate AI systems under the supervision of competent regulatory authorities. These authorities will provide guidance on compliance with the requirements of the regulation, aiding developers in properly complying with the complex rules and regulations set forth. Regulatory sandboxes have been largely applied in the finance and fintech industries [380], providing controlled environments for technical innovation. In the context of the AIA, few sandbox frameworks have been proposed [284, 345]. At this time, no federated regulatory environment exists in the context of the AIA.

The AIA has been criticized for its broad definition of AI and the significant burden that will be placed on providers and developers, particularly for high-risk systems. Many aspects of the AIA can be misinterpreted or misapplied, some even lacking enough clarity or functional tools for direct and practical application [351]. While these ethical and regulatory principles are vital for encouraging responsible and trustworthy AI, applying these principles during the technical development and deployment of such systems comes without guidance. Each of these elements requires a specific metric for compliance analysis, and while many tools exist to analyze trustworthiness, explainability, transparency, and data governance, there are no such metrics that match all requirements laid out for compliance [328]. Further, while regulatory sandboxes are meant to encourage innovation for start-ups and small-scale innovators, compliance with strict regulations may widen the gap between small- and large-sized entities in terms of AI development.

6.3 Federated Regulatory Sandbox

In this chapter, it is proposed that FL can be leveraged as a sandbox approach for appropriate regulation of the AIA. Here, only the requirements for developers and providers regarding the regulation of high-risk AI are considered. Those AI systems classified as unacceptable risk are out-of-scope for this regulatory sandbox, as they are strictly prohibited by the AIA.

Limited- and minimal-risk systems can also utilize the federated regulatory sandbox, benefiting from increased compliance and diligence during development and deployment.

FL is primarily leveraged to allow collaborative ML development without the requirement of sharing private data. An implementation of a centralized FL system involves integration with a central server, where participating 'clients', and users of the federated system send model updates to the central server without sharing private data. Within the context of a regulatory sandbox, the national regulators would take on the role of a central management server maintaining a repository of AI models, and taking in training updates from approved users. A federated regulatory sandbox can be leveraged in several ways. Here, two potential applications are considered:

1. The use of a federated regulatory sandbox to create use-case-specific AI systems.
2. The use of a federated regulatory sandbox to create isolated and private individual testing environments for model testing and compliance assessment.

In the first case, under a federated environment, regulatory agencies could provide a regulated central model (specific to each use case). To encourage compliance with the rules, developers could join the federated sandboxing environment and provide their data or model weights for training, receiving a trained model in return. This shared model will be the result of all developers and providers in the region, leading to an accurate and robust model that has continuous regulatory oversight. The use of a federated system essentially produces one central model that is subject to regulatory requirements, reducing the need for regulatory assessments for potentially thousands of entities.

In the second case, a federated regulatory sandbox can be used as a training, testing, and validation environment that does not require the sharing of private data. Individual FL environments can be opened by the regulatory agency, allowing developers and providers

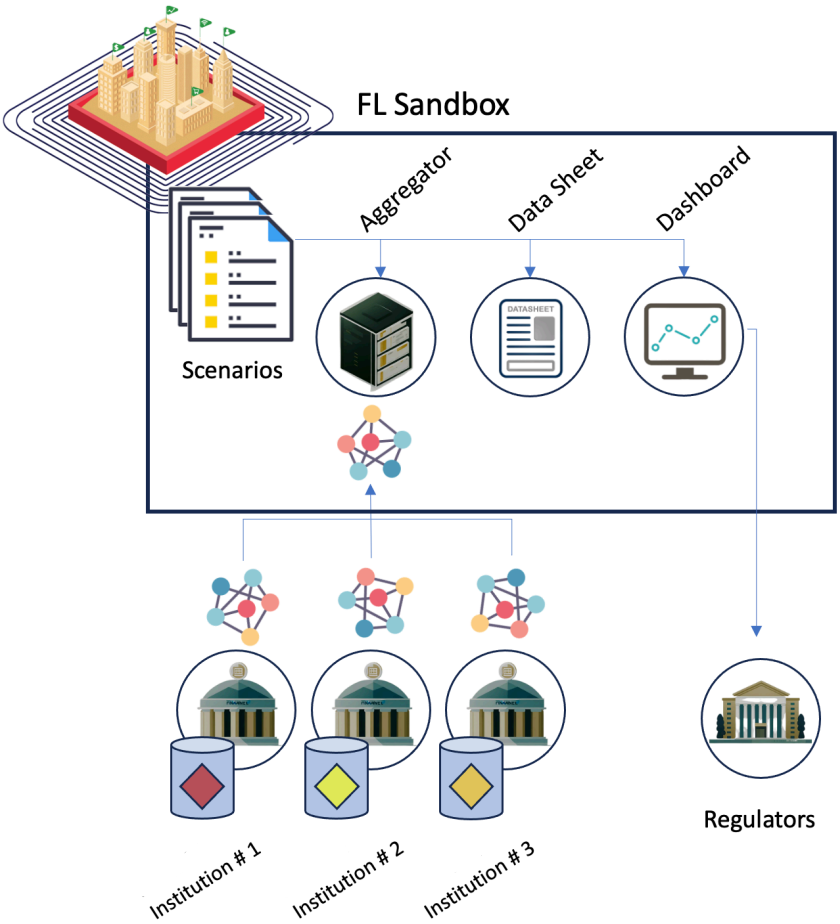

Figure 6.1: A potential architecture for a use-case specific FL system. The FL sandbox would include an aggregator, datasheet, and dashboard available to both the developer and regulatory agencies. Each institution or client would train a local model using their data, which is then sent to the sandbox and aggregated into a global model. All the relevant data can then be verified by regulatory agencies.

to have a direct connection with the competent authorities for the assessment of their high-risk systems.

In both cases, trust can be increased by providing a transparent environment for developers and regulators to assess the AI system. Regulators can assess the elements of trustworthiness disclosed by the creators, and developers can safely and securely demonstrate the robustness of their AI systems. Further, in either case, FL can be leveraged to simplify

the regulatory requirements in a sandbox environment. The AIA requires the following features: risk management systems, data protection and governance, technical documentation, record-keeping, and automatic logging, transparency of information to users, human oversight, accuracy, robustness, cybersecurity, and quality management. Each of these will be discussed with relevant context on how FL can ease the compliance burden.

The following sections detail how the above regulatory requirements can be addressed with an FL regulatory sandbox. Please note that the proposed regulatory sandbox only applies to the above two scenarios. A federated regulatory sandbox would be particularly advantageous in small, individual, or use-case-specific scenarios, but has limited feasibility in generalized regulatory environments.

6.3.1 Risk Management system (Article 9)

The AIA calls for a risk management system applied to the entire lifecycle of a high-risk AI system, including the identification and analysis of risks, estimation of risks that may occur when the AI system is placed into production, evaluation of other potential risks, and the adoption of risk management measures.

A centralized federated regulatory sandbox could be leveraged to decrease the need for individual risk management systems. In the case that several systems have similar use cases or components, it is reasonable that the risk management systems would be similar or only differ in small, very specific ways. The competent authority could develop risk management systems by use-case and share risk management assets with end-users of the federated systems. This will provide the opportunity for the competent regulatory authorities to define precise risk management systems and reduce the need for continuous approval of such systems when significant differences are not noted.

6.3.2 Data protection and data governance (Article 10)

FL has been leveraged to improve data protection and data governance. Inherently, FL protects user data by its design: data is not shared among users, ensuring data protection and privacy in a multitude of domains and use-cases [202]. The federated regulatory sandbox ensures that user data is kept private and not shared among other participants. Developers can leverage this functionality to ensure adequate data provenance and governance of their data by reducing the requirements for the movement or transfer of data from their original source.

Consider as an example a situation where several ISPs seek to develop an AI model to better address customer needs. In a non-federated system, the participating ISPs share their data, leading to a situation where the ISP that is the *data controller* for a client's data has to elevate every other participating ISP to the role of a *data processor*. If a data point has to be removed from the model, such as the right to be forgotten, each of the clients must communicate that this data point must be removed, collaboratively remove that data point from the data set, and then update the model training. Meanwhile, in a federated system only the *data controller* ISP needs to ever interact with the customer's data. They are a singular point of contact for privacy concerns, vastly simplifying the exclusion of data.

From a regulator's perspective, federated systems make data validation considerably easier to process as each *data controller* can be evaluated individually, and issues can be corrected on a client-by-client basis without requiring dataset audits.

6.3.3 Technical documentation, Record-Keeping, and Logs (Articles 11, 12, 20)

The federated regulatory sandbox would act as a repository of AI use cases. By maintaining a register of regulator-vetted datasheets containing detailed information on the technical

documentation requirements, all stakeholders would benefit by providing precedence. This would effectively mean sharing the burden.

By collaborating on a federated model, participants can effectively reduce the overall amount of data stored for historical and audit purposes, aligning with the data minimization principle of the General Data Protection Regulation (GDPR). It is important to maintain an audit trail that records the necessary information while minimizing the storage of personal data. In a FL context, the focus is on tracking the model updates rather than storing individual-level data. The audit trail can include information such as participant IDs, timestamps, and aggregated statistics about the model updates. This allows for accountability and transparency at every stage of the learning lifecycle without compromising the privacy of individual data points.

The Central Server can store documentation alongside models. Models can be tracked and reverted to address issues.

6.3.4 Transparency and provision of information to users (Article 13)

Transparency is a significant challenge in many AI systems. This article calls for the transparency and provision of information to users. However, FL has few impacts on the transparency of information to users. It is feasible that a federated regulatory environment greatly improves the transparency between developers and the regulatory agency, and that may increase the ability of developers to create transparent systems.

As a FL regulatory sandbox will provide an environment for collaboration between the developer and the regulatory agency, and transparency in the model development, training, and validation process. Further, the model weights may be available to regulatory agencies, depending on the settings of the federated environment. However, it is noted that

the inherent data privacy mechanisms of FL decrease the transparency regarding any data visibility and validity checks that the regulatory agency could conduct. Further, changes to the AI system can be easily tracked between updates, allowing the regulatory agent to rapidly ensure compliance.

Further, while there is no direct impact of transparency on end-users, it is possible that developers in a use-case-specific federated environment would benefit from a more interpretable and generalizable model. When defining a scenario in the regulatory sandbox, participants will define the model to be collaboratively learned. During this process, different parties can propose and decide on a model that achieves the desired level of interpretability for the end users. A transparency solution can be developed by one developer, or by the regulatory agency, and applied to all developers or providers in the regulatory sandbox. Regulators will have direct access to models to verify transparency standards.

6.3.5 Human Oversight (Article 14)

A federated framework with a human oversight layer added during the model training would provide a four-eye principle into the collaborative model learning process, ensuring enhanced transparency and accountability. The four-eye principle refers to the concept of having two or more individuals, involved in critical decision-making processes to minimize errors and increase security. In the context of model learning, participants of the federation would be able to monitor and audit the different stages of the collaborative learning process, preventing potential cases of internal fraud.

6.3.6 Accuracy, Robustness, and Cybersecurity (Article 15)

While a federated system will not provide robustness in the running of an AI system or directly improve the accuracy of an AI it still can provide benefits to both accuracy and cybersecurity of models.

In all cases, the regulatory agency can confirm accuracy after users submit model updates to the central server. In cases where there is collaboration, accuracy will likely be improved with the increased quantity of training data available for model development. As multiple actors collaborate, models can be developed with increased quantities of data, increasing not only accuracy but also potentially providing opportunities to mitigate bias and discrimination by improving data heterogeneity and robustness [3].

A federated system with a centralized entity and model repository requires a specific security focus. However, significant attention has been given to addressing cybersecurity in FL, including the development of Secure Aggregation [202], Differential Privacy [202], and Homomorphic Encryption [202]. A careful design with sufficient security mechanisms and good security hygiene will address this risk. Additionally, it is feasible for cooperation with EU Security Operation Centers for security assessments [2].

6.3.7 Cost

The increased security needs and the exchange of information among users required by the proposed solution would result in an increase in overall costs when compared with a traditional data lake approach; however recent work has shown that there are cost-effective designs that can be deployed which provide similar convergence speeds of the federated models, while reducing overall costs [221]. In addition, these solutions can reduce the overall carbon footprint, particularly when deploying Deep Learning solutions [279].

6.4 Discussion

This chapter proposes that FL can be leveraged as an AI regulatory sandboxing environment. It can be used as a regulatory sandbox in individual and use-case-specific scenarios, where creators and developers can be connected with regulators efficiently and securely.

The proposed sandboxing environment would allow individual and use-case-specific federated environments to be created. Individual sandboxes would ensure isolated and private individual testing environments. Use-case-specific federated environments would allow regulators to provide a registered central model (specific to each use-case), which creators can ensure is compliant with the current regulations. In both cases, working alongside regulators allows for developers to have a "seal of approval" for their models. This not only lowers the risk of deployment but the exchange and review of documents can be done faster. In the use-case environment, regulators do not need to undergo the process for each company, but can rather benefit from the federated sandboxing environment and ensure that all those involved are compliant.

In the future, additional regions will be deploying their regulation regarding AI, varying in levels of stringency. While this chapter has focused on FL amongst EU participants only, it would be feasible to achieve a collaborative model across regulatory borders. These groups would consist of participants from specific jurisdictions or regions that share similar compliance obligations, ensuring that all data used in the learning process is mutually compliant. As participants from less demanding regulatory environments wish to participate, different groupings would be formed, allowing stricter regulatory zones to still monetize their models, without infringing local regulation.

Further consideration should be given to collaboration with other EU and European Commission services, groups, and entities. It is feasible that such a federated regulatory sandbox can integrate with other EU-provided services. Namely, AI training and testing can be implemented using computational services within the Union, such as those governed by the European High-Performance Computing Joint Undertaking (EuroHPC JU) [1]. Integration with existing services will simplify the introduction of regulatory sandboxes. Further, it should be noted the FL regulatory sandbox would not aid regulatory agencies in broad, generalized settings. It is not feasible for a regulatory agency to train a federated model on

greatly different topics, as the generalized model would not be useful to any participating developer. Further consideration would be required to apply these concepts to generalized FL scenarios.

Implementation of federated regulatory sandboxes by competent authorities allows for providing a safe and secure sandboxing environment. This environment can leverage FL for private and secure data sharing, removing any need for data sharing with regulatory agencies. Further, FL and associated tools can be leveraged to reduce the compliance burden regarding to the regulatory rules for high-risk systems.

In turn, this sandboxing environment also has the potent to increase trust in AI systems via an increased transparency in design, development, and deployment. Developers have increased opportunities to be transparent about their systems, particularly focused on transparency in how accountability, developmental processes, robustness, risk management systems, data protection and governance, transparency of information to users, human oversight, safety, and security issues are addressed. The AIA builds the stage for trustworthy AI development, and by leveraging FL in hand with regulatory agencies, trustworthy AI systems can be fostered in an innovative way.

Chapter 7

Conclusions and Future Work

In this dissertation, the concept of trustworthiness in AI systems and models was defined, and the challenges associated with designing, developing, and deploying trustworthy AI systems were analyzed considering the greater ethical, technical, and legal implications. Methods to increase trustworthiness by leveraging privacy-preserving collaborative ML methods, specifically FL, were explored to assess the impact on trustworthy principles and concepts.

The key contributions of this dissertation included a critical assessment of the design, development, and deployment of trustworthy AI systems and models. The main principles associated with Trustworthy AI were identified as accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security. These key principles were identified based on commonality in Trustworthy AI texts, including academic literature and legal texts alike. These principles were leveraged to create a simplified framework for the design, development, and deployment of trustworthy AI, the proposed Know Your Model framework. This contribution provides a concise

and straightforward framework to assess trustworthiness during all stages of AI development.

These elements of trust were further explored in the context of privacy-preserving methods. Specifically, FL was explored for implications on the trustworthy principles of privacy, accountability, robustness & reliability, and safety & security. Privacy was established as a significant benefit of FL, enabling collaboration between participants such as financial institutions without the sharing of private data. The security of FL was explored via targeted attacks by malicious actors, demonstrating the ability of FL systems to defend against such attacks. Further, FL was leveraged to comply with proposed regulatory requirements, revealing that FL can be utilized in a privacy-preserving manner to increase trust.

In the following sections, the contributions of this dissertation are summarized.

7.1 Principles of Trustworthy AI

Trustworthy AI literature and texts were critically analyzed for the key components associated with the design, development, and deployment of trustworthy AI. This analysis combined academic literature and legal texts, including an assessment of the High-Level Expert Group (HLEG) reports, the Artificial Intelligence Act (AIA) in the European Union [142, 338], and the *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* in the United States of America [43]. These texts revealed six key trustworthy principles: accountability, explainability & interpretability, fairness & non-discrimination, privacy, robustness & reliability, and safety & security. While Trustworthy AI literature is plentiful, this dissertation provides a clear analysis of this area to identify the most vital concepts associated with trust in AI systems.

7.2 Trustworthy AI Design, Development, and Deployment

The defined trustworthy principles were composed of complex, abstract, and difficult-to-apply concepts. Few guidelines were identified for the technical application of these principles in real-world applications. The second contribution of this thesis applies these trustworthy principles to define guidelines for the design, development, and deployment of AI systems. Twenty key guidelines toward trustworthy AI were defined with the KYM framework. Concise guidelines were established and clear examples of each guideline were provided.

7.3 Leveraging Privacy-Preserving Methods to Increase Trust

This dissertation particularly focused on leveraging FL to increase trust in AI. In Chapter 4, FL was described along with its applications and challenges. To analyze the benefits of FL, a case study was provided to explore utilizing FL for transaction monitoring in Anti-Money Laundering applications. It was demonstrated that FL has strengths in privacy, enabling collaboration between multiple actors without the need for sharing data. Further, it has implications for model robustness, allowing for models to be developed on larger volumes of data and resulting in better-performing, more robust models.

7.4 Security Implications of Federated Learning

In Chapter 5, the security challenges of FL were assessed. The robustness of FL against adversarial attacks, specifically model poisoning backdoor attacks, was explored. A model poisoning attack was demonstrated on FL, revealing that it is vulnerable to attacks that can remain hidden from participants while injecting a specific attacker-chosen behavior.

Various Byzantine-tolerant aggregation mechanisms were analyzed for their effectiveness in preventing and defending against attacks, including an assessment of the fairness of the resulting robustness of models when the aggregation mechanisms are deployed.

7.5 Leveraging Federated Learning in Regulatory Environments

In Chapter 6, FL was leveraged to act as a regulatory sandbox to fulfill the requirements of the regulatory framework proposed in the AIA. This chapter proposes FL as a regulatory sandbox, leveraging the inherent trustworthy elements and privacy-preserving benefits of FL to propose an environment for collaboration between developers and regulators. This sandboxing environment may increase trust in AI in a private and secure manner by increasing transparency in all areas of trustworthiness.

7.6 Future Work

This dissertation explores a wide variety of elements of Trustworthy AI and explores the benefits and challenges associated with FL. While many elements of trust are explored, there are several areas that may be of interest for future research.

For example, while FL enables increased privacy, reliability, and robustness, and has implications to increase accountability and transparency associated with other trustworthy principles, the question of incentivization to participate in an FL network was not explored. The field of incentivization for FL is a busy domain, with many researchers proposing various incentive schemas to encourage participation [379]. These aspects may be of interest for additional research for their role in trust.

This dissertation primarily concerns centralized FL networks that utilize a centralized server for model aggregation. The trust implications of a decentralized FL system were not ex-

plored. There may be different implications on trust for decentralized systems that warrant additional research.

Further, a significant question associated with FL regards the individual contributions of each participant. These contributions, and the implications on trust, were only limitedly explored here. Free-rider attacks, where a participant aims to benefit from membership in the federation without contributing [117], are an area of research that may have different implications on trust. Trustworthy participation may be an area of interest for further analysis.

Finally, there is still much work to be done to move from requirements to tools that monitor the trustworthiness of an AI system. This dissertation lays the groundwork and provides a directional guide for future research in this field.

References

[1] The european high performance computing joint undertaking (eurohpc ju). URL https://eurohpc-ju.europa.eu/index_en.

[2] Cybersecurity: Eu launches first phase of deployment of the european infrastructure of cross-border security operations centres. URL https://digital-strategy.ec.europa.eu/en/news/cybersecurity-eu-launches-first-phase-deployment-european- infrastructure-cross-border-security.

[3] A. Abay, Y. Zhou, N. Baracaldo, S. Rajamoni, E. Chuba, and H. Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

[4] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[5] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[6] S. Akter, Y. K. Dwivedi, K. Biswas, K. Michael, R. J. Bandara, and S. Sajib. Addressing algorithmic bias in ai-driven customer management. *Journal of Global Information Management (JGIM)*, 29(6):1–27, 2021.

[7] G. Alicioglu and B. Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 102:502–520, 2022.

[8] A. A. S. Alsuwailem and A. K. J. Saudagar. Anti-money laundering systems: a systematic literature review. *Journal of Money Laundering Control*, 23(4):833–848, 2020.

[9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[10] V. W. Anelli, Y. Deldjoo, T. Di Noia, and A. Ferrara. Towards effective device-aware federated learning. In *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian*

Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18, pages 477–491. Springer, 2019.

[11] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[12] T. T. Anh, N. C. Luong, D. Niyato, D. I. Kim, and L.-C. Wang. Efficient training management for mobile crowd-machine learning: A deep reinforcement learning approach. *IEEE Wireless Communications Letters*, 8(5):1345–1348, 2019.

[13] Y. Aono, T. Hayashi, L. Wang, S. Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE transactions on information forensics and security*, 13(5):1333–1345, 2017.

[14] C. A. Ardagna, M. Cremonini, E. Damiani, S. De Capitani di Vimercati, and P. Samarati. Location privacy protection through obfuscation-based techniques. In *IFIP annual conference on data and applications security and privacy*, pages 47–60. Springer, 2007.

[15] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

[16] D. W. Arner, J. N. Barberis, and R. P. Buckley. The emergence of regtech 2.0: From know your customer to know your data. 2016.

[17] M. Arnold, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 2019.

[18] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[19] S. Avin, H. Belfield, M. Brundage, G. Krueger, J. Wang, A. Weller, M. Anderljung, I. Krawczuk, D. Krueger, J. Lebensold, et al. Filling gaps in trustworthy development of ai. *Science*, 374(6573): 1327–1329, 2021.

[20] S. Awan, F. Li, B. Luo, and M. Liu. Poster: A reliable and accountable privacy-preserving federated learning framework using the blockchain. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2561–2563, 2019.

[21] R. Aziz, S. Banerjee, S. Bouzefrane, and T. Le Vinh. Exploring homomorphic encryption and differential privacy techniques towards secure federated learning paradigm. *Future internet*, 15(9):310, 2023.

[22] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020.

[23] Y. Balagurunathan, R. Mitchell, and I. El Naqa. Requirements and reliability of ai in the medical context. *Physica Medica*, 83:72–78, 2021.

[24] X. Bao, C. Su, Y. Xiong, W. Huang, and Y. Hu. Flchain: A blockchain for auditable federated learning with trust and incentive. In *2019 5th International Conference on Big Data Computing and Communications (BIGCOM)*, pages 151–159. IEEE, 2019.

[25] N. Baracaldo, A. Anwar, M. Purcell, A. Rawat, M. Sinn, B. Altakrouri, D. Balta, M. Sellami, P. Kuhn, U. Schopp, et al. Towards an accountable and reproducible federated learning: A factsheets approach. *arXiv preprint arXiv:2202.12443*, 2022.

[26] S. Barocas and A. D. Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

[27] P. Basu, T. S. Roy, R. Naidu, and Z. Muftuoglu. Privacy enabled financial text classification using differential privacy and federated learning. *arXiv preprint arXiv:2110.01643*, 2021.

[28] R. E. Bawack, S. F. Wamba, K. D. A. Carillo, and S. Akter. Artificial intelligence in e-commerce: a bibliometric study and literature review. *Electronic markets*, 32(1):297–338, 2022.

[29] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pages 217–228. IEEE, 2005.

[30] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[31] G. Beigi and H. Liu. A survey on privacy in social media: Identification, mitigation, and applications. *ACM Transactions on Data Science*, 1(1):1–38, 2020.

[32] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[33] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[34] E. T. M. Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 2023.

[35] E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[36] H. Berghel. Equifax and the latest round of identity theft roulette. *Computer*, 50(12):72–76, 2017.

[37] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[38] T. J. Berkmans and S. Karthick. Credit card fraud detection with data sampling. In *2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, pages 1–6. IEEE, 2022.

[39] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão, et al. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

[40] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pages 634–643. PMLR, 2019.

[41] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020.

[42] X. Bi, A. Gupta, and M. Yang. Understanding partnership formation and repeated contributions in federated learning: An analytical investigation. *Management Science*, 2023.

[43] J. R. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023.

[44] G. Bilali. Know your customer-or not. *U. Tol. L. Rev.*, 43:319, 2011.

[45] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[46] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[47] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.

[48] K. Boyd, K. H. Eng, and C. D. Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 451–466. Springer, 2013.

[49] S. Bradner. Rfc2119: Key words for use in rfcs to indicate requirement levels, 1997.

[50] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[51] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

[52] D. Broeders, E. Schrijvers, B. van der Sloot, R. Van Brakel, J. de Hoog, and E. H. Ballin. Big data and security policies: Towards a framework for regulating the phases of analytics and use of big data. *Computer Law & Security Review*, 33(3):309–323, 2017.

[53] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

[54] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR, 2019.

[55] M. E. Bunnage, A. M. Gilbert, L. H. Jones, and E. C. Hett. Know your target, know your molecule. *Nature chemical biology*, 11(6):368–372, 2015.

[56] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[57] M. Busuioc. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81(5):825–836, 2021.

[58] D. Byrd and A. Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–9, 2020.

[59] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

[60] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.

[61] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[62] L. Cao. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55 (3):1–38, 2022.

[63] X. Cao, M. Fang, J. Liu, and N. Z. Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020.

[64] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[65] R. Chalapathy and S. Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[66] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ quality & safety*, 2019.

[67] H. Chen, S. U. Hussain, F. Boemer, E. Stapf, A. R. Sadeghi, F. Koushanfar, and R. Cammarota. Developing privacy-preserving ai systems: The lessons learned. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–4. IEEE, 2020.

[68] L. Chen, P. Chen, and Z. Lin. Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278, 2020.

[69] P. Chen, X. Du, Z. Lu, J. Wu, and P. C. Hung. Evfl: An explainable vertical federated learning for data-oriented artificial intelligence systems. *Journal of Systems Architecture*, 126:102474, 2022.

[70] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[71] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[72] Y. Chen, L. Su, and J. Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2): 1–25, 2017.

[73] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93, 2020.

[74] Z. Chen, L. D. Van Khoa, E. N. Teoh, A. Nazir, E. K. Karuppiah, and K. S. Lam. Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, 57:245–285, 2018.

[75] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, and Q. Yang. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, 36(6):87–98, 2021.

[76] Y. Cheng, Y. Liu, T. Chen, and Q. Yang. Federated learning for privacy-preserving ai. *Communications of the ACM*, 63(12):33–36, 2020.

[77] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[78] C. Collins, D. Dennehy, K. Conboy, and P. Mikalef. Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, 60:102383, 2021.

[79] E. Commission. White paper on artificial intelligence: A european approach to excellence and trust, 2020.

[80] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[81] D. Cox. *Handbook of anti-money laundering*. John Wiley & Sons, 2014.

[82] B. d'Alessandro, C. O'Neil, and T. LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134, 2017.

[83] T. K. Dang, X. Lan, J. Weng, and M. Feng. Federated learning for electronic health records. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(5):1–17, 2022.

[84] Y. Dang, Q. Lin, and P. Huang. Aiops: real-world challenges and research innovations. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 4–5. IEEE, 2019.

[85] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

[86] H. B. Desai, M. S. Ozdayi, and M. Kantarcioglu. Blockfla: Accountable federated learning via hybrid blockchain architecture. In *Proceedings of the eleventh ACM conference on data and application security and privacy*, pages 101–112, 2021.

[87] W. DeSombre, J. SHIRES, J. WORK, R. MORGUS, P. H. O'NEILL, L. ALLODI, and T. HERR. *Countering Cyber Proliferation: Zeroing in on Access-as-a-Service*. Atlantic Council., 2021.

[88] T. G. Dietterich. Steps toward robust artificial intelligence. *Ai Magazine*, 38(3):3–24, 2017.

[89] T. G. Dietterich and E. J. Horvitz. Rise of concerns about ai: reflections and directions. *Communications of the ACM*, 58(10):38–40, 2015.

[90] Y. Djebrouni, N. Benarba, O. Touat, P. De Rosa, S. Bouchenak, A. Bonifati, P. Felber, V. Marangozova, and V. Schiavoni. Bias mitigation in federated learning for edge computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–35, 2024.

[91] R. Doku and D. B. Rawat. Mitigating data poisoning attacks on a federated learning-edge computing network. In *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–6. IEEE, 2021.

[92] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, et al. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[93] H. Du, M. Shen, R. Sun, J. Jia, L. Zhu, and Y. Zhai. Malicious transaction identification in digital currency via federated graph deep learning. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE, 2022.

[94] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms*, pages 13–22, 2001.

[95] W. Du, Y. S. Han, and S. Chen. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 222–233. SIAM, 2004.

[96] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[97] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.

[98] J. Dumford and W. Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.

[99] C. Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[100] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

[101] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

[102] A. El Ouadrhiri and A. Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380, 2022.

[103] D. F. Engstrom and D. E. Ho. Algorithmic accountability in the administrative state. *Yale J. on Reg.*, 37: 800, 2020.

[104] E. Esoimeme. Curbing employee fraud and corruption in financial institutions with an effective know your employee program. *KYC Global Technologies*, 2018.

[105] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and A. S. Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7494–7502, 2023.

[106] M. Fang, X. Cao, J. Jia, and N. Gong. Local model poisoning attacks to byzantine-robust federated learning. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1605–1622, 2020.

[107] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad. Federated learning for covid-19 screening from chest x-ray images. *Applied Soft Computing*, 106:107330, 2021.

[108] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[109] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6):3333–3361, 2020.

[110] J. Ferwerda. The economics of crime and money laundering: does anti-money laundering policy reduce crime? *Review of Law & Economics*, 5(2):903–929, 2009.

[111] U. Fiege, A. Fiat, and A. Shamir. Zero knowledge proofs of identity. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 210–217, 1987.

[112] J. Fiosina. Explainable federated learning for taxi travel time prediction. In *VEHITS*, pages 670–677, 2021.

[113] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, 2020.

[114] L. Floridi and J. Cowls. A unified framework of five principles for ai in society. *Issue 1.1, Summer 2019*, 1(1), 2019.

[115] L. Floridi, J. Cowls, T. C. King, and M. Taddeo. How to design ai for social good: seven essential factors. *Ethics, Governance, and Policies in Artificial Intelligence*, pages 125–151, 2021.

[116] F. A. T. Force. What is money laundering. *Policy Brief July 1999*, 1999.

[117] Y. Fraboni, R. Vidal, and M. Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1846–1854. PMLR, 2021.

[118] Y. Fu, H. Wang, K. Xu, H. Mi, and Y. Wang. Mixup based privacy preserving mixed collaboration learning. In *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)*, pages 275–2755. IEEE, 2019.

[119] C. Fung, C. J. Yoon, and I. Beschastnikh. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.

[120] D. Gao, C. Ju, X. Wei, Y. Liu, T. Chen, and Q. Yang. Hhhfl: Hierarchical heterogeneous horizontal federated learning for electroencephalography. *arXiv preprint arXiv:1909.05784*, 2019.

[121] S. Garfinkel et al. *De-identification of Personal Information:*. US Department of Commerce, National Institute of Standards and Technology, 2015.

[122] A. Gascón, P. Schoppmann, B. Balle, M. Raykova, J. Doerner, S. Zahur, and D. Evans. Secure linear regression on vertically partitioned datasets. *IACR Cryptol. ePrint Arch.*, 2016:892, 2016.

[123] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[124] G. Gharibi, V. Walunj, R. Nekadi, R. Marri, and Y. Lee. Automated end-to-end management of the modeling lifecycle in deep learning. *Empirical Software Engineering*, 26(2):1–33, 2021.

[125] I. Giacomelli, S. Jha, M. Joye, C. D. Page, and K. Yoon. Privacy-preserving ridge regression with only linearly-homomorphic encryption. In *Applied Cryptography and Network Security: 16th International Conference, ACNS 2018, Leuven, Belgium, July 2-4, 2018, Proceedings 16*, pages 243–261. Springer, 2018.

[126] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[127] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[128] Google. Know your data. `https://knowyourdata.withgoogle.com`, 2023. Accessed: 2023-02-01.

[129] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

[130] O. E. Gundersen and S. Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[131] D. Gunning and D. Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2): 44–58, 2019.

[132] J. Guo, Z. Liu, S. Tian, F. Huang, J. Li, X. Li, K. K. Igorevich, and J. Ma. Tfl-dt: A trust evaluation scheme for federated learning in digital twin for mobile networks. *IEEE Journal on Selected Areas in Communications*, 2023.

[133] P. Guo, P. Wang, J. Zhou, S. Jiang, and V. M. Patel. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2423–2432, 2021.

[134] J. Hamer, M. Mohri, and A. T. Suresh. Fedboost: A communication-efficient algorithm for federated learning. In *International Conference on Machine Learning*, pages 3973–3983. PMLR, 2020.

[135] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[136] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[137] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[138] A. Hawken and G. L. Munck. Do you know your data? measurement validity in corruption research. *Unpublished typescript, Pepperdine University and University of Southern California, Malibu, CA, and Los Angeles*, 2009.

[139] A. Henelius, K. Puolamäki, and A. Ukkonen. Interpreting classifiers through attribute interactions in datasets. *arXiv preprint arXiv:1707.07576*, 2017.

[140] M. Hengstler, E. Enkel, and S. Duelli. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105:105–120, 2016.

[141] M. Herschel, R. Diestelkämper, and H. B. Lahmar. A survey on provenance: What for? what form? what from? *The VLDB Journal*, 26(6):881–906, 2017.

[142] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Report, European Commission, Brussels, Apr. 2019. URL `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`.

[143] High-Level Expert Group on AI. The assessment list for trustworthy ai (altai) for self assessment. Report, European Commission, Brussels, July 2020. URL `https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment`.

[144] M. Hind, S. Houde, J. Martino, A. Mojsilovic, D. Piorkowski, J. Richards, and K. R. Varshney. Experiences with improving the transparency of ai models and services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[145] B. Hitaj, G. Ateniese, and F. Perez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 603–618, 2017.

[146] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[147] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[148] Y. Hu, W. Kuang, Z. Qin, K. Li, J. Zhang, Y. Gao, W. Li, and K. Li. Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[149] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

[150] W. Hummer, V. Muthusamy, T. Rausch, P. Dube, K. El Maghraoui, A. Murthi, and P. Oum. Modelops: Cloud-based lifecycle management for reliable and trusted ai. In *2019 IEEE International Conference on Cloud Engineering (IC2E)*, pages 113–120. IEEE, 2019.

[151] O. Ibitoye, R. Abou-Khamis, A. Matrawy, and M. O. Shafiq. The threat of adversarial attacks on machine learning in network security–a survey. *arXiv preprint arXiv:1911.02621*, 2019.

[152] S. Idowu, D. Strüber, and T. Berger. Asset management in machine learning: A survey. *arXiv preprint arXiv:2102.06919*, 2021.

[153] M. Iezzi. Practical privacy-preserving data science with homomorphic encryption: an overview. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3979–3988. IEEE, 2020.

[154] A. Imteaj and M. H. Amini. Leveraging asynchronous federated learning to predict customers financial distress. *Intelligent Systems with Applications*, 14:200064, 2022.

[155] C. Irti. Personal data, non-personal data, anonymised data, pseudonymised data, de-identified data. *Privacy and Data Protection in Software Services*, pages 49–57, 2022.

[156] ISO24028:2020. Information Technology–Artificial Intelligence–Overview of Trustworthiness in Artificial Intelli- gence. Standard, International Organization for Standardization, Geneva, CH, May 2020.

[157] Z. Ji, Z. C. Lipton, and C. Elkan. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*, 2014.

[158] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[159] P. Johnston and R. Harris. The boeing 737 max saga: lessons for software organizations. *Software Quality Professional*, 21(3):4–12, 2019.

[160] C. Ju, D. Gao, R. Mane, B. Tan, Y. Liu, and C. Guan. Federated transfer learning for eeg signal classification. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pages 3040–3045. IEEE, 2020.

[161] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[162] M. E. Kaminski and G. Malgieri. Multi-layered explanations from algorithmic impact assessments in the gdpr. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 68–79, 2020.

[163] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[164] F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, volume 1. Citeseer, 2010.

[165] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[166] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, pages 869–874. IEEE, 2010.

[167] T. Kamishima, S. Akaho, H. Asoh, and J. Sukama. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012*, pages 35–50, 2012.

[168] S. Kanamori, T. Abe, T. Ito, K. Emura, L. Wang, S. Yamamoto, T. P. Le, K. Abe, S. Kim, R. Nojima, et al. Privacy-preserving federated learning for detecting fraudulent financial transactions in japanese banks. *Journal of Information Processing*, 30:789–795, 2022.

[169] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani. Reliable federated learning for mobile networks. *IEEE Wireless Communications*, 27(2):72–80, 2020.

[170] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[171] A. F. Karr, X. Lin, A. P. Sanil, and J. P. Reiter. Privacy-preserving analysis of vertically partitioned data using secure matrix products. *Journal of Official Statistics*, 25(1):125–138, 2009.

[172] B. Kasasbeh, B. Aldabaybah, and H. Ahmad. Multilayer perceptron artificial neural networks-based model for credit card fraud detection. *Indonesian Journal of Electrical Engineering and Computer Science*, 26(1):362–373, 2022.

[173] B. S. Kašin. Diameters of some finite-dimensional sets and classes of smooth functions. *Izvestiya: Mathematics*, 11(2):317–333, 1977.

[174] D. Kaur, S. Uslu, and A. Durresi. Requirements for trustworthy artificial intelligence–a review. In *Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020) 23*, pages 105–115. Springer, 2021.

[175] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

[176] D. Kawa, S. Punyani, P. Nayak, A. Karkera, and V. Jyotinagar. Credit risk assessment from combined

bank records using federated learning. *International Research Journal of Engineering and Technology (IRJET)*, 6(4):1355–1358, 2019.

[177] D. Kemp and F. Vanclay. Human rights and impact assessment: clarifying the connections in practice. *Impact Assessment and Project Appraisal*, 31(2):86–96, 2013.

[178] I. Kevin, K. Wang, X. Zhou, W. Liang, Z. Yan, and J. She. Federated transfer learning based cross-domain prediction for smart manufacturing. *IEEE Transactions on Industrial Informatics*, 18(6):4088–4096, 2021.

[179] F. A. KhoKhar, J. H. Shah, M. A. Khan, M. Sharif, U. Tariq, and S. Kadry. A review on federated learning towards image processing. *Computers and Electrical Engineering*, 99:107818, 2022.

[180] A. Kiayias, M. Kohlweiss, and A. Sarencheh. Peredi: Privacy-enhanced, regulated and distributed central bank digital currencies. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1739–1752, 2022.

[181] B. Kim and F. Doshi-Velez. Machine learning techniques for accountability. *AI Magazine*, 42(1):47–52, 2021.

[182] B. Kim, J. Park, and J. Suh. Transparency and accountability in ai decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134:113302, 2020.

[183] A. A. Kodiyan. An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool. *Researchgate Preprint*, pages 1–19, 2019.

[184] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[185] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[186] F. Königstorfer and S. Thalmann. Ai documentation: A path to accountability. *Journal of Responsible Technology*, 11:100043, 2022.

[187] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[188] H. Kusetogullari, A. Yavariabdi, A. Cheddad, H. Grahn, and H. Johan. Ardis: a swedish historical handwritten digit dataset. *Neural computing & applications (Print)*, 32(21):16505–16518, 2020.

[189] C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50(Suppl):S82, 2012.

[190] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[191] S. Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16, 2018.

[192] K. Lerman. Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science*, 1(1):49–58, 2018.

[193] K. Lerman and T. Hogg. Leveraging position bias to improve peer recommendation. *PloS one*, 9(6): e98914, 2014.

[194] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau. Federated learning for keyword spotting. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6341–6345. IEEE, 2019.

[195] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. 2015.

[196] N. G. Leveson. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

[197] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[198] B.-h. Li, B.-c. Hou, W.-t. Yu, X.-b. Lu, and C.-w. Yang. Applications of artificial intelligence in intelligent manufacturing: a review. *Frontiers of Information Technology & Electronic Engineering*, 18:86–96, 2017.

[199] J. Li, T. Cui, K. Yang, R. Yuan, L. He, and M. Li. Demand forecasting of e-commerce enterprises based on horizontal federated learning from the perspective of sustainable development. *Sustainability*, 13 (23):13050, 2021.

[200] L. Li, Y. Fan, M. Tse, and K.-Y. Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, page 106854, 2020.

[201] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2006.

[202] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[203] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[204] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[205] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[206] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

[207] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

[208] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[209] H. Liu, X. Zhang, X. Shen, and H. Sun. A federated learning framework for smart grids: Securing power traces in collaborative learning. *arXiv preprint arXiv:2103.11870*, 2021.

[210] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59, 2022.

[211] M. Liu, H. Jiang, J. Chen, A. Badokhon, X. Wei, and M.-C. Huang. A collaborative privacy-preserving deep learning system in distributed mobile environment. In *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 192–197. IEEE, 2016.

[212] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. Leung. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access*, 6:12103–12117, 2018.

[213] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35(4):70–82, 2020.

[214] S. Lockey, N. Gillespie, D. Holm, and I. A. Someh. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. 2021.

[215] G. Long, Y. Tan, J. Jiang, and C. Zhang. Federated learning for open banking. In *Federated learning*, pages 240–254. Springer, 2020.

[216] E. Lopez-Rojas, A. Elmir, and S. Axelsson. Paysim: A financial mobile money simulator for fraud detection. In *28th European Modeling and Simulation Symposium, EMSS, Larnaca*, pages 249–255. Dime University of Genoa, 2016.

[217] E. A. Lopez-Rojas and S. Axelsson. A review of computer simulation for fraud detection research in financial datasets. In *2016 Future technologies conference (FTC)*, pages 932–935. IEEE, 2016.

[218] M. Y. Lu, R. J. Chen, D. Kong, J. Lipkova, R. Singh, D. F. Williamson, T. Y. Chen, and F. Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical image analysis*, 76:102298, 2022.

[219] S. Lu, Y. Zhang, and Y. Wang. Decentralized federated learning for electronic health records. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5. IEEE, 2020.

[220] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[221] B. Luo, X. Li, S. Wang, J. Huang, and L. Tassiulas. Cost-effective federated learning design. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

[222] J. Luo, X. Wu, Y. Luo, A. Huang, Y. Huang, Y. Liu, and Q. Yang. Real-world image datasets for federated learning. *arXiv preprint arXiv:1910.11089*, 2019.

[223] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510, 2011.

[224] W. Luping, W. Wei, and L. Bo. Cmfl: Mitigating communication overhead for federated learning. In *2019 IEEE 39th international conference on distributed computing systems (ICDCS)*, pages 954–964. IEEE, 2019.

[225] L. Lyu, H. Yu, and Q. Yang. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020.

[226] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and S. Y. Philip. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*, 2022.

[227] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[228] T. Mahoney, K. Varshney, and M. Hind. *AI fairness*. O'Reilly Media, Incorporated, 2020.

[229] S. Mäkinen, H. Skogström, E. Laaksonen, and T. Mikkonen. Who needs mlops: What data scientists seek to accomplish and how can mlops help? In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 109–112. IEEE, 2021.

[230] G. Malgieri and F. A. Pasquale. From transparency to justification: Toward ex ante accountability for ai. *Brooklyn Law School, Legal Studies Paper*, (712), 2022.

[231] P. Malik, M. Pathania, V. K. Rathaur, et al. Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*, 8(7):2328, 2019.

[232] D. M. Manias and A. Shami. Making a case for federated learning in the internet of vehicles and intelligent transportation systems. *IEEE Network*, 35(3):88–94, 2021.

[233] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[234] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[235] N. Mehrabi, Y. Huang, and F. Morstatter. Statistical equity: A fairness classification objective. *arXiv preprint arXiv:2005.07293*, 2020.

[236] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[237] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[238] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

[239] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci. Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, 58(3):401–415, 2016.

[240] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 735–746, 2021.

[241] Microsoft Research Blog. Ai models vs. ai systems: Understanding units of performance assessment. `https://www.microsoft.com/en-us/research/blog/ai-models-vs-ai-systems-understanding-units-of-performance-assessment/`, 2022. Accessed: 2013-17-01.

[242] B. S. Miguel, A. Naseer, and H. Inakoshi. Putting accountability of ai systems into practice. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 5276–5278, 2021.

[243] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[244] N. Möller and S. O. Hansson. Principles of engineering safety: Risk and uncertainty reduction. *Reliability Engineering & System Safety*, 93(6):798–805, 2008.

[245] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal. From what to how. an overview of ai ethics tools, methods and research to translate principles into practices. *arXiv preprint arXiv:1905.06876*, 2019.

[246] V. Mugunthan, R. Rahman, and L. Kagal. Blockflow: An accountable and privacy-preserving solution for federated learning. *arXiv preprint arXiv:2007.03856*, 2020.

[247] D. Myalil, M. Rajan, M. Apte, and S. Lodha. Robust collaborative fraudulent transaction detection using federated learning. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 373–378. IEEE, 2021.

[248] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[249] S. Naz, K. T. Phan, and Y.-P. P. Chen. A comprehensive review of federated learning for covid-19 detection. *International Journal of Intelligent Systems*, 37(3):2371–2392, 2022.

[250] G. S. Nelson. Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification. In *SAS global forum proceedings*, pages 1–23, 2015.

[251] D. C. Nguyen, M. Ding, Q.-V. Pham, P. N. Pathirana, L. B. Le, A. Seneviratne, J. Li, D. Niyato, and H. V. Poor. Federated learning meets blockchain in edge computing: Opportunities and challenges. *IEEE Internet of Things Journal*, 8(16):12806–12825, 2021.

[252] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(3):1–37, 2022.

[253] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.

[254] T. Nishio and R. Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications (ICC)*, pages 1–7. IEEE, 2019.

[255] R. Nock, S. Hardy, W. Henecka, H. Ivey-Law, G. Patrini, G. Smith, and B. Thorne. Entity resolution and federated learning get a federated resolution. *arXiv preprint arXiv:1803.04035*, 2018.

[256] C. Novelli, M. Taddeo, and L. Floridi. Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY*, pages 1–12, 2023.

[257] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[258] F. Nuding and R. Mayer. Data poisoning in sequential and parallel federated learning. In *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*, pages 24–34, 2022.

[259] B. Nushi, E. Kamar, and E. Horvitz. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 126–135, 2018.

[260] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[261] OECD. Recommendation of the council on artificial intelligence, May 2019. URL `https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449`.

[262] OECD. Advancing accountability in ai governing and managing risks throughout the lifecycle for trustworthy ai. *OECD DIGITAL ECONOMY PAPERS*, 2023.

[263] A. Oseni, N. Moustafa, H. Janicke, P. Liu, Z. Tari, and A. Vasilakos. Security and privacy for artificial intelligence: Opportunities and challenges. *arXiv preprint arXiv:2102.04661*, 2021.

[264] B. Oztas, D. Cetinkaya, F. Adedoyin, M. Budka, H. Dogan, and G. Aksu. Perspectives from experts on developing transaction monitoring methods for anti-money laundering. In *2023 IEEE International Conference on e-Business Engineering (ICEBE)*, pages 39–46. IEEE, 2023.

[265] K. S. Paithankar and E. F. Garman. Know your dose: Raddose. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):381–388, 2010.

[266] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[267] P. Papadopoulos, W. Abramson, A. J. Hall, N. Pitropakis, and W. J. Buchanan. Privacy and trust redefined in federated machine learning. *Machine Learning and Knowledge Extraction*, 3(2):333–356, 2021.

[268] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[269] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[270] C. Percy, S. Dragicevic, S. Sarkar, and A. d'Avila Garcez. Accountability in ai: From principles to industry-specific accreditation. *AI Communications*, 34(3):181–196, 2021.

[271] L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[272] S. Pichai. Ai at google: our principles. *The Keyword*, 7:1–3, 2018.

[273] K. Pillutla, S. M. Kakade, and Z. Harchaoui. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.

[274] K. Pillutla, S. M. Kakade, and Z. Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

[275] F. Pinelli, G. Tolomei, and G. Trappolini. Flirt: Federated learning for information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3472–3475, 2023.

[276] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.

[277] A. Pyrgelis, C. Troncoso, and E. De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145*, 2017.

[278] PyTorch. Basic mnist example. `https://github.com/pytorch/examples/tree/master/mnist`, 2016.

[279] X. Qiu, T. Parcollet, J. Fernandez-Marques, P. P. B. de Gusmao, D. J. Beutel, T. Topal, A. Mathur, and N. D. Lane. A first look into the carbon footprint of federated learning, 2021.

[280] A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

[281] K. J. Rahman, F. Ahmed, N. Akhter, M. Hasan, R. Amin, K. E. Aziz, A. M. Islam, M. S. H. Mukta, and A. N. Islam. Challenges, applications and design aspects of federated learning: A survey. *IEEE Access*, 9:124682–124700, 2021.

[282] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.

[283] M. A. Ramirez, S.-K. Kim, H. A. Hamadi, E. Damiani, Y.-J. Byon, T.-Y. Kim, C.-S. Cho, and C. Y. Yeun. Poisoning attacks and defenses on artificial intelligence: A survey. *arXiv preprint arXiv:2202.10276*, 2022.

[284] S. Ranchordas. Experimental regulations for ai: sandboxes for morals and mores. *University of Groningen Faculty of Law Research Paper*, (7), 2021.

[285] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[286] M. H. u. Rehman, A. M. Dirir, K. Salah, E. Damiani, and D. Svetinovic. Trustfed: A framework for fair and trustworthy cross-device federated learning in iiot. *IEEE Transactions on Industrial Informatics*, 17 (12):8485–8494, 2021.

[287] M. H. u. Rehman, W. Hugo Lopez Pinaya, P. Nachev, J. T. Teo, S. Ourselin, and M. J. Cardoso. Federated learning for medical imaging radiology. *The British Journal of Radiology*, 96(1150):20220890, 2023.

[288] A. Renda, P. Ducange, F. Marcelloni, D. Sabella, M. C. Filippou, G. Nardini, G. Stea, A. Virdis, D. Micheli, D. Rapone, et al. Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking. *Information*, 13(8):395, 2022.

[289] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[290] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[291] P. Richtárik and M. Takáč. Distributed coordinate descent method for learning with big data. *The Journal of Machine Learning Research*, 17(1):2657–2681, 2016.

[292] J. Riffi, M. A. Mahraz, A. El Yahyaouy, H. Tairi, et al. Credit card fraud detection based on multilayer perceptron and extreme learning machine architectures. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–5. IEEE, 2020.

[293] S. Ritchie. Privacy impact assessment system and associated methods, Sept. 21 2017. US Patent App. 15/459,909.

[294] M. Robnik-Šikonja and I. Kononenko. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600, 2008.

[295] I. Rohlfing. Do you know your data? criteria for dataset quality, 2008.

[296] D. Roselli, J. Matthews, and N. Talagala. Managing bias in ai. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 539–544, 2019.

[297] M. Roszel, R. Norvill, and R. State. An analysis of byzantine-tolerant aggregation mechanisms on model poisoning in federated learning. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 143–155. Springer, 2022.

[298] M. Roszel, B. Fiz, and R. State. Flairs: Federated learning ai regulatory sandbox. In *Machine Learning and Knowledge Discovery in Databases: Workshop on ML, Law, and Society: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023*. Springer Nature, 2023.

[299] M. Roszel, R. Norvill, B. Fiz, J. Hilger, and R. State. Know your model (kym): Increasing trust in ai and machine learning. *Deployable AI (DAI) Workshop of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023)*, 2023.

[300] S. Ruggieri et al. Using t-closeness anonymity to control for non-discrimination. *Trans. Data Priv.*, 7 (2):99–129, 2014.

[301] S. J. Russell and P. Norvig. *Artificial intelligence a modern approach*. London, 2010.

[302] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[303] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala. The price of fair pca: One extra dimension. *Advances in neural information processing systems*, 31, 2018.

[304] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah. Distributed federated learning for ultra-reliable low-latency vehicular communications. *IEEE Transactions on Communications*, 68(2):1146–1159, 2019.

[305] H. Sanders and J. Saxe. Garbage in, garbage out: How purport-edly great ml models can be screwed up by bad data. *Proceedings of Blackhat 2017*, 2017.

[306] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357, 2014.

[307] A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving regression modelling via distributed computation. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 677–682, 2004.

[308] Y. M. Saputra, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, M. D. Mueck, and S. Srikanteswara. Energy demand prediction with federated learning for electric vehicle networks. In *2019 IEEE global communications conference (GLOBECOM)*, pages 1–6. IEEE, 2019.

[309] S. Saria and A. Subbaswamy. Tutorial: safe and reliable machine learning. *arXiv preprint arXiv:1904.07204*, 2019.

[310] I. H. Sarker. Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, 3(2):158, 2022.

[311] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

[312] S. Schelter, J.-H. Boese, J. Kirschnick, T. Klein, and S. Seufert. Automatically tracking metadata and provenance of machine learning experiments. In *Machine Learning Systems Workshop at NIPS*, pages 27–29, 2017.

[313] S. Sharma, C. Xing, Y. Liu, and Y. Kang. Secure and efficient federated transfer learning. In *2019 IEEE international conference on big data (Big Data)*, pages 2569–2576. IEEE, 2019.

[314] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364, 2020.

[315] D. Shin. User perceptions of algorithmic decisions in the personalized ai system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, 64(4):541–565, 2020.

[316] D. Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021.

[317] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[318] K. Siau and W. Wang. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2):47–53, 2018.

[319] S. Silva, B. A. Gutman, E. Romero, P. M. Thompson, A. Altmann, and M. Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 270–274. IEEE, 2019.

[320] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[321] B. Singh. Federated learning for envision future trajectory smart transport system for climate preservation and smart green planet: Insights into global governance and sdg-9 (industry, innovation and infrastructure). *National Journal of Environmental Law*, 6(2):6–17, 2023.

[322] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

[323] K. Sokol and P. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

[324] S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

[325] Y.-Y. Song and L. Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[326] R. Souza, L. Azevedo, V. Lourenço, E. Soares, R. Thiago, R. Brandão, D. Civitarese, E. Brazil, M. Moreno, P. Valduriez, et al. Provenance data in the machine learning lifecycle in computational science and engineering. In *2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*, pages 1–10. IEEE, 2019.

[327] R. Souza, L. Azevedo, R. Thiago, E. Soares, M. Nery, M. A. Netto, E. Vital, R. Cerqueira, P. Valduriez, and M. Mattoso. Efficient runtime capture of multiworkflow data using provenance. In *2019 15th International Conference on eScience (eScience)*, pages 359–368. IEEE, 2019.

[328] F. Sovrano, S. Sapienza, M. Palmirani, and F. Vitali. Metrics, explainability and the european ai act proposal. *J*, 5(1):126–138, 2022.

[329] J. Stremmel and A. Singh. Pretraining federated text models for next word prediction. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 477–488. Springer, 2021.

[330] E. Strumbelj and I. Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

[331] Z. Su, Y. Wang, T. H. Luan, N. Zhang, F. Li, T. Chen, and H. Cao. Secure and efficient federated learning for smart grid with edge-cloud collaboration. *IEEE Transactions on Industrial Informatics*, 18(2):1333–1344, 2021.

[332] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

[333] T. Suzumura, Y. Zhou, N. Baracaldo, G. Ye, K. Houck, R. Kawahara, A. Anwar, L. L. Stavarache, Y. Watanabe, P. Loyola, et al. Towards federated graph learning for collaborative financial crimes detection. *arXiv preprint arXiv:1909.12946*, 2019.

[334] T. Suzumura, Y. Zhou, R. Kawahara, N. Baracaldo, and H. Ludwig. Federated learning for collaborative financial crimes detection. In *Federated Learning: A Comprehensive Overview of Methods and Applications*, pages 455–466. Springer, 2022.

[335] A. Z. Tan, H. Yu, L. Cui, and Q. Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[336] K. Tan, D. Bremner, J. Le Kernec, and M. Imran. Federated machine learning in vehicular networks: A summary of recent applications. In *2020 international conference on UK-China emerging technologies (UCET)*, pages 1–4. IEEE, 2020.

[337] X. Tan, K. Xu, Y. Cao, Y. Zhang, L. Ma, and R. W. Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 30:9085–9098, 2021.

[338] The European Commission. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. The European Commission, 2021.

[339] The European Commission. *ANNEXES to the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. The European Commission, 2021.

[340] The European Commission. *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. The European Commission, 2023.

[341] S. Thiebes, S. Lins, and A. Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31:447–464, 2021.

[342] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu. Data poisoning attacks against federated learning

systems. In *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25*, pages 480–501. Springer, 2020.

[343] R. Toorajipour, V. Sohrabpour, A. Nazarpour, P. Oghazi, and M. Fischl. Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, 122:502–517, 2021.

[344] M. Y. Topaloglu, E. M. Morrell, S. Rajendran, and U. Topaloglu. In the pursuit of privacy: the promises and predicaments of federated learning in healthcare. *Frontiers in Artificial Intelligence*, 4:746497, 2021.

[345] J. Truby, R. D. Brown, I. A. Ibrahim, and O. C. Parellada. A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation*, 13(2):270–294, 2022.

[346] U. G. Union. Top 10 principles for ethical artificial intelligence. *The future world of work*, 2017.

[347] A. Vaid, S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. De Freitas, T. Wanyan, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with covid-19: machine learning approach. *JMIR medical informatics*, 9(1):e24207, 2021.

[348] P. Vanhaesebrouck, A. Bellet, and M. Tommasi. Decentralized collaborative learning of personalized models over networks. In *Artificial Intelligence and Statistics*, pages 509–517. PMLR, 2017.

[349] K. R. Varshney and H. Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3):246–255, 2017.

[350] P. Vassilakopoulou. Sociotechnical approach for accountability by design in ai systems. In *ECIS*, 2020.

[351] M. Veale and F. Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021.

[352] E. F. Villaronga, P. Kieseberg, and T. Li. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2):304–313, 2018.

[353] P. Voigt and A. v. d. Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017. ISBN 3319579584.

[354] R. T. Vought. Guidance for regulation of artificial intelligence applications, January 2020. URL `https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on -Regulation-of-AI-1-7-19.pdf`.

[355] S. Wachter, B. Mittelstadt, and L. Floridi. Transparent, explainable, and accountable ai for robotics. 2017.

[356] O. A. Wahab, G. Rjoub, J. Bentahar, and R. Cohen. Federated against the cold: A trust-based federated

learning approach to counter the cold start problem in recommendation systems. *Information Sciences*, 601:189–206, 2022.

[357] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *arXiv preprint arXiv:2007.05084*, 2020.

[358] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

[359] M. Wen, R. Xie, K. Lu, L. Wang, and K. Zhang. Feddetect: A novel privacy-preserving federated learning framework for energy theft detection in smart grid. *IEEE Internet of Things Journal*, 9(8):6069–6080, 2021.

[360] D. Wright. A framework for the ethical impact assessment of information technology. *Ethics and information technology*, 13:199–226, 2011.

[361] D. Wright, M. Friedewald, and R. Gellert. Developing and testing a surveillance impact assessment methodology. *International Data Privacy Law*, 5(1).

[362] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer, 2019.

[363] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5:1–19, 2021.

[364] Z. Xu, F. Yu, J. Xiong, and X. Chen. Helios: Heterogeneity-aware federated learning with dynamically balanced collaboration. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 997–1002. IEEE, 2021.

[365] R. V. Yampolskiy and M. Spellchecker. Artificial intelligence safety and cybersecurity: A timeline of ai failures. *arXiv preprint arXiv:1610.07997*, 2016.

[366] C. Yang, Q. Wang, M. Xu, Z. Chen, K. Bian, Y. Liu, and X. Liu. Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In *Proceedings of the Web Conference 2021*, pages 935–946, 2021.

[367] F. Yang, M. Z. Abedin, and P. Hajek. An explainable federated learning and blockchain-based secure credit modeling method. *European Journal of Operational Research*, 2023.

[368] H. Yang, H. He, W. Zhang, and X. Cao. Fedsteg: A federated transfer learning framework for secure image steganalysis. *IEEE Transactions on Network Science and Engineering*, 8(2):1084–1094, 2020.

[369] L. Yang, B. Tan, V. W. Zheng, K. Chen, and Q. Yang. Federated recommendation systems. *Federated Learning: Privacy and Incentive*, pages 225–239, 2020.

[370] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[371] W. Yang, Y. Zhang, K. Ye, L. Li, and C.-Z. Xu. Ffd: A federated learning based method for credit card fraud detection. In *Big Data–BigData 2019: 8th International Congress, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25–30, 2019, Proceedings 8*, pages 18–32. Springer, 2019.

[372] S. Yanisky-Ravid and S. K. Hallisey. Equality and privacy by design: A new model of artificial intelligence data transparency via auditing, certification, and safe harbor regimes. *Fordham Urb. LJ*, 46:428, 2019.

[373] D. Yin, Y. Chen, R. Kannan, and P. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

[374] X. Yin, Y. Zhu, and J. Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6):1–36, 2021.

[375] V. Yogarajan, B. Pfahringer, and M. Mayo. A review of automatic end-to-end de-identification: Is high accuracy the only metric? *Applied Artificial Intelligence*, 34(3):251–269, 2020.

[376] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

[377] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, 41(4):39–45, 2018.

[378] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, , and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[379] R. Zeng, C. Zeng, X. Wang, B. Li, and X. Chu. A comprehensive survey of incentive mechanism for federated learning. *arXiv preprint arXiv:2106.15406*, 2021.

[380] D. A. Zetzsche, R. P. Buckley, J. N. Barberis, and D. W. Arner. Regulating a revolution: From regulatory sandboxes to smart regulation. *Fordham J. Corp. & Fin. L.*, 23:31, 2017.

[381] Y. Zhan, J. Zhang, Z. Hong, L. Wu, P. Li, and S. Guo. A survey of incentive mechanism design for federated learning. *IEEE Transactions on Emerging Topics in Computing*, 10(2):1035–1044, 2021.

[382] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[383] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu. Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 2020)*, 2020.

[384] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[385] D. Y. Zhang, Z. Kou, and D. Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060. IEEE, 2020.

[386] J. Zhang, C. Li, J. Ye, and G. Qu. Privacy threats and protection in machine learning. In *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, pages 531–536, 2020.

[387] J. M. Zhang, M. Harman, L. Ma, and Y. Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1):1–36, 2020.

[388] L. Zhang, Y. Wu, and X. Wu. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1335–1344, 2017.

[389] M. Zhang, Y. Zhang, L. Zhang, C. Liu, and S. Khurshid. Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pages 132–142, 2018.

[390] S. Zhang, A. E. Choromanska, and Y. LeCun. Deep learning with elastic averaging sgd. *Advances in neural information processing systems*, 28, 2015.

[391] S. Zhang, J. Li, L. Shi, M. Ding, D. C. Nguyen, W. Tan, J. Weng, and Z. Han. Federated learning in intelligent transportation systems: Recent applications and open problems. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[392] T. Zhang, Q. Li, C.-s. Zhang, H.-w. Liang, P. Li, T.-m. Wang, S. Li, Y.-l. Zhu, and C. Wu. Current trends in the development of intelligent unmanned autonomous systems. *Frontiers of information technology & electronic engineering*, 18:68–85, 2017.

[393] W. Zhang and X. Li. Federated transfer learning for intelligent fault diagnostics using deep adversarial networks with data privacy. *IEEE/ASME Transactions on Mechatronics*, 27(1):430–439, 2021.

[394] Y. Zhang, Q. V. Liao, and R. K. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

[395] Z. Zhang, E. R. Sparks, and M. J. Franklin. Diagnosing machine learning pipelines with fine-grained

lineage. In *Proceedings of the 26th International Symposium on High-Performance Parallel and Distributed Computing*, pages 143–153, 2017.

[396] B. Zhao, K. R. Mopuri, and H. Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

[397] S. Zhao, F. Blaabjerg, and H. Wang. An overview of artificial intelligence applications for power electronics. *IEEE Transactions on Power Electronics*, 36(4):4633–4658, 2020.

[398] F. Zheng, K. Li, J. Tian, X. Xiang, et al. A vertical federated learning method for interpretable scorecard and its application in credit scoring. *arXiv preprint arXiv:2009.06218*, 2020.

[399] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang. Federated meta-learning for fraudulent credit card detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4654–4660, 2021.

[400] Z. Zheng, Y. Zhou, Y. Sun, Z. Wang, B. Liu, and K. Li. Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges. *Connection Science*, 34(1):1–28, 2022.

[401] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[402] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[403] L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

[404] Z. Zhu, J. Shu, X. Zou, and X. Jia. Advanced free-rider attacks in federated learning. In *the 1st NeurIPS Workshop on New Frontiers in Federated Learning Privacy, Fairness, Robustness, Personalization and Data Ownership*, 2021.