IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.

IJSRA

International Journal of Science and Research Archive

Research Journal Archive, INDIA

(Review Article)

Check for updates

Towards trustworthy AI: An analysis of the relationship between explainability and trust in AI systems

Vibhuti Choubisa [1, *] and Divyansh Choubisa [2]

[1] *Department of Computer Science and Engineering, Pacific Academy of Higher Education and Research University, India.*
[2] *Department of Informatics, Wilfrid Laurier University, Canada.*

International Journal of Science and Research Archive, 2024, 11(01), 2219–2226

Publication history: Received on 05 January 2024; revised on 14 February 2024; accepted on 17 February 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.11.1.0300

Abstract

As artificial intelligence (AI) becomes increasingly integral to our lives, ensuring these systems are trustworthy and transparent is paramount. The concept of explainability has emerged as a crucial element in fostering trust within AI systems. Nevertheless, the dynamics between explainability and trust in AI are intricate and not fully comprehended. This paper delves into the nexus between explainability and trust in AI, offering perspectives on crafting AI systems that users can rely on. Through an examination of existing literature, we investigate how transparency, accountability, and human oversight influence trust in AI systems and assess how various explainability approaches contribute to trust enhancement. Utilizing a set of experiments, our research examines how different explanatory models impact users' trust in AI systems, revealing that the nature and quality of explanations have a significant influence on trust levels. Additionally, we scrutinize the balance between explainability and accuracy in AI systems, discussing its implications for the development of reliable AI. This study underscores the critical role of explainability in engendering trust in AI systems, providing guidance on the development of AI systems that are both transparent and trustworthy, thereby fostering confidence among users.

Keywords: AI; Robustness; Fairness; Explainability; Privacy; Accountability; Trust; Trustworthiness; Machine Learning; Neural Networks

1. Introduction

The potential applications of artificial intelligence (AI) range from healthcare and education to transportation and entertainment. However, there is growing scepticism regarding the reliability of AI systems as they proliferate. In order for AI to be trusted by individuals and society as a whole, it must be transparent, explainable, accountable, and ethical. Achieving trustworthy AI is a complex and multifaceted challenge that requires addressing issues related to data privacy, bias, fairness, explainability, accountability, and more. In this paper, we provide a comprehensive review of the challenges and solutions related to building trustworthy AI systems [1].

People are reaping major economic and societal benefits from the growing development of artificial intelligence (AI). There is a rising public awareness that we need to make these systems trustworthy as a result of the extensive application of AI in sectors including entertainment, transportation, finance, medical, and security. This is because, given how commonplace these AI systems are, a breach of stakeholder confidence could have negative social effects. These flaws can render current AI systems useless and raise doubts about their dependability, which has become a major barrier for AI to overcome in order to grow as a field, get wider acceptance, and provide more economic value. As a result, both academics and business are now concentrating on how to create trustworthy AI systems. XAI or Explainable Artificial Intelligence constitutes a subset of AI dedicated to crafting AI models capable of presenting comprehensible and lucid explanations for their decision-making procedures. The primary objective of XAI revolves around enhancing

[*] Corresponding author: Vibhuti Choubisa

the reliability, accountability, and transparency of AI systems. By furnishing explicit elucidations of the reasoning behind an AI model's specific decision, XAI endeavours to foster trust in the interaction between humans and AI.

The establishment of trust stands as a pivotal element in the widespread adoption and acceptance of AI systems. Its significance lies in ensuring that AI technologies are employed in manners consistent with human values and ethical standards. Without a foundation of trust, individuals might hesitate to utilize AI systems or harbor skepticism towards their decision-making capabilities. Therefore, comprehending the intricate connection between explainability and trust within AI systems emerges as crucial for advancing their acceptance and integration.

2. Trustworthy AI - Basic Outlook

Trustworthy AI denotes to the expansion and employment of artificial intelligence systems that are translucent, explainable, fair, safe, secure, and accountable. In other words, it refers to AI systems that can be trusted to make decisions and take actions that align with human values and ethical principles.[3] Here are some key elements of trustworthy AI:

- Transparency: AI systems must be created so that people may comprehend how they function and how they make judgements.
- Explainability: AI systems must be able to explain their choices and actions in a way that is both obvious and understandable.
- Fairness: AI systems ought to be created without prejudice or discrimination and to treat all people and groups equally.
- Safety: AI systems should be made to work safely and without endangering people or the environment.
- Security: AI systems must be created to be safe from hacking or other unauthorised access.
- Responsibility: Mechanisms for supervision and responsibility should be implemented for AI systems to make sure they are utilised in ways that are consistent with ethical and legal norms.

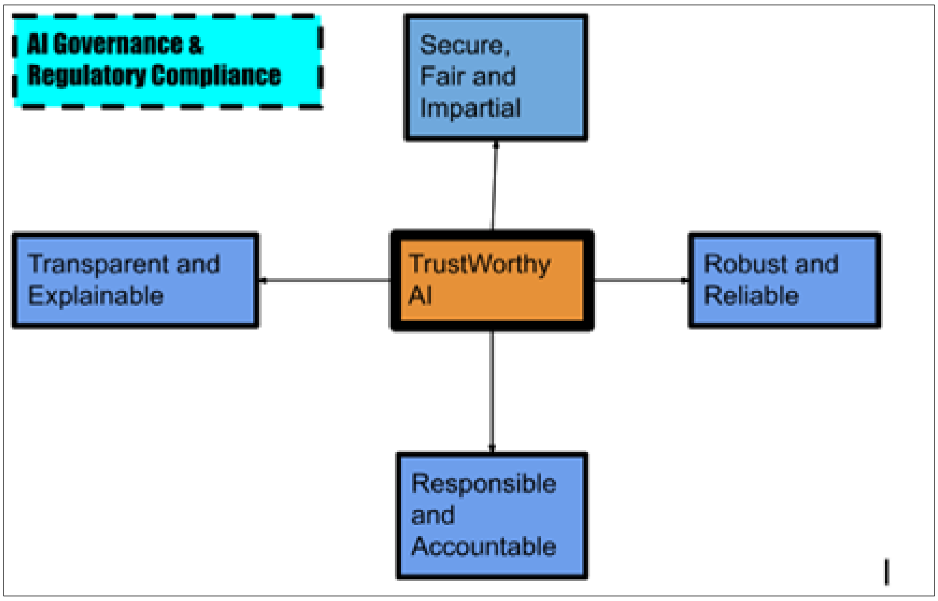
AI Governance & Regulatory Compliance
Secure, Fair and Impartial
Transparent and Explainable
TrustWorthy AI
Robust and Reliable
Responsible and Accountable

Figure 1 Elements of Trustworthy AI

By incorporating these elements into the development and deployment of AI systems, ensuring the trustworthiness and alignment of AI systems with human values is imperative. As AI becomes increasingly ingrained in our daily lives, its impact extends across various domains, encompassing decisions made by autonomous vehicles and recommendations offered by healthcare systems. Consequently, establishing confidence in the responsible and ethical conduct of AI systems is critical for their reliable operation and acceptance.

3. Explainable AI - A Subjective Evaluation And Theoretical Approaches

Artificial intelligence (AI) has rapidly evolved in recent years, and its applications now span across various fields, including healthcare, finance, and transportation. However, the black-box nature of many AI models is a significant challenge to their adoption. Many AI models are opaque, making it difficult to understand how they make decisions and which factors they consider. This lack of transparency is a significant concern, particularly in critical applications such as healthcare, where the decisions made by AI models can have significant consequences. Explainable AI (XAI) seeks to address this challenge by making AI models more transparent and understandable.[4]

A branch of artificial intelligence called explainable AI aims to create AI models that can explain their decision-making procedures in simple and intelligible terms. In other words, explainable AI attempts to improve the human ability to interpret AI models. Enhancing the reliability, accountability, and transparency of AI systems is the aim of XAI. XAI can aid in increasing trust between people and AI by clearly articulating how an AI model arrived at a given conclusion.

There are several methods for making AI models more comprehensible. Using rule-based systems, which base choices on a set of predetermined rules, is one strategy.These systems are relatively easy to understand and explain since the rules used to make decisions are explicitly defined. However, rule-based systems have limited flexibility and may not perform well in complex environments.

Another approach is to use model-based systems that rely on mathematical models to make decisions. These models can be more complex than rule-based systems, but they offer greater flexibility and can handle more complex environments. However, model-based systems can be difficult to interpret, making it challenging to understand how they arrived at their decisions.

Recent progress in machine learning, especially in the realm of deep learning, has enabled the creation of increasingly sophisticated AI models that learn directly from data. Despite their capabilities, the complexity inherent in deep learning models often makes their decision-making processes opaque, due to the intricate mathematical operations they perform. In response to this issue, a variety of methods have been devised to make these models more interpretable, including the use of saliency maps, determining feature significance, and providing counterfactual narratives. Saliency maps are utilized to pinpoint the specific areas within an image that significantly influence the decision of a deep learning model.

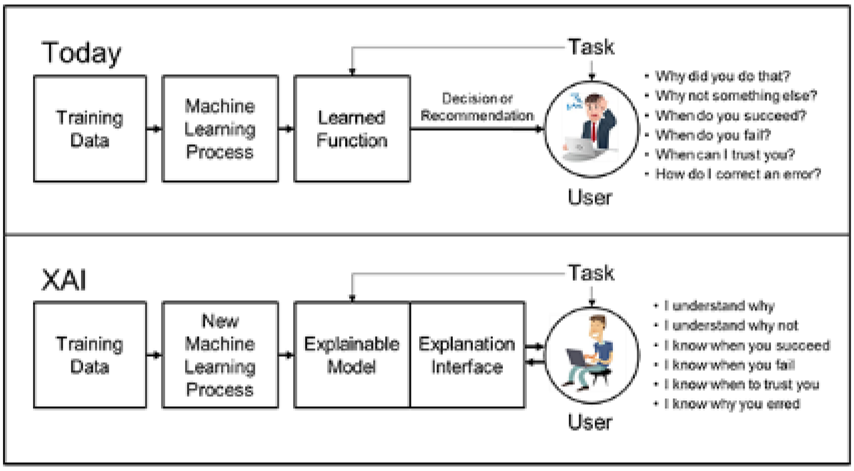

Figure 2 Explainable AI

Methods assessing feature significance aim to identify the crucial elements that a model considers when making a decision. Meanwhile, counterfactual narratives illustrate how alterations in the model's input could lead to different outcomes. Moreover, several metrics have been established to evaluate the efficacy of explainable AI (XAI) frameworks. These metrics, which include measures of accuracy, completeness, consistency, and sufficiency, are vital in appraising the quality of an AI model's explanations. Accuracy measures how correct an explanation is, while completeness gauges the depth of detail in an explanation. Consistency evaluates how well an explanation matches with existing knowledge, and sufficiency assesses the quantity of information an explanation provides.

Explainable AI is an essential field of study aimed at demystifying AI models, making them more accessible and understandable to humans. This endeavor to elucidate AI decisions is crucial for fostering trust between humans and artificial intelligence systems. Although various strategies have been proposed to enhance AI explainability, propelled by advancements in machine learning, the journey towards developing robust evaluation metrics for XAI systems that deliver precise, detailed, consistent, and ample explanations is ongoing.

4. Trust and Explanation - Computational Models and Dimensional Conflicts

Artificial intelligence (AI) has revolutionised various industries, from healthcare to finance, and transportation. However, the adoption of AI has been limited by the black-box nature of many AI models, which makes it difficult to understand how they make decisions and which factors they consider. The lack of transparency is a significant concern, particularly in critical applications such as healthcare, where the decisions made by AI models can have significant consequences. To address this issue, researchers have developed two critical subfields of AI: Trustworthy AI and Explainable AI.

Trustworthy AI (TAI) refers to the development and deployment of AI systems that are transparent, explainable, fair, safe, secure, and accountable. In other words, it refers to AI systems that can be trusted to make decisions and take actions that align with human values and ethical principles. TAI seeks to ensure that AI models can be trusted to operate in a responsible and ethical manner while minimising the risk of harm to humans and the environment.[6]

On the other hand, Explainable AI (XAI) focuses on creating AI models that can explain its decision-making procedures in a straightforward and intelligible manner. XAI aims to increase the accountability, openness, and trustworthiness of AI systems by making AI models more transparent and intelligible. The purpose of XAI is to foster trust between people and AI by clearly articulating how an AI model arrived at a given conclusion.

TAI and XAI have similar objectives, although they address distinct aspects of the same issue. TAI concentrates on the moral and social aspects of AI, whereas XAI concentrates on the technological aspect. TAI works to make sure that AI models are developed and used in a way that is consistent with moral and ethical standards. XAI seeks to provide clear explanations of how AI models arrive at their decisions.

However, there are potential conflicts between these two dimensions. For example, ensuring the safety of an AI model may require limiting its accuracy or complexity, which may reduce its performance. Similarly, ensuring the transparency of an AI model may require simplifying its architecture, which may reduce its accuracy or complexity. These conflicts highlight the need for a balanced approach to building trustworthy and explainable AI models.

To address these conflicts, researchers have developed computational models that combine TAI and XAI. One such model is the Fairness, Accountability, and Transparency (FAT) model, which seeks to ensure that AI models are transparent, explainable, and fair while minimising the risk of harm to humans and the environment. The FAT model emphasises the importance of considering the social and ethical implications of AI models during their development and deployment.

Another model is the Human-in-the-Loop (HITL) model, which involves human experts in the decision-making process of AI models. The HITL model seeks to combine the strengths of AI models and human experts to arrive at more accurate and reliable decisions. By involving human experts in the decision-making process, the HITL model can provide clear explanations of how AI models arrive at their decisions while ensuring that these decisions align with human values and ethical principles.

4.1. XAI models

There are several XAI (Explainable AI) models, each with its own strengths and limitations. Some of the most commonly used XAI models are:

- The LIME (Local Interpretable Model-Agnostic Explanations) method can explain the results of any machine learning model. By roughly simulating the behaviour of the model in a constrained, comprehensible region surrounding each prediction, LIME generates local explanations for specific predictions.
- SHAP (SHapley Additive exPlanations) is another method that is model-independent and may be used to interpret the results of any machine learning model. Each feature in a forecast is given a value by SHAP,

indicating how much it contributed to the prediction. The Shapley values idea from cooperative game theory serves as the foundation for SHAP.

- Decision Trees: A traditional XAI model that can describe a model's decision-making process is the decision tree. Decision trees build a tree-like model of decisions and their possible consequences, making it easy to understand how a model arrived at a particular decision.
- Rule-based models: Rule-based models are another XAI model that builds a set of rules that describe the decision-making process of a model. Rule-based models are often used in expert systems to explain the reasoning behind a decision.
- Neural Networks with Attention Mechanisms: Neural networks with attention mechanisms are a class of XAI models that can explain which features the model is paying attention to when making a prediction. Attention mechanisms assign weights to different parts of the input, indicating their relative importance to the prediction.
- Counterfactual Explanations: Counterfactual explanations are XAI models that generate explanations for individual predictions by identifying the minimal changes to the input that would change the prediction. Counterfactual explanations can be used to explore "what-if" scenarios, making them valuable for decision-making.
- Prototypes and Criticisms: Prototypes and criticisms are XAI models that generate explanations for individual predictions by comparing them to prototypes and criticisms. Prototypes are examples that represent the decision boundary of the model, while criticisms are examples that are misclassified by the model. By comparing a prediction to prototypes and criticisms, these XAI models can provide insights into how the model arrived at the prediction.

However, It is challenging to list all trustworthy AI and Explainable AI models as there are many models available, each with its own strengths and limitations. However, here are a few examples of trustworthy AI and XAI models, along with some observations on their use:

- Random Forest: A trusted machine learning technique, Random Forest is frequently employed in AI systems. It is an approach to ensemble learning that assembles several decision trees and aggregates their forecasts. Random Forest is renowned for being reliable and adaptable. It might, however, give up some interpretability in the name of accuracy.
- Gradient Boosting Machines (GBMs): Another machine learning approach frequently utilised in reliable AI systems is the GBM. GBMs construct a number of decision trees sequentially, with each tree learning from the mistakes of the one before it. GBMs are renowned for their reliability and accuracy. Similar to Random Forest, they might give up some interpretability for accuracy, though.
- Explainable Neural Networks (XNNs): XNNs are a class of neural networks that are designed to be more interpretable than traditional neural networks. XNNs use techniques such as attention mechanisms and sparse activations to highlight the most critical features in the input. XNNs are useful in applications where interpretability is critical, such as healthcare and finance.
- Local Interpretable Model-Agnostic Explanations (LIME): LIME is a XAI model that can explain any machine learning model's predictions. By roughly simulating the behaviour of the model in a constrained, comprehensible region surrounding each prediction, LIME generates local explanations for specific predictions. LIME is helpful in fields like healthcare where model explainability is important.
- SHapley Additive exPlanations (SHAP): SHAP is an additional XAI model that can explain any machine learning model's predictions. Each feature in a forecast is given a value by SHAP, indicating how much it contributed to the prediction. The Shapley values idea from cooperative game theory serves as the foundation for SHAP. In applications like banking where feature importance is crucial, SHAP is helpful.

Observations

Here are Some following Observations made on different Trustworthy AI and Explainable Al Models:-

- Reliable AI and XAI models are necessary to make sure that AI systems are open, accountable, and consistent with moral and ethical standards.
- The issue domain, the volume and complexity of the data, and the level of interpretability required all play a role in the decision of which AI model to use.
- There is often a trade-off between interpretability and performance, and researchers and practitioners must carefully balance these factors when choosing an AI model.
- AI models must be designed and evaluated with fairness and bias mitigation in mind to ensure that they produce fair and unbiased outcomes.

- The development of trustworthy AI and XAI models requires collaboration between AI researchers, domain experts, policymakers, and end-users to ensure that the models align with human values and ethical principles.

5. Conflicts Between Trustworthy AI and Explainable AI

Trustworthy AI and Explainable AI are both critical concepts in ensuring that AI systems are transparent, accountable, and aligned with human values and ethical principles. However, there can be conflicts between these two concepts when designing and developing AI systems.

A notable challenge stems from the inherent trade-off between interpretability and performance in machine learning models. Certain models, like deep neural networks, can attain impressive accuracy levels but prove challenging to interpret. On the flip side, models such as decision trees and linear models offer enhanced interpretability but might compromise on accuracy. Striking the appropriate equilibrium between interpretability and performance becomes paramount in the development of AI systems that are both trustworthy and explainable [7].

Another potential conflict arises from the complexity of the data and the models used in AI systems. Large and complex data sets may require more sophisticated models to achieve high accuracy. However, these models may be more challenging to interpret, making it difficult to understand how the system arrived at a particular decision. In such cases, it is essential to use XAI techniques to generate explanations for the model's predictions.

A third potential conflict arises from the need to protect user privacy and proprietary information. In some cases, providing explanations for an AI system's decision-making process may reveal sensitive information about the user or the organisation. In such cases, it may be necessary to use XAI techniques that provide only partial or aggregated explanations to preserve privacy and confidentiality.

Also, conflicts between trustworthy AI and Explainable AI can arise from the trade-off between interpretability and performance, the complexity of the data and models used in AI systems, and the need to protect user privacy and proprietary information. Addressing these conflicts requires careful consideration of the problem domain, the data, the models, and the stakeholders involved in the AI system's development and deployment.

6. Solutions for Achieving Trustworthy AI

To address the challenges associated with achieving trustworthy AI, a variety of solutions have been proposed. These solutions include:

- Data Anonymization: Techniques for removing personal information from the data used to train AI models are available. This can aid in addressing concerns with bias and data quality.
- Model Interpretability: To make AI systems more understandable, model interpretability techniques can be applied. These methods can assist users in comprehending the reasoning behind a given choice made by an AI system.
- Algorithmic Fairness: Techniques for ensuring algorithmic fairness can be used to make sure AI systems are impartial and fair. These methods can assist in addressing difficulties with bias and fairness.
- Ethical Principles: AI system development and implementation can be governed by ethical principles.

These guidelines can help to ensure that AI systems are designed in an ethical and responsible manner.

6.1. Future Prospects

A reliable AI system should steer clear of discriminating actions in interactions with humans and make sure that decisions are made fairly for all persons and groups. More and more evidence suggests that AI systems exhibit human-like discriminating prejudice or make unjust judgements as they are rapidly becoming a part of our daily life.

Different evaluation measures are important for examining explanation techniques. However, determining whether the explanations are plausible and accurate with relation to specific predictions has become impossible due to a lack of information and human subjectivity.

The outlook for explainable AI (XAI) and reliable AI appears very promising. As AI technologies evolve and become more integrated into various aspects of our lives, there's an increasing emphasis on the need for AI systems to be ethical, responsible, and trustworthy.

In the domain of trustworthy AI, we anticipate ongoing initiatives aimed at creating AI systems that are transparent, accountable, equitable, and respectful of privacy and human dignity. This will include continuous research into innovative methods to guarantee the reliability and safety of AI technologies, alongside the formulation of new standards and best practices for the development and implementation of AI.

In the realm of XAI, we can expect to see continued progress in developing techniques for explaining the decision-making processes of AI algorithms in a way that is understandable to humans. This will involve ongoing research into new approaches for visualizing and explaining complex data structures and algorithms, as well as the development of new tools and platforms for enabling human-AI collaboration.

In general, the potential for both trustworthy AI and XAI is deeply intertwined with the broader social and ethical implications of AI technology. As AI becomes more pervasive across different sectors of society, the imperative to develop and deploy AI systems in an ethical, responsible, and dependable manner grows ever more critical. By adhering to these principles, we aim to harness AI's potential to benefit society at large, rather than a select few.

7. Conclusion

In conclusion, Trustworthy AI (TAI) and Explainable AI (XAI) stand as crucial subfields within the realm of artificial intelligence, both aiming to address the opaque nature of numerous AI models. TAI is dedicated to exploring the ethical and societal dimensions of AI, while XAI focuses on the technical aspects. However, potential conflicts between these dimensions underscore the necessity for a balanced approach in constructing AI models that are both trustworthy and explainable. Computational models like FAT (Fairness, Accountability, and Transparency) and HITL (Human-in-the-Loop) can play a pivotal role in harmonizing TAI and XAI. These models contribute to ensuring that AI systems are transparent, explainable, fair, secure, and accountable, while simultaneously minimizing risks to humans and the environment. The significance of Trustworthy AI and XAI extends to the ongoing evolution and implementation of artificial intelligence. Trustworthy AI involves the creation and deployment of AI systems that prioritize reliability, transparency, fairness, and respect for privacy and human values. In contrast, XAI focuses on developing techniques that clarify the decision-making processes of AI algorithms in a comprehensible manner for humans.

Both Trustworthy AI and XAI play indispensable roles in guaranteeing the ethical, responsible, and trustworthy development and deployment of AI. Adhering to the principles of trustworthy AI and implementing XAI techniques are essential steps in ensuring that AI serves the greater good of society rather than catering to a select few. As AI continues to advance and integrate into various aspects of our lives, prioritizing these principles and techniques becomes increasingly vital to ensure the safe, transparent, and value-conscious development and deployment of AI.

Compliance with ethical standards

Disclosure of conflict of interest

The authors report no declarations of interest.

References

[1] Nitin Agrawal, Ali ShahinShamsabadi, Matt J. Kusner, and AdriàGascón. 2019. QUOTIENT: Two-party secure neural network training and prediction. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. 1231–1247.

[2] AI HLEG, European Commission. 2021. Ethics guidelines for trustworthy AI.

[3] AI HLEG, European Commission. 2020. Assessment list for trustworthy artificial intelligence (altai) for self assessment.

[4] Alberto Aimar. 1998. Introduction to software documentation.

[5] Naveed Akhtar and AjmalMian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. Ieee Access 6 (2018), 14410–14430.

[6] Arjun R. Akula, KezeWang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, SinisaTodorovic, Joyce Chai, and Song-Chun Zhu. 2022. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. Iscience 25, 1 (2022), 103581. 2021.

[7] Federated AI Technology Enabler. https://fate.fedai.org/. 2022. LEAF: A Benchmark for Federated Settings. https://leaf.cmu.edu/. 2022. A list of Homomorphic Encryption libraries, software or resources. https://github.com/jonaschn/awesome-he. 2023. A list of MPC software or resources. https://github.com/rdragos/awesome-mpc. 2023. OpenDP: Open Source Tools for Differential Privacy. https://opendp.org/. 2023. Opacus: Train PyTorch models with Differential Privacy. https://opacus.ai/. 2023. Paddle Federated Learning. https://github.com/PaddlePaddle/PaddleFL.