Blockchain-Enabled Multi-stage Incentive Framework for Federated Learning

by **Han Xu**

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dr. Priyadarsi Nanda and Dr. Jie Liang

School of Electrical and Data Engineering

Faculty of Engineering and IT

University of Technology Sydney

January 31, 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Han Xu, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: January 31, 2025

Abstract

In federated learning scenarios where both data and model owners make significant contributions, traditional incentive models often fail to fairly assess the value of these various inputs – especially intangible efforts. In this research, we address this critical gap by introducing a novel framework that combines a multi-stage incentive mechanism, a blockchain-based clearing protocol, and a contribution buy-back method. The multi-stage incentive mechanism optimises compensation based on both quantifiable and unquantifiable contributions from data and model owners. At the same time, the blockchain-based clearing protocol facilitates trustless reward distribution, model ownership transfer, and cost-effective settlements. In addition, the framework is also compatible with various existing contribution assessment mechanisms through the contribution buyback method, mitigating risks arising from data and model incompatibility and promoting reliable collaboration. This research significantly advances federated learning by promoting fair compensation, security, and ethical practices, enabling broader adoption across various domains. For example, in healthcare, this approach can enable secure and equitable collaboration between hospitals, healthcare facilities, and machine learning experts, advancing goals like predictive analytics while ensuring data privacy and regulatory compliance.

Acknowledgements

In my acknowledgments I would like to extend my deepest gratitude to individuals, organizations, and sponsors whose support and contributions have been invaluable to my research journey.

To Dr. Priyadarsi Nanda and Professor Sean He, my current and past supervisors, whose encouragement and wisdom helped me navigate through periods of self-doubt. I will always be grateful for the pivotal role you've played in my academic and personal growth. The confidence you two have placed in me gave me the courage to embark on this journey, and your guidance and mentorship have been invaluable.

To A/Professor Christy Liang, my co-supervisor, for being a source of comfort and guidance during the most challenging of times. Your anchoring theory resonated deeply and helped me find my anchor point. I'm forever grateful for the wisdom and insight you shared with me throughout the past 2 years.

Not to forget the University of Technology Sydney (UTS) and the Australia Research Council for their generous sponsorship and support of my research. Thank you for your contributions, which have provided me with the resources and opportunities necessary to pursue and complete this work.

And of course, to my family, my wife Charmaine and my two kids Michael and Justin, for being the cornerstones in my journey. I can't thank you enough for providing me with courage and strength even during the toughest of times.

Han Xu

January 31, 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

With the rapid advancement of machine learning and artificial intelligence technologies, there is a growing reliance on big data, the cornerstone for building high-accuracy models. Data-driven approaches have proven their effectiveness and potential in various fields, including image recognition, natural language processing, medical diagnostics, and financial analytics. However, there is increasing tension between the demand for extensive data and the stringent data privacy protection regulations worldwide, such as the General Data Protection Regulation (GDPR)[1] in the European Union and the Consumer Privacy Act (CCPA)[2] in California, USA. These laws reflect the public's concern for privacy protection and impose limitations on the free access to and use of personal data, presenting new challenges to data-driven research and applications.

In this context, traditional centralized machine learning methods, where data is typically aggregated at a single point for processing, may become infeasible or breach regulations, especially in sectors dealing with sensitive information. Consequently, there is a clear need for an innovative machine-learning approach that balances the vast data requirements with individual privacy protection.

Federated learning has emerged as an innovative learning paradigm designed to address the complex privacy issue in data-driven models. Initiated by McMahan et

al.[3], [4] at Google to solve the problem of optimizing keyboard input for individual users, this approach enables collaboration between multiple data owners to collectively train shared models without requiring direct data exchange, thereby preserving sensitive information. In federated learning, the model owner provides the initialized global model, sets the training agenda, and oversees the overall learning process. Data owners collaboratively train a model without having to reveal their sensitive data. They receive an initialized global model from the model owner, utilize their private data to improve the model locally, and then upload their updated model parameters to the federated learning server managed by the model owner. The model owner is responsible for collecting these updated parameters from all participating data owners, integrating them into a new, enhanced global model, and then redistributing this updated model back to the data owners for further refinement and training. This cyclical process continues until the model's accuracy reaches a standard satisfactory to all participants.

Through its unique structure, federated learning effectively addresses the issue of data isolation inherent in traditional methodologies while maintaining adherence to the highest privacy protection standards. This method has shown significant promise in various sectors, including healthcare, finance, and insurance. For example, in the healthcare sector, federated learning enables the aggregation of insights from diverse institutional data, improving diagnostic accuracy and treatment outcomes while meticulously protecting patient privacy.

However, the success of federated learning heavily relies on the active and qualitative participation of all parties involved, including the provision of high-quality and ample training data from data owners and the model owners' efforts in model design and parameter tuning, along with significant computation, network, and power resources consumed by the training. Behind these contributions are substantial costs. Furthermore, participating in federated learning can expose individuals to security risks, such as the potential for intermediate gradients to reveal important training data information, as demonstrated by Song et al.[5]. There is also the risk that curious parameter servers might uncover private data details using generative adversarial networks.

Due to these financial, computational, and security-related concerns, participants may hesitate to contribute effectively to federated learning tasks unless they receive adequate rewards to justify their involvement. Therefore, developing well-designed and fair incentive mechanisms is crucial to ensure that all parties are appropriately compensated for their contributions and risks. This is especially important as the different types of federated learning may require tailored ways to evaluate the value of data and computation. Without fair incentives, data owners may be reluctant to share high-quality data, and model owners could lose motivation to engage fully in the process, impacting the effectiveness of federated models. This thesis explores the challenges related to incentives in federated learning. It proposes strategies to address these challenges, aiming to encourage broader adoption and successful implementation of this collaborative learning approach.

1.2 Research Problems

Current incentive mechanisms in federated learning often presuppose a dominant model owner who dictates reward structures. However, in open and competitive markets, this model breaks down. Consider a scenario where healthcare providers lacking comprehensive machine learning expertise band together to form the 'Data Owners Alliance.' They aim to develop AI-powered predictive analytics for personalized care delivery while safeguarding individual patient data privacy. To achieve this, they partner with a machine learning professional services entity — either a research organization or a consulting firm, referred to as the 'Model Owner'— specializing in models robust to varying data streams (wearables, in-house diagnostics, etc.) that enable rapid health pattern detection. Aiming to gain a competitive edge in their field, both parties strive to maximize profits by pooling resources, including sensitive patient data, clinical knowledge, and capital outlay - all vital to successful predictive health applications. The data owners contribute training data and capital while seeking model owners capable of not only handling diverse data but integrating it into real-time health monitoring under strict regulatory limitations. Upon project success, the model owner receives a predetermined training incentive, funded by the data owners, and transfers ownership of the collaboratively trained

model to the alliance following a predetermined equity distribution structure.

The risk-averse nature of healthcare mandates a federated learning scenario with nuanced incentives to offset the uncertainty of achieving the desired model performance. The lack of technical expertise on the part of the data owner further fuels the need to justly incentivize the model owner, whose optimization efforts may not be fully visible during the collaboration, while data owners confront the challenge of demonstrating the value of their data contributions match their financial investment. This multi-faceted goal structure makes designing fair and sustainable incentives incredibly complex.

To succeed, both parties must pool sensitive patient data, clinical knowledge, capital, and technical expertise. However, without the right incentives, participants may hesitate to share their most valuable resources, jeopardising the project's potential. In such multi-faceted collaborations, the timing of rewards becomes crucial. Multi-stage incentives are particularly important because they address the risks faced by both parties. Data owners can assess model quality incrementally, reducing the risk of investing resources in a biased or inaccurate model. Model owners receive compensation throughout the process, mitigating the risk of non-payment after significant effort. This framework formed the foundation for my research into optimal incentive contracts.

The thesis addresses the following interlinked research problems:

- **Problem 1:** Valuing Strategic Contributions and Mitigating Moral Hazard: How can incentive mechanisms be designed to fairly evaluate and incentivize the unquantifiable strategic contributions of data owners and model owners in multi-stage federated learning? This must consider the ethical implications of the valuation method and the risks associated with collaborative decision-making.
- **Problem 2:** Designing a Fair and Efficient Liquidation Protocol: How to achieve transparent and secure reward liquidation and transfer model ownership in multi-stage federated learning? The protocol must ensure fairness for all participants while minimizing computational overhead in resource constrained environments.

- **Problem 3:** Addressing Data-Model Mismatches: Due to objective data heterogeneity, there may be a gap between the data owner's prior estimate of the contribution value of their data and the actual contribution value during the model training process. How can we manage the risks of such data-model mismatches? These risk management mechanisms should promote transparency within collaborative projects and enable fair course corrections when pre-agreed performance indicators are not met.

1.3 Contributions

This thesis advances the field of federated learning incentives by developing and analysing mechanisms that prioritise fairness, efficiency, and sustained participation in open collaboration scenarios. Specific contributions include:

- **Theoretical Mechanism for Valuing Strategic Contributions:** In multi-stage federated learning, where incentive contracts are established upfront, we delve into the dynamic between data owners and model owners. We employ a multi-stage game theory model that accounts for moral hazard to understand this relationship. Federated learning is rife with information asymmetries. "Hidden information" makes it difficult for data owners to discern a model owner's true capabilities, leading to adverse selection. "Hidden actions" obscure the model owner's work ethic, creating moral hazard. Data owners can combat moral hazard by breaking federated learning into stages, each with its reward. We focus on this moral hazard: how model owners might shirk their responsibilities and how data owners can use staged incentives to promote diligence. This research is crucial for successful federated learning outcomes. We analyse the interplay between data and model owners when the latter's efforts cannot be directly observed. Our multi-stage game model assumes a contract established at the start, in contrast to repeated or single-stage interaction models commonly used in federated learning frameworks. This captures the 'pay-it-after' nature of instalment-based incentives. The critical insight of this contribution is that, within a federated learning scenario led by the data owner, the optimal incentive scheme is one where as much of the incremental

value of the model created by the model owner as possible is paid back to it in the later stages. This ensures that success in later training stages depends on success in earlier ones, incentivising effort throughout. (Chapter 3)

- **Secure Two-Party Clearing for Multi-Stage Federated Learning:** This thesis also introduces a novel, efficient blockchain-based clearing protocol that leverages smart contracts to implement the proposed multi-stage incentive model. A key challenge for fair settlement in multi-stage federate learning is to ensure consensus on the choice of test dataset for model performance validation and to ensure that validation results on this dataset are acceptable to both parties at each stage. Considering privacy protection issues, finding a trusted third-party verifier to run the trained federated learning model on the verification data set and settle the rewards and the ownership of the model is costly and sometimes impossible. This protocol addresses these challenges by enabling efficient two-party verification (data owner and model owner). It strategically combines smart contracts and cryptographic techniques to protect the trained model's confidentiality until fair payment is confirmed. The protocol minimizes the computational load on the blockchain by offloading validation to the model and data owners. In the absence of complaints, the blockchain stores minimal information; in the presence of disputes, only the necessary components are recalculated. This design makes deployment on public blockchains (like Ethereum) cost-effective for diverse federated learning scenarios. A complete security analysis demonstrates the protocol's robustness against potential adversarial behavior. (Chapter 4)

- **Reciprocal Federated Learning Framework (RFLF) and Equitable Contribution Valuation:** This thesis introduces the Reciprocal Federated Learning Framework (RFLF), a pioneering approach to address the challenges of fair compensation and continued motivation in collaborative federated learning projects. It is particularly well-suited for scenarios where privacy concerns necessitate federated learning and where participants have varying levels of resources and expertise, such as in healthcare collaborations. The RFLF tackles information asymmetry and the risk-averse nature of certain domains by dynamically aligning incentives with contributions throughout the project. Crit-

ically, it incorporates both data contributions and financial investments, promoting balanced participation and recognising the value of diverse resource types. The RFLF includes a unique compensation mechanism that promotes both equitable reward distribution and enhanced project outcomes. If a data owner discovers their data is underperforming, they can opt-out to avoid further expenses. Their pre-deposited buy-back funds then compensate other data owners who increase their contributions. This creates positive feedback where capital offsets data shortfalls, ensuring fair pay for all participants while maintaining overall project incentives. By fostering trust in the fairness of the system, the RFLF encourages honest data quality assessment and maximal effort, ultimately improving the federated learning process's social utility towards the goals of privacy, reduced data silos, and overall model robustness. (Chapter 5)

1.4 Publications

This thesis builds upon and expands the research presented in the following publications, where I am listed as the author:

1. H. Xu *et al.*, "Designing incentive mechanisms for fair participation in federated learning," in *2023 IEEE International Conference on High Performance Computing and Communications*, IEEE, 2023, pp. 357–373 (Chapter 2)

2. H. Xu *et al.*, "The force of compensation, a multi-stage incentive mechanism model for federated learning," in *International Conference on Network and System Security*, Springer, 2022, pp. 357–373 (Chapter 3)

3. H. Xu *et al.*, "Fch, an incentive framework for data-owner dominated federated learning," *Journal of Information Security and Applications*, vol. 76, p. 103 521, 2023 (Chapter 4)

4. H. Xu *et al.*, "Reciprocal federated learning framework: Balancing incentives for model and data owners," *Future Generation Computer Systems*, vol. 161, pp. 146–161, 2024 (Chapter 5)

1.5 Thesis Outline

This thesis investigates the design of effective incentive mechanisms to address key challenges in federated learning. It is structured as follows:

- *Chapter 1:* This chapter sets the stage by exploring the evolution of machine learning and the increasing need for privacy-preserving approaches to leverage large, diverse datasets. It introduces federated learning as a solution but highlights its unique challenges in designing fair and effective incentives. The chapter presents the thesis's research questions, exemplifies the limitations of current incentive mechanisms through a case study, and outlines the core research contributions with their corresponding publications.

- *Chapter 2:* This chapter establishes a strong theoretical foundation for the subsequent contributions. It introduces core federated learning concepts and examines how blockchain technology facilitates secure reward distribution within collaborative projects. The chapter then critically reviews existing incentive mechanisms in federated learning, identifying limitations in addressing scenarios with multi-stage interactions, strategic contributions, and the need for fairness and transparency.

- *Chapter 3:* This chapter addresses the crucial issue of incentivising the model owner's often unquantifiable contributions in federated learning scenarios. The ethical risks of information asymmetry between data and model owners are highlighted, as these can lead to opportunistic behaviour that undermines project outcomes. To tackle this, a novel multi-stage incentive mechanism is proposed. Inspired by the Stackelberg game framework, the chapter establishes a multi-stage contract-theoretic model focused on quantifying and mitigating the impact of implicit efforts. Theoretical analysis demonstrates that strategically delaying most of the model owner's compensation until later project stages effectively promotes fairness and optimal effort from all participants. This chapter begins by defining the problem and highlighting the importance of ethical incentives within federated learning. It then introduces the theoretical model that underpins the proposed multi-stage incentive mechanism. Mathematical analysis and simulation examples are used to validate

the framework's effectiveness and ability to promote fair outcomes. The chapter concludes with a summary of key contributions and insights.

- *Chapter 4:* This chapter addresses the critical need for fair and transparent settlement processes within federated learning, where participants' contributions can be difficult to quantify. A significant risk is the lack of guarantees that rewards will be distributed as promised. To tackle this, the chapter introduces a novel framework integrating a blockchain-based two-party clearing protocol within the multi-stage incentive structure established in Chapter 3. This protocol leverages smart contracts to create a secure and auditable settlement system. Trust is increased as the model owner gains assurance of payment upon completing milestones, while the data owner is guaranteed a model that meets the agreed-upon standards. The chapter begins by outlining essential background concepts and terminology necessary for understanding the proposed framework. It then presents the novel clearing protocol, its design, and an in-depth analysis of its security and fairness properties. Subsequently, the chapter focuses on integrating this protocol within a federated learning architecture. Finally, it demonstrates the framework's effectiveness through rigorous implementation and performance evaluation, showcasing its practical applications.

- *Chapter 5:* This chapter addresses the challenges of designing fair and effective incentives in federated learning scenarios where multiple data owners contribute varying quality and quantity datasets. It introduces the Reciprocal Federated Learning Framework (RFLF), designed to balance the power between data and model owners while mitigating the risks of unpredictable data-model compatibility. The RFLF features a unique self-correcting mechanism, providing flexibility for data owners experiencing dataset underperformance while maintaining overall project fairness. Theoretical models are used to analyse the framework's incentive design, with empirical evaluations demonstrating its effectiveness in improving model quality and promoting equitable participation. The chapter begins by highlighting the core concepts and terminology relevant to the RFLF. It then dives into the framework's specific challenges, providing theoretical analysis to support its design principles.

The practicality of the RFLF is demonstrated through its contractual framework, followed by an in-depth analysis of the contract optimality. To prove its efficacy, rigorous empirical evaluations on established datasets showcase the framework's benefits. The chapter concludes with a summary of key insights, the significance of the research, and potential avenues for future exploration.

- *Chapter 6:* summarises the main contents and contributions of this work and provides recommended directions for the continuation of this work in the future.

Chapter 2

Literature Review

2.1 Introduction

Federated learning has emerged as a powerful paradigm for privacy-preserving collaborative machine learning. By enabling model training on decentralised data sources, federated learning addresses the critical challenges of data privacy and sensitivity that often hinder traditional centralised approaches. However, ensuring fair incentives for all participants, including the model and data owners, remains a core challenge for unlocking the full potential of federated learning. These mechanisms are even more critical in long-term multi-stage federated learning scenarios with multiple data owners, where dynamic data quality and quantity changes during the cooperation can introduce additional complexities. To address these challenges, robust mechanisms focused on designing fair and effective incentives are essential.

This chapter provides a comprehensive literature review of the core concepts, techniques, and research advances that form the basis of our work. We begin by introducing the fundamental principles of federated learning in Chapter 2.2. Chapter 2.3 explores the potential of integrating federated learning with blockchain technology to enhance security and transparency. In Chapter 2.4, we shift focus to the latest developments in federated learning incentive mechanisms. Chapter 2.5, through a critical analysis of existing approaches, outlines their strengths and limitations. This chapter highlights the challenges our research aims to address, setting the stage for the novel contributions proposed in subsequent chapters.

2.2 Federated Learning Foundations

2.2.1 Basic Introduction of Federated Learning

In traditional machine learning, model training often requires collecting and centralising data from different sources. Assuming N data owners, let's denote the dataset held by data owner i as $D_i = \{(x_i^{(j)}, y_i^{(j)}) \mid j = 1, 2, ..., n_i\}$. Here, $x_i^{(j)}$ represents the j-th data sample (often referred to as a feature vector), and $y_i^{(j)}$ represents the corresponding label (or target value) in data owner i's dataset D_i. The size of this local dataset is denoted by n_i. The notation $n = \sum_{i=1}^{N} n_i$ refers to the total number of data samples across all data owners. The centralised dataset is then the aggregation of these individual datasets: $D = \bigcup_{i=1}^{N} D_i$. The server aims to minimise a loss function $L(\theta, D)$ with respect to the model parameters θ, updating them iteratively. While efficient, this approach raises privacy concerns with sensitive data.

To address these privacy concerns, federated learning was developed[3], [4]. This approach enables collaborative model training without the need to centralise sensitive data. When data is distributed across multiple owners, federated learning protects privacy by allowing each owner to train a model locally on their private dataset and only share updated model parameters with the model owner (Server).

In the federated learning framework, the key roles are the data owners (also known as clients), who possess the training data and the model owner (also known as the federated learning server), who owns and controls the model updates. The federated learning training process typically involves several rounds (iterations), up to a maximum of T rounds. The training can terminate before reaching T rounds if the model meets predefined performance metrics (e.g., accuracy). Here's a breakdown of a typical federated learning round: The key roles in federated learning are the data owners (clients), who possess the training data and local computation and network resources, and the model owner (federated learning server), who is responsible for model updates. Training typically involves several rounds (iterations), up to a maximum of T rounds or until predefined performance metrics are met. Similar to traditional machine learning algorithms, the federated learning participants first work together to define the problem to solve. then, the model owner defines the

global model type $W(\theta_0)$, and θ_0 represents the initial model parameters. The typical federated learning training process includes several steps as showed in Figure 2.1:

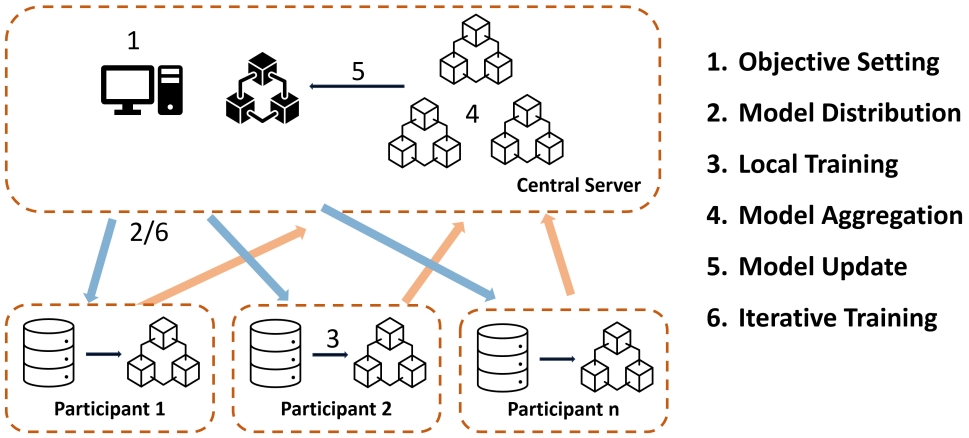

Figure 2.1: Federated Learning Process Overview

1. Objective Setting: The model owner defines the global model type $W(\theta_0)$, where θ_0 represents the initial model parameters. and specifies the overall learning objectives, such as the target accuracy or loss function.

2. Model Distribution: The model owner distributes the current global model parameters θ_{t-1} to all participating data owners.

3. Local Training: Each data owner i uses their private dataset D_i to train the model locally. This involves minimising a loss function $L(\theta, D_i)$ with respect to the model parameters, using techniques like gradient descent: $\theta_{t,i} \leftarrow \theta_{t-1,i} - \eta \nabla L(\theta_{t-1,i}, D_i)$ where η is the learning rate. This process ensures private information is not leaked.

4. Model Aggregation: Each data owner sends their local model updates ($\Delta\theta_t^{(i)} = \theta_{t,i} - \theta_{t-1,i}$) to the model owner. The model owner aggregates these updates, often using a weighted average based on dataset size: $\theta_t \leftarrow \theta_{t-1} + \sum_{i=1}^{N} \frac{n_i}{n} \Delta\theta_t^{(i)}$ Where N is the number of data owners, n_i is the size of data owner i 's dataset, and n is the total number of data samples across all data owners (i.e.,

$n = \sum_{i=1}^{N} n_i$).

5. Model Update: The aggregated global model $W(\theta_t)$ is updated based on the aggregated parameters. The model owner evaluates the updated model to ensure progress toward the predefined training objectives. If the objectives are not yet met, the updated global model is prepared for another iteration.

6. Iterative Training: If the training objectives are unmet, Steps 2 through 5 are repeated iteratively until a stopping criterion is satisfied (e.g., reaching a target accuracy or completing a fixed number of iterations).

We've explored the steps involved in a typical federated learning round. However, for this collaborative training process to function effectively, certain underlying conditions are generally assumed about the federated learning environment:

- Multi-party Participation: Two or more parties collaborate to train a federated model. Each party makes decisions based on their own interests.

- Data Locality: All parties value their own data privacy and security. During the collaborative training process, no participant's original data leaves their local environment, in whole or in part.

- Trusted Transmission: Information such as gradients and parameters must be transmitted securely during the federated learning process. The federated learning server is considered trustworthy and will complete the training tasks according to the requirements of federated learning without stealing any participant's data; meanwhile, the participants are considered semi-honest, meaning they will upload genuine gradient updates based on their own data but remain curious about other participants' private information. This assumption underscores the need for privacy-preserving techniques within the federated learning framework, as it prevents complete trust between parties[10].

By enabling collaborative model training on decentralised data sources, federated learning offers several advantages compared to traditional centralised machine learning approaches:

- Enhanced Data Privacy: By keeping data local to the owner, federated learning minimizes the risk of data exposure or breaches. This is crucial for applications

involving sensitive personal or proprietary information, as it ensures that raw data is not transmitted to a central server.

- Improved Scalability: By keeping data local to the owner, federated learning minimizes the risk of data exposure or breaches. This is crucial for applications involving sensitive personal or proprietary information, as it ensures that raw data is not transmitted to a central server.

- Reduced Communication Overhead: By keeping data local to the owner, federated learning minimizes the risk of data exposure or breaches. This is crucial for applications involving sensitive personal or proprietary information, as it ensures that raw data is not transmitted to a central server.

- Regulatory Compliance: By ensuring that data never leaves its original location, federated learning helps organizations adhere to stringent data privacy regulations. This approach enables companies to derive insights from data while maintaining compliance with applicable privacy laws and standards.

2.2.2 Federated Learning Methods and Frameworks

Our exploration begins with the core methods used in federated learning. Notably, the flexibility of federated extends beyond theoretical concepts. Its processes readily adapt to diverse machine learning models driven by the Stochastic Gradient Descent (SGD) method. This adaptability makes federated learning a practical solution for scenarios involving Support Vector Machines (SVMs) [11], neural networks, or even linear regression [12].

A core component of federated learning is the global model aggregation algorithm. This algorithm determines how model updates received from multiple devices are combined to update the shared global model. Three widely used algorithms are FedAvg[13], FedProx[14], and FedOpt[15].

FedAvg is a foundational algorithm that calculates a weighted average of model updates from participating devices. Weights are often determined by the size of each device's local dataset. Due to its simplicity, robustness, and broad applicability, FedAvg is a popular choice for federated learning, and we will utilise it in our subse-

quent experiments. FedProx extends FedAvg by introducing a proximal term to the local update calculation, aiming to improve convergence, especially with heterogeneous data. For greater flexibility, FedOpt offers a framework that accommodates a variety of server-side and client-side optimisers, extending beyond simple averaging methods.

We've explored the core methods used in federated learning, providing a foundational understanding of how model updates are aggregated across participating devices. Now, let's delve into the practical side of federated learning by examining several powerful open-source frameworks that facilitate the development and implementation of these methods:

- *TensorFlow Federated (TFF)[16]*: Developed by Google, TFF builds upon TensorFlow, offering specialised functionalities for federated learning environments. It provides flexibility for defining computations and simulations and includes tools for training models on diverse devices, including mobile and embedded systems. TFF's compatibility with the familiar TensorFlow ecosystem makes it a natural choice for researchers and developers already in that space. We will leverage TensorFlow Federated as the foundation for our subsequent experiments.

- *PySyft[17]*: Created by OpenMined, PySyft strongly emphasises privacy preservation within its federated learning framework. It seamlessly integrates techniques like secure multi-party computation (SMPC), differential privacy, and homomorphic encryption to safeguard sensitive data. PySyft's Pythonic syntax makes it relatively user-friendly and particularly well-suited for use cases where data confidentiality is of paramount importance.

- *FedML[18]*: FedML is a comprehensive and versatile open-source library for federated learning. It supports various federated learning algorithms, model architectures, and training processes across diverse device types. FedML prioritises ease of use, modularity, and the ability to handle real-world deployment challenges. Its features make it suitable for both research and production environments.

- *FATE[19]*: Developed by AI specialists at WeBank, FATE (Federated AI Tech-

nology Enabler) is an industrial-grade federated learning platform. Designed for production-level environments, it boasts a scalable architecture, supports various secure computation protocols, and offers a range of machine-learning algorithms. Organisations aiming to deploy privacy-sensitive federated learning applications at scale would find FATE a suitable choice.

2.2.3 Categorizations of Federated Learning

Federated learning offers a powerful paradigm for collaborative machine learning, enabling entities to train models on distributed datasets without directly sharing sensitive information. This approach is particularly valuable when data privacy is paramount or when data is geographically dispersed across multiple devices or institutions. To best leverage federated learning, it's crucial to understand how it can be categorised based on the datasets' characteristics and the nature of the participating entities. In this section, we'll explore two key ways to categorise federated learning: by the data distribution among participants and by the type of participants themselves.

Based on the datasets involved in training clients, federated learning can be classified into the following categories [20].

- Horizontal Federated Learning: In horizontal federated learning, participants have different data samples (rows) but share the same feature space (columns). Think of multiple banks operating with the same data models: each bank has information on different customers, but the data collection type (features) remains consistent. A typical application is the next-word prediction model trained on smartphone user input data. Various mobile phone users contribute unique data samples, but the underlying feature space is the same. HFL allows training a shared global model across millions of devices while preserving individual privacy. Figure 2.2[20]: A diagram illustrating horizontal federated learning, where each participant has a subset of data with the same feature set.

- Vertical Federated Learning: In vertical federated learning, participants share the same sample space (e.g., customers) but possess different feature sets (data

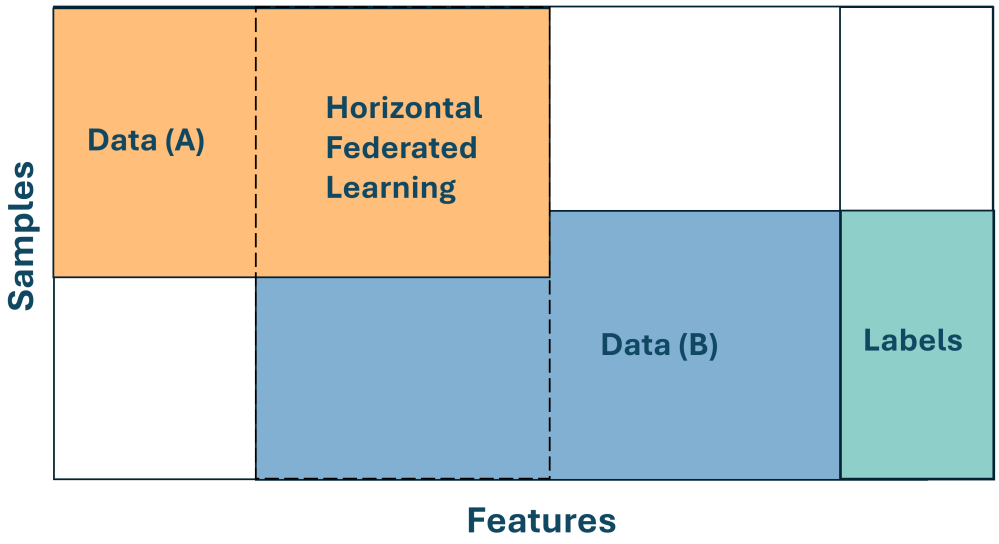

Figure 2.2: Horizontal Federated Learning

columns). Consider the collaboration between a bank and an e-commerce platform in the same region. The bank holds financial data on customers, whereas the e-commerce company has their purchase behaviour data. VFL enables these entities to combine their different insights on the same customers for tasks like credit risk assessment or personalised services without directly exchanging sensitive data. Figure 2.3[20]: A diagram illustrating vertical federated learning, where each participant has different features for the same set of data samples.

- Federated Transfer Learning: Federated transfer learning tackles situations where the participants' feature spaces and sample spaces differ. A prime use case is when a hospital wants to enhance its medical image analysis model by leveraging a pre-trained model from a different domain, like natural images. Federated transfer learning helps adapt this existing knowledge to the hospital's specific task despite the need for the data to be a better match. Figure 2.4[20]: A diagram illustrating federated transfer learning, emphasizing the transfer of knowledge between domains with differing feature and sample spaces.

Federated learning scenarios are further classified based on the nature of the partici-

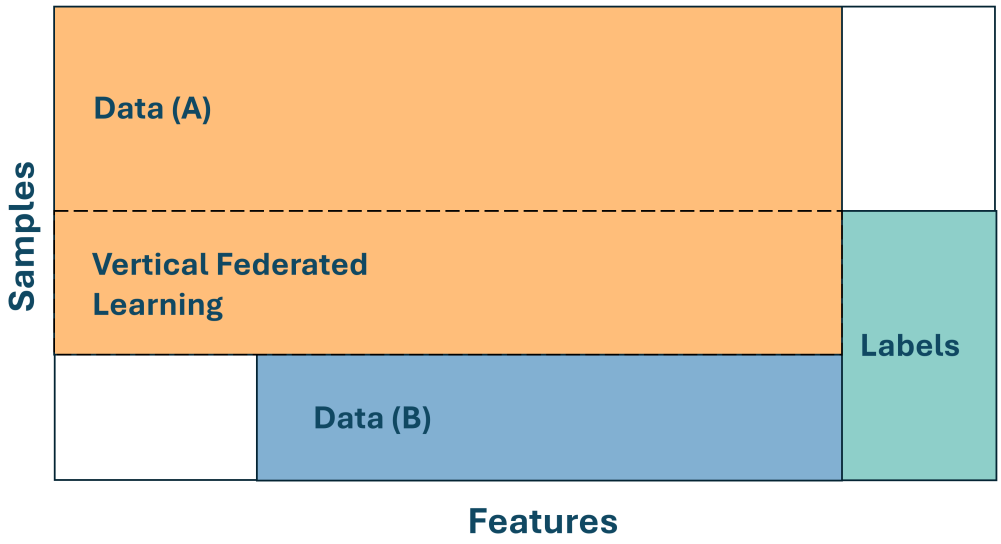

Figure 2.3: Vertical Federated Learning

pants. This distinction is crucial because it influences the scale of the collaboration, the typical data characteristics, and the primary motivations of the involved entities. Here, we'll delve into these two primary categories of federated learning[21]:

- Cross-Device Federated Learning (B2C FL): In this type, individual users with personal devices (e.g., smartphones, IoT devices) collaborate in the federated learning process. The focus is often on consumer-oriented applications like next-word prediction models. Cross-device federated learning typically involves large numbers of participants, each with limited data and computational resources and a potential need for incentives to encourage participation.

- Cross-Institutional Federated Learning (B2B FL): Here, the participants are organisations or institutions, such as banks, hospitals, or companies. Participants might leverage specialized platforms like the FATE[19] federated learning solution, designed for secure collaboration in industries like finance. Cross-institutional federated learning often has fewer participants but with more data per participant. Primary goals might include enhancing model performance through collaboration or obtaining monetary rewards based on data contributions. Security and fairness in assessing contributions become crucial in this scenario.

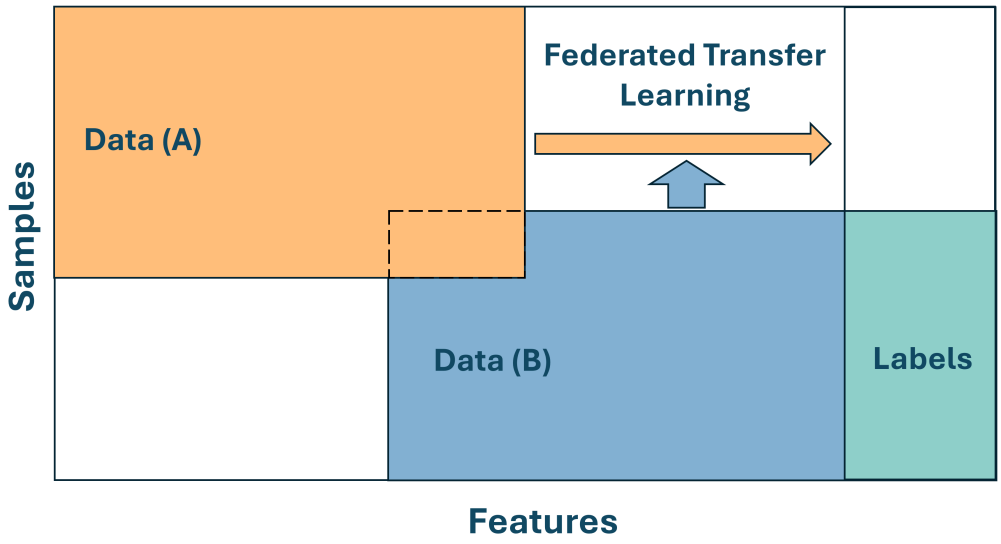

Figure 2.4: Federated Transfer Learning

Our research specifically focuses on Cross-Institutional Federated Learning (B2B FL). This focus is driven by the importance of promoting fairness and robust security in collaborations involving sensitive institutional data. As B2B scenarios commonly involve specialised model owners, our work aims to design incentive mechanisms that motivate these owners to drive federated learning initiatives while ensuring fair distribution of benefits among participants.

2.3 Blockchain in Federated Learning

Federated learning establishes a compelling framework for privacy-preserving collaborative model training in modern, data-driven environments. Inter-institutional data alliances could leverage federated learning to create powerful analytical models as we envision the Internet of Everything. However, ensuring robust participation and trust within such alliances poses challenges. Without effective incentive mechanisms, attracting sufficient high-quality training data becomes difficult. Additionally, traditional federated learning needs a mechanism to establish client reputation, potentially hindering the selection of reliable participants and subsequently impacting model accuracy.

Researchers have explored integrating blockchain technology to address these fed-

erated learning challenges. Blockchain, the distributed ledger underpinning Bitcoin[22], offers a secure, immutable, and auditable platform for addressing data storage and trust concerns. By leveraging blockchain within a federated learning framework, model updates can be recorded transparently, enhancing accountability and resilience against tampering. Moreover, blockchain's inherent incentive mechanisms can reward clients based on their contributions to the model, encouraging active and valuable participation[23].

We will now discuss two types of blockchain-enhanced federated learning systems, classified according to their level of interdependence with the federated learning process[24].

2.3.1 Loosely Coupled Paradigm

The core federated learning process maintains the traditional server-client architecture in this paradigm. The blockchain's primary role focuses on managing client reputation and providing an auditable record of contributions. The loosely coupled paradigm incentivises honest participation and helps mitigate malicious behaviour by verifying model updates and tracking reputation.

Data owners begin by training models on their local datasets. They then upload local model updates to the blockchain for verification. Miners on the blockchain verify these updates and evaluate client reputations based on their contributions. Miners then compete to generate new blocks containing validated model updates and reputation-related data, adding these newly created blocks to the distributed ledger. The federated learning aggregator (model owner) collects verified updates and executes the global model aggregation algorithm. Finally, rewards and penalties are distributed to clients based on their reputation information recorded on the blockchain.

Several research efforts have explored the design and implementation of loosely coupled blockchain federated learning. Bao et al. propose a reputation and payment-based incentive mechanism within the Flchain framework[25]. Clients establish their reputation based on the quality of their model updates and earn rewards based on their contributions. The blockchain facilitates secure information storage and en-

ables clients to evaluate model quality, enhancing trust and incentivising participation.

Building upon the concept of reputation-driven incentives, Toyoda et al. present a model where federated learning tasks and participants are coordinated on the blockchain[26]. Smart contracts facilitate task selection, client voting, and reward distribution based on model performance. This system introduces transparency and a degree of decentralisation to the participant selection process.

Kang et al. employ a multi-weight subjective logic model and contract-based approach for client reputation management to mitigate gradient poisoning attacks[27]. The blockchain provides a platform to track reputation, which influences client selection and reward distribution. This system prioritises reliable clients with high-quality data but lacks privacy protection mechanisms for client gradient updates. Additionally, it does not explicitly address the fairness of consistent client performance over time.

Lo et al. shift the focus to fairness and accountability[28]. They propose a scheme where data and model versions are hashed and registered on the blockchain. Data sampling weights are adjusted to address imbalances, aiming to improve the fairness of the training data distribution and enhance overall model performance.

Finally, Rückel et al. address privacy concerns while promoting fairness and integrity[29]. Their work combines zero-knowledge proofs and local differential privacy. Zero-knowledge proofs allow clients to verify compliance without revealing private data, and differential privacy adds noise to model updates. Rewards are allocated based on actual model contributions, measured on a public test dataset.

These works demonstrate the ongoing development of sophisticated reputation management, secure verification, and incentive optimisation techniques within the loosely coupled blockchain federated learning paradigm.

This loosely coupled approach offers several advantages. Firstly, it maintains a degree of separation between blockchain operations and core federated learning processes, preserving essential privacy features. Secondly, the reputation mechanism promotes honest participation and facilitates the identification and exclusion of ma-

licious actors, fostering trust and improving model accuracy.

However, there are also disadvantages to consider. The blockchain's role in this paradigm primarily focuses on verification and reputation management, leaving the federated learning process vulnerable to centralisation risks and data leakage. Maintaining separate systems for federated learning and the blockchain can also result in less efficient resource utilisation. Despite these challenges, the loosely coupled paradigm offers valuable advancements in reputation management and incentive mechanisms within federated learning environments.

2.3.2 Tightly Coupled Paradigm

The tightly coupled blockchain federated learning paradigm addresses several challenges traditional federated learning faces by integrating blockchain technology directly into the learning processes:

1. The blockchain's decentralised nature removes the reliance on a central aggregator, replacing it with a peer-to-peer system. Global model aggregation can occur through distributed blockchain nodes, mitigating the single-point-of-failure risk.

2. The blockchain provides a mechanism for verifying model updates and removing unqualified or malicious updates before aggregation.

3. The blockchain enables effective reward distribution, encouraging active and honest participation by clients.

Two main approaches exist for global model aggregation in this paradigm:

- Selected Clients: Chosen clients (nodes) collect verified updates and execute the aggregation algorithm.

- All Clients: Every client participates in global model aggregation.

Within this paradigm, the distributed ledger plays a crucial role by storing verified local and global model updates and other essential data generated throughout the training process. The general workflow proceeds: clients first train models on their local datasets. Subsequently, designated clients (nodes) are responsible for verifying

these updates. Following verification, selected clients execute the aggregation algorithm, generating a new global model. New blocks containing the verified updates are then added to the distributed ledger. Finally, rewards are distributed to participating clients based on the incentive mechanism implemented within the blockchain framework.

Several research efforts have explored the tightly coupled paradigm. Kuo et al. proposed GloreChain, where clients take turns aggregating the global model using a Proof of Equity (PoE) consensus mechanism[30]. Unfortunately, GloreChain lacks malicious attack defence mechanisms and model update privacy protection.

Weng et al. introduced DeepChain[31], where clients transmit gradient updates, and miners execute model parameter aggregation. The fastest miner becomes the leader, updates the model with aggregated gradients, and receives token rewards. DepChain uses the Paillier homomorphic encryption algorithm for privacy. Although its incentive mechanism enhances the robustness and collaborative fairness of federated learning, ensuring consistent client model performance fairness and dealing with the fairness of sensitive attributes is challenging.

Gao et al. proposed FGFL[32], employing a multi-center federated learning network structure for client coordination. Partially reliable clients handle local training and aggregation, while unreliable clients only train local models. FGFL uses digital signatures and blockchain-managed smart contracts to track reputation and contribution and distribute rewards through five core modules. This approach enhances privacy and robustness, but focusing on indirect reward distribution may only partially optimise client model performance.

This paradigm offers several advantages. Firstly, its decentralised nature eliminates single points of failure. Secondly, removing the need for a central server enhances privacy and potentially reduces communication costs. However, this paradigm also has disadvantages. Clients manage local training and global model integration, demanding greater computing power. Additionally, the limited bandwidth of blockchain networks may create challenges for the tightly coupled paradigm's deployment[33].

Despite these challenges, the tightly coupled paradigm represents a significant step towards decentralised, secure, and incentivised federated learning, offering the po-

tential to overcome limitations within traditional federated learning frameworks.

2.4 Fair Incentive Mechanisms in Federated Learning

In federated learning, where participants collaborate while their data remains private, designing fair and effective incentive mechanisms is crucial for fostering trust, continuous participation, and the system's overall success. Federated fairness, a well-established concept in the literature, underscores the need for equitable treatment of all contributors, whether in data, computational resources, or other forms of participation[20]. Imbalances in contributions or rewards could deter participation and undermine the entire federated learning process[34].

To design effective incentive mechanisms, it's essential to understand the key motivations driving collaboration in federated learning. Here, we explore three primary drivers:

1. **Sustained Development and Participant Enthusiasm:** A key consideration for long-term success is sustainability. Mechanisms that equitably recognise and incentivise the ongoing and multifaceted contributions of both data owners and the model owner are crucial for the evolution and effectiveness of the learning process. Data owners contribute valuable data, computational resources for local training, and the network resources necessary for model exchange. The model owner provides the initial model, computational power for model aggregation, expertise in model design, tuning, knowledge sharing, and overall project guidance. These complementary contributions form the foundation for sustained model improvement and the federation's success.

2. **Incentives for Self-Interested Participants:** Most participants in federated learning act out of self-interest, participating in the hope of reaping specific benefits. These might range from improved performance of their local models, access to cutting-edge predictive models, or direct capital rewards for joining the training process. Recognising and offering clear advantages to these participants can foster more active and consistent engagement. Furthermore,

incentives beyond purely monetary rewards can be powerful motivators, such as the ability to leverage the expertise of other participants or gain access to valuable anonymised datasets for internal research.

3. **Ethical Considerations:** With federated learning pooling resources from many contributors, ensuring nondiscrimination is imperative. Equity and social ethics dictate that all participants perceive the learning process as fair and that no resultant model unduly disadvantages any participant. Beyond individual benefits, fair incentive mechanisms are crucial for building public trust in federated learning as a technology.

The motivations above highlight the emergence of two distinct incentive paradigms within federated learning:

- **Contribution Fairness:** Incentive mechanisms grounded in contribution fairness seek to assess each participant's value and reward them proportionally accurately. This might involve using marginal contribution calculations to quantify value, resource allocation mechanisms to track computational costs, or reputation systems to acknowledge past contributions and reliability. (We'll explore these mechanisms in detail in the section2.4.1.)

- **Equilibrium Fairness:** Incentive mechanisms based on equilibrium fairness aim to ensure all participants feel they benefit from the collaboration regardless of the size of their contributions. These mechanisms could involve fair participant selection to ensure diverse representation, weight redistribution to balance outcomes, or personalisation strategies to tailor benefits to individual needs. (We'll delve deeper into these specific approaches in the subsection2.4.2.)

A harmonious balance between technical viability, ethical standards, and participant preferences is essential to crafting effective incentives for federated learning. While achieving perfect equilibrium and contribution-based fairness simultaneously can be exceedingly difficult, we can promote an environment ripe for active collaboration by accounting for these factors. This, in turn, leads to more precise and resilient federated models.

Based on the motivations and fairness paradigms discussed, we propose a taxonomy for fairness in federated learning collaborations, depicted in Figure 2.5. This taxonomy serves as a framework to guide our subsequent analysis of existing fair incentive mechanisms.

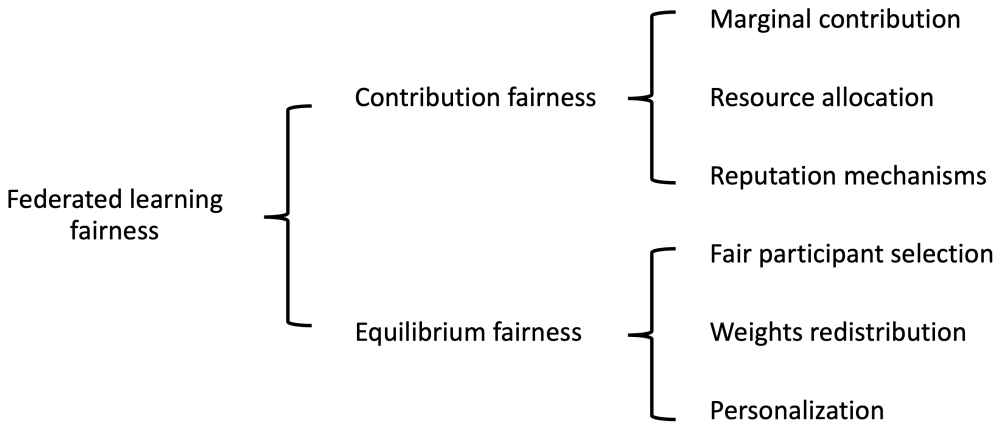

Figure 2.5: Proposed Taxonomy for Federated Learning Fairness

2.4.1 Incentives for Contribution Fairness

While equilibrium fairness, demanded by vulnerable participants, focuses on achieving consistent outcomes across various devices in the final model and avoids inequitable distribution, it is noteworthy that different contributors eventually receive the same federated learning model. This equitability can result in dissatisfaction among higher contributors, potentially leading to a "free-riding" problem. Such problems can considerably hamper the sustainable development of federated learning. Therefore, incentive methods based on contribution fairness will be more widely adopted in scenarios where participants are evenly matched, or the model's effectiveness heavily relies on key contributors. Central to contribution fairness is the equitable assessment of each participant's contribution to the global model while safeguarding their sensitive data.

Various techniques for measuring participant contributions exist[35], such as Shapley values, blockchain, contract mechanisms, reputation systems, game theory, auction

mechanisms, and advanced technologies like reinforcement learning. This section delves into *marginal contribution*, *resource allocation*, and *reputation mechanisms*.

Marginal contribution

The prevailing method to measure participants' marginal impact in federated learning is through Shapley values. Introduced in 1953, Shapley values aim to solve cooperative game problems[36] and have been extensively employed to evaluate each participant's contribution to a game.

$$\begin{aligned} \phi_i &= \mathbb{E}\pi \in \Pi[\nu(S^i\pi \cup \{i\}) - \nu(S^i_\pi)] \\ &= \frac{1}{n!}\sum_{\pi\in\Pi}[\nu(S^i_\pi \cup \{i\}) - \nu(S^i_\pi)] \end{aligned} \tag{2.1}$$

In this context, $\pi \in \Pi$ represents permutations of all participants, S^i_π signifies the collection of participants ranked prior to i in the permutation π, and ν denotes a value function, often associated with the market value of the machine learning model. The Shapley value of participant i can be interpreted as the anticipated incremental contribution of i across all possible joining sequences in federated learning. To calculate Shapley values, one can simplify the process by listing all potential joining sequences that exclude participant i, determining the anticipated value increase introduced by participant i, and then averaging these incremental contributions, considering the likelihood of occurrence for these sub-combinations.

Shapley values satisfy the following properties: Group Rationality, Symmetry, Zero Contribution and Additivity. Group rationality ensures that the assessment of contributions effectively reflects the proportion of each participant's contribution to the federation's value metric, such as the test accuracy in federated learning. The combination of symmetry and zero contribution properties ensures that the assessment of contributions is objectively based on value metrics and does not differentiate between participants. Additivity guarantees that there is no need to recompute the value metrics for completed evaluations in a linear combination of multiple objectives in subsequent multi-objective optimisation scenarios. By considering all possible joining orders for federated learning participants, Shapley values satisfy fairness in evaluating contributions among participants.

Early adopters of Shapley values for fair evaluation in this context were Jia et al.[37]. However, the inherent computational demands of Shapley values, with its $O(N^2)$ complexity, can be prohibitive in real-world settings. Addressing this, Jia et al. proposed an approximate computation that reduces model training volume but maintains a strong correlation between the approximations and actual values.

Ghorbani et al. tackled data quality concerns, such as label errors[38]. By applying Shapley values to ascertain individual dataset contributions, they presented a Monte Carlo sampling method as an efficient approximation for Shapley values. Their technique effectively flagged low-quality training data with limited model retraining.

Further innovations include the Contribution Index (CI) by Song et al.[39] and a multi-dimensional contribution method based on stepwise computation by Nishio et al.[40].

Resource allocation

While Shapley values consider the contribution differences among participants from different federated alliances, they assume equal initial contributions from all participants before evaluating marginal contributions. However, this assumption can lead to unequal initial contributions among participants for specific learning tasks such as classification or regression, potentially resulting in imbalanced rewards or incentives.

To address this issue, Zhang et al. introduced the Hierarchically Fair Federated Learning (HFFL) framework, which utilises publicly verifiable factors like data quality, quantity, and collection cost, to classify participating clients into various tiers [41]. Participants within the same level are considered equal contributions, with higher contributions corresponding to higher levels. Participants of different levels will converge to different models. During the training of a lower-level model, Participants at higher tiers provide an equivalent volume of data as their counterparts at lower tiers. Conversely, when engaged in training higher-level federated learning models, Participants at lower tiers are required to contribute all their local data.

In contrast to HFFL, which trains models for each level, Lyu et al. presented a Fair and Privacy-Preserving Deep Learning (FPPDL) framework to encourage

participants to earn points by sharing their information with others, which they can then exchange for information from other participants[42]. Participants earn more points by uploading more gradient information, which they can use to obtain more information from other participants. All transaction records are transparently recorded on the blockchain, and a three-tier onion encryption scheme is proposed to protect gradient privacy. Every participant's contribution results in variant models of different levels of the global model.

Various other methods have also been proposed, with each offering distinct advantages. For instance, Kang et al. introduced an incentive mechanism based on contract theory[27]. Higher-quality local data lead to faster training of local models, allowing participants to receive greater rewards. Similarly, Sarikaya et al. proposed a Stackelberg game model between devices and models[43]. In this model, model owners motivate workers with devices to allocate more CPU computational resources for local training to achieve faster convergence. Le et al. presented an auction game between base stations and multiple mobile users[44]. In this scheme, mobile users act as sellers, making optimal decisions based on their resources and local accuracy to minimise energy consumption. Based on users ' bidding information, base stations select the most suitable candidates to maximise social welfare. A primal-dual greedy algorithm is proposed to solve such NP problems.

Furthermore, Zeng et al. introduced the FMore incentive mechanism, which is grounded in the concept of a multi-dimensional auction [45]. This approach involves the aggregator transmitting bidding requests to participants. Upon receiving these requests, participants evaluate their resources and projected budgets to determine whether to submit a bid. Subsequently, the aggregator identifies K winners using scoring mechanisms. FMore is a lightweight and compatible framework with minimal computational and communication overhead.

In addition, Deng et al. devised a quality-aware auction technique [46]. This method frames the problem of selecting winners as an NP-hard task of maximising learning quality. The proposal involves the creation of a greedy algorithm based on Myerson's theorem, serving the purpose of real-time task allocation and equitable reward distribution.

In conclusion, resource allocation in Federated Learning is an active research area with many methodologies being proposed. The choice of method often depends on the specific requirements and constraints of the federated learning setup.

Reputation mechanisms

Reputation mechanisms have gained traction to evaluate a participant's contribution to federated learning, ensuring fairness and promoting trustworthy collaboration. This assessment is typically based on a participant's historical reliability and engagement in federated learning tasks. Two primary categories emerge in this context: *direct* and *indirect* reputation.

Direct Reputation: This metric evaluates participants based on their trained local models' quality and activity level. Direct reputation provides a real-time assessment, considering recent contributions and engagements. Lyu et al. introduced the Collaborative Fairness in Federated Learning (CFFL) framework[47]. Within this framework, the server evaluates the accuracy of gradients uploaded by participants and calculates their reputations for each round through normalisation. The reputation of each participant undergoes iterative updates based on both the reputation from the current round and their historical reputation. This iterative process results in participants converging towards different models through reputation adjustments, thereby promoting fairness. While CFFL demonstrates a noteworthy level of fairness, it does not explicitly address considerations related to the system's robustness.

Indirect Reputation: This metric takes a longer view, assessing a participant's reputation across multiple federated learning tasks. It offers a safeguard against malicious activities by cross-referencing consistency in reputation feedback. Zhao et al. presented a reputation-based system that leverages blockchain technology[48]. Initially, all clients possess identical reputation values. As clients successfully contribute models, their reputation values increase. However, uploading malicious parameters results in a reduction of reputation values. The server employs these reputation values to select dependable clients, favouring those with higher reputations that are more likely to be chosen and rewarded. Rehman et al. proposed a reputation system based on blockchain[49]. It establishes a collaborative framework involving

three tiers: edge devices, fog nodes acting as data arbitrators, and cloud servers owned by model creators. The cloud server updates models to fog nodes, distributing updated local models to edge devices. Smart contracts facilitate the aggregation, computation, and recording of participant reputations in federated learning. This system ensures privacy and security, assuring the authenticity of users' provided data. However, it also involves trade-offs such as heightened model complexity, increased computational costs, and more significant communication expenses.

However, many reputation scoring mechanisms are subjective and require comprehensive quality assessment schemes. It leaves the door open for malicious rating manipulation. Kang et al. introduced a multi-weight subjective logic model to address this issue[50]. This model calculates reputation based on a participant's historical performance and recommendations from other participants. This approach aims to design a blockchain-based system that manages and records data owners' reputations. The individual participant's reputation calculation method uses a multi-weight subjective logic model to balance various reputation assessments comprehensively. It ensures a holistic evaluation of participants' contributions to federated learning.

In conclusion, while reputation mechanisms offer a promising avenue for evaluating contributions in federated learning, they are full of challenges. Striking a balance between objective evaluation and preventing manipulations remains a pertinent concern.

2.4.2 Incentives for Equilibrium Fairness

As we navigate the complex landscape of federated learning, data heterogeneity emerges as a pivotal challenge. With participating clients demonstrating diverse data distributions, ensuring optimal performance across the board becomes intricate, especially in real-world scenarios, where dominant parties armed with substantial data can overshadow the contributions of more vulnerable entities.

Central to this challenge is the concept of equilibrium fairness, a beacon guiding us towards more balanced outcomes. Under this paradigm, an intuitive way to approach this model is to have the server adopt a random selection strategy during

the federated learning training process, choosing participants solely for local updates and model uploads. This server-centric model aggregation, which emphasises assigned weights, marks a step towards inclusivity. However, championing fairness goes beyond this; it requires addressing under-representations by actively engaging vulnerable parties and ensuring impartiality in weight distribution or a more personalised approach, a crucial step to prevent inadvertent biases.

Different metrics, such as Standard Deviation, Gini Coefficient, and Jain's fairness index, are employed for measuring fairness. By implementing these measures, participants can engage in federated learning more equitably, receiving fair weight allocations and contributing based on their unique characteristics. This approach facilitates the achievement of genuinely balanced fairness.

Fair Participant Selection

Federated learning organisers often favour selecting participants with high data quality and abundant resources when orchestrating the training process. This tendency results in the stronger data factions being more likely to be chosen, which subsequently influences the final globally trained model to exhibit characteristics of the dominant factions. While this approach aids in maximising overall gains, it may disregard participants with limited resources, leading to unfairness.

To mitigate biases faced by participants with lower computational capabilities or smaller datasets in federated learning, the solution proposed by Yang et al. introduces the concept of participation frequency[51]. It allows less frequently selected participants to engage in training more often. Furthermore, Huang et al. presented the RBCS-F algorithm, which requires that a participant's selection probability stays within a threshold in the long term to ensure fairness[52].

However, here is a conundrum: How do we factor in disparities in resources and capabilities? Nishio et al. proposed the FedCS approach to address the selection challenge of resource-constrained participants[53]. This approach mandates participants to disclose their resource information during the selection phase, followed by selecting based on it to encompass a diverse range of participants. This strategy aims to balance participant opportunity fairness and outcome fairness.

Furthermore, considering that participants with slower internet speeds might frequently encounter data retransmissions, leading to additional training delays in the federated learning model, Zhou et al. highlight another dimension - introducing a resilient framework called "Throw Right Away" (TRA)[54]. This framework suggests that discarding some data packets in suitable scenarios is only sometimes detrimental. By reporting network conditions during participant selection, intentionally disregarding some lost data packets becomes possible. It facilitates the acceptance of data uploads from devices with lower bandwidth, thus expediting the federated learning training process. However, this hinges on accurate assessment and truthful reporting of resource conditions by the participants.

An alternative to undersampling participants with insufficient contributions is to employ a local compensation approach. In this regard, Wang et al. proposed an innovative Pulling Reduction with Local Compensation (PRLC) method[55]. This method enables end-to-end communication in federated learning. Participants not selected are empowered to perform local updates through PRLC to reduce the gap between their local and global models. This method's participant selection aims to maximise utility and primarily hinges on optimising dynamic resource allocation issues among diverse participants.

Hu et al. also employed game theory to model the utility maximisation problem for servers and users in federated learning as a two-stage Stackelberg game[56]. Through this approach, utility maximisation for servers and users is considered separately. Solving for Stackelberg equilibrium yields the optimal strategies for servers and users, facilitating selection of users most likely to provide reliable privacy data for compensation.

While each method above has its unique appeal, they collectively guide us towards fairness in participant selection and ensure balanced outcomes in real-world applications.

Weight Redistribution

Various notable approaches have emerged in the realm of blending fairness with model optimisation. Mohri et al. delved into the challenges posed by worst-

performing devices, crafting an Agnostic Federated Learning (AFL) approach based on the min-max loss function, which acts as a deterrent to model overfitting to specific customers[57]. However, this approach best fits smaller customer scales due to its concentrated focus on underperforming ones.

Similarly, with a lens on the underachievers, Hu et al. formulated the FedMGDA+ strategy to balance fairness with robustness harmoniously[58]. Their approach refines the fairness of federated models by adjusting participant gradient merging weights. They employed Pareto-stable solutions, emphasising on universally beneficial model outcomes.

In contrast, Cui et al. took a broader perspective with their Fair and Consistent Federated Learning (FCFL) technique. By leveraging gradient-constrained multi-objective optimization, they sought to iron out disparities and inconsistencies that arose due to varying preference directions[59]. Their approach is inclusive, considering the objectives of all participants and fostering uniform participant performance.

Drawing inspiration from AFL, Li et al. developed the q-Fair Federated Learning (q-FFL) method. This method intriguingly utilises q-parameterized weights, pivoting attention to devices grappling with higher losses, thus ensuring fair distribution[60]. The dynamic nature of the q parameter offers a versatile solution, but determining its optimal value in diverse data environments remains a hurdle.

Recognizing the constraints posed by q-FFL, Tian et al. introduced the innovative α-FedAvg algorithm. This approach elegantly weaves in Jain's index to balance fairness and utility, with the α parameter fine-tuned by the algorithm even before training begins[61].

Meanwhile, Zhao et al. presented an alternative to the q-FFL's loss amplification mechanism by proposing a direct weight redistribution methodology[62]. This strategy emphasizes penalizing higher-loss clients with more significant weight allocations. On a similar thread, Li et al. ventured into modifying device weights with empirical risk minimisation to facilitate a fluid balance between fairness and accuracy[63].

However, all of the approaches above assume that participants are honest. If partic-

ipants maliciously exaggerate their losses, this can degrade the overall performance of the global model. To address this concern, some scholars have introduced the concept of blockchain to mitigate the potential malicious actions of dishonest users. Ur et al. proposed employing blockchain as a decentralised training entity in the network, presenting TrustFed, a fully decentralised cross-device federate learning system[64].

Personalization

The data heterogeneity significantly impacts the performance distribution of the global model, rendering it arduous to maintain consistent performance across diverse clients. This variance can sometimes lead to discriminatory behaviour by the federated learning model towards specific attributes within the sample population. To tackle the challenge, the personalisation federated learning approach becomes imperative to optimise the global model for each client[65].

Data-based personalization approaches aim to address the uneven distribution of client data, often characterized by statistical heterogeneity in federated learning environments. To ensure a comprehensive representation of the overall data, federated learning setups may require data-sharing strategies or the acquisition of virtual datasets. Zhao et al. proposed a data-sharing approach that equitably allocates a small portion of global data to individual clients based on categorical balance[66]. Empirical results from these studies demonstrate that even minimal data addition can significantly improve model accuracy. Jeong et al., on the other hand, developed a federated augmentation technique (FAug) that utilizes generative adversarial networks (GANs) trained on a federated learning server[67]. This method uploads data samples from underrepresented groups to the server for GAN training. The trained GAN models are then distributed to clients, allowing them to generate additional data to supplement their local datasets and create a more balanced distribution.

However, while data-based approaches enhance the global federated learning model's convergence by mitigating client data drift, they often necessitate certain refinements in the local data distribution. Such adjustments might lead to the loss of crucial information related to the diversity of client behaviours, which is instrumental in constructing personalised global models. Another model-based approach to global

model personalisation aims to cultivate a robust global federated learning model adaptable to each customer's needs in the future or to enhance the local model's adaptability. Li et al. proposed FedProx, enabling each client to perform partial training based on available resources[14]. It introduces a regularisation term composed of the squared distance between the local and global models. This term encourages local updates to align with the global model, leading to higher-quality local updates and enhancing training stability. Conversely, Li et al. introduced FedMD, an federated learning framework that employs Transfer Learning (TL) and Knowledge Distillation (KD), allowing clients to develop autonomous models utilising their private data[68]. Before federated training and KD phases, the TL phase employs a pre-trained model on a public dataset, which is subsequently fine-tuned by each client using their private data.

Using a personalised approach fully integrates each customer's local data, making the global model better suited to address different customers' unique data characteristics and needs. This method promotes not only a more balanced and effective optimisation of the global model but also ensures data privacy and security.

2.5 Conclusion and Chapter Discussion

This chapter provides a tutorial on federated learning, followed by an exploration of blockchain's role in enhancing federated learning collaborations, and concludes with a comprehensive survey of associated incentive mechanisms. This review lays the groundwork for my thesis, which seeks to enhance fairness and trust in federated learning by developing novel incentive mechanisms tailored to the complexities of real-world collaborations. The concept of fairness is integral to widespread participation and overall success. A lack of fairness jeopardises the long-term viability of federated learning systems. However, current research faces several challenges hindering the realisation of a truly fair and equitable federated learning landscape.

Key Challenges and My Thesis Contributions:

- **Comprehensive Fairness Definitions:** While metrics like the Gini Coefficient and Jain's Fairness index provide valuable tools, there is no universally

accepted standard for measuring fairness in federated learning. Often, fairness is defined narrowly within a technical context, disregarding its broader legal, regulatory, and social implications. This underscores the need for a robust understanding of fairness that accounts for such nuances, enabling the evaluation of incentive mechanisms beyond simple metrics. My thesis will contribute to this by developing a theoretical framework for valuing diverse contributions under information asymmetry, promoting context-aware definitions of fairness.

- **Realistic Application Scenarios:** Current research often assumes a dominant model owner who dictates the parameters of the incentive mechanism. However, open and competitive markets necessitate the exploration of scenarios where power dynamics are more balanced. My thesis addresses this challenge by introducing a novel Reciprocal Federated Learning Framework (RFLF) and a secure multi-stage clearing protocol. These innovations specifically empower data owners and promote competition between model owners, fostering a fairer distribution of influence in model aggregation and reward allocation processes.

- **Adapting to Dynamic Environments:** Fairness is not a static concept; definitions should evolve alongside the project, its outcomes, and changes in the broader technological landscape. Existing mechanisms often need more adaptability to respond to real-time feedback or changing circumstances. My thesis addresses this need through the RFLF, incorporating dynamic incentive mechanisms guided by verifiable data quality and contribution assessments. This approach ensures that rewards remain equitable and aligned with demonstrated effort throughout the project, fostering fairness even in the face of evolving conditions.

- **Interpretability of Fairness:** Participants are likelier to engage in a system where fairness principles are transparent. Explainable metrics and transparent decision-making processes are crucial for building trust. These should elucidate how decisions impact each participant's interests while respecting individual privacy and maintaining model confidentiality. My thesis research contributes towards this effort by utilising transparent data valuation methods

and incorporating fairness assessments throughout project milestones.

- **Robust Fairness in Adversarial Settings:** Current methods often presuppose a level of trust between participants. However, participants might possess diverse and sometimes conflicting motives in real-world situations. Safeguarding fairness amidst adversarial behaviours is paramount. This includes a specific need to consider potential privacy attacks, fraudulent contributions, and their impact on equitable outcomes. My thesis research, using cryptographic techniques and leveraging secure, intelligent contracts, contributes towards this goal through the proposed multi-stage clearing protocol and integrated dispute resolution mechanisms.

This review's challenges and research directions underscore the need for innovative incentive mechanisms that promote fairness while addressing scalability, privacy concerns, and dynamic model environments. My thesis explores these issues and provides related solutions throughout subsequent chapters.

Chapter 3

Dynamic Multi-Stage Incentives for The Model Owner

3.1 Introduction

In recent years, advances in wearable devices, medical sensors, and pervasive communication technologies have fueled the growth of the Internet of Medical Things (IoMT). It enables applications in patient monitoring, disease diagnosis, and personalised treatment. The vast amount of data IoT devices collect offers immense potential for AI-driven models to transform healthcare delivery. For example, remote sensors can monitor vital signs, collect physiological data, and provide image-based diagnostics. This information can improve disease management, facilitate early interventions, and ultimately enhance patient health outcomes.

While representation-learning models like Deep Learning offer groundbreaking potential in healthcare, they often require vast amounts of data to excel. A single healthcare provider may need help collecting enough diverse cases to build a high-accuracy and generalizability model. Likewise, even research institutions specialising in machine learning may need more access to the breadth of real-world medical data necessary to develop robust models. For instance, building a robust image classifier to detect rare diseases might require a much larger dataset than any single hospital or research institute possesses. Collaboration and data sharing are logical solutions, but healthcare organisations are understandably hesitant to disclose raw

patient data due to privacy regulations and concerns about competitive advantage. It is where Federated Learning can play a transformative role, enabling collaborative model training without compromising the confidentiality of sensitive patient information.

As illustrated in Figure 3.1, such a federated learning system could involve a group of data owners and a single model owner. In this scenario, participants are healthcare organisations and research institutions. The model owner is a research institute specialising in machine learning. The data owners are the healthcare organisations that begin by collecting healthcare data through various means, such as wearable sensors or electronic health records. This data stays on-device or within the participants' local infrastructure. When the research institute initiates model training, they transmit a request and a starting set of model parameters to the participants. Participants then leverage their local datasets to train the model locally, updating only the model parameters, not the raw data. These updated parameters are then securely transmitted to the research institute for aggregation. This iterative process of local training and parameter aggregation continues until the model achieves the desired accuracy. The research institute's payoff is tied to achieving preset performance targets in this unique collaboration. Upon reaching those targets, they receive rewards, while the data owners gain ownership of the trained model for their applications.

Our research topic has significant implications for the widespread adoption of federated learning beyond its current experimental stage. The success of federated learning depends on a foundation of trust and equitable compensation for all participants. However, the inherent information asymmetry within the process creates ethical risks for data and model owners. For example, a data owner might provide low-quality data, or a model owner might exert minimal effort in refining the aggregated model. If left unchecked, these risks can lead to opportunistic behaviour, undermining collaboration and jeopardising the project's overall outcome. This lack of trust poses a significant barrier to the widespread adoption of federated learning, especially in sensitive domains like healthcare, where data integrity and model reliability are paramount. Our research aims to pave the way for broader federated learning use in these sensitive domains by designing incentive mechanisms that di-

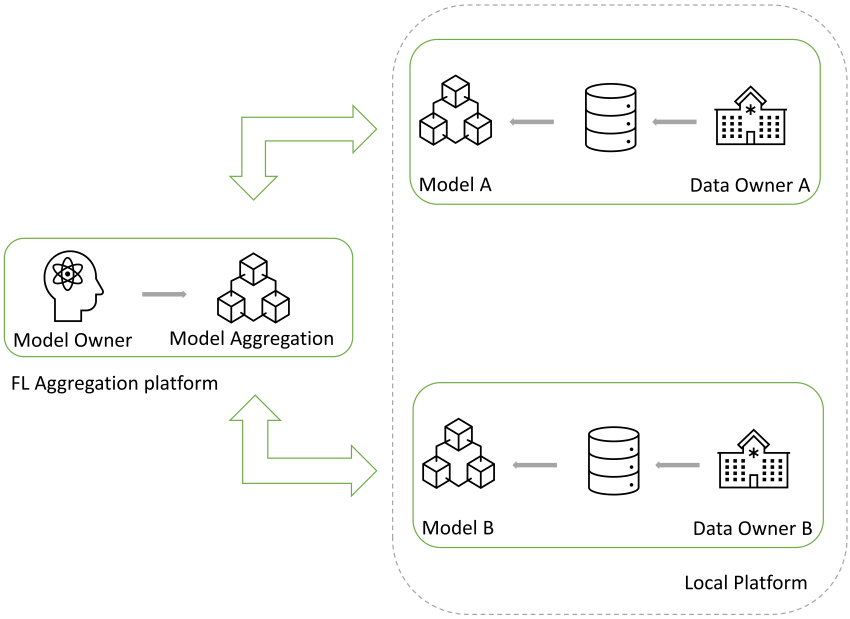

Figure 3.1: Federated Learning Structure

rectly address these ethical considerations. It can unlock advancements in medical diagnostics, financial fraud detection, and other areas where privacy and accuracy are equally crucial. To effectively address these challenges, we employ a Stackelberg game framework, where data owners act as leaders.

This chapter introduces an incentive mechanism that directly tackles this dual ethical risk, focusing on quantifying implicit efforts and mitigating their impact through multi-stage game theory. We examine the interaction between data and model owners where efforts are unobservable, designing a multi-stage incentive contract established before training.

This approach allows for the design of a pre-training incentive contract that promotes fairness and transparency. Using a multi-stage approach, we can tailor the incentives to the evolving project dynamics. This structure ensures that compensation and effort remain aligned, mitigating opportunistic behaviour and fostering optimal outcomes.

This research directly confronts the 'double ethical risk' inherent in federated learning, where both data and model owners face uncertainty about each other's efforts. Our contribution provides a mechanism for designing multi-stage incentive contracts that mitigate this risk and promote optimal effort. Our analysis reveals that optimal

incentive contracts from the data owner's perspective prioritize late-stage rewards for the model owner, strongly linking compensation to the model's increasing value throughout the collaboration. This fosters sustained high-quality model development efforts.

The remainder of this chapter is structured as follows. Section 3.2 presents the incentive mechanism model used in our research and the results, and Section 3.3 provides a simulation example to validate the model. Finally, conclusions and future work are drawn in Section 3.4.

3.2 The Multi-stage Incentive Mechanism Model

To ensure the success in federated learning and allow for the best training result, it is crucial to implement an effective incentive mechanism that minimises the possibility of dual ethical risk. Based on the discussion in the previous section, no existing incentive mechanism has suitably addressed this issue. This section introduces a multi-stage incentive mechanism model based on contract theory. It addresses the dual ethical risks associated with federated learning while incentivising both parties to cooperate successfully. Note that, for simplicity, the game assumes one data owner and one model owner. A contract-theoretic solution for federated learning scenarios with more than one data owner is left to future work.

3.2.1 The Model

The two participants in our model, the data owner and the model owner, are risk-neutral. Both parties agree that the entire training process will be conducted in K stages, with both parties jointly checking the training results at the end of each stage to confirm that the training was successful. Additionally, both parties agree that the contract cannot be ended earlier than these K stages unless the training fails. We assume that the effort value committed by the data owner at stage k is De_k, and the effort value committed by the model owner at stage k is Me_k. De_k and Me_k are both uncorrelated variables. Furthermore, $De_k \geq 0, Me_k \geq 0$.

Table 3.1 lists the key notations commonly used in this chapter for ease of reference.

Notation	Description
k	Training stages, $k = 1, \cdots K$.
Me_k	The effort committed by the model owner at stage k
De_k	The effort committed by the data owner at stage k
$P_k(Me_k, De_k)$	The probability of successful training at stage k
$C(Me_k)$	The effort cost of the model owner at stage k
$C(De_k)$	The effort cost of the data owner at stage k
V_k	The incremental value of the model after stage k
M_k	The market value of the model at stage k
I_k	The data owner's costs at stage k
DR_k	Total expected revenue of the data owner from stage k to K
MR_k	Total expected revenue of the model owner from stage k to K
R_k	The reward for model owner if training success at stage k
$X_k(Me_k, De_k)$	The model's performance at stage k
ϕ, ν	The weight parameters of the model at stage k

Table 3.1: Glossary of Key Mathematical Notations (Chapter 3)

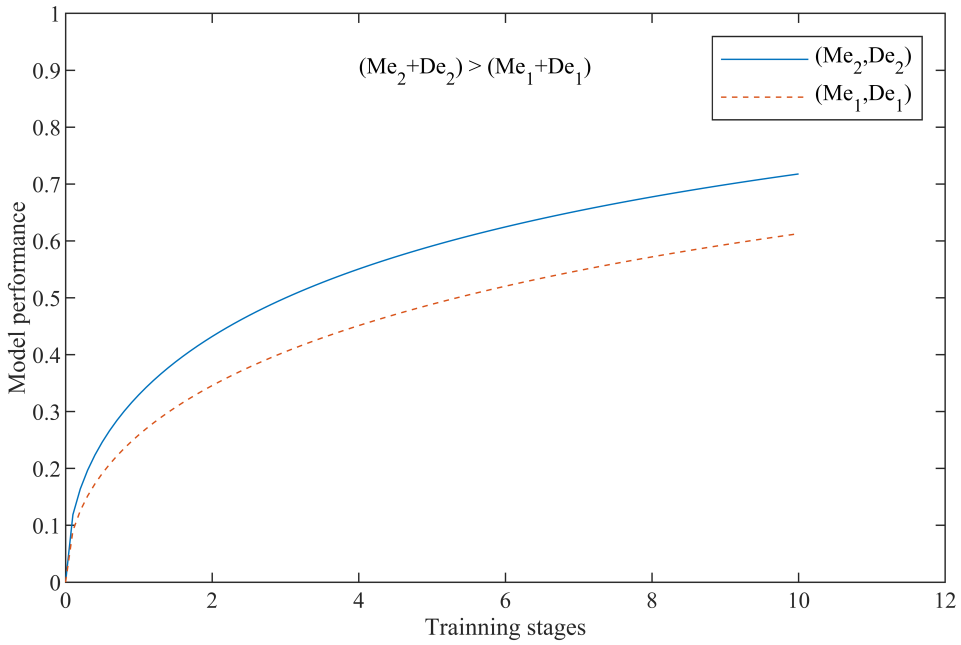

Figure 3.2: Federated Learning Performance

Naturally, the performance of a model, e.g., the accuracy of its inferences, will be higher if the data owner contributes more effort to providing more and higher quality data. Similarly, if the model owner puts in more effort, such as improving the algorithm, model performance will also increase. The model's performance is assumed to be

$$X_k(Me_k, De_k) = 1 - e^{-\phi(Me_k, De_k)^{\nu}},$$

where ϕ and ν are the weight parameters.

Fig. 3.2 shows the relationship between the performance of a typical federation learning model and the effort values Me and De of the training participants.

The following assumptions are made over the probability that training at stage k will be successful:

$$P_k(Me_k, De_k),$$

and

$$1 \geq P_k(Me_k, De_k) \geq 0, \frac{\partial P_k(Me_k, De_k)}{\partial Me_k} > 0, \frac{\partial P_k(Me_k, De_k)}{\partial De_k} > 0,$$

$$\frac{\partial^2 P_k(Me_k, De_k)}{\partial {Me_k}^2} < 0, \frac{\partial^2 P_k(Me_k, De_k)}{\partial {De_k}^2} < 0, (k = 1, \cdots, K).$$

Thus, there is a positive correlation between the probability of successful training and the efforts contributed by the data and model owners. The probability of success increases as Me_k and De_k increase with diminishing marginal returns.

The cost of inputting effort by the two parties in the training stage k are $C(Me_k)$ and $C(De_k)$. Obviously, these costs increase as the effort increases, i.e., $C'(Me_k) > 0$, $C'(De_k) > 0$. Similarly, the marginal cost of effort increases as well, i.e., $C''(Me_k) > 0$, $C''(De_k) > 0$.

Suppose that federated learning is successful in stage k. In that case, the data owner receives the incremental value of the upgraded model as V_k (V_k is a constant agreed upon by both participants before the contract), and the training continues into stage $k+1$. Assuming that the model's market value at the end of stage k is M_k and the data owner's cost at stage k is I_k, we have $V_k = M_k - I_k$. After all K stages of training have been completed, the data owner receives the final value of the model as $\sum_{k=1}^{K} V_k = \sum_{k=1}^{K}(M_k - I_k)$.

DR_k and MR_k are defined as the total expected revenues of the data owner and model owner from stage k to K. Logically, the data owner will only participate in training if they believe that the total expected revenue will be positive. If the total expected revenue in stages k to K turns out to be a loss, the data owner will drop out at any stage from $k+1$ to K and terminate the contract. Therefor, we can assume that $V_k + DR_{k+1} > 0$ and $DR_k \geq 0$. This assumption is reasonable because it assumes that the parties have some opportunity to argue success or failure at each stage. If the data owner expects a negative payoff, they will claim failure to get out of the contract. It is assumed that before a particular point in the training $V_k < 0$, i.e., the data owner's contribution is more significant than the benefit. After that point, the data owner's payoff becomes positive. This assumption ensures that the

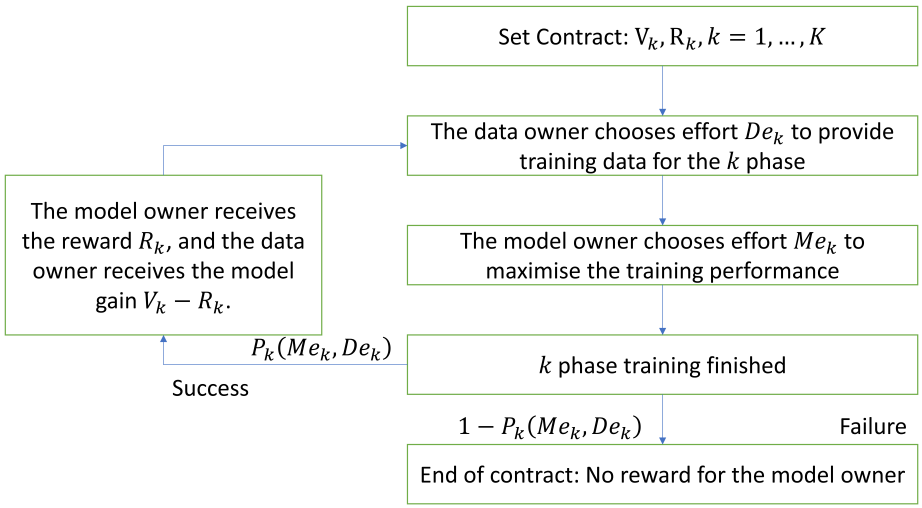

Figure 3.3: Federated Learning Contract Execution Stages

data owner agrees to cooperate with the model owner for the purposes of training the model. R_k represents the reward given by the data owner to the model owner if the training is successful at stage k. The event sequence in the contract is shown in Fig. 3.3.

Before entering the federated learning scheme, the data owner and the model owner need to agree on the reward $R_k > 0 \quad (k = 1, \cdots, K)$ and set up the contract. The model owner receives R_k from the data owner after training is confirmed to be successful in stage k. According to the contract, the model owner commits the optimal level of effort ${Me_k}^*$ to maximise their expected return MR_k. At the same time, the data owner also to commit the optimal level of effort ${De_k}^*$ to maximise DR_k. If the training result is successful at the end of stage k, the value of the updated model held by the data owner increases by V_k, and the model owner receives the reward R_k from the data owner. Training then proceeds to the next stage. If stage k training fails, both the model owner and the data owner gain nothing for that stage. Note that the optimal strategy for the Stackelberg game leader is to not reward the follower for failure at each stage of the game [69], [70]. Both parties will pay $C(Me_k)$ and $C(De_k)$ regardless of success or failure. Thus, the following recursive equation describes the profit of the data owner and the model owner,

$$MR_k = P_k(Me_k, De_k)[R_k + MR_{k+1}] - C(Me_k) \tag{3.1}$$

and

$$DR_k = P_k(Me_k, De_k)[V_k - R_k + DR_{k+1}] - C(De_k), k = 1, \cdots, K. \tag{3.2}$$

In our model, the contract is set before the first phase. The relevant payoffs in the first phase are DR_1 for the data owner and MR_1 for the model owner. Note that the payoff for stage k is directly effected by the payoffs for stage $k+1$. Expanding the above recursive equations, we have:

$$\begin{aligned} MR_m = &\sum_{k=m}^{K} \left\{ \prod_{j=m}^{k} P_j(Me_j, De_j) R_k \right\} \\ &- \sum_{k=m}^{K} \left\{ \prod_{j=m}^{k-1} P_j(Me_j, De_j) C(Me_k) \right\} \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} DR_m = &\sum_{k=m}^{K} \left\{ \prod_{j=m}^{k} P_j(Me_j, De_j)(V_k - R_k) \right\} \\ &- \sum_{k=m}^{K} \left\{ \prod_{j=m}^{k-1} P_j(Me_j, De_j) C(De_k) \right\}. \end{aligned} \tag{3.4}$$

3.2.2 Research Findings

In this section, we outline the findings of the above model, beginning with the optimal effort ${De_k}^*$ of the data owner.

The derivative of the data owner's payoff with respect to their effort De_k from Equation 3.2 is

$$\begin{aligned} \frac{dDR_k}{dDe_k} =& \frac{dP_k(Me_k, De_k)}{dDe_k}(V_k - R_k + DR_{k+1}) - \frac{dC(De_k)}{dDe_k} \\ =& 0 \quad (k = 1, \cdots, K), \end{aligned} \tag{3.5}$$

where

$$\frac{dP_k(Me_k, De_k)}{dDe_k}(V_k - R_k + DR_{k+1}) = \frac{dC(De_k)}{dDe_k} \quad (k = 1, \cdots, K). \tag{3.6}$$

Thus, the optimal effort ${De_k}^*$ of the data owner is:

$${De_k}^* = {De_k}^*(V_k - R_k + DR_{k+1}). \tag{3.7}$$

Corollary 1. *The optimal effort of the data owner is a function of the incremental value of the model, the reward to the model owner, and the data owner's expectation of future payoffs. Reducing the reward to the model owner and increasing the incremental value of the model and the data owner's expectations for the future should motivate the data owner to put in more effort and reduce their ethical risk.*

In the same way, we can solve the optimal effort ${Me_k}^*$ of the model owner. The derivative of the model owner's payoff with respect to it's effort Me_k from equation 3.1 is

$$\begin{aligned} \frac{dMR_k}{dMe_k} =& \frac{dP_k(Me_k, {De_k}^*)}{dMe_k}(R_k + MR_{k+1}) - \frac{dC(Me_k)}{dMe_k} \\ =& 0 \quad (k = 1, \cdots, K). \end{aligned} \tag{3.8}$$

Thus, the optimal effort ${Me_k}^*$ of the model owner is:

$${Me_k}^* = {Me_k}^*(R_k + MR_{k+1}). \tag{3.9}$$

Corollary 2. *The optimal effort level of the model owner is positively correlated with the reward and their expected future payoff. Higher rewards from the data owner and increasing the model owner's future expectations should motivate the model owner to work harder and reduce any ethical risks.*

Based on Corollaries 1 and 2, we have the following conditions:

$$\begin{cases} \frac{dP_k(Me_k, De_k)}{dDe_k}(V_k - R_k + DR_{k+1}) = \frac{dC(De_k)}{dDe_k}; \\ \frac{dP_k(Me_k, De_k)}{dMe_k}(R_k + MR_{k+1}) = \frac{dC(Me_k)}{dMe_k} \quad (k = 1, \cdots, K). \end{cases} \tag{3.10}$$

Corollary 3. *An optimal incentive mechanism should be such that the marginal benefit of each participant's effort equals their marginal cost.*

Given the optimal level of effort ${Me_k}^*$ and ${De_k}^*$ for the model owner and data owner, MR_m in Equation 3.3 satisfies the following conditions:

$$\frac{\partial MR_m}{\partial R_k} = \prod_{j=m}^{k} P_j({Me_j}^*, {De_j}^*) \quad (k = 1, \cdots, K; m \leq k). \tag{3.11}$$

From Equation 3.11,

$$\begin{aligned} \frac{\frac{\partial MR_1}{\partial R_k}}{\frac{\partial MR_1}{\partial R_{k+1}}} = \frac{\prod_{j=1}^{k} P_j({Me_j}^*, {De_j}^*)}{\prod_{j=1}^{k+1} P_j({Me_j}^*, {De_j}^*)} = \frac{1}{P_{k+1}({Me_{k+1}}^*, {De_{k+1}}^*)} > 1 \\ (k = 1, \cdots, K-1). \end{aligned} \tag{3.12}$$

Then

$$\left.\frac{\partial MR_1}{\partial R_k}\right|_{{Me_j}^*, {De_j}^*} > \left.\frac{\partial MR_1}{\partial R_{k+1}}\right|_{{Me_j}^*, {De_j}^*} \quad (k = 1, \cdots, K-1). \tag{3.13}$$

Corollary 4. *The marginal utility of the rewards diminishes for the model owner over time. Therefore, to encourage the model owner to increase their effort, the rewards for the model owner in the incentive mechanism should be gradually increased as training continues. This should mean the incentive mechanism stays effective in motivating the model owner to work hard.*

The optimal incentive ${R_k}^* > 0 \quad (k = 1, \cdots, K)$ for the model owner is determined before starting the first stage of training. Therefore, the optimal payoff ${R_k}^*$ of the data owner can also be solved. The first-order condition of data owner with respect to payoff R_k from Equation 3.2 is

$$\begin{aligned} \left.\frac{\partial DR_1}{\partial R_k}\right|_{{R_i}^*, i=1,\cdots,K} = \Bigg[& {P_1}'({Me_1}^*, {De_1}^*) {Me_1}^{*\prime} \frac{\partial MR_2}{\partial R_k} \\ & + {P_1}'({Me_1}^*, {De_1}^*) {De_1}^{*\prime} \frac{\partial DR_2}{\partial R_k} \Bigg] (V_1 - R_1 + DR_2) \\ & + P_1({Me_1}^*, {De_1}^*) \frac{\partial DR_2}{\partial R_k} - C'(De_1^*) De_1^{*\prime} \frac{\partial DR_2}{\partial R_k} = 0. \end{aligned} \tag{3.14}$$

From Corollary 4, we can derive $\frac{\partial MR_2}{\partial R_k} = \prod_{j=2}^{k} P_j({Me_j}^*, {De_j}^*)$ and from Corollary 1, we can derive $P_1'({Me_1}^*, {De_1}^*)(V_1 - R_1 + DR_2) - C'({De_1}^*) = 0, V_1 - R_1 + DR_2 > 0$. Substituting both of these into Equation 3.14 and rearranging the terms yield:

$$\begin{aligned}&\left\{De_1{}^{*\prime}[P_1{}'(Me_1{}^*, De_1{}^*)(V_1 - R_1 + DR_2) - C'(De_1{}^*)] + P_1(Me_1{}^*, De_1{}^*)\right\}\\&\frac{\partial DR_2}{\partial R_k} + P_1{}'(Me_1{}^*, De_1{}^*)Me_1{}^*\left[\prod_{j=2}^{k} P_j(Me_j{}^*, De_j{}^*)\right](V_1 - R_1 + DR_2) = 0.\end{aligned} \tag{3.15}$$

Then,

$$\begin{aligned}\left.\frac{\partial DR_2}{\partial R_k}\right|_{R_i{}^*, i=1,\cdots,K} =& -\frac{1}{P_1(Me_1{}^*, De_1{}^*)}P_1{}'(Me_1{}^*, De_1{}^*)Me_1{}^{*\prime}\\&\left[\prod_{j=2}^{k} P_j(Me_j{}^*, De_j{}^*)\right](V_1 - R_1 + DR_2) < 0.\end{aligned} \tag{3.16}$$

Thus, if $R_k{}^* > 0$ and $R_j{}^* > 0$,$j > k$, then

$$\begin{aligned}\left.\frac{\partial DR_2}{\partial R_j}\right|_{R_i{}^*, i=1,\cdots,K} &= \left(\prod_{i=k+1}^{j} P_i(Me_i{}^*, De_i{}^*)\right)\left.\frac{\partial DR_2}{\partial R_k}\right|_{R_i{}^*, i=1,\cdots,K}\\&> \left.\frac{\partial DR_2}{\partial R_k}\right|_{R_i{}^*, i=1,\cdots,K}.\end{aligned} \tag{3.17}$$

Corollary 5. *The expected payoff to the model owner increases marginal utility for the data owner over time. Intuitively, the data owner always wants to delay the reward to the model owner, while the model owner wants to receive the reward as early as possible. For the data owner, the later the reward is given to the model owner, the more likely it is for ethical risk to be avoided.*

From Corollary 5, for $k > 1$,

$$\begin{aligned}\frac{\partial DR_1}{\partial R_k} =& \left[P_1(Me_1{}^*, De1*)'Me_1{}^{*\prime}\frac{\partial MR_2}{\partial R_k} + P_1(Me_1{}^*, De1*)'De_1{}^{*\prime}\frac{\partial DR_2}{\partial R_k}\right]\\&(V_1 - R_1 + DR_2) + P_1(Me_1{}^*, De1*)\frac{\partial DR_2}{\partial R_k} - C(De_1{}^*)'De_1{}^{*\prime}\frac{\partial DR_2}{\partial R_k}.\end{aligned} \tag{3.18}$$

For every $m < k$,

$$\begin{aligned}\frac{\partial DR_m}{\partial R_k} =& \left[P_m(Me_m{}^*, De_m*)'Me_m{}^{*\prime}\frac{\partial MR_{m+1}}{\partial R_k}\right.\\&\left.+\ P_m(Me_m{}^*, De_m*)'De_m{}^{*\prime}\frac{\partial DR_{m+1}}{\partial R_k}\right](V_m - R_m + DR_{m+1})\\&+ P_m(Me_m{}^*, De_m*)\frac{\partial DR_{m+1}}{\partial R_k} - C(De_m{}^*)'De_m{}^{*\prime}\frac{\partial DR_{m+1}}{\partial R_k},\end{aligned} \tag{3.19}$$

and for every k,

$$\begin{aligned}\frac{\partial DR_k}{\partial R_k} =&[P_k({Me_k}^*, De_k*)'{Me_k^*}' - P_k({Me_k}^*, De_k*)'{De_k^*}'](V_k - R_k + DR_{k+1}) \\ &- P_k({Me_k}^*, De_k*) + C({De_m}^*)'{De_k^*}'.\end{aligned} \tag{3.20}$$

From Corollary 1, we can derive $P_k({Me_k}^*, De_k*)'(V_k - R_k + DR_{k+1}) - C(De_k)^{*\prime} = 0$, and substituting this into the three equations above, we have

$$\begin{aligned}\frac{\partial DR_1}{\partial R_k} =& \left(\prod_{j=1}^{k} P_j({Me_j}^*, {De_j}^*)\right) \sum_{i=1}^{k} \frac{1}{P_i({Me_i}^*, {De_i}^*)} \\ &{P_i}'({Me_i}^*, {De_i}^*){Me_i}^*[V_i - R_i + DR_{i+1}] - \prod_{j=1}^{k} P_j({Me_j}^*, {De_j}^*),\end{aligned} \tag{3.21}$$

and

$$\begin{aligned}\frac{\partial DR_1}{\partial R_{k+1}} =&\frac{\partial DR_1}{\partial R_k} P_{k+1}({Me_{k+1}}^*, {De_{k+1}}^*) \\ &+ \left(\prod_{j=1}^{k} P_j({Me_j}^*, {De_j}^*)\right) {P_{k+1}}'({Me_{k+1}}^*, {De_{k+1}}^*) \\ &{Me_{k+1}^*}'[V_{k+1} - R_{k+1} + DR_{k+2}] = 0 \quad (k = 1, \cdots, K-1).\end{aligned} \tag{3.22}$$

Since $\frac{\partial DR_1}{\partial R_k}\Big|_{R_i{}^*, i=1,\cdots,K} = 0$, from Equation 3.22, we can derive $V_{k+1} - R_{k+1} + DR_{k+2} = 0$. It is known that $DR_{K+1} = 0$, so it follows that ${R_K}^* = V_K$, so $DR_K = 0$. Similarly, for any δ, there is $1 \le \delta \le K-1$. If ${Me_\delta}^* > 0$ and ${R_\delta}^* > 0$, then:

$$\begin{cases} {R_k}^* = V_k \quad (k = \delta + 1, \cdots, K). \\ DR_k = 0 \quad (k = \delta + 1, \cdots, K). \end{cases} \tag{3.23}$$

Then,

$$\begin{aligned}DR_1 =& \sum_{j=1}^{\delta-1} \left(\prod_{i=1}^{j} P_i({Me_i}^*, {De_i}^*)(V_j - C({De_i}^*))\right) \\ &+ \left(\prod_{i=1}^{\delta-1} P_i({Me_i}^*, {De_i}^*)\right) P_\delta({Me_\delta}^*, {De_\delta}^*)[V_\delta - R_\delta].\end{aligned} \tag{3.24}$$

theorem 1. *The data owner can receive their optimal payoff at point δ during training such that*

$$\begin{cases} {R_k}^* = 0 \quad (k < \delta), \\ {R_k}^* = {V_k}^*, {DR_k}^* = 0 \quad (k > \delta), \end{cases} \tag{3.25}$$

and

$$\begin{cases} DR_1 \geq \sum_{j=1}^{\delta-1} \left(\prod_{i=1}^{j} P_i({Me_i}^*, {De_i}^*)\big(V_j - C({De_j}^*)\big) \right), \\ DR_1 \leq \sum_{j=1}^{\delta} \left(\prod_{i=1}^{j} P_i({Me_i}^*, {De_i}^*)\big(V_j - C({De_j}^*)\big) \right). \end{cases} \tag{3.26}$$

Theorem 1 shows an optimal payoff point for the data owner, where the data owner receives the total payoff from the federated learning process and the reward given to the model owner is zero in phases $1 - \delta$. However, after that point, the data owner does not have any profit, the expected future payoffs are zero, and the benefit goes entirely to the model owner. Thus, point δ is the optimal payoff point for the data owner. Essentially, what Theorem 1 indicates is that, for a federated learning scenario initiated by the data owner, the optimal incentive scheme is one where as much of the incremental value of the model as possible is paid to the model owner. Therefore, success in the later stages of training is based on the success in the earlier stages and, in turn, rewards in the later stages incentivise effort in the earlier stages. Overall, giving back as much of the value created by the model owner's efforts as possible in the later stages is the least costly incentive scheme for the data owner.

3.3 Experimental Evaluation

To complement the analytical findings and evaluate the performance of our incentive mechanism for federated learning, we create a multi-stage contract simulator for the data and model owners. The simulator evaluates the impact of different reward settings on the level of effort contributed by each participant and gives the total payoff for both parties.

3.3.1 Experiment Settings

Assume that the incremental model values are $V_1 = 1, V_2 = 2$ and $V_3 = 3$, where federated learning is carried out in 3 stages (i.e., $K = 3$) and the functional expres-

sion for the probability of success at each stage is $P_k(Me_k, De_k) = MIN(0.6(Me_k + De_k), 1)$. As we will see later, the equilibrium effort satisfies $0.6({Me_k}^* + {De_k}^*) < 1$, so we can count $P_k(Me_k, De_k) = 0.6(Me_k + De_k)$. We also assume that the effort cost of the model owner's function is $C(Me_k) = {Me_k}^2$, and the effort cost of the data owner's function is $C(De_k) = {De_k}^2$, such that the utility function of the model owner is

$$mr_k = 0.6(Me_k + De_k)(R_k + mr_{k+1}) - {Me_k}^2, \quad k = 1, 2, 3,$$

$$mr_4 = 0.$$

Taking the utility function for each stage and deriving it to its effort level determines the optimal effort yield for the model owner:

$${Me_k}^* = \frac{\partial mr_k}{\partial Me_k} = 0.3(R_k + mr_{k+1}) \quad k = 1, 2, 3.$$

Repeating the same approach, and its based on Equation 3.24, we can derive the utility function of the data owner and their optimal effort:

$$dr_k = 0.6(Me_k + De_k)(V_k - R_k + dr_{k+1}) - {De_k}^2, \quad k = 1, 2, 3,$$

$$dr_3 = 0, dr_4 = 0.$$

$${De_k}^* = \frac{\partial dr_k}{\partial De_k} = 0.3(V_k - R_k + dr_{k+1}) \quad k = 1, 2, 3,$$

$$De_3 = 0.$$

The utility functions and the optimal efforts of the two parties in different stages are listed in Table 3.2.

3.3.2 Experimental Result and Discussion

Fig. 3.4 shows the optimal rewards yielded for the model owner, calculated by recurring the above equations in Table 3.2 and the derivative of the data owner's payoff dr_1 with respect to the reward ${R_2}^*$:

$${R_1}^* = 0, {R_2}^* = 0.4085, {R_3}^* = 3,$$

K	**Data Owner**	**Model Owner**
1	$dr_1 = 0.6(Me_1 + De_1)(V_1 - R_1 + dr_2) - {De_1}^2$ $De_1 = 0.3(V_1 - R_1 + dr_2)$	$mr_1 = 0.6(Me_1 + De_1)(R_1 + mr_2)$ $Me_1 = 0.3(R_1 + mr_2)$
2	$dr_2 = 0.6(Me_2 + De_2)(V_2 - R_2 + dr_3) - {De_2}^2$ $De_2 = 0.3(V_2 - R_2 + dr_3)$	$mr_2 = 0.6(Me_2 + De_2)(R_2 + mr_3)$ $Me_2 = 0.3(R_2 + mr_2)$
3	$dr_3 = 0$ $De_3 = 0$	$mr_3 = 0.6(Me_3 + De_3)(R_3 + mr_4)$ $mr_4 = 0, Me_3 = 0.3(R_3 + mr_3)$

Table 3.2: Optimal Efforts and Corresponding Utility Functions for Model and Data Owners

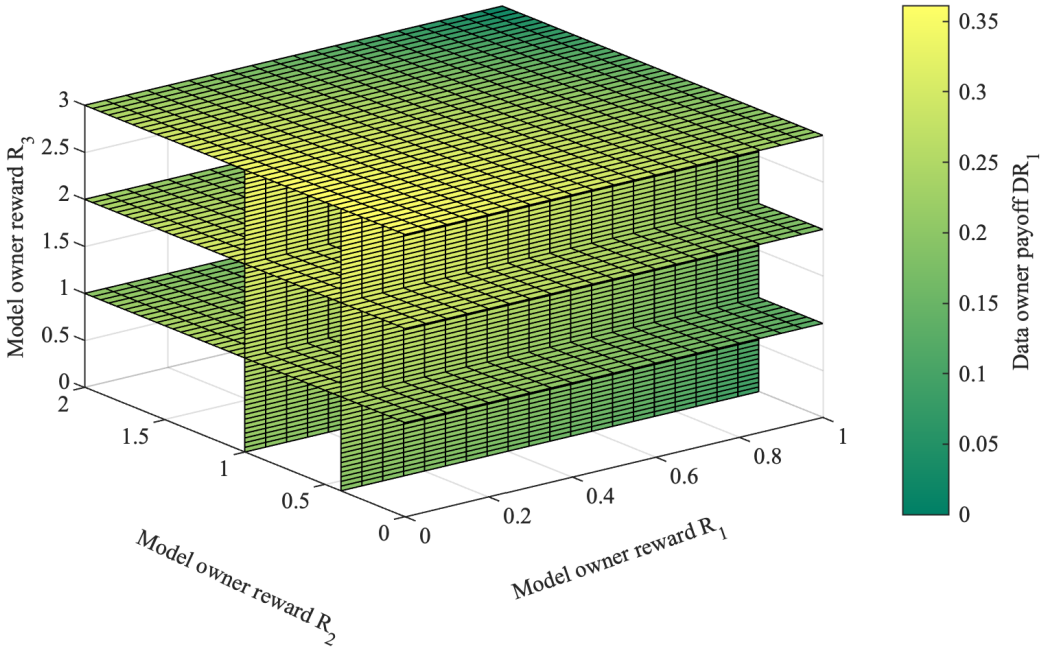

Figure 3.4: Optimal Rewards Yielded for Model Owner

where the probabilities of successes are $P_1({Me_1}^*, {De_1}^*) = 0.3707, P_2({Me_2}^*, {De_2}^*) = 0.5058, P_3({Me_3}^*, {De_3}^*) = 0.54$. As predicted by Theorem 1, the optimal payoff point for the data owner is $\delta = 2$, and ${R_1}^* = 0, R_3* = V_3$, and $0 < R_2 < V_2$. The data owner's expected payoff is $dr_1 = 0.3608$, which is consistent with Theorem 1,

$$\begin{cases} dr_1 \geq P_1({Me_1}^*, {De_1}^*)(V_1 - {De_1}^2) = 0.2878, \\ dr_1 \leq P_1({Me_1}^*, {De_1}^*)(V_1 - {De_1}^2) + P_2({Me_2}^*, {De_2}^*)(V_2 - {De_2}^2) = 1.1841. \end{cases}$$

Reward settings	**DO expected payoff** dr_1	**Stg1 BEs** $Me_1 + De_1$	**Stg2 BEs** $Me_2 + De_2$	**Stg3 BEs** $Me_3 + De_3$
$R_1 = 0, R_2 = 0.2, R_3 = 3$	0.3508	0.6114	0.843	0.9
${R_1}^* = 0, {R_2}^* = 0.4085, {R_3}^* = 3$	***0.3608***	***0.6179***	***0.843***	***0.9***
$R_1 = 0, R_2 = 1, R_3 = 3$	0.3386	0.6109	0.843	0.9
$R_1 = 0.5, R_2 = 0.4085, R_3 = 3$	0.2949	0.6179	0.843	0.9

Table 3.3: Optimal Efforts at Each Stage Under Varying Reward Settings

We have taken some relevant data from the simulator to make it easier to understand, as shown in Table 3.3. This table shows the effects of the reward value settings at different stages on the efforts of the participants and the expected payoff for the data owner in the incentive contract. Some settings around the optimal one have been selected as comparisons: ${R_1}^* = 0, {R_2}^* = 0.4085, {R_3}^* = 3$. From the results, we can see that:

1. Any deviation from the optimal value of ${R_2}^* = 0.4085$ negatively impacts the efforts of both participants and the expected training payoff for the data owner. This means that any reward setting that deviates from the optimal value ${R_2}^*$ will increase the ethical risk of the participants.

2. If the data owner keeps $R_2 = {R_2}^*$ and increases the reward R_1 for stage 1, this scenario is identical to the optimal incentive scenario in terms of the effort values at each stage. However, the data owner's expected training payoffs will be significantly lower. From a self-interested perspective by the data owner, as the leader of the incentive contract, there is no incentive to increase the reward given to the model owner at Stage 1.

Thus, we can conclude that our model is able to reduce the dual ethical risk of federated learning due to information asymmetry. It can motivate the participants to exert an optimized effort to training, confirming the intuition behind our model that the success in the later stages is based on success in the earlier stages. Thus, rewards in the later stages incentivise efforts in the earlier stages. Moreover, giving back as much of the value created by the model owner's efforts in the later stages is the least costly incentive scheme for the data owner.

3.4 Conclusion and Chapter Discussion

In this chapter, we have used the framework of a dynamic game to investigate the dual ethical risk problem between model owners and data owners in federated learning. The model used is novel, and it has derived optimal incentive payoff contracts for the data and model owners through two sets of analyses: one for a multi-stage incentive payoff game and the other for the dual ethical risk affecting the contract design. The output is an optimal payoff point for the data owners. Our approach has provided insights into the characteristics of optimal incentive contracts between data owners and model owners in federated learning schemes, including their endogenous optimality. Specifically, our study has shown that, within a data owner-initiated federated learning scenario, a significant portion of the model's incremental value can be optimally allocated to the model owner in the later stages, fostering participation and collaboration.

This research can influence how federated learning incentive mechanisms are designed in real-world applications, particularly in sensitive domains where trust and equitable compensation are paramount. This chapter focused on the dual ethical risk problem in a data owner-led federated learning scenario utilising a multi-stage incentive model. To further advance this research, subsequent chapters will explore the integration of this model into federated learning frameworks and its implications. Chapter 4 will delve into the model's adaptation to various scenarios, providing a comprehensive comparison with existing state-of-the-art frameworks. Chapter 5 will expand this research to address the complexities introduced by multiple data owners joining the federated learning collaboration, building upon the foundational model

presented in this chapter.

Chapter 4

Fair Clearing House Framework for Secure and Trustworthy Federated Learning

4.1 Introduction

The Industrial Internet of Things (IIoT) proliferation has ushered in a new era of data-driven insights that have transformed traditional sectors such as agriculture[71], supply chain management[72], and logistics[73]. By deploying a network of interconnected sensors and devices, IIoT systems gather a wealth of data that can be harnessed to develop powerful AI-based models[74], unlocking the potential for increased efficiency and productivity.

In smart agriculture[75], for instance, IIoT devices strategically placed in crop fields can capture valuable environmental data and images[76]. This information can then be used to train sophisticated machine-learning models capable of rapidly identifying pest infestations[77], optimising resource allocation, and enhancing crop yields.

However, the efficacy of these models often hinges on the availability of extensive and diverse training datasets. An IIoT network deployed by a single farm may not encompass the breadth and depth of data required to train models that generalize well across various scenarios. For example, an individual farm may lack sufficient

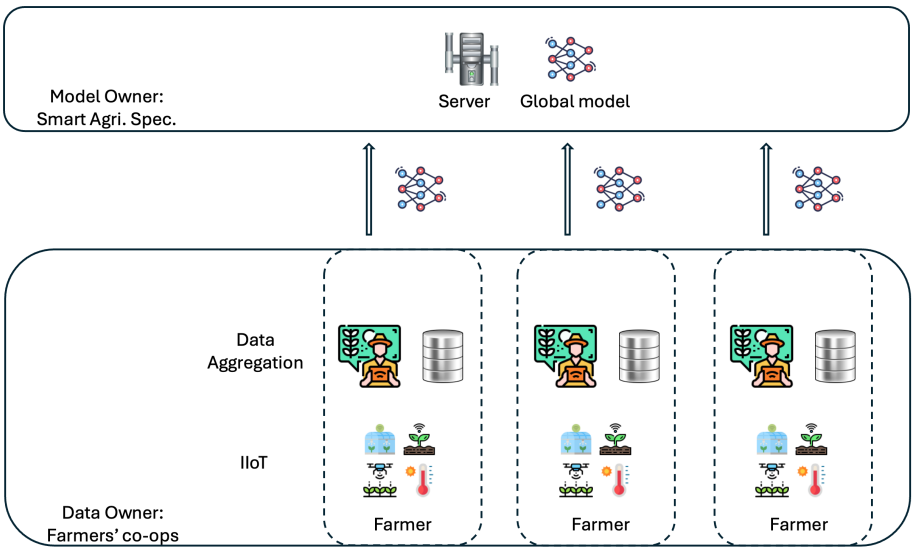

Figure 4.1: Federated Learning for Smart Agricultural

samples of diseased crop images to train a model that can reliably detect and diagnose a wide range of plant pathologies[78].

To address this data scarcity challenge, agricultural cooperatives, also known as farmers' co-ops, emerge as a viable solution for collaborative model development. By pooling the data from multiple farms, these cooperatives can aggregate a much larger and more diverse dataset, enabling the training of more robust and accurate models. However, direct sharing of sensitive raw data within the cooperative still raises concerns regarding data privacy and potential misuse. Federated learning presents a promising paradigm to overcome these concerns.

In our proposed federated learning system (Figure 4.1), multiple farmers, represented by an agricultural cooperative, collaborate with a model owner, such as a company specializing in smart agriculture applications. Farmers collect and store data through their IIoT devices on their respective servers. When the model owner initiates a training request, a set of model parameters is directly distributed to each farmer. Subsequently, farmers leverage their local data to train the model and send only the updated parameters back to the model owner for aggregation. This iterative process of local training and global aggregation continues until the desired model accuracy is achieved.

By adopting this federated learning approach, multiple data owners can collaboratively train a robust and accurate model for their innovative industrial activities while safeguarding sensitive data. Moreover, the model owner can receive the training reward after the cooperation, fostering innovation and knowledge sharing within the IIoT ecosystem and opening new avenues for addressing challenges and optimizing processes in various industrial sectors.

However, despite its potential, federated learning in the IIoT domain faces critical challenges in incentive alignment and trust-building. Due to differing knowledge structures and priorities, data and model owners often encounter divergent interests, a phenomenon known as information asymmetry. This asymmetry and the inherent uncertainties of machine learning outcomes can lead to suboptimal scenarios. Data owners may hesitate to invest in federated learning projects and dedicate resources to data collection due to concerns about fair compensation for their investment and potential exploitation of their data. Conversely, model owners may prioritise their gains, potentially neglecting the collective benefit of the collaboration.

Additionally, the data owner-led nature of this type of federated learning, where the data owners ultimately own the final model, further complicates the establishment of trust and equity. Traditional model owner-led mechanisms, in which the model owner retains ownership of the final model, are ill-suited for this environment. It necessitates the development of innovative frameworks to ensure fair model-reward settlements between the model and data owners based on their agreements.

In this chapter, we introduce the Fair Clearing House (FCH) framework to address these challenges. Leveraging the immutable and transparent nature of blockchain technology and the self-executing capabilities of smart contracts, the FCH creates a decentralised and trustless platform for federated learning collaborations. By incorporating a multi-stage incentive mechanism and a secure two-party clearing protocol, the FCH framework aims to foster trust, fairness, and cooperation while mitigating the risks associated with data sharing and reward distribution. we make several significant contributions to the field of federated learning:

- We present a secure and efficient clearing protocol designed explicitly for two-party federated learning transactions. This protocol offers strong privacy guar-

antees, reduces reliance on third parties, and utilises a lightweight smart contract.

- Our framework demonstrates how the two-party clearing protocol can seamlessly integrate within a federated learning architecture built upon the multi-stage incentive mechanisms from Chapter 3. This integration fosters trust, transparency, and adaptability and significantly reduces real-world settlement risks.

- We empirically demonstrate that our framework outperforms conventional federated learning frameworks in various metrics.

The remainder of the chapter is structured as follows. In Section 4.2, we highlight the preliminaries of our work. Section 4.3 describes our framework and analysis, and Section 4.4 presents our implementation and performance evaluation. We will conclude with future work in Section 4.5.

4.2 Preliminaries

This section introduces the notations, cryptographic building blocks and smart contracts used in this chapter.

Cryptographic building blocks The FCH framework leverages cryptographic hash functions and Merkle trees as its primary cryptographic primitives. A hash function, denoted as $H : \{0,1\}^* \rightarrow \{0,1\}^\mu$, is a one-way mathematical algorithm that maps data of arbitrary size to a fixed-length binary string (a hash value) of length μ. To ensure the security and integrity of the FCH framework, the hash function must satisfy specific properties:

- **Collision Resistance:** It should be computationally infeasible to find two different inputs that produce the same hash value. This property is essential to prevent tampering and ensure data authenticity within the FCH framework. Specifically, it ensures that the data owner cannot forge a fake dataset with the same hash value as the validation dataset locked by the model owner, thereby guaranteeing the integrity of the model validation process.

- **Hiding:** The hash value should not reveal any information about the original input data, protecting the confidentiality of sensitive information. In the context of the FCH, this property ensures that the models and validation datasets shared between the data owner and the model owner remain private and cannot be inferred from the hash value. This means that even if an attacker obtains the hash value, they cannot determine the original data used to generate it.

- **Binding:** Once a hash value of a message is committed, it should be computationally infeasible to find a different message that produces the same hash value. This property guarantees the integrity of the commitments made in the FCH framework. For example, if a data owner commits to a training dataset using its hash value, they cannot later deny or change their commitment. This ensures that all parties in the federated learning process can trust the integrity of the data and the model.

In practice, these security requirements are typically met by well-established hash functions such as SHA-256 or SHA-3 [79]. Our security analysis assumes that the hash function H is modelled as a global random oracle H [80]. Let H be a family of (t, ε), collision-resistant hash functions. A binding commitment scheme is then constructed to input messages x. The algorithm to return $h \underset{\mu}{\leftarrow} H$ is set as a public parameter. So, to submit x, $h(x)$ needs to be returned as the digest, and x needs to be returned as the decryption string. If $h(x) = c$, the message x is a valid opening for c. A cryptographically secure commitment scheme must satisfy the *hiding* and *binding* properties. Hiding guarantees that c and c' in any two messages x, x' and $c = Commit(x)$ and $c' = Commit(x')$ are computationally indistinguishable. The binding property requires that it is computationally hard to find a c such that $Open(c, x) = 1$, $Open(c, x') = 1$, and $x \neq x'$[81].

A Merkle tree[82], also known as a hash tree, is a fundamental cryptographic building block in the FCH framework. It enables efficient and secure verification of data integrity and consistency. A Merkle tree is a binary tree where each leaf node is labelled with the cryptographic hash of a data block, and

each non-leaf node is labelled with the cryptographic hash of its child nodes' labels. This structure verifies the integrity of the entire dataset by checking only a small portion of it. We utilise three key algorithms related to Merkle trees in this chapter:

- **Merkle Tree Hash (Algorithm 1):** This algorithm recursively computes the hash values of data blocks, ultimately producing a single root hash representing the entire dataset.

Algorithm 1 Merkle Tree Hash MerkleTreeHash

```
1: function MerkleTreeHash(dataBlocks)
2:     leafNodes ← [hash(block) for block in dataBlocks]
3:     while len(leafNodes) > 1 do
4:         if len(leafNodes) % 2 == 1 then
5:             leafNodes.append(leafNodes[-1])    ▷ Duplicate last node if odd number
6:         end if
7:         intermediateNodes ← []
8:         for i in range(0, len(leafNodes), 2) do
9:             intermediateNodes.append(hash(leafNodes[i] + leafNodes[i + 1]))
10:        end for
11:        leafNodes ← intermediateNodes
12:    end while
13:    return leafNodes[0]                          ▷ Merkle root
14: end function
```

- **Merkle Tree Proof (Algorithm 2):** This algorithm generates a sequence of hash values (a Merkle proof) that can be used to verify the presence and integrity of a specific data block within the tree.

Algorithm 2 Merkle Tree Proof MerkleTreeProof

```
1: function MerkleTreeProof(dataBlocks, blockIndex)
```

```
    proof ← []
    currentIndex ← blockIndex
    leafNodes ← [hash(block) for block in dataBlocks]
    while len(leafNodes) > 1 do
        if len(leafNodes) % 2 == 1 then
            leafNodes.append(leafNodes[-1])    ▷ Duplicate last node if odd number
        end if
        if currentIndex % 2 == 1 then
            proof.append(leafNodes[currentIndex - 1])
        else
            if currentIndex + 1 < len(leafNodes) then
                proof.append(leafNodes[currentIndex + 1])
            end if
        end if
        currentIndex ← currentIndex // 2
        leafNodes ← [hash(leafNodes[i] + leafNodes[i + 1]) for i in range(0, len(leafNodes), 2)]
    end while
    return proof
end function
```

- **Merkle Tree Proof Verification (Algorithm 3):** This algorithm takes a root hash, a data block, and its corresponding Merkle proof to ascertain whether the data block is indeed part of the dataset represented by the root hash.

Algorithm 3 Merkle Tree Proof Verification MerkleTreeProofVerify

```
function MerkleTreeProofVerify(rootHash, dataBlock, proof)
    blockHash ← hash(dataBlock)
    for siblingHash in proof do
        if blockHash < siblingHash then
            blockHash ← hash(blockHash + siblingHash)
```

```
        else
            blockHash ← hash(siblingHash + blockHash)
        end if
    end for
    return blockHash == rootHash
end function
```

By leveraging Merkle trees and their associated algorithms, the FCH framework ensures the integrity and authenticity of data throughout the federated learning process, strengthening trust and fairness within the system.

Smart contracts Smart contracts are self-executing programs stored on a blockchain that automatically enforce the terms of an agreement when predetermined conditions are met. In the FCH framework, smart contracts play a pivotal role in facilitating secure and transparent transactions between data owners and model owners. These contracts automate the settlement process, ensuring that rewards are disbursed only when the agreed-upon performance criteria are met. By leveraging smart contracts, the FCH framework minimises the need for intermediaries, reduces transaction costs, and enhances federated learning collaborations' overall efficiency and trustworthiness.

Ethereum[83], a prominent public blockchain ecosystem, supports smart contracts. It provides a robust infrastructure for deploying and executing smart contracts. Ethereum's Solidity programming language enables the development of flexible and secure smart contracts tailored to the specific requirements of the FCH framework. Executing smart contracts on the Ethereum network requires the payment of transaction fees, also known as "gas." This gas compensates the miners the miners who validate and add transactions to the blockchain. Each operation within a smart contract consumes a certain amount of gas, and the total gas cost of a transaction depends on the complexity of the contract's logic and the current network congestion. The exchange rate between Ether (ETH), the native cryptocurrency of Ethereum, and gas fluctuates based on market demand.

By leveraging the decentralized and trustless nature of the Ethereum block-

Notation	Description
k	Training stages, $k = 1, \cdots K$.
Me_k	The effort committed by the model owner at stage k
De_k	The effort committed by the data owner at stage k
$P_k(Me_k, De_k)$	The probability of successful training at stage k
$C(Me_k)$	The effort cost of the model owner at stage k
$C(De_k)$	The effort cost of the data owner at stage k
V_k	The incremental value of the model after stage k
M_k	The market value of the model at stage k
I_k	The data owner's costs at stage k
DR_k	Total expected revenue of the data owner from stage k to K
MR_k	Total expected revenue of the model owner from stage k to K
R_k	The reward received by the model owner if training success at stage k
O_k	The training objects at stage k

Table 4.1: Glossary of Key Mathematical Notations (Chapter 4)

chain and the flexibility of smart contracts, the FCH framework ensures the secure and transparent execution of federated learning collaborations, mitigating the risks associated with traditional centralized systems.

Notation. Table 4.1 lists the notations commonly used in this chapter for ease of reference.

4.3 The Fair Clearing House Framework

Our proposed framework introduces a multi-stage federated learning scheme incorporating a secure two-party reward-clearing protocol. This approach diverges from existing methods by dynamically allocating training incentives to federated learning participants at different stages. These incentives, defined in the training contract, serve to minimize unethical behaviour and optimize training results. Simultaneously, based on a smart contract, the secure two-party reward-clearing protocol ensures fair settlements between the model and data owners upon training completion. In this

context, we define *Fair Clearing* for federated learning as follows:

Definition *Fair Clearing refers to an incentive mechanism that discourages unethical behaviour among participants in a federated system. Predetermined rewards motivate participants to contribute their best efforts and achieve optimal training results. At the same time, participants can settle for optimized rewards based on their contributions and efforts within the program.*

Our framework must ensure the following security properties to meet this fair clearing definition:

i Data owner fairness Honest data owners must be confident that rewards are issued only for models that meet the agreed-upon performance criteria at each stage of the process.

ii Model owner fairness Honest model owners must be assured that they will receive the agreed-upon rewards for models that successfully meet the performance criteria at each stage.

iii Termination If at least one party is honest, the contract can be terminated after a limited number of rounds, ensuring the release of any remaining funds locked in the smart contract.

To achieve fair clearing in federated learning, our FCH framework leverages a smart contract as a neutral arbitrator. This arbitrator oversees the settlement process to determine whether the trained model satisfies the contract metrics. The protocol operates as follows: if the trained model passes a test on a random validation dataset, the model owner receives the contract rewards, and the data owner obtains the trained model. If not, the data owner does not receive the updated model, but they recover the unallocated rewards, and the federated learning contract automatically terminates. We do not perform the validation process to minimise the cost of executing the smart contract on the blockchain. Instead, we delegate the task of validating the model's results on the validation dataset to the model and data owners. The arbitrator only needs to compare the test result provided by the model owner with the true value provided by the data owner to determine the outcome. The workings of our framework are illustrated in Figure 4.2.

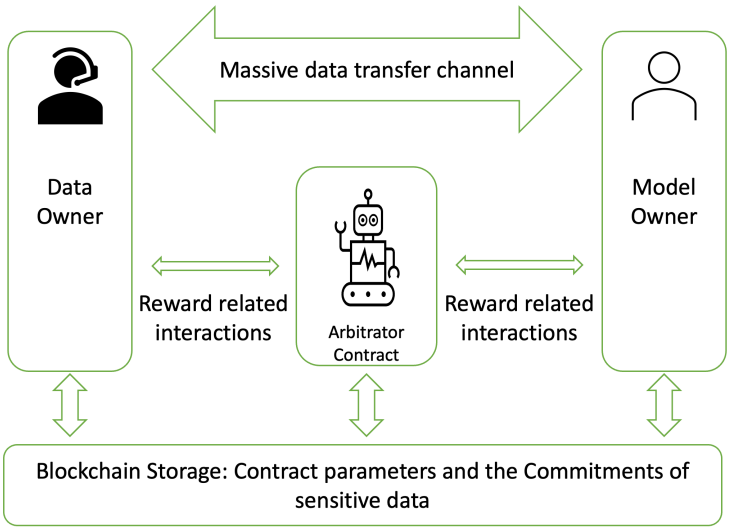

Figure 4.2: FCH Framework Workings

Before collaborating, the data and model owners must agree on critical initialisation information for the training contract. This information includes the number of training stages K, rewards for each stage R_k, model performance goals O_k, and the validation dataset $Data_k$ used for model acceptance.

The above parameters are stored in the arbitrator contract during the initialisation stage. The arbitrator also locks the $\sum_1^K R_k$ coins in the data owner's account.

The data owner and the model owner then start the federated learning process. At the end of each stage, both participants complete a fair settlement of the latest model, where rewards are issued for a successful training stage via a lightweight two-party clearing protocol. This clearing protocol, managed by the arbitrator, will exit after K loops or a failed training stage.

4.3.1 The Multi-stage Incentive Mechanism

The foundation of the FCH framework rests upon a multi-stage incentive mechanism introduced and analyzed in Chapter 3. This mechanism addresses the critical challenge of aligning incentives between data owners and model owners to foster a fair and trustworthy collaborative environment in federated learning.

In this context, the data owner acts as the principal, seeking to leverage the expertise of the model owner (the agent) for model training. However, information asymmetry,

where the data owner cannot directly observe the model owner's effort levels, creates opportunities for the model owner to shirk responsibility or submit subpar models.

To counteract this risk, the mechanism strategically divides the federated learning process into distinct stages. Each stage is associated with specific, verifiable training objectives and corresponding rewards. By tying compensation to the successful achievement of these objectives, the mechanism incentivizes the model owner to invest optimal effort throughout the entire training process. This dynamic reward structure aligns the interests of both parties, promoting collaboration and discouraging opportunistic behavior.

Moreover, the mechanism incorporates verifiable contribution measurements, allowing for the objective assessment of the model owner's contributions at each stage. This ensures fair and equitable reward allocation, mitigating the risk of free-riding and bolstering trust between participants.

The analysis conducted in Chapter 3 reveals a key insight: in a federated learning scenario led by the data owner, the optimal incentive scheme involves allocating a significant portion of the incremental value generated by the model to the model owner. This front-loading of rewards not only incentivizes the model owner to exert greater effort in the earlier stages but also ensures the success of subsequent stages, ultimately leading to a higher-quality final model for the data owner. By prioritizing the return on investment for the model owner in later stages, the data owner minimizes the overall cost of incentivization while maximizing the value derived from the collaboration.

To formalize this mechanism, we model the interaction between the data owner and the model owner as a multi-stage federated learning game. The data owner acts as the buyer of the trained model, and the model owner is the seller. Both participants agree to K training stages, with the understanding that the contract concludes if training fails at any stage.

At each stage k, the data owner and model owner invest efforts De_k and Me_k respectively. The success of the training at each stage, represented by the probability $P_k(Me_k, De_k)$. And the probability of success increases with higher effort from both parties but exhibits diminishing returns. The cost of efforts $C(Me_k)$ and $C(De_k)$

increases as the effort levels rise.

In the event of a successful training outcome at stage k, the data owner receives an improved model with an incremental value of V_k (a pre-agreed constant), and the process continues to the next stage. The total expected profit for the data owner and model owner from stage k to K is denoted as DR_k and MR_k respectively. A successful training at stage k results in a reward R_k paid by the data owner to the model owner.

Prior to training, the data owner and model owner agree on key parameters like the number of stages K, stage objectives O_k, and rewards R_k. At the end of each stage, they jointly verify the training outcome. If the outcome meets or exceeds the objective O_k, the data owner rewards the model owner with R_k and receives the updated model. Otherwise, neither party gains for that stage.

Under the contract, the model owner seeks to maximize their expected payoff $MR_k$by selecting the optimal effort ${Me_k}^*$, while the data owner aims to maximize their gain DR_k by choosing the optimal effort ${De_k}^*$. These efforts incur costs $C(Me_k)$ and $C(De_k)$, regardless of training success.

The profit of the data owner and the model owner can be formulated as follows:

$$DR_k = P_k(Me_k, De_k)[V_k - R_k + DR_{k+1}] - C(De_k), k = 1, \cdots, K \tag{4.1}$$

$$MR_k = P_k(Me_k, De_k)[R_k + MR_{k+1}] - C(Me_k), k = 1, \cdots, K \tag{4.2}$$

For a comprehensive understanding of the mathematical model underpinning this mechanism and its intricate details, including contract feasibility analysis and optimality conditions, please refer to Chapter 3.

The Two-party Clearing Protocol

To facilitate the fair and secure settlement of rewards and model exchange within the FCH framework, we introduce a novel two-party clearing protocol that leverages the underlying blockchain infrastructure and smart contracts. This protocol, illustrated

in Figure 4.3, is executed at the end of each training stage to clear the updated model and reward. The arbitrator assesses the information from both participants against the pre-defined contract conditions and finalizes the clearing process as agreed.

The protocol comprises the following rounds, which either loop K times or end if the training fails:

Round 1: The model owner encrypts the model using secret key z_i and sends the ciphertext to the data owner. Simultaneously, the model owner sends the commitments of key z_i to the arbitrator.

Round 2: The data owner transmits the validation dataset D_i to the model owner and the true labels of D_i to the arbitrator.

Round 3: The model owner checks the integrity of D_i. If valid, the model owner runs the latest model on the validation dataset, obtains the validation labels of D_i, and sends them to the arbitrator. Otherwise, a "Complain" message is sent to the arbitrator.

Round 4: The arbitrator compares the true and validation labels to verify the cooperation results. If successful, the arbitrator sends reward R_i to the model owner. If not, the arbitrator unfreezes the unallocated rewards and terminates the contract.

Round 5: Upon receiving the R_i coins, the model owner reveals key z_i to the data owner, completing the model clearing and reward distribution for the current stage.

Additional rounds are included to handle exceptions within the protocol. Suppose a participant aborts during the protocol's execution. In that case, the counterparty can submit a "finalize" request to the arbitrator, who will either unfreeze the data owner's unallocated rewards or allocate compensation to the model owner, depending on the exit point of the departing participant. The arbitrator can also address complaints about incorrect data by comparing the proof of misbehaviour provided by the model owner with the data owner's data integrity commitments.

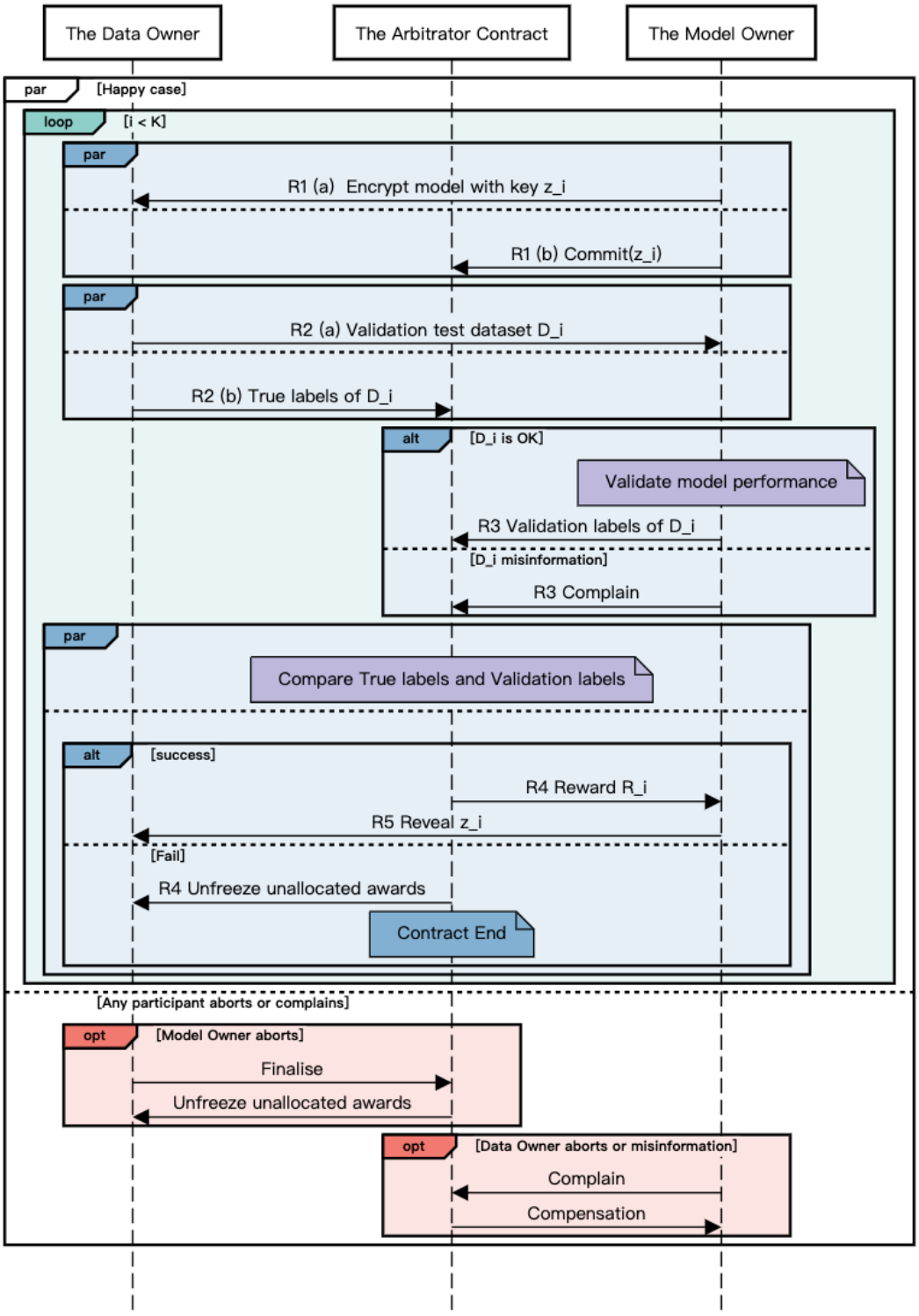

Figure 4.3: Outline of Two-party Clearing Protocol

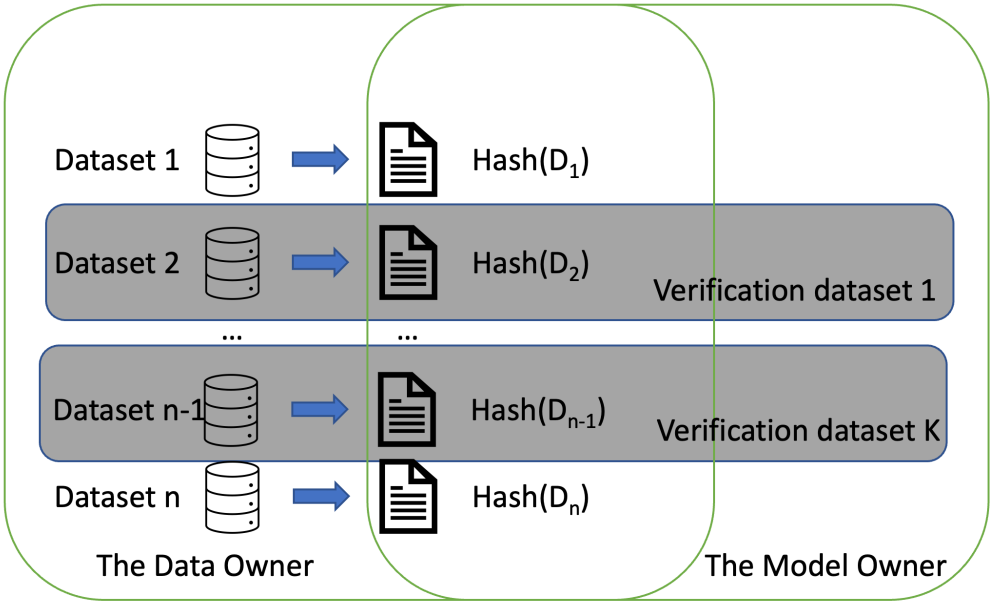

Figure 4.4: Validation Datasets Selection

The Validation Datasets

A fair clearing process requires the model and data owners to reach a consensus on the test datasets used to validate the model's performance. Two conditions are vital: the datasets must be 'known but fresh' to the model owner, who must trust the datasets but only access them during the validation phase. Additionally, the datasets must be 'read-only' to the data owner, who knows the details of the test datasets but cannot modify them or decide which datasets are used in the model acceptance phase.

Figure 4.4 provides a high-level design for selecting the validation datasets, with the protocol specifications below:

The protocol describes the behaviour of the honest model owner (MO) and data owner (DO).

Initialise

DO**:** The data owner randomly divides the training data into m groups, removing the actual labels. Each group consists of an un-

labelled training dataset $Data_m$ and a corresponding label set $TrueLabel_m$. The data owner then creates an array $D = \{Commit(Data_i) \parallel Commit(TrueLabel_i), i \in (1, 2, ..., m)\}$ containing all dataset commitments and proceeds to the *Consensus* phase.

Consensus

DO: The data owner sends array D to the model owner and proceeds to the *Finalisation* phase.

MO: Upon receiving array D, the model owner randomly selects K commitment sets to create a commitment array $VD_i = \{Commit(D_i) \parallel Commit(TD_i), i \in (1, 2, ..., K)\}$ of validation datasets. The model owner sends array VD to the data owner and proceeds to the *Finalisation* phase.

Finalisation

DO: Upon receiving array VD, the data owner removes the selected test data from the training data, ensuring the model owner is not exposed to the test data in early training phases.

MO: The model owner posts array VD to the blockchain.

The method establishes test datasets that both the data owner and model owner can accept. The data owner can trust the validation results since they initially provided the datasets. The method also ensures that the model owner cannot access the data until the validation period. The model owner is comfortable since the test datasets come from the original training datasets and contain the same categories. Upon receipt, both parties can verify dataset integrity by comparing them to the pre-existing commitment information in the smart contract. If the model owner can prove the received test datasets are incorrect, they can use the complaint process to obtain compensation.

4.3.2 Security Analysis

In the beginning of this section, we defined the security properties of the framework. Analyzing this security protocol provides insight into why neither party should be able to break the other's fairness properties.

Model owner fairness: Fairness for the model owner means that the data owner should only be able to access the latest model after rewarding an honest model owner. From the secrecy property of the encoding scheme and the hiding property of the hash commitment, it is clear that the data owner cannot decrypt the latest model from the ciphertext until the model owner reveals the decryption key. When the model owner publishes the decryption key, the reward coins R will be frozen for clearing by the arbitrator. At this point, the contract's settlement process has been initiated. An honest model owner is guaranteed to be rewarded if the training is successful, even if the data owner aborts. Furthermore, we need to show that the data owner cannot forge commitments so that the data owner's sending of the wrong verification dataset appears honest and that the forged commitments are acceptable to the arbitrator. Informally, forging such a hash commitment on an erroneous validation dataset is impossible unless the data owner finds a collision in the hash function H.

Data owner fairness: After the data owner activates the training contract, the reward coins are frozen for further reward allocation. To prove the Fairness of an honest data owner, we must show that the model owner cannot send a forged set of verification labels to satisfy the contract conditions. To successfully execute such an attack, the model owner must know the contents of real labels from the hash commitments. The probability that the model owner can forge the validation labels that match the real ones is negligible, as it would require the model owner to break the collision resistance of the underlying hash function. Thus, the data owner can be guaranteed that, once the validation process is underway, they will either receive a successfully trained model or be able to retrieve the unallocated rewards.

Termination The clearing protocol will terminate after four or five rounds per

stage, depending on whether the training succeeds or fails. With an active arbitrator contract and at least one honest participant, we can get an overview of the protocol termination by analyzing the following cases:

No abort: This case occurs when both participants are acting in good faith. In this case, if the training is successful in every stage, the clearing protocol ends after the fifth round of stage K when the model owner sends the key Z_K to the data owner. Otherwise, the protocol ends after the fourth round of the failed stage and unfreezes the data owner's unallocated reward coins. Additionally, if the model owner complains about the integrity of the validated datasets, the protocol allows the model owner to complain in the third round. Suppose the arbitrator contract confirms that the complaint is valid based on proof provided by both participants. In that case, the data owner's frozen reward coins will be allocated to the model owner as compensation, and the protocol will then be terminated. Otherwise, the complaint is rejected, and the settlement process continues.

The model owner aborts: After the previous clearing stage has been completed, or the data owner has sent out the validation datasets for the current stage, the model owner has no further response. The data owner can send a "Finalise" message to the arbitrator contract. If the contract status is confirmed to meet the termination conditions, the data owner's undistributed reward coins will be unfrozen, and the protocol will be terminated.

The data owner aborts: If the data owner fails to respond after the model owner sends the encrypted model for validation, the model owner has the option to send a "Complain" message to the arbitrator contract. Once the arbitrator verifies that the conditions for the complaint are satisfied, the frozen reward coins belonging to the data owner will be released to the model owner as compensation. Following this, the protocol will be terminated.

4.4 Implementation and Performance

In this section, we evaluate our framework and compare its performance to other existing frameworks.

4.4.1 Contract Optimality and Baseline

Suppose a data owner and a model owner engage in a federated learning task after an initial negotiation where they anticipate implementing the entire learning goal in 3 stages (i.e. $K = 3$). If successful, each stage assumes that the model's incremental value to be $V_1 = 1, V_2 = 2$ and $V_3 = 3$. It is also assumed that the functional expression for the probability of success in each stage is $P_k(Me_k, De_k) = MIN(0.6(Me_k + De_k), 1)$. As we will see later, the equilibrium effort satisfies $0.6({Me_k}^* + {De_k}^*) < 1$, so we can count $P_k(Me_k, De_k) = 0.6(Me_k + De_k)$. We also assume that the effort cost of the model owner's function is $C(Me_k) = {Me_k}^2$ and the effort cost of the data owner's function is $C(De_k) = {De_k}^2$.

In considering contract optimality, we must first derive the utility function for the model owner:

$$mr_k = 0.6(Me_k + De_k)(R_k + mr_{k+1}) - {Me_k}^2, \quad k = 1, 2, 3,$$

$$mr_4 = 0.$$

Taking the utility function for each stage and deriving it to its effort level determines the optimal effort yield for the model owner:

$${Me_k}^* = \frac{\partial mr_k}{\partial Me_k} = 0.3(R_k + mr_{k+1}) \quad k = 1, 2, 3.$$

Repeating the same approach, we can derive the utility function of the data owner and their optimal effort:

$$dr_k = 0.6(Me_k + De_k)(V_k - R_k + dr_{k+1}) - {De_k}^2, \quad k = 1, 2, 3,$$

$$dr_3 = 0, dr_4 = 0.$$

$${De_k}^* = \frac{\partial dr_k}{\partial De_k} = 0.3(V_k - R_k + dr_{k+1}) \quad k = 1, 2, 3,$$

$$De_3 = 0.$$

We can get the optimal rewards yielded for the model owner, calculated by recurring the above equations and the derivative of the data owner's payoff dr_1 with respect to the reward ${R_2}^*$:

$${R_1}^* = 0, {R_2}^* = 0.41, {R_3}^* = 3,$$

Here, the probabilities of success are

$$P_1({Me_1}^*, {De_1}^*) = 0.37, P_2({Me_2}^*, {De_2}^*) = 0.51, P_3({Me_3}^*, {De_3}^*) = 0.54.$$

The total incentive the data owner pays to the model owner is $\sum_{k=1}^{3} {R_k}^* = 3.41$. We set this is the baseline for comparison with existing frameworks.

4.4.2 Comparison with Existing Frameworks

We take the optimal incentive value discussed in the previous section as a precondition and observe the differences in the value of the model owner's effort, the probability of a successful federated learning process, and the expected payoff of the data owner between the frameworks, conditional on the data owner paying the same total incentive.

Offer Rewards framework In this framework, the data owner rewards the model owner prior to training, and the reward has no relationship to the success of the training. In this simulation, we set the rewards the model owner receives **before** the training stages to 1[1], 1.205, and 1.205.

Share Profits framework In this framework, the data owner shares the incremental value of the trained model with the model owner after successful training. The total reward the data owner provides, i.e., 3.41, is assumed to be exchanged for the model with a total increment value of 6. In other words, the total reward the model owner receives is 57% of the total increment model value. Hence, we set the rewards received by the model owner after each successful training stage to 0.57, 1.14, and 1.71.

[1]Since the stage 1 added model value is 1, it is logical that the maximum reward offered by the data owner is 1 before the first stage.

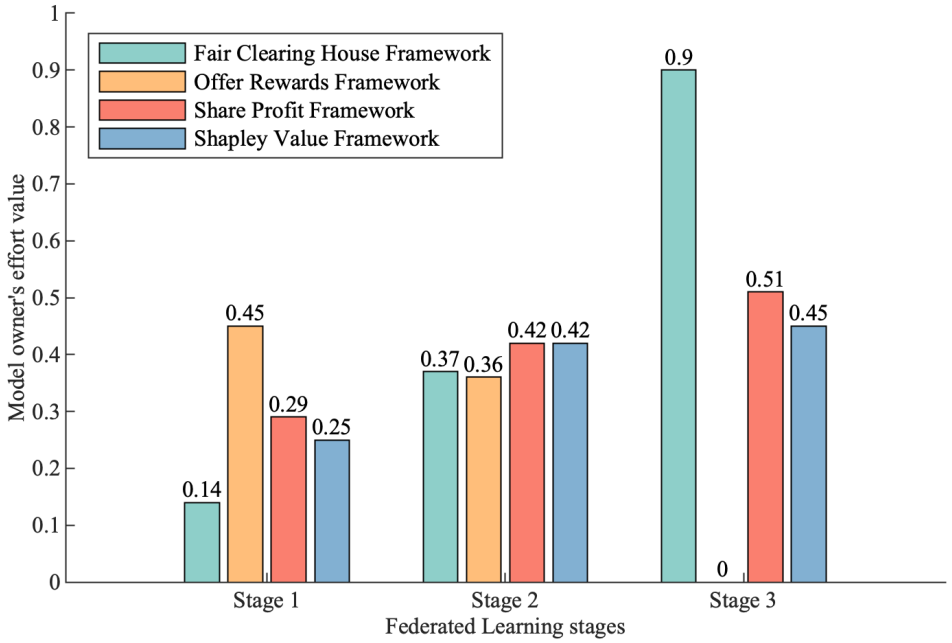

Figure 4.5: Comparison of Model Owner Effort Value Across Frameworks

Sharply Value framework A Shapley value is a utility assessment method based on marginal contributions, where the rewards are distributed purely based on the contributions provided by the participants. Since our scenario simplifies the participants to two parties, the model owner and the data owner each receive the same Shapley value. In other words, the data owner rewards half of the increment value of the model to the model owner. In the simulation, we set the model owner to receive the rewards of $0.5, 1$, and 1.5 after each successful training stage.

We obtained the following results from our simulations:

1. Figure 4.5 displays the variation in unobservable effort values of model owners affected by rewards in different frameworks. The distribution of effort values across stages demonstrates that the FCH framework has the advantage of reducing the impact of unethical behaviour on an incentivized objector. The model owner's effort continues to grow despite the gradual increase in training difficulty stage by stage. Most importantly, in the third stage, which is the most challenging and where the model's value increases the most, the model owners' effort far exceeds those of the other frameworks. It has a crucial impact

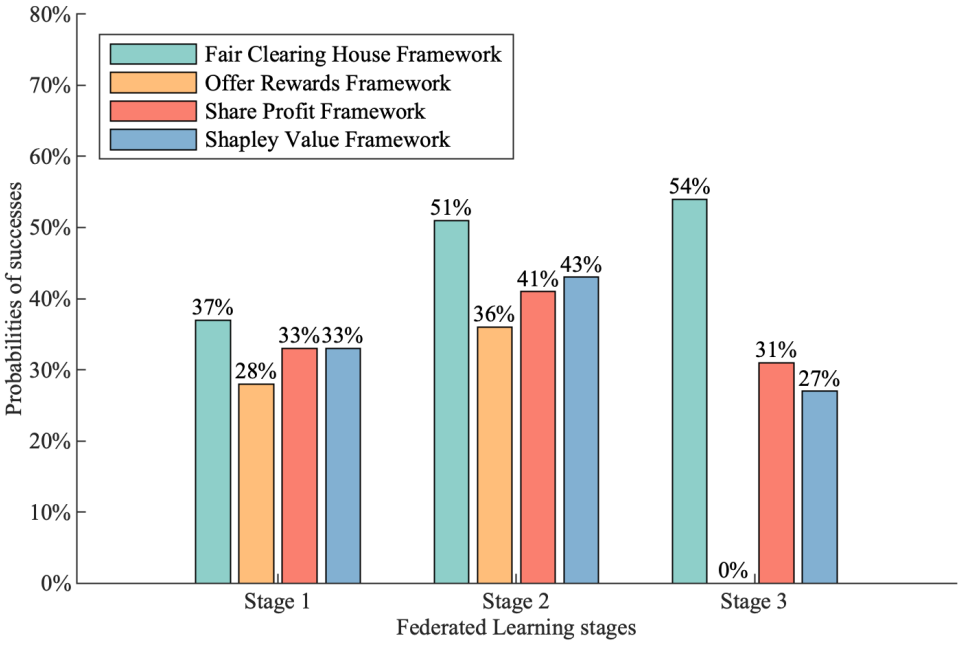

Figure 4.6: Comparative Analysis of Training Success Probabilities Across Different Frameworks

Figure 4.7: Comparative Analysis of Expected Payoffs for Data Owners Across Different Frameworks

on ensuring that the federated learning scheme produces optimal results.

2. Figure 4.6 presents the reward influenced success probability for different frameworks. The FCH framework outperforms the other frameworks regarding success probability in each stage, especially in the most challenging third stage. This result directly relates to the model owners' strong efforts in the final training stage. We observe that the Offer Rewards framework flattens in the final stage. We believe this is because the model owner receives all her rewards before that stage. Thus, a self-interested and opportunistic model owner will no longer invest much effort, which may prevent an optimal result from the scheme.

3. Figure 4.7 illustrates the expected payoffs for the data owner in different frameworks. The FCH framework yields the highest return for the data owner. The potential return to the data owner is the lowest in the Offer Rewards framework. Federated learning is a high-risk collaboration. Paying rewards in advance weakens the model owner's contributions and can lead to failed training.

Our framework mitigates unethical behaviour associated with information asymmetries, such as opportunism and self-interested acts, within federated learning systems. Moreover, it motivates participants to invest optimal efforts into training, confirming the intuition behind our model: success in later stages relies on success in earlier stages. Thus, rewards in the later stages motivate efforts in the earlier stages. Additionally, our experimental data demonstrate that the data owner needs to return as much value created by the model owner's effort in the later stages. It is the least costly incentive scheme.

4.4.3 Potential Applications of The FCH Framework

The rise of federated learning techniques has enabled more small and medium-sized organizations to collaborate on modelling through an open federated learning marketplace, which was once the prerogative of mega-companies with vast amounts of data. This marketplace fosters collaboration between data and model owners, allowing them to work on machine-learning projects while preserving data privacy. In

this marketplace, data owners and model owners exchange rewards and trained models under the FCH framework via smart contract technology, ensuring secure and transparent transactions. In manufacturing and financial sector case studies, multiple companies and financial institutions partner with specialized firms to develop predictive maintenance and fraud detection models. Organizations can leverage shared expertise, maintain data privacy and intellectual property protection, and receive rewards through secure and automated smart contracts by participating in the federated learning market.

4.5 Conclusion and Chapter Discussion

This chapter has presented the Fair Clearing House (FCH) framework, a novel approach designed to address the challenges of incentive alignment and trust-building in data-owner-led federated learning scenarios. Our framework builds upon the multi-stage incentive mechanism introduced in Chapter 3, which leverages contract theory to analyze unethical behavior and derive optimal incentive contracts. This mechanism, characterized by allocating the incremental value of the model primarily to the model owner in later stages, fosters collaboration and minimizes costs for the data owner.

In addition to incorporating this incentive mechanism, the FCH framework introduces a secure two-party clearing protocol built on a smart blockchain contract. This protocol ensures fair and transparent settlement between the model and data owners, guaranteeing the secure exchange of models and rewards. Through extensive experimental evaluation, we demonstrated the superior performance of the FCH framework compared to conventional federated learning approaches, such as "Offer-Rewards," "Share-Profits," and "Shapley-Value" frameworks. Our results highlight the effectiveness of the FCH in optimizing participant efforts and reducing unethical behavior in data-owner-led federated learning systems.

While this chapter presents a significant advancement in addressing challenges in federated learning, several avenues for future research remain. Our current work assumes that all participants act honestly. However, in real-world scenarios, malicious actors may attempt to exploit the system for their own gain. Future research

should investigate strategies to detect and mitigate malicious behavior in federated learning environments, ensuring the robustness and security of the FCH framework even in the presence of adversarial participants.

Additionally, our focus has been on data-owner-led federated learning. It would be valuable to explore the performance and characteristics of the FCH framework in model-owner-led scenarios. Comparing the effectiveness of different leadership structures in federated learning could lead to a deeper understanding of optimal strategies for various applications.

In conclusion, the Fair Clearing House (FCH) framework offers a promising solution for addressing the challenges of incentive design and unethical behavior in data-owner-led federated learning scenarios. By leveraging contract theory and blockchain technology, our proposed framework provides a robust, secure, and fair solution for optimizing participant efforts and reducing unethical behavior. Future research efforts should build upon this foundation to explore the aforementioned extensions, ultimately advancing the field of federated learning and its real-world applications.

Chapter 5

Reciprocal Federated Learning: Balanced Incentives for Model & Data Owners

5.1 Introduction

In an era increasingly dominated by artificial intelligence, machine learning has emerged as a transformative force across various domains, including the rapidly advancing fields of Web 3.0 and digitalized industrial applications within the 5G/6G communication landscape. While these next-generation technologies offer enhanced security, decentralized data management, and seamless connectivity, they create a unique paradox. These technologies often demand vast amounts of data to fuel the potential of artificial intelligence and machine learning. However, Web 3.0's decentralized architecture, coupled with the high-speed, low-latency communication of 5G/6G, fosters an environment where data is produced at unprecedented volumes and distributed across many devices and networks. This landscape complicates traditional data collection approaches for centralized machine learning and raises significant privacy concerns. This clashes with the heightened privacy concerns and decentralized data ownership paradigms that characterize Web 3.0's ethos[84]. To fully harness machine learning's transformative potential within this new technological paradigm while respecting privacy, federated learning[85] provides a compelling

solution. It allows models to be trained collaboratively on distributed data without compromising individual data ownership.

The key to the success of federated learning is to ensure the active participation of all parties, including data owners who provide training data and model owners who develop and refine models. Despite extensive research on incentives in federated learning, a significant gap remains in developing a system that fairly and effectively incentivizes both model and data owners. Traditionally, incentives were primarily designed by the model owner, who possesses extensive modelling experience and capabilities, focusing on their perspectives and interests. While these incentives have effectively promoted contributions, performance, and privacy preservation from data owners, their crucial perspectives are often overlooked. Such approaches struggle to quantify the model owner's value-added in refining the model, especially in scenarios where data owners significantly influence incentive allocation and have strong bargaining power over the terms of the process and potential capital investment. Furthermore, these frameworks fail to properly address the alignment between participants' expected and actual returns, leading to potential disparities in incentive distribution and participation equity.

In Chapter 1, we presented a scenario where healthcare providers, often lacking deep machine learning expertise, formed a data owners alliance. This alliance collaborated with specialized machine learning entities, or model owners, to develop advanced predictive analytics for personalized care. The alliance contributed funds to reward model owners and also bore the cost of collecting and processing vast amounts of data. In exchange, model owners provided expertise to create powerful models capable of processing diverse data streams and detecting health patterns.

Due to the cautious nature of healthcare, a federated learning approach with carefully structured incentives was necessary to mitigate the uncertainty of achieving optimal model outcomes. Given the limited technical expertise among data owners, fair incentives for model owners were crucial, even though their efforts might not be immediately apparent. Data owners also faced the challenge of proving that the value of their data matched their expected contributions.

Addressing these multifaceted challenges, This chapter introduces the Reciprocal

Federated Learning Framework (RFLF), a groundbreaking approach designed to motivate model and data owners equitably within federated learning environments. This framework, the centrepiece of our contribution, pioneers a novel reciprocal incentive design that dynamically aligns contributions with rewards, designed to motivate contributions across the collaborative process. It tackles information asymmetry and directly integrates financial investments into the dynamic rewards structure. Critically, the RFLF accommodates the reality of heterogeneous data contribution among collaborating data owners, recognizing that capital can play a vital role in ensuring equity and participation. The contribution of this chapter is as follows.

- **Pioneering Equitable Incentives:** Introduced RFLF to fundamentally reshape incentive structures in collaborative machine learning. This groundbreaking framework balances rewards and contributions, motivating the model and data owners within the federated learning paradigm.

- **Unraveling Incentive Complexity:** Employed analytical models to dissect the complexities of optimizing incentives within the RFLF. This analysis deepened the understanding of the relationships between data quality, quantity, and the balance of power between model and data owners.

- **Empirical Validation:** Proved the RFLF's effectiveness through rigorous evaluations with the MNIST and CIFAR-10 datasets. This practical demonstration shows the framework's ability to enhance model performance and foster strong, equitable participation in real-world federated learning scenarios.

This chapter offers a comprehensive exploration of our contributions and findings. Section 5.2 provides the necessary background and terminology for understanding our framework. Section 5.3 clearly defines the problems our solution tackles. Section 5.4 demonstrates the real-world practicality of our contractual framework. Section 5.5 analyzes the optimality of our proposed contracts in RFLF. In Section 5.6, empirical evaluations rigorously test our framework's performance on established datasets. Finally, Section 5.7 summarizes key takeaways, discusses the significance of our work, and suggests areas for future investigation.

5.2 Preliminaries and Notations

5.2.1 Definitions of Key Notations

The table 5.1 below outlines key notations used in this chapter, crucial for understanding the mathematical modelling and subsequent analysis of the RFLF.

5.2.2 Proposed Framework

Our framework adopts a structured, multi-stage federated learning approach, thoroughly evaluating outcomes at each stage to allow for efficient resource allocation. This design addresses challenges commonly associated with large-scale machine learning projects, such as misalignment of objectives and resource inefficiencies, and distributes the associated risks among all participants.

Given the prevalent privacy considerations in federated learning, involving a direct third-party mediator for settlements may be costly or impractical. Therefore, the smart contract is employed within the RFLF as a decentralized, transparent, and automated solution for secure and fair transactions. This contract eliminates the need for a third-party intermediary by directly managing settlements and dispute resolutions within the framework, ensuring that all transactions and interactions are conducted equitably and in compliance with the privacy needs inherent to federated learning. The operational role of the smart contract, including its facilitation of settlements and the resolution of disputes, will be detailed in subsequent sections of this chapter.

To illustrate the operational mechanics of the RFLF, Figure 5.1 highlights the pivotal role of a smart contract within the framework's workflow.

- *Training Contract Initialization:* The alliance of data owners, represented by D and consisting of N members, collaborates with the model owner M, a federated machine learning expert. They establish a smart contract to orchestrate the K -stage federated learning. This contract outlines key training aspects, including the predefined number of collaboration stages K, the performance benchmarks for each stage, and the set reward for the model owner at each stage R_k. When activated, the smart contract secures the funds deposited by

Symbol	Description
N	Total data owners in the alliance.
i	Index for individual data owners, $i \in N$.
K	Total planned stages in federated learning.
k	Specific stage in the process, $1 \leq k \leq K$.
$U_{D,k}$	Data owners' alliance utility projection from stage k to K.
$U_{M,k}$	Model owner's cumulative utility forecase from stage k to K.
$U_{i,k}$	Expected utility for data owner i at stage k.
$P_k(Me_k)$	Probability of successful training at stage k, based on model owner's effort Me_k.
R_k	Reward for model owner upon successfully completing stage k.
$C(Me_k)$	Operational costs for model owner at stage k.
V_k	Additional value generated for the model after successful training at stage k.
b_i	Data owner i's stake in the collective model.
S_i^k	Data contribution of data owner i during stage k on model improvement.
BBF_k	Allocated funds for buying back data contributions at stage k.
$c(s_i^k)$	Incurred cost by data owner i for data contribution at stage k.
w_i	Weight parameter representing the relative importance of data owner i.

Table 5.1: Glossary of Key Mathematical Notations (Chapter 5)

the data owners. These funds are reserved to remunerate the model owner and buy back the data contributions from each data owner at every stage. The proportion of the deposited funds each data owner provides aligns with their respective shares in the trained model.

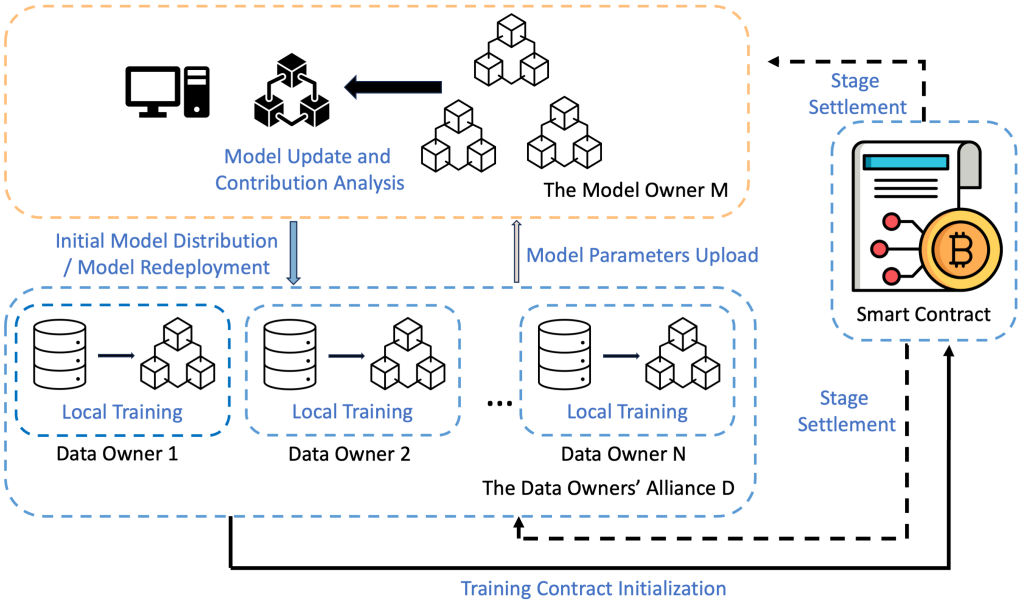

Figure 5.1: Reciprocal Federated Learning Framework

- *Initial Model Distribution:* As each stage commences, the model owner disseminates the foundational model to all participating data owners, ensuring a consistent starting point.
- *Local Training:* Leveraging their individual data sets, data owners enhance the given model on their respective local systems. This distributed approach ensures data confidentiality.
- *Model Parameters Upload:* Data owners forward their refined model parameters to a central repository after local training. The model owner consolidates these submissions to achieve an integrated model perspective.
- *Model Update and Contribution Analysis:* The model owner updates the model using the combined parameters. Concurrently, the contributions of each data owner are assessed, highlighting their specific contributions.
- *Model Redeployment:* Ready for the next rounds, the model owner disseminates the augmented model to all data owners. This repetitive methodology promotes consistent model improvement.
- *Stage Settlement:* The stage is declared complete upon completing the designated objectives. The smart contract facilitates the settlement between the

collaborating parties: the data owners' alliance and the model owner. The updated model, enriched with the added value V_k, is transferred to the alliance, while the model owner receives the previously agreed-upon rewards R_k. Additionally, compensation for data contributions BBF_k is distributed to the alliance members based on a buy-back mechanism that is proportional to the extent of their individual contributions S_i^k, as outlined in the contribution analysis. If the established criteria are met, and the current stage does not represent the final phase, the process is designed to return to the initial steps for a new cycle. However, if the criteria are not satisfied or the final stage has been reached, the collaborative effort is concluded, any remaining deposits are refunded, and the smart contract is subsequently terminated.

5.2.3 Assumptions

The proposed framework is guided by several fundamental assumptions designed to simplify the understanding of its key components. These assumptions facilitate a straightforward analysis of the framework's performance by delineating clear conditions and behaviours expected from participants. By establishing these assumptions, we aim to clarify the operational dynamics of the RFLF and highlight the mechanisms that drive reciprocal benefits among participants:

1. **Risk Neutrality and Non-malicious Behavior:**

   It is assumed that all participants, both data owners and the model owner, are risk-neutral and act without malicious intent. This simplification allows the analysis to focus on the strategic decisions and interactions that underpin the framework without the need to model complex behaviours related to risk aversion or malicious actions. This assumption ensures a focused exploration of how incentives and cooperative dynamics contribute to the framework's goals.

2. **Stage-Based Training with Success Criteria:**

   The training process is structured into Kstages, with each stage's completion being contingent on meeting predefined success criteria. This structured approach ensures that the model's development progresses systematically and

that specific performance benchmarks are achieved at each stage.

3. **Model Owner's Effort and Impact:**

   The effort exerted by the model owner in stage k is represented as Me_k. Assuming this effort is an independent and non-negative variable indicates that any increase in the model owner's algorithmic refinement is expected to impact the model's performance positively. This assumption highlights the direct correlation between the model owner's effort and the model's improvement.

4. **Success Probability and Diminishing Marginal Returns:**

   The probability of successful training in stage k is represented as $P_k(Me_k)$, which is bounded between 0 and 1. The success probability function satisfies $1 \geq P_k(Me_k) \geq 0$, with a positive first derivative $\frac{\partial P_k(Me_k)}{\partial Me_k} > 0$ and a negative second derivative $\frac{\partial^2 P_k(Me_k)}{\partial^2 Me_k} < 0$. This implies that while increased effort Me_k enhances the likelihood of success, it offers diminishing returns.

5. **Effort-Related Costs and Marginal Costs for Model and Data Owners:**

   - The Model Owner: Costs incurred due to the model owner's efforts are represented by $C(Me_k)$. We assume:
     - Increasing Marginal Costs: $\frac{\partial C(Me_k)}{\partial Me_k} > 0$, meaning the cost increases with higher levels of effort.
     - Diminishing Returns on Effort: $\frac{\partial^2 C(Me_k)}{\partial^2 Me_k} < 0$, indicating that the rate of cost increase slows down as effort increases. This captures the intuition that initial effort may have a large impact, but subsequent increases in effort yield smaller gains.
   - Data Owner i: We consider a broad notion of effort for data owners, encompassing costs associated with data collection, preparation, and potentially transmission. For data owner i, these costs are represented by $c(S_i^k)$. We assume:
     - Increasing Marginal Costs: $\frac{\partial c(S_i^k)}{\partial S_i^k} > 0$, meaning cost increases with greater contributions.

- Diminishing Returns on Contributions: $\frac{\partial^2 c(S_i^k)}{\partial^2 S_i^k} < 0$, indicating that the rate of cost increase slows down as contributions increase. This can reflect scenarios like acquiring initial data being easier than obtaining increasingly rare or specialized data.

6. **Stage Completion and Model Value Increment:**

   Assuming successful completion of the k^{th} stage of federated learning, the market value increment of the model post-training is indicated by V_k. This assumption captures the financial gains associated with successful model development, highlighting the economic incentives driving the federated learning process.

7. **Pre-defined Rewards and Buy-Back Funding:**

   The reward R_k and the buy-back funding BBF_k for each stage k are predetermined at the project's onset. This assumption ensures financial predictability and stability within the framework, allowing for consistent planning and budgeting throughout the project.

By elucidating these assumptions, the framework's foundational principles are made accessible, supporting an intuitive grasp of the RFLF's design and its intended benefits. This approach ensures that the innovative aspects of the framework, particularly those related to fostering equitable and reciprocal collaborations, are communicated and understood.

5.2.4 Reciprocal Relationship between Model and Data Owners

The RFLF fosters a 'reciprocal' relationship between data and model owners, wherein both parties are incentivized to improve model performance X_k in stage k through their respective contributions. This performance is mathematically represented as a function of data owner contribution D_k and model owner contribution M_k[20], [86]:

$$X_k = 1 - e^{-\phi(D_k, M_k)^\nu}$$

Here, Parameters ϕ and ν regulate the impact of the participant's contribution on

model performance, reflecting the diminishing returns of additional contributions. Improved performance directly translates to higher added economic value V_k, calculated as:

$$V_k = \theta \times (X_k - X_{k-1})$$

where θ is the scaling parameter converting model performance into economic value. Thus, the contributes of both parties synergistically enhance the model's performance and value, fostering a mutually beneficial environment.

Within the RFLF framework, data owners obtain the total economic value of the final model through the model ownership transfer mechanism, directly benefiting from the value created by the federated learning process. Therefore, the model owner's effort, which directly impacts model performance, indirectly influences the economic value added to the model and, consequently, the data owners' payoff.

The model owner's reward (R_k) in the RFLF framework is contingent on achieving the pre-defined performance goal for each stage. Optimized data contribution efforts by data owners aid the model owner in improving the model's performance, thus increasing the likelihood of reaching the stage goal. Furthermore, the model owner's reward (R_k) within the RFLF framework is closely related to the increase in the model's economic value (V_k)[1]. This strong incentive encourages model owners to exert optimal efforts, reinforcing the mutual dependence between data owner contributions and model owner rewards.

The subsequent sections elaborate on this inherent 'reciprocity' in the RFLF framework through mathematical modelling and quantitative demonstration of the relationship between the contributions of all participants in federated learning, the economic returns of the model, and training incentives. This reciprocal environment fosters collaboration, trust, and equitable distribution of benefits in the federated learning process.

[1]This relationship is formally established in Theorem 3, Section 6.

5.3 Problem Formulation

5.3.1 Description of the RFLF Process

Within the established federated learning framework, the formation of a data owners' alliance serves as the fulcrum for not only enhancing the contribution of training data but also encouraging the model owner to offer optimal effort during training. This alliance, comprising data owners, assumes all financial responsibilities, including the payment of stage rewards R_k to the model owner and the provisioning of stage-specific data contribution buy-back funds BBF_k to incentivize high-quality data contributions. Through consensus, the data owners agree upon their capital contribution ratios b_i, which align with their financial input and ensure that their share in the post-training model corresponds to their investments. Operating autonomously, the alliance internally manages cost-sharing and profit allocation, effectively mitigating free-riding issues and the uncertainties associated with the heterogeneous value of data contributions. This autonomy in managing financial and data contributions ensures that data owners are more inclined to fully engage in the training regimen, confident that their investments are equitably managed and that their contributions are meaningfully valued. At the same time, the model owner can concentrate on optimizing the training process, ensuring that the data owners' alliance is committed to providing high-quality data without influencing the model owner's strategic decisions.

As delineated in Figure 5.2, a multi-level Stackelberg game model encapsulates the strategic interactions within this framework. In the initial phase, the RFLF establishes two pivotal contracts: a training rewards contract between the data owners' alliance and the model owner, and a data contribution buy-back contract among the data owners themselves. These contracts lay the foundation for the framework's operation by defining the terms of engagement and compensation for the training stages. The terms are formalized through a smart contract, referred to as the arbitrator contract within our framework, which serves as an impartial enforcer of the agreed-upon conditions. By securing capital commitments through these contracts, the framework ensures that all participants, including data and model owners, have vested interests aligned with the successful outcomes of the federated learning pro-

cess.

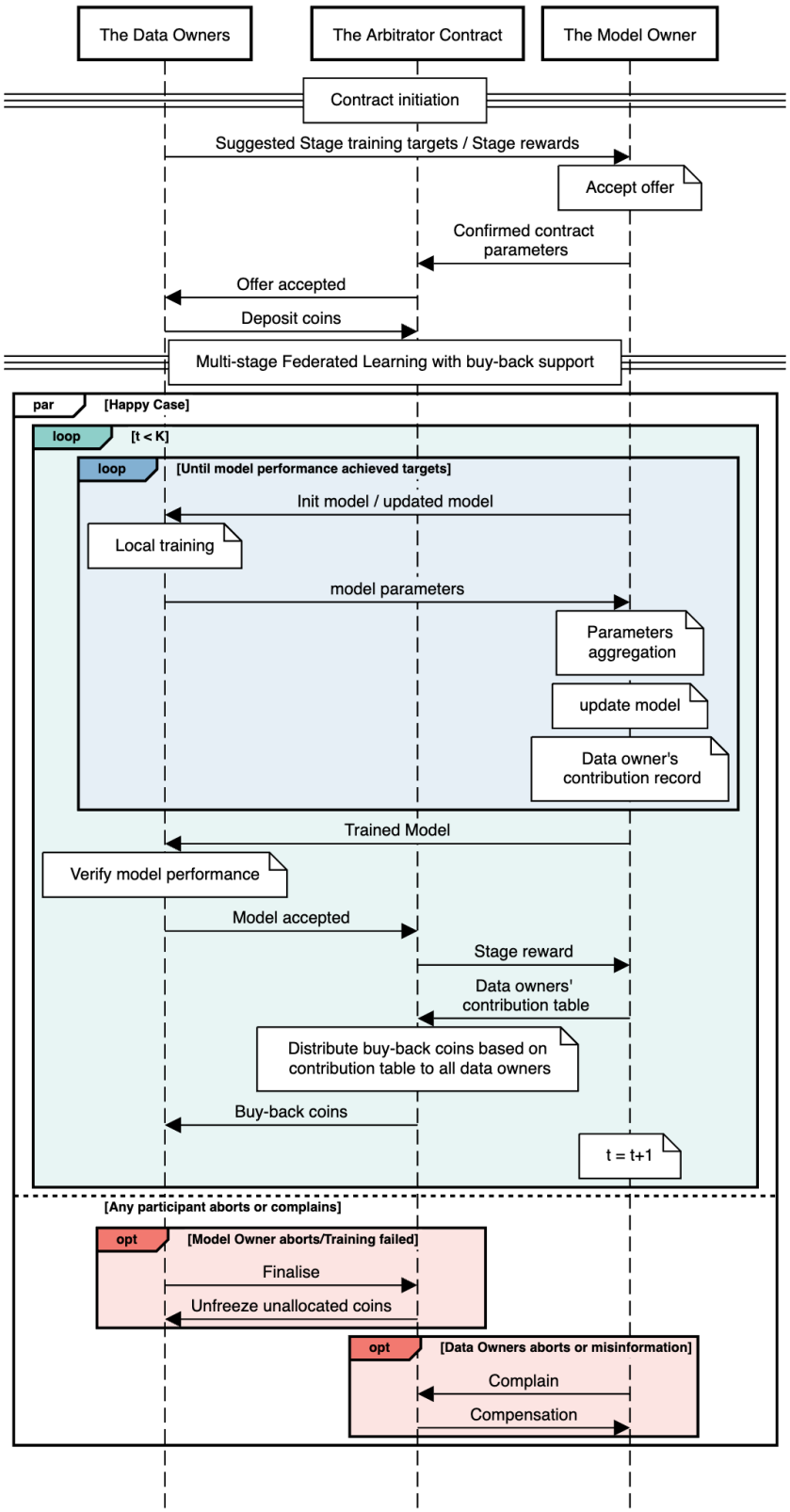

Figure 5.2: RFLF Events Sequence

In the subsequent stages of this iterative process, a cyclical exchange is established

between the data owners and the model owner. Data owners locally enhance the model with their datasets, while the model owner integrates these improvements into a centralized model. Following each stage, the data owners' alliance assesses and validates the model's enhanced performance. Upon validation, the alliance communicates with the arbitrator contract, responsible for managing the settlements. The arbitrator contract then releases the stage-specific training rewards to the model owner and allocates the data contribution buy-back coins to individual data owners based on their contributions.

This process is intended to persist until either the stage targets are not met or the maximum number of stages K is reached. The arbitrator contract stands as the mechanism for resolution in disruptions or disputes. If the model owner withdraws or a staging target is not achieved, data owners have the right to approach the arbitrator contract to seek termination of the engagement and the release of funds. On the other hand, if the data owners' alliance fails to meet their obligations or submits performance verification data that does not meet the agreed-upon standards, the model owner is entitled to call upon the arbitrator contract for suitable compensation. This mechanism ensures a fair and impartial verification of the model's performance, upholding the integrity of the training process. Such a structured and methodical approach ensures a harmonious, secure, and equitable training process, bolstering collective cooperation and guaranteeing fair compensation for all participants.

5.3.2 Formulation of the Utility Functions

The strategic interactions and decision-making processes within the RFLF are fundamentally driven by the utility functions of the involved parties. These functions articulate the economic motivations and constraints that guide the behaviour of the model owner, the data owners' alliance, and individual data owners. By formalizing these utility functions, we can analyze how each stakeholder balances their contributions, investments, and expected returns throughout the federated learning process. This subsection introduces and defines the utility functions of each key participant, laying the groundwork for a deeper exploration of their strategic equilibria and the mechanisms that foster collaboration and ensure the equitable distribution

of benefits across the RFLF.

Definition 1 (Utility Function of the Model Owner M): The utility function $U_{M,k}$ represents the model owner's expected net benefit throughout the federated learning stages, from k to K. We assume a zero interest rate (i.e., no discounting of future returns), a simplification justified by the added complexity and lack of additional insights gained from incorporating a positive interest rate, as demonstrated in Section 5.5. It incorporates the probability of successful training $P_k(Me_k)$, the rewards R_k received, future anticipated payoffs $U_{M,k+1}$, and operational costs $C(Me_k)$, capturing the balance model owners seek between their efforts and the rewards of their participation:

$$U_{M,k} = P_k(Me_k) \times [R_k + U_{M,k+1}] - C(Me_k), \quad k = 1, \cdots, K. \tag{5.1}$$

The utility function for the model owner reflects their primary motivation to maximize net benefit across all stages of the federated learning process. It balances the probability of success, potential rewards, future returns, and the costs of their effort. This captures the economic trade-offs the model owner must consider when deciding how much effort to invest in improving the model.

Definition 2 (Utility Function of the Data Owners' Alliance D): The data owners' alliance plays a pivotal role in the RFLF by acting as the intermediary that facilitates collaboration between individual data owners and the model owner. The alliance's utility function, $U_{D,k}$, represents the collective interests of the data owners and is primarily focused on maximizing the long-term benefits derived from the cooperation.

While the utility function is expressed in terms of revenue, this serves as a proxy for the overall value accruing to the data owners. The alliance's revenue is directly tied to the value of the trained model, which ultimately becomes the property of the data owners upon successful project completion. Therefore, maximizing the alliance's revenue through strategic decision-making aligns with the ultimate goal of maximizing benefits for the individual data owners.

Formally, the utility function takes into account the probability of successful training

$P_k(Me_k)$, the incremental value V_k added to the model, anticipated future benefits $U_{D,k+1}$, and the rewards paid to the model owner R_k. This interplay of factors is encapsulated as follows:

$$U_{D,k} = P_k(Me_k) \times [V_k - R_k + U_{D,k+1}], \quad k = 1, \cdots, K. \tag{5.2}$$

This formula delineates the economic dynamics of the alliance, emphasizing its instrumental role in negotiating and managing the contributions and rewards that drive the federated learning effort. It reflects the alliance's strategy to enhance the model's value while ensuring fair compensation for the model owner's efforts, ultimately benefiting the individual data owners.

Definition 3 (Utility Function of Individual Data Owner i): The utility function, denoted as $U_{i,k}$, assesses the expected payoff for individual data owner i at each stage k, incorporating the interplay between their stake in the model, their data contributions, the associated costs, and their relative importance within the system. Specifically, b_i represents the stake or share percentage held by data owner i in the collective model, determined by their capital contribution. S_i^k quantifies the contribution of data owner i during stage k, reflecting their input's impact on model improvement. BBF_k denotes the allocated funds for buying back data contributions from data owners at stage k, and $c(S_i^k)$ accounts for the costs incurred by data owner i for contributing data at stage k, including collection and processing expenses. To account for potential asymmetries between different data owners, a weight parameter w_i is introduced, representing data owner i's relative importance or contribution potential. The utility function is thus defined as:

$$\begin{aligned} U_{i,k} = b_i \times U_{D,k} + w_i \times \left(\frac{S_i^k}{\sum_{j=1}^{N} S_j^k} \right) \times BBF_k - c(S_i^k), \\ k = 1, \cdots, K \quad i = 1, \cdots, N. \end{aligned} \tag{5.3}$$

This equation elegantly incorporates the consideration of each data owner's significance within the federated learning process, particularly acknowledging the potential for unequal contributions among participants. It emphasizes the balance between

their contributions, the financial incentives received, the costs undertaken, and their relative value within the collaborative environment, further enhancing the RFLF's capacity to incentivize and reward engagement equitably.

The subsequent analysis, leveraging the backward induction method, will explore the equilibrium strategies that emerge from these interactions, shedding light on the optimal paths for collaboration and benefit sharing within the RFLF.

5.4 Contracts Feasibility

This section comprehensively analyses the factors governing the practicability and efficacy of contractual agreements binding the model and individual data owners within the RFLF. A cornerstone of this analysis involves assessing the contracts' feasibility through the lens of Individual Rationality (IR) and Incentive Compatibility (IC). These economic principles ensure that the contracts not only motivate participation but also align the interests of all parties towards the collective success of the federated learning process.

- **Individual Rationality (IR)** ensures that participating in the contract benefits all parties, guaranteeing that the expected utility or payoff from participating is at least as good as not participating. For the RFLF, IR must satisfy the model and data owners, ensuring that the rewards and buy-back mechanisms sufficiently cover their costs and incentivize their contributions.

- **Incentive Compatibility (IC)** guarantees that the optimal strategy for all parties, according to the contract, aligns with the overall objectives of the federated learning process. This means that following the contract's terms will naturally lead participants to behaviors that maximize their utility, which in coherence, maximizes the collective outcome of the RFLF.

Understanding the necessity of IR and IC in the contractual framework facilitates a detailed exploration of the model owner's incentive feasibility and the data owner's buy-back mechanism's feasibility. These conditions provide the theoretical underpinning for sustainable and effective collaboration within the RFLF, ensuring that contracts are not only theoretically sound but also practically viable.

5.4.1 Model Owner Incentive Feasibility

In the context of RFLF, the efficacy of the learning process is heavily reliant on the contributions of the model owner. As such, it is imperative to establish a reward system that not only encourages but also ensures the model owner is adequately compensated for their efforts. The following definitions outline the foundational conditions for a feasible reward system within the RFLF, where the data owners' alliance provides rewards to incentivize the model owner's contribution. These conditions focus on individual rationality and incentive compatibility, ensuring that both the model owner and the data owners' alliance engage in the federated learning process under mutually beneficial terms.

Definition 4 (IR for the Model Owner and Data Owners' Alliance): Individual rationality within the RFLF is met when the expected utility of the model owner, as well as that of the data owners' alliance, across all K stages, remains non-negative. This consideration accounts for potential payoffs and costs associated with each stage, as expressed mathematically in Eq. 5.4:

$$\begin{cases} P_k(Me_k) \times [R_k + U_{M,k+1}] - C(Me_k) \geq 0 \\ P_k(Me_k) \times [V_k - R_k + U_{D,k+1}] \geq 0, \end{cases} \qquad k = 1, \cdots, K. \tag{5.4}$$

These conditions require that the utilities of federated learning participants, factoring in success rates $P_k(Me_k)$, rewards R_k, anticipated future benefits $U_{M,k+1}$, $U_{D,k+1}$, and costs $C(Me_k)$, must remain non-negative for rational participation across all stages.

Definition 5 (IC for the Model Owner and Data Owners' Alliance): Incentive compatibility within the RFLF is achieved when the contract maximizes the utility of both the model and the data owners' alliance by adhering to the specified effort levels, Me_k^*, for all stages, as shown in Eq. 5.5:

$$\begin{cases} U_{M,k}(Me_k^*) \geq U_{M,k}(Me_k) \\ U_{D,k}(Me_k^*) \geq U_{D,k}(Me_k) \end{cases} \text{subject to} \quad Me_k \neq Me_k^*, \forall k. \tag{5.5}$$

Here, $U_{M,k}$ and $U_{D,k}$ represent the utilities of the model owner and data owners'

alliance at stage k, which are functions of the effort level Me_k exerted by the model owner. Me_k^* denotes the optimal effort level prescribed by the contract for stage k, and the maximization is performed over the choice of effort levels Me_k across all stages.

A contract achieves incentive compatibility when the utility of both the model owner and data owners' alliance is maximized by conforming to the specified effort levels Me_k^*, ensuring no incentive to deviate from these levels that would result in a higher utility. It underpins the commitment of both parties to adhere to the contract terms, contributing optimally to the success of federated learning.

These feasibility conditions provide theoretical foundations and are essential in practice for the sustainable operation of RFLF. They align the interests of both the model owner and data owners' alliance with the objectives of the learning process, thus fostering a cooperative and productive environment for all stakeholders involved.

5.4.2 Data Owner Buy-Back Feasibility

When considering the feasibility conditions for an individual data owner in the RFLF, it is crucial to define the conditions under which data owners find participation beneficial and are incentivized to act according to the contract terms. Here are the definitions of individual rationality and incentive compatibility in this context:

Definition 6 (IR for Data Owner i): Individual rationality is met for data owner i if their expected utility from participating in the federated learning process across all K stages is greater than or equal to their utility from not participating, which is their reservation utility. Mathematically, this can be expressed as:

$$U_{i,k} \geq b_i \times U_{D,k} \quad \forall k, \tag{5.6}$$

where $U_{i,k}$ is the utility of data owner i at stage k, and according to Eq. 5.3, $b_i \times U_{D,k}$ is the reservation utility of data owner i.

Definition 7 (IC for Data Owner i): Incentive compatibility for data owner i is satisfied when the contract is structured so that the data owner maximizes their utility by truthfully revealing their data quality and quantity, aligning with the

optimal data contribution specified by the contract. This ensures that the data owner has no incentive to misreport their capability or contribution. The incentive compatibility condition for data owner i is given by:

$$U_{i,k}(S_i^{k*}) \geq U_{i,k}(S_i^k) \quad \text{for all } S_i^k \neq S_i^{k*} \quad \forall k. \tag{5.7}$$

Here, S_i^{k*} is the optimal data contribution level for data owner i at stage k, and S_i^k represents any other level of data contribution. The utility function $U_{i,k}$ should be structured such that it maximizes at S_i^{k*}, ensuring that the data owner is best off when they contribute the level of data as per the contract terms.

These feasibility conditions ensure that each data owner's participation in the federated learning process is rational from an individual perspective and that their contributions are aligned with the collective objectives of the learning framework. These conditions motivate data owners to provide high-quality data in the required quantity.

5.4.3 Feasibility of Integrated Contracts

Establishing a successful federated learning project within the RFLF necessitates ensuring that contractual agreements for the model and individual data owners harmoniously coexist. This synergy is crucial to mitigate common challenges, such as misalignment of incentives and conflicts of interest, which can derail project objectives. This section explores the joint feasibility of these contracts, elucidating how they are designed to synergize, thereby incentivizing and compensating participants equitably. Central to this analysis are two pivotal theorems: *Budget Balance* and *Fair Reward Allocation*, which anchor the equilibrium between individual interests and collective goals.

Theorem 1: Budget Balance. *The RFLF ensures financial sustainability by maintaining that the total rewards to the model owner and the buy-back funds for data owners across all stages do not exceed the overall budget allocated by the data owners' alliance for the federated learning project.*

To formalize the Budget Balance condition, we have:

$$\sum_{k=1}^{K}(R_k + BBF_k) \leq B, \tag{5.8}$$

where B represents the total budget allocated for the project, R_k is the reward for the model owner, and BBF_k is the buy-back funding used to compensate data owners' data contribution at each stage k.

Proof: To establish this condition, we consider the strategic design of the reward and buy-back contracts, ensuring that their cumulative allocations over all stages do not exceed B. The budget allocation for each component is $\sum_{k=1}^{K} R_k \leq B_R$ and $\sum_{k=1}^{K} BBF_k \leq B_{BBF}$.

Here, B_R and B_{BBF} designate the budget portions for the model owner's rewards and the data owners' buy-back payoffs, respectively, with $B_R + B_{BBF} = B$.

By summing these allocations, we get: $\sum_{k=1}^{K} R_k + \sum_{k=1}^{K} BBF_k \leq B_R + B_{BBF}$. Since $B_R + B_{BBF} = B$, the Eq. 5.8 can be concluded.

This demonstrates that the budget is balanced, ensuring that total expenditures remain within the confines of the allocated budget B. The theorem thus confirms the financial feasibility of the reward and buy-back contracts and emphasizes the prudent financial planning by the data owners' alliance to keep the project within its financial parameters.

Theorem 2: Fair Reward Allocation. *The RFLF ensures that rewards are allocated equitably within the framework. The model owner receives compensation proportional to their effort, while data owners are rewarded based on their contributions, with their relative importance within the system taken into account. This fosters a fair and motivating environment for all participants.*

Proof: To establish the theorem's claim of fair reward allocation, we demonstrate that $\frac{\partial U_{M,k}}{\partial Me_k} > 0$ and $\frac{\partial U_{i,k}}{\partial S_i^k} > 0$. Recall from Section 5.2 the assumptions related to success probability and cost functions that are essential for establishing this theorem.

Model Owner: The derivative of the utility function for the model owner, $U_{M,k}$ (as defined in Eq. 5.1), with respect to the effort variable Me_k is:

$$\frac{\partial U_{M,k}}{\partial Me_k} = \frac{\partial P_k(Me_k)}{\partial Me_k} \times [R_k + U_{M,k+1}] - \frac{\partial C(Me_k)}{\partial Me_k}. \tag{5.9}$$

Evaluating $\frac{\partial U_{M,k}}{\partial Me_k}$ for $Me_k < Me_k^*$ reveals:

1. Positive Derivative of Success Probability: An increase in effort Me_k leads to a higher success probability, as indicated by $\frac{\partial P_k(Me_k)}{\partial Me_k} > 0$.

2. Diminishing Marginal Returns: The rate of increase in success probability diminishes with escalating effort Me_k , denoted by $\frac{\partial^2 P_k(Me_k)}{\partial^2 Me_k} < 0$, up to the optimal effort level Me_k^*.

3. Increasing Marginal Cost: Each additional unit of effort incurs higher costs, as evidenced by $\frac{\partial C(Me_k)}{\partial Me_k} > 0$.

4. Optimal Effort Level Me_k^*: At Me_k^*, the marginal benefit and cost balance each other. For $Me_k < Me_k^*$, the marginal benefit surpasses the marginal cost.

Consequently, for $Me_k < Me_k^*$, it is reasonable to infer $\frac{\partial U_{M,k}}{\partial Me_k} \geq 0$ based on these assumptions, indicating the fairness in reward allocation for the model owner under the stipulated conditions.

Data Owner: Differentiating $U_{i,k}$ for data owner i (as defined in Eq. 5.3) with respect to S_i^k yields:

$$\frac{\partial U_{i,k}}{\partial S_i^k} = \frac{\partial}{\partial S_i^k}\left(b_i \times U_{D,k} + w_i \times \sum_{m=k}^{K}\left(\frac{S_i^k}{\sum_{j=1}^{N} S_j^k}\right) \times BBF_k - c(S_i^k)\right) \tag{5.10}$$

Simplifying, we find:

$$\frac{\partial U_{i,k}}{\partial S_i^k} = w_i \times \sum_{m=k}^{K} \frac{BBF_k}{\sum_{j=1}^{N} S_j^k} - \frac{\partial c(S_i^k)}{\partial S_i^k}. \tag{5.11}$$

Evaluating $\frac{\partial U_{i,k}}{\partial S_i^k}$ for $S_i^k < S_i^{k*}$ reveals:

1. Increasing Marginal Cost: Each additional unit of effort incurs higher costs, as evidenced by $\frac{\partial c(S_i^k)}{\partial S_i^k} > 0$.

2. Optimal Effort Level S_i^{k*}: At S_i^{k*}, the marginal benefit and cost balance each other. For $S_i^k < S_i^{k*}$, the marginal benefit surpasses the marginal cost.

Given the positivity of BBF_k and the sum $\sum_{j=1}^{N} S_j^k$, consequently, for $S_i^k < S_i^{k*}$, it is reasonable to infer $\frac{\partial U_{i,k}}{\partial S_i^k} > 0$ based on these assumptions. This result substantiates the assertion that an incremental increase in the data owner's contribution S_i^k leads to an increase in their utility $U_{i,k}$, affirming the fairness of the reward allocation for the data owner i within the established utility framework.

5.5 Contract Optimality

5.5.1 Reward Contract Optimization

Central to the RFLF is resolving the optimization problem inherent in the model owner's reward contract. This requires a model encompassing the owner's effort levels and reward mechanisms to ensure they're incentivized for optimal contributions at each stage. This is crucial for project success, economic viability, and fairness.

The RFLF establishes the contract at the first stage $k = 1$, with stage-specific payoffs defined as $U_{D,1}$ for the data owners' alliance and $U_{M,1}$ for the model owner. Payoffs are cumulative, with each stage incorporating the value generated in subsequent stages. This dynamic reflects the evolving nature of federated learning, where model owner effort and strategic decisions by both parties are impacted by stage rewards and the growing value of the model. The data owners' alliance D acts as a Stackelberg leader, carefully structuring the training rewards R_k across time, anticipating the model owner's self-interested actions.

Given the constraints and the previously delineated equations, the optimization problem for the data owners' alliance D can be reformulated as follows:

$$\max_{R_k} U_{D,1} = P_1(Me_1)[V_1 - R_1] + \sum_{k=2}^{K} \left\{ \prod_{j=1}^{k-1} P_j(Me_j) \right\} P_k(Me_k)[V_k - R_k],$$
$$k = 2, \cdots, K. \tag{5.12}$$

Conversely, the model owner engages in the following recursive optimization process:

$$\max_{Me_k} U_{M,k} = P_k(Me_k)[R_k + U_{M,k+1}] - C(Me_k) \quad k = 1, \cdots, K. \tag{5.13}$$

To maximize the value generated through the RFLF, we must carefully analyze the interplay of factors impacting model owner motivation and the corresponding optimization problem faced by the data owners' alliance.

Proposition 1: Reward, Future Prospect Incentivization, and Optimal Effort. *An increase in the rewards offered by the data owners, coupled with an enhancement in the model owner's prospects, is expected to motivate them to elevate their level of effort. To maximize this motivation, the model owner's marginal benefit from effort must equal their marginal cost. Striking this balance at each stage ensures optimal contribution, balancing immediate rewards, long-term gains, and costs incurred. This interplay of factors is pivotal in ensuring the model owner engages in diligent and ethical work practices.*

Proof: To analyze the model owner's optimal effort under RFLF, we consider the partial derivative of the model owner's utility function (Equ. 5.2) with respect to their effort level Me_k. This analysis helps identify the effort level that maximizes the model owner's utility at each stage k.

Recalling the utility function of the model owner, the optimal effort is derived by setting the partial derivative of $U_{M,k}$ with respect to Me_k to 0. Considering the relaxed constraints discussed in Section 5.4, we arrive at a crucial condition that signifies the peak of utility, corresponding to the optimal effort level:

$$Me_k^* = Me_k^*(R_k + U_{M,k+1}) \quad k = 1, \cdots, K. \tag{5.14}$$

This equation demonstrates the direct influence of both current reward R_k and future expected model utility $U_{M,k+1}$ on the model owner's optimal effort Me_k^*.

To ensure maximal impact from this relationship, the model owner's marginal benefit from effort must equal their marginal cost:

$$\frac{\partial P_k(Me_k)}{\partial Me_k}(R_k + U_{M,k+1}) = \frac{\partial C(Me_k)}{\partial Me_k} \quad k = 1, \cdots, K. \tag{5.15}$$

This condition ensures the model owner optimally balances immediate and future incentives against the cost of their effort.

Building upon the principles in Proposition 1, which demonstrated the importance of combining immediate and future facing rewards to maximize model owner effort, we now focus on a critical aspect of reward allocation: the temporal dynamics. This framework aspect is intuitive and has been rigorously determined through mathematical analysis.

Proposition 2: Divergence in Reward Timing Preferences between Model and Data Owners. *Within the RFLF, a mathematically demonstrable disparity exists in the temporal preferences regarding the timing of rewards between the model owner and the data owners. The model owner's marginal utility of rewards diminishes over time, indicating a preference for receiving rewards earlier. In contrast, the dynamic reward mechanism enhances the data owner's marginal utility over time, leading to their preference for delaying rewards to the model owner.*

The mathematical expression of this temporal dynamic is as follows:

$$\begin{cases} \frac{\partial U_{M,1}}{\partial R_k}\Big|_{Me_j{}^*} > \frac{\partial U_{M,1}}{\partial R_{k+1}}\Big|_{Me_j{}^*} & (k = 1, \cdots, K-1). \\ \frac{\partial U_{D,2}}{\partial R_j}\Big|_{R_i{}^*, i=1,\cdots,K} > \frac{\partial U_{D,2}}{\partial R_k}\Big|_{R_i{}^*, i=1,\cdots,K}, & R_k{}^* > 0, R_j{}^* > 0, j > k. \end{cases} \tag{5.16}$$

Proof: Given the optimal level of effort $Me_k{}^*$ for the model owner, $U_{M,k}$ in Equ. 5.3 satisfies the following conditions:

$$\frac{\partial U_{M,k}}{\partial R_k} = \prod_{j=m}^{k} P_j(Me_j{}^*). \tag{5.17}$$

From Equ. 5.17,

$$\frac{\frac{\partial U_{M,1}}{\partial R_k}}{\frac{\partial U_{M,1}}{\partial R_{k+1}}} = \frac{\prod_{j=1}^{k} P_j(Me_j{}^*)}{\prod_{j=1}^{k+1} P_j(Me_j{}^*)} = \frac{1}{P_{k+1}(Me_{k+1}{}^*)} > 1 \quad (k = 1, \cdots, K-1). \tag{5.18}$$

Then

$$\left.\frac{\partial U_{M,1}}{\partial R_k}\right|_{Me_j{}^*} > \left.\frac{\partial U_{M,1}}{\partial R_{k+1}}\right|_{Me_j{}^*} \quad (k = 1, \cdots, K-1). \tag{5.19}$$

This inequality states that the change in the model owner's utility due to a reward in stage k is greater than the change from the same size reward in stage $k+1$. In simpler terms, a reward received today has a stronger positive impact on the model owner's utility than one received later.

The optimal incentive ${R_k}^* > 0 \quad (k = 1, \cdots, K)$ for the model owner is determined before starting the first stage of training. Therefore, the optimal utility ${U_{D,k}}^*$ of the data owners' alliance can also be solved. The first-order condition of the utility of data owners' alliance with respect to rewards R_k from Eq. 5.1 is

$$\begin{aligned}\left.\frac{\partial U_{D,1}}{\partial R_k}\right|_{R_i{}^*, i=1,\cdots,K} =& {P_1}'(Me_1{}^*){Me_1}^{*\prime}\frac{\partial U_{M,2}}{\partial R_k}(V_1 - R_1 + U_{D,2}) \\ &+ P_1(Me_1{}^*)\frac{\partial U_{D,2}}{\partial R_k} = 0.\end{aligned} \tag{5.20}$$

From eq.5.19, we can derive $\frac{\partial U_{M,2}}{\partial R_k} = \prod_{j=2}^{k} P_j(Me_j{}^*)$ and rearranging the terms yield:

$$\begin{aligned}\left.\frac{\partial U_{D,2}}{\partial R_k}\right|_{R_i{}^*, i=1,\cdots,K} =& -\frac{1}{P_1(Me_1{}^*)}{P_1}'(Me_1{}^*){Me_1}^{*\prime} \\ &\left[\prod_{j=2}^{k} P_j(Me_j{}^*)\right](V_1 - R_1 + U_{D,2}) < 0.\end{aligned} \tag{5.21}$$

Thus, if ${R_k}^* > 0$ and ${R_j}^* > 0, j > k$, then

$$\begin{aligned}\left.\frac{\partial U_{D,2}}{\partial R_j}\right|_{R_i{}^*, i=1,\cdots,K} &= \left(\prod_{i=k+1}^{j} P_i(Me_i{}^*)\right)\left.\frac{\partial U_{D,2}}{\partial R_k}\right|_{R_i{}^*, i=1,\cdots,K} \\ &> \left.\frac{\partial U_{D,2}}{\partial R_k}\right|_{R_i{}^*, i=1,\cdots,K}.\end{aligned} \tag{5.22}$$

This means that later-stage rewards (at stage j, when $j > k$) have a stronger positive impact on the data owners' utility than earlier rewards (at stage k).

This proposition is a testament to the analytical depth of the RFLF, highlighting the necessity for strategic planning and execution in the temporal allocation of rewards. It underscores the importance of achieving an equilibrium that satisfies the model's and data owners' distinct utility maximisation strategies. Furthermore, it mathematically substantiates that this divergence points to an optimal payoff point within the framework, effectively balancing these contrasting preferences.

Theorem 3: Optimal Payoff Point for the Model Owner in RFLF. *Within a feasible multi-stage federated learning contract, there often comes a point where the model's value has increased so substantially that the most efficient strategy is to allocate all subsequent profits to the model owner. The RFLF framework allows us to identify this optimal payoff point, denoted as δ , through rigorous mathematical analysis. It is defined as follows:*

$$\begin{cases} {R_k}^* = 0 \quad (k < \delta), \\ {R_k}^* = {V_k}^*, {U_{D,k}}^* = 0 \quad (k > \delta). \end{cases} \tag{5.23}$$

Proof : Proposition 2 established that the data owners' preference is to delay rewards to the model owner. To determine when this delay is no longer advantageous, we must analyze how their utility changes across training stages. Let's begin with the first-order condition of their utility with respect to rewards:

$$\begin{aligned} \frac{\partial U_{D,1}}{\partial R_k} =& {P_1}'({Me_1}^*){Me_1}^{*\prime}\frac{\partial U_{M,2}}{\partial R_k}(V_1 - R_1 + U_{D,2}) \\ &+ P_1({Me_1}^*)\frac{\partial U_{D,2}}{\partial R_k} \end{aligned} \tag{5.24}$$

For every m and k such that $k > m$,

$$\begin{aligned} \frac{\partial U_{D,m}}{\partial R_k} =& {P_m}'({Me_m}^*){Me_m}^{*\prime}\frac{\partial U_{M,m+1}}{\partial R_k}(V_m - R_m + U_{D,m+1}) \\ &+ P_m({Me_m}^*)\frac{\partial U_{D,m+1}}{\partial R_k} \end{aligned} \tag{5.25}$$

and for every k,

$$\frac{\partial U_{D,k}}{\partial R_k} = P_k{}'(Me_k{}^*)Me_k{}^{*\prime}(V_k - R_k + U_{D,k+1}) - P_k(Me_k{}^*) \tag{5.26}$$

The derivative of the model owner's utility with respect to their effort is

$$\frac{\partial U_{M,k}}{\partial Me_k} = P_k{}'(Me_k)(R_k + U_{M,k+1}) - C'(Me_k) = 0, K = 1, \cdots, K \tag{5.27}$$

To see how the utility of data owners' alliance changes with respect to rewards across stages, we derive the following expression:

Substituting Equ. 5.24 into the above three equations, we have

$$\begin{aligned}\frac{\partial U_{D,1}}{\partial R_k} = &\left(\prod_{j=1}^{k} P_j(Me_j{}^*)\right)\sum_{i=1}^{k}\frac{1}{P_i(Me_i{}^*)} \\ &P_i{}'(Me_i{}^*)Me_i{}^*[V_i - R_i + U_{D,i+1}] - \prod_{j=1}^{k} P_j(Me_j{}^*),\end{aligned} \tag{5.28}$$

and

$$\begin{aligned}\frac{\partial U_{D,1}}{\partial R_{k+1}} = &\frac{\partial U_{D,1}}{\partial R_k}P_{k+1}(Me_{k+1}{}^*) + \left(\prod_{j=1}^{k} P_j(Me_j{}^*)\right) \\ &P_{k+1}{}'(Me_{k+1}{}^*)Me_{k+1}{}^*[V_{k+1} - R_{k+1} + U_{D,k+2}] = 0\end{aligned} \tag{5.29}$$

Since $\frac{\partial U_{D,1}}{\partial R_k}\Big|_{R_i{}^*, i=1,\cdots,K} = 0$, from Equ. 5.20, we can derive $V_{k+1} - R_{k+1} + U_{D,k+2} = 0$. It is known that $U_{D,K+1} = 0$, so it follows that $R_K{}^* = V_K$, so $U_{D,K} = 0$. Similarly, for any δ, there is $1 \leq \delta \leq K-1$. If $Me_\delta{}^* > 0$ and $R_\delta{}^* > 0$, then the Equ. 5.23 is proofed.

Theorem 3 plays a crucial role in the strategic design of the RFLF. It delineates a clear demarcation point in the reward structure, beyond which the incentive system is wholly skewed in favour of the model owner. This determination of δ is vital for ensuring fairness and motivation, particularly in sustaining the model owner's engagement and contribution throughout the federated learning project.

Essentially, theorem 3, as presented within the RFLF, elucidates a critical aspect of the incentive structure. It posits that the most efficient incentive scheme is one

where the incremental value generated by the model owner's efforts is predominantly returned to them, particularly in the later stages of the training process. This theorem underscores a foundational principle of the RFLF: the success achieved in the latter stages of training is inherently contingent upon the accomplishments of the earlier stages. Consequently, the reward system is designed such that the incentives provided in the later stages act as a driving force for exerting effort in the earlier stages.

This structure of incentive allocation is not only strategic but also economically prudent. By aligning the rewards more significantly with the later stages—where the model's value has presumably been enhanced due to the cumulative efforts of the model owner—the framework ensures that the model owner is adequately compensated for their contributions. This approach minimizes the cost of incentivization while maximizing the model owner's motivation throughout the project lifecycle.

5.5.2 Data Contribution Buy-back Contract Optimization

The RFLF prioritises optimizing both the reward contract for model owners and the buy-back contract for data owners. This buy-back mechanism governs compensation for data contributions and is carefully designed to incentivize substantial, consistent contributions that directly correlate with the expected financial rewards for each participant. As data can hold multifaceted value, the contract is adaptable to various valuation techniques. Achieving the ideal balance within this contract involves fostering continuous, high-volume contributions while remaining equitable throughout the federated learning process to address varying types and amounts of data.

During the initial formation of the collaborative project, data owners contribute capital proportional to their nominal share value b_i. This forms the backbone of the project's financial structure. This investment underpins the data contribution buy-back fund pool $BBF_k, k = 1, \ldots, K$ while financing the stage rewards awarded to the model owner. This upfront financial commitment helps create clear responsibilities and incentives from the outset, fostering a balanced and effective federated learning environment. Dynamically allocating buy-back funds across different train-

ing stages allows for incentivizing high-quality data and fair compensation for all contributors. While returns across data owners are consistent, each participant's cost will naturally vary according to the amount of data they contribute. Aligning returns with contributions while considering these cost differences promotes overall fairness.

The optimization problem for data owner contributions under the buy-back contract incorporates the utility $U_{i,k}$ for each data owner i at stage k. It is formulated as follows:

$$\begin{aligned} \max_{S_i^k} U_{i,k} = b_i \times U_{D,k} + \left(\frac{S_i^k}{\sum_{j=1}^{N} S_j^k} \right) \times BBF_k - c(s_i^k), \\ k = 1, \cdots, K \quad i = 1, \cdots, N. \end{aligned} \tag{5.30}$$

The guiding principle for optimizing data owner contributions under the buy-back contract is captured in a fundamental theorem. It establishes the conditions that encourage individual data owners to contribute their data optimally to the federated learning project. This theorem highlights the effectiveness of the buy-back contract and how it aligns participant incentives and rewards with individual data contributor needs and overall project goals for data diversity and abundance.

Theorem 4: Optimal Data Contribution under Buy back Contract. *Within the RFLF, the buy-back contract is designed to incentivize each individual data owner to maximize their data contribution S_i^k by ensuring that the marginal benefits from additional data contributions, in terms of buy-back funding, are aligned with the overarching goals of the federated learning process. This alignment encourages data owners to actively participate by contributing the maximum high-quality data.*

Proof: Consider the utility function for an individual data owner i at stage k, as defined in Eq. 5.3

To maximize $U_{i,k}$, we examine the condition where the marginal increase in utility from contributing additional data is at least as great as any marginal costs associated with such contributions. This condition is met when the derivative of $U_{i,k}$ with respect to S_i^k suggests that increasing S_i^k continues to provide net positive utility to the data owner: $\frac{\partial U_{i,k}}{\partial S_i^k} \geq 0$.

This derivative condition ensures that the buy-back contract must be structured such that the incremental utility from additional data contributions (S_i^k)—accounting for both the compensation from BBF_k and any associated costs—is maximized, thus incentivizing the data owner to contribute as much data as possible.

Theorem 4 demonstrates the effectiveness of the buy-back contract in aligning data owners' incentives with the federated learning process's objectives. By ensuring that data owners are compensated in a manner that reflects the true value of their contributions, the RFLF promotes a cooperative and productive environment conducive to the success of federated learning initiatives.

5.6 Performance Evaluation

To rigorously assess the RFLF's efficacy in incentivizing the model and data owners, we conducted a series of experiments simulating a federated learning environment. These experiments incorporated variations in data distribution and rewards across different stages. In this section, we describe the experimental setup, including datasets, comparison baselines, and evaluation metrics. We then present and analyze the RFLF's experimental results.

5.6.1 Experiment Setup

All experiments were conducted using a system equipped with 2 × Intel(R) Xeon(R) CPUs @ 2.30GHz, 12GB memory, and one Tesla T4 GPU, utilizing Python 3.11 and TensorFlow 2.15 for implementation.

Table 5.2 provides a comprehensive overview of the experimental setup, including the numerical simulation parameters, the federated learning environment and hyperparameters for the datasets used.

Datasets: We selected the MNIST [2] and CIFAR-10 [3] data-sets due to their wide-

[2]MNIST[89] consists of 70,000 grayscale images of handwritten digits (0-9), divided into a training set of 60,000 images and a test set of 10,000 images. Each image is a 28x28 pixel square centred around a single digit.

[3]CIFA-10[90] contains 60,000 32x32 colour images in 10 classes, with 6,000 images per class. The dataset is split into a training set of 50,000 images and a test set of 10,000 images featuring

Parameter/Function	**Value/Formula**	
Numerical Simulation Parameters:		
K	3	
N	4	
V_k	$[1, 2, 3]$	
$P_k(Me_k)$	$min(0.8 \times Me_k, 1)$	
$C(Me_k)$	${Me_k}^2$	
$c(s_i^k)$	sample number × data unit cost	
Data Unit Cost (per 100 units)	0.06 for MNIST 0.006 for CIFAR-10	
BBF_k	$[1, 2, 3]$	
w_i	$[1, 1, 1, 1]$	
Federated Learning Hyperparameters/Setting:		
	MNIST	CIFAR-10
Local Epochs	2	2
Local learning Rate	0.15	- (Adam optimizer used)
Local Batch Size	16	
Learning Rate Decay	0.977	
Communication Rounds per Trainning Stage	30	30
Optimizer	SGD[87]	Adam[88]
Algorithm	FedAvg [13]	

Table 5.2: RFLF Simulation Setup and Configurations

spread use in federated learning benchmarking.

We simulated real-world data heterogeneity and potential fluctuations in data contributions by using structured random assignment with fixed ratios to allocate varying dataset sizes to each data owner (Table 5.3). This approach captured realistic data distribution disparities and potential under-contributions, and pre-allocated buy-back funding was included to address under-contributions, reflecting practical

various objects and animals.

commercial considerations.

		Equal Reward Framework			Share Profit Framework			RFLF			Standalone Framwork
Stage		1	2	3	1	2	3	1	2	3	N/A
Data Owner 1	MNIST	150	150	150	150	150	150	150	150	150	150
	Cifar10	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500
Data Owner 2	MNIST	150	300	300	150	300	300	150	300	300	300
	Cifar10	1500	3000	3000	1500	3000	3000	1500	3000	3000	3000
Data Owner 3	MNIST	150	300	600	150	375	800	150	375	800	800
	Cifar10	1500	3000	6000	1500	3750	8000	1500	3750	8000	8000
Data Owner 4	MNIST	150	300	600	150	375	1150	150	375	1150	1150
	Cifar10	1500	3000	6000	1500	3750	11500	1500	3750	11500	11500

Table 5.3: Datasets Assignment for Performance Evaluation

Before experimentation, both datasets underwent standard preprocessing steps, including normalization and augmentation, to ensure a fair and consistent input for all models. Furthermore, 10% of the training examples were randomly selected to form a validation set, facilitating periodic assessment of model performance during the training phase.

Evaluation Metrics: The performance of the RFLF was measured using several metrics: Firstly, cooperation performance metrics will include the model owner's effort, probabilities of success, and model accuracies. Secondly, we'll examine participants' utility, evaluating data owner and model owner utilities. Finally, we'll focus on social utility metrics, specifically the framework's ability to prevent free-riding behaviour and analyze its inherent compensation mechanisms for data owners. Results across these metrics will be compared for the RFLF, a 'Standalone' framework (where no rewards are shared), an 'Equal Reward' framework[91] , and a 'Share Profit' framework[92].

The following sections detail the results of these experiments, providing insights into the framework's capability to foster an equitable and efficient learning environment.

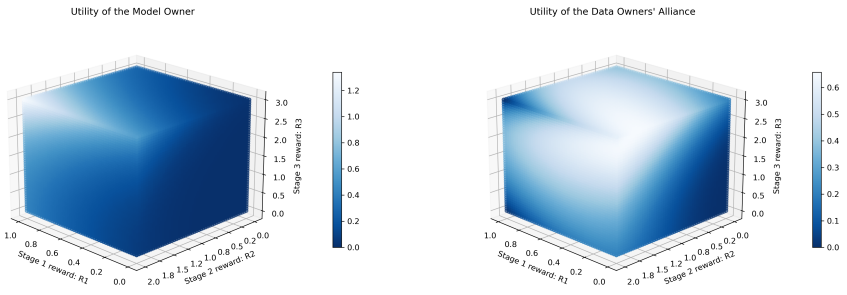

Figure 5.3: Impact of Reward Structures on Utilities in Model Owner-Data Owner Alliances

5.6.2 Contract Optimality

Based on the aforementioned settings, we first delve into the optimality of the incentive contract under RFLF, with an initial focus on the contract for the model owner. This approach is rooted in the framework's dynamics, where data owners, forming a collective interest group, engage in an incentive game with the model owner. It's important to note that how data owners divide rewards internally among themselves takes place after the settlement with the model owner is reached. Since this internal distribution is based on control rights over the delivered model, it ensures our analysis can focus on finding a model owner reward structure optimal for the project as a whole.

Fig.5.3 reveals steadily increasing variations in the model owner's utility for different reward settings. This trend indicates a positive correlation between the model owner's utility and the reward structure offered by the data owners' alliance. Consequently, the data owners' alliance achieved peak utility at $R_1 = 0$, $R_2 = 1.648$, and $R_3 = 3$, signifying an optimized status at these reward levels. This optimized status aligns perfectly with the Incentive Compatibility (IC) constraint, confirming the reward mechanism's effectiveness in motivating the model owner's contributions (Fig.5.4).

Fig.5.5 demonstrates that all participants, including the model owner and the data owners' alliance, enjoy positive utilities, satisfying the Individual Rationality (IR)

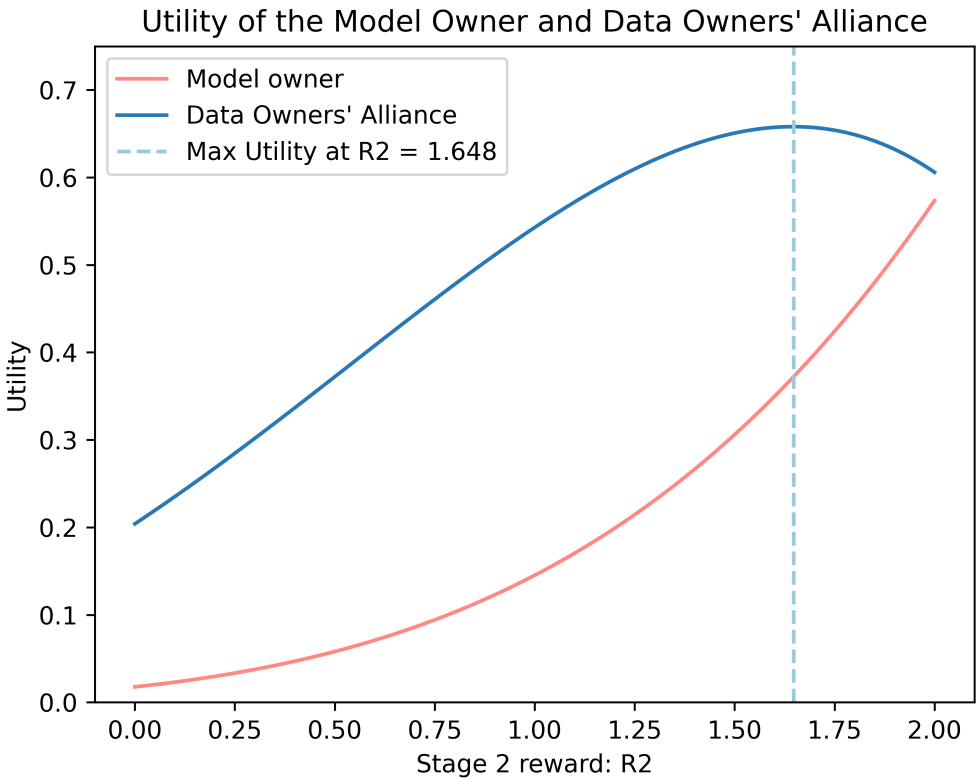

Figure 5.4: Optimal Rewards Setting

constraint. This ensures that participation in the federated learning process is rational and beneficial for everyone involved.

These observations collectively demonstrate our RFLF contract's efficacy in leading to an incentive structure consistent with a Nash Equilibrium - meaning neither the model owner nor the data owners are incentivized to unilaterally change their behaviours based on this reward framework.

5.6.3 Baselines

To comprehensively assess the performance of the RFLF, we compare its outcomes against three baseline frameworks, each representing a distinct approach to model training and incentive mechanisms within federated learning environments:

Standalone: In this approach, individual standalone models are trained on each data owner's local dataset without collaboration or data sharing. This baseline highlights the potential limitations of non-collaborative work, potentially harming

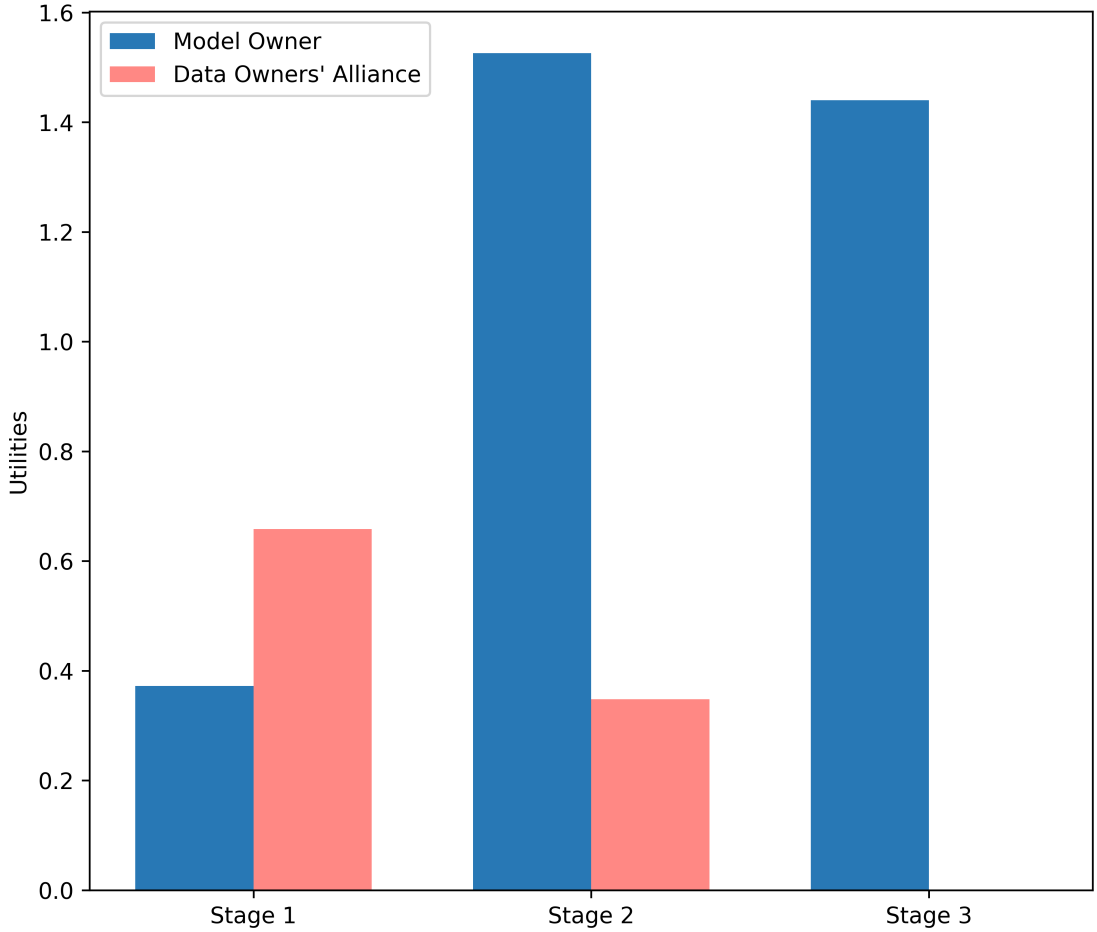

Figure 5.5: Evolution of Utilities in Model Owner-Data Owner Alliances Across Different Stages

model performance due to limited data volume and variety for each owner.

Equal Reward Framework[91]: Rewards given to participants are prearranged at the start of training, independent of the model's value and success. Specifically, we set the rewards received by model owners at the first stage to 1 (representing the complete stage 1 model value) and 1.824 for subsequent stages, totalling 4.648. This upfront approach incentivizes participation but doesn't link rewards to evolving model quality. The first stage value constraint ensures the Individual Rationality (IR) constraint is met for the data owners' alliance - awarding higher rewards at this early stage would make participation infeasible for them.

Share Profit Framework[92]: This framework links a model owner's incentives to the quality of the model delivered to the data owners, with benefits similarly distributed

amongst data owners in proportion to their contribution. Given that the total reward discussed in the previous section is 4.648 (representing 77.5% of the total incremental value of the model), we set the rewards received by model owners after each successful training phase to 0.775, 1.55 and 2.325, respectively. Data owners first pre-deposit the same proportion of the project margin as they expect to earn and then receive a proportionate payoff based on the actual contribution at the end of each phase, including contribution buyback allowance and model ownership.

By comparing these baselines, we aim to demonstrate the advantages of the RFLF in fostering a more equitable, efficient, and effective federated learning environment. We'll next examine results concerning cooperation performance, followed by participant utilities and social utility within each framework.

5.6.4 Experimental Results

Experimental evaluation of the RFLF demonstrates significant improvements in cooperation, predictability of returns, and overall model performance compared to baseline federated learning frameworks. Our analysis delves into several key metrics to comprehensively assess RFLF's effectiveness in addressing the challenges identified in existing incentive frameworks.

Cooperation Performance

RFLF notably increases the model owner's effort, particularly in later stages of the federated learning process. In both Stage 2 and 3, the effort exerted under RFLF surpassed that observed in the Equal Reward and Share Profit frameworks. Specifically, compared to Share framework, model owner effort increased by 28.47% and 29.03% in Stage 2 and 3, respectively. This increase was even more pronounced when compared to the Equal Reward framework, with increases of 92.19% and 93.55% (Figure 5.6). This significant boost in effort directly addresses the research gap of incentivizing active model owner participation, demonstrating the efficacy of RFLF's reward structure in motivating sustained contribution.

Success probabilities further highlight the effectiveness of RFLF's incentivization strategy. Under RFLF, Stage 2 and 3 probabilities reached 98.8% and 96%, respec-

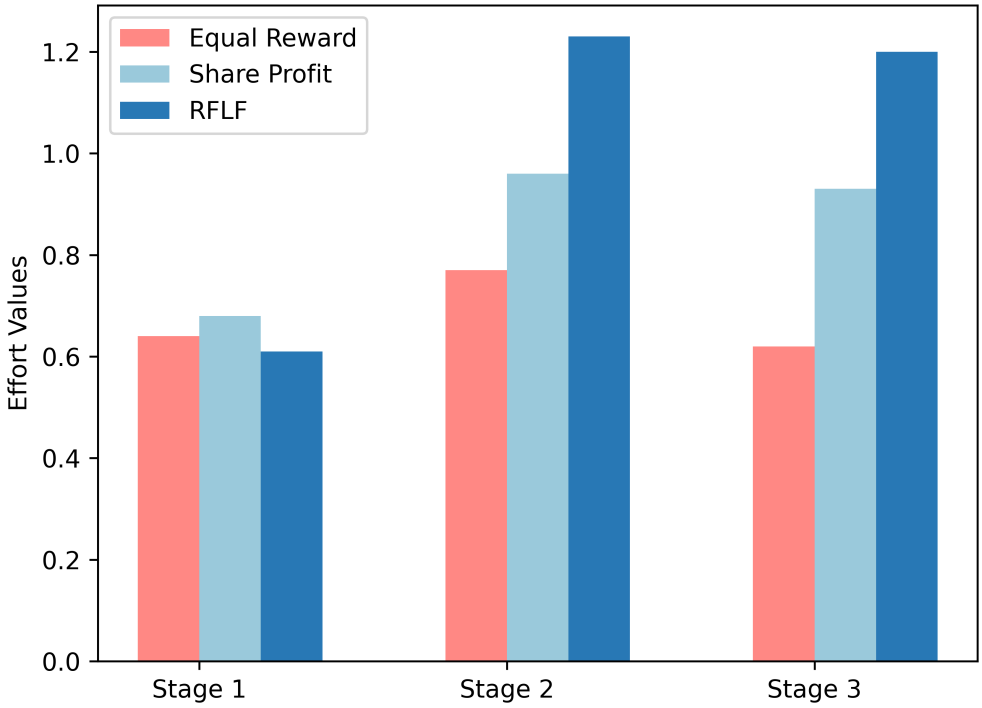

Figure 5.6: Model Owners' Effort Across Different Frameworks

tively, showcasing significant improvements over both the Share Profit (21.66% in Stage 2, 47% in Stage 3) and Equal Reward frameworks (47.67% in Stage 2, 46.43% in Stage 3) (Figure 5.7). These results underscore the importance of aligning rewards with both performance and contribution, a key feature of RFLF that sets it apart from existing models.

Additionally, we compared the final model accuracy achieved under different frameworks. While all federated learning approaches outperformed standalone models, RFLF and Share Profit framework achieved the highest accuracy, surpassing the Equal Reward framework (Table 5.4). This result further reinforces the effectiveness of incentive structures that link rewards to both performance and contribution, a principle that RFLF uniquely extends by considering the evolving nature of contributions throughout the federated learning process.

RFLF's emphasis on strategically allocating the majority of incentives in the final stages proves highly effective in driving collaboration. This incentivization approach

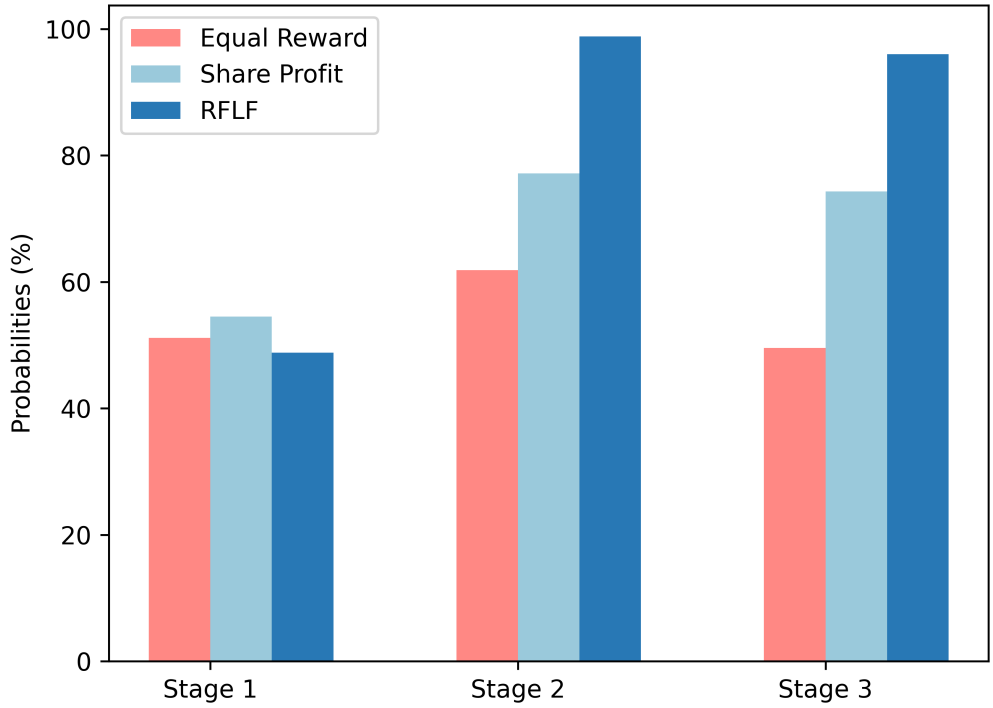

Figure 5.7: Probabilities of Success Training Across Different Frameworks

		MNIST			CIFAR-10		
	Stages	1	2	3	1	2	3
Standalone Framework	Data Owner 1	N/A	N/A	81.87%	N/A	N/A	43.44%
	Data Owner 2	N/A	N/A	86.90%	N/A	N/A	51.41%
	Data Owner 3	N/A	N/A	90.81%	N/A	N/A	59.65%
	Data Owner 4	N/A	N/A	93.28%	N/A	N/A	61.81%
Equal Reward Framework		93.89%	94.72%	96.15%	55.24%	58.22%	62.43%
Share Profit Framework		93.99%	95.27%	96.94%	55.31%	60.14%	64.40%
RFLF		93.90%	95.30%	**97.06%**	56.09%	61.27%	**64.67%**

Table 5.4: Model Accuracies within Different Frameworks

significantly improves success rates and model accuracy compared to the Share Profit and Equal Reward frameworks.

Participants' Utility

Examining data owner and model owner utilities across Equal Rewards, Share Profit, and RFLF frameworks reveals how different incentive mechanisms impact outcomes.

Framework	Data Owner	Individual Utility	Total Utility	Data Contribution	Share of the Model
Equal Reward	1	0.92	3.04	10%	25%
	2	0.83		18%	25%
	3	0.65		36%	25%
	4	0.65		36%	25%
Share Profit	1	0.08	3.40	6%	6%
	2	0.41		13%	13%
	3	1.20		33%	33%
	4	1.72		48%	48%
RFLF	1	**0.20**	**4.56**	**6%**	**25%**
	2	**0.60**		**13%**	**25%**
	3	**1.57**		**33%**	**25%**
	4	**2.20**		**48%**	**25%**

Table 5.5: Utilities of Data Owners within Different Frameworks

First, we provide a comprehensive analysis of the utilities of data owners within each framework (see Table 5.5). Under Equal Rewards, the total utility for all data owners is 3.04. Despite contribution levels ranging from 9% to 36%, each data owner receives an equal 25% model share. While simple, this distribution disregards individual effort, potentially lowering efficiency.

The Share Profit framework, yielding a total data owners' utility of 3.40, improves motivation with a direct link between utilities, model ownership, and contributions. For example, data owner 4, with a higher contribution of 48%, has the greatest utility of 1.72 and the largest model share of 48%. This incentivizes more effort, as rewards depend on success - though note that final model shares may differ from the pre-agreed distribution depending on individual contributions throughout the training process.

RFLF stands apart with the highest total data owners' utility of 4.56, suggesting significant cumulative benefit. While model shares remain consistent with the pre-agreed distribution, the individual data owner's utility distribution within RFLF mirrors the Share Profit framework, rewarding proportionally to contribution.

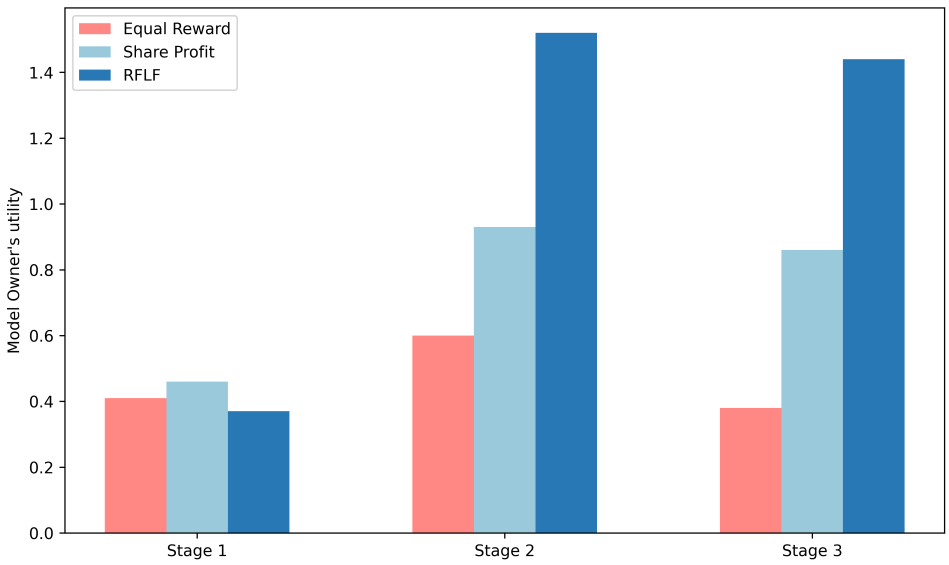

Figure 5.8: Model Owner's Utility Across Different Frameworks and Stages

Second, the model owner's utility showcases RFLF's strategic incentives. Starting low in stage 1, the model owner's utility rises significantly in stages 2 and 3 - a pattern made clear in Figure 5.8. This outpaces both the Share Profit and Equal Rewards frameworks. Backloading incentives align with evolving model value, promoting optimal, continuous effort from the model owner.

RFLF's model excels because it sustains motivation on all sides. Equal Rewards framework, while simple, may under-motivate; Share Profit framework addresses this by linking rewards to success. However, RFLF's strategic timing fosters a more cooperative and productive federated learning environment with the potential for high engagement from both data owners and model owner.

Social Utility

A crucial component of social utility in federated learning is preventing free-riding, where participants gain unduly without proportional contribution. In Equal Rewards, this poses a serious threat. Despite contribution levels ranging from 9% to 36.5%, data owner utilities were tightly clustered (0.92, 0.83, 0.65, 0.65). These near-equal rewards discourage high effort for two reasons: minimal additional return regardless of input and contributors potentially subsidizing low-effort peers due to

differing data costs.

By contrast, both Share Profit framework and RFLF incorporate features mitigating free-riding risk. In Share Profit framework, the final model share directly tracks data contribution size. RFLF addresses contribution differences through its compensation mechanism balancing capital and data inputs. This ensures larger contributors aren't inadvertently exploited for their efforts, aligning with equity and motivation principles.

Comparing Share Profit framework and RFLF further, we find similar utility distributions but divergent final model shares . While Share Profit framework's proportional payout may encourage larger contributions, it raises the spectre of imbalance if one participant dominates; this imbalance is counteracted by RFLF's pre-agreed, equal split of the final model. Beyond direct payouts, the RFLF offers an interesting additional feature: Should a data owner discover their data is underperforming, they can opt-out to avoid further expense. Their pre-deposited buy-back funds then compensate other data owners for stepping up contributions. This creates positive feedback: capital offsets any data shortfalls, ensuring both fair pay and that the project as a whole remains incentivized.

This unique compensation mechanism within RFLF goes beyond curbing free-riding to promote equitable reward distribution. Participants trust they'll be fairly compensated, encouraging honest data quality assessment and maximal effort. Ultimately, this enhances the federated learning process's social utility, furthering goals of privacy, reduced data silos, and overall model robustness.

Comparision of Different Incentive Frameworks

Table 5.6 summarizes the comparisons between RFLF, Equal Reward, and Share Profit frameworks across various metrics, highlighting the distinct advantages of each model.

The Equal Reward framework, while providing high predictability of returns, suffers from low model owner effort and is vulnerable to free-riding. The Share Profit framework improves upon this by incentivizing higher effort and partially mitigating free-riding, but it can lead to unequal distribution of the final model among data

Metric		Equal Reward	Share Profit	RFLF
Cooperation Performance	Model Owner Effort	Low	Medium	High
	Success Probability	Low	Medium	High
	Model Accuracy	Medium	High	High
Participants' Utility	Total Data Owner Utility	Low	Medium	High
	Individual Data Owner Utility Distribution	Equal	Proportional	Proportional
Social Utility	Mitigates Free-riding	No	Partial	Yes
	Adapts to Dynamic Contributions	No	Partial	Yes
	Predictability of Returns	High	Medium	High

Table 5.6: Comparison of Incentive Frameworks

owners.

RFLF, by contrast, achieves the best overall performance across most metrics. It incentivizes active model owner participation throughout the entire federated learning process, particularly in later stages where effort is crucial. Additionally, RFLF's multi-stage incentives and unique model ownership transfer mechanism effectively reduce free-riding and ensure equitable distribution of both rewards and the final model.

Importantly, RFLF maintains the high predictability of returns offered by Equal Reward, providing a more reliable and secure collaborative environment for all participants. This highlights the framework's potential to foster trust and encourage long-term participation in federated learning projects.

5.6.5 Potential Applications in Web 3.0 and 5G/6G Communication Domains

The RFLF framework is well-suited for the decentralized and privacy-conscious environments of Web 3.0 and 5G/6G technologies, where data is increasingly decentralized and held by smaller entities. It pioneers a unique approach by actively incentivizing both data and model owners within a unified federated learning environment. Its critical technological innovation lies in its multi-stage incentive alignment mechanism, which actively engages all participants throughout the federated learning process, lowering barriers to entry for individual machine learning experts and small to medium-sized enterprises (SMEs).

As the scenario described in Section 5.1 illustrates, in healthcare, the RFLF framework is particularly significant for small to medium-sized enterprises (SMEs) healthcare providers, who often possess valuable patient data but lack the machine learning expertise to develop sophisticated models. Under the RFLF, healthcare providers, acting as data owners, can collaborate with specialized machine learning experts/researchers (model owners) to leverage their respective strengths The RFLF's multi-stage incentive mechanism encourages both parties to contribute optimally throughout the project lifecycle. For instance, the model owner is motivated to invest significant effort in refining the model, as their rewards are directly tied to the model's incremental performance improvements at each stage. This alignment is crucial in healthcare, where the model's accuracy and reliability directly impact patient care. Simultaneously, RFLF's data contribution buy-back mechanism addresses data heterogeneity in healthcare settings, where data sources can vary significantly in quality and quantity. It incentivizes healthcare providers to contribute high-quality data by allowing them to opt out while retaining their expected revenue from their locked initial capital investment if their data underperforms. The remaining data contribution buy-back funds are then used to incentivize other providers to cover the contribution gap. The weight parameter also ensures fair compensation based on each provider's relative contribution. This dual approach benefits the overall data contribution and ensures equitable compensation for the healthcare providers.

Beyond healthcare, RFLF can be applied to various industries like smart agriculture, supply chain management, and logistics. In these sectors, which are also rapidly adopting Web 3.0 and 5G/6G technologies, RFLF's ability to incentivize collaboration and leverage decentralized data sources can lead to significant advancements in areas such as pest control, demand forecasting, inventory optimization, and route planning.

5.7 Conclusion and Chapter Discussion

This chapter addressed the crucial challenge of designing a unified incentive framework for model and data owners in dynamic contribution-driven federated learning

cooperation. We introduced the Reciprocal Federated Learning Framework (RFLF), a novel approach that fosters fairness, accountability, and sustainable collaboration. RFLF uniquely advances the field of federated learning incentive mechanisms in several vital ways.

Firstly, RFLF builds a mutually beneficial payoff structure through a dynamic, reciprocal approach that considers the evolving contributions of both data and model owners. This approach incentivizes both parties to exert optimal effort throughout the federated learning process. Secondly, RFLF employs multi-stage incentives, which ensure rewards and trained models are settled based on verifiable metrics at each stage. It mitigates investment and payoff risks for data and model owners, fostering a more secure, controlled and sustainable collaborative environment. Thirdly, integrating cryptographic and blockchain technologies allows RFLF to guarantee the secure, transparent, and automated execution of incentive contracts and model ownership transfer. It fosters trust and accountability among participants, essential for long-term collaborative success. Finally, empirical evaluations on established datasets (MNIST, CIFAR-10) demonstrated RFLF's effectiveness in real-world settings. These evaluations demonstrated that RFLF incentivizes cooperation, mitigates free-riding tendencies, and significantly enhances the overall efficacy of the federated learning process.

Building upon the promising foundation of RFLF, several avenues hold great potential for future research. Firstly, a crucial line of inquiry is evaluating RFLF's behaviour and impact within large-scale federated learning networks across diverse real-world application domains. It would involve analyzing its influence on model precision, training efficiency, and overall system sustainability in complex settings. Secondly, identifying potential limitations of RFLF under specific conditions (e.g., highly heterogeneous data distributions, complex model architectures, churn in data or model ownership) is vital. Such exploration paves the way for developing adaptation strategies to optimize RFLF's robustness and effectiveness across various use cases. Finally, examining how RFLF can be tailored and optimized for specific application areas within the target technologies mentioned (Web 3.0, 5G/6G, and digitalized industries) is paramount. It could involve exploring integration with blockchain technologies for enhanced security and trust in incentive management

or investigating its suitability for privacy-preserving federated learning in sensitive domains such as healthcare and finance.

Chapter 6

Conclusions and Future Works

In this chapter, we summarise the contributions of this thesis and propose potential future research directions for further exploration.

6.1 Conclusion

This thesis commences with a comprehensive overview of existing incentives for federated learning, presented in Chapter 2. In this chapter, we provide a foundational understanding of federated learning and the integration of blockchain and survey existing incentives for federated learning. After that, we discuss the shortcomings and limitations of existing incentives. In particular, the challenges of under-researched evaluation and incentives for model owners' contributions in federated learning scenarios, inappropriate incentive allocation, and insufficient incentives due to potential mismatches between data and models in federated learning implementations must be addressed before a scalable implementation of federated learning.

Given the breadth of these pressing challenges, we begin Chapter 3 with a discussion of technical work that addresses the design of incentives for model owners in federated learning. In contrast to existing works in the literature, we propose a dynamic multi-stage incentive mechanism that strategically allocates rewards to model owners based on their contributions and the resulting incremental value generated by the model. The mechanism is rigorously analysed using a mathematical model based on contract theory, revealing its ability to mitigate unethical behaviour, optimise

participant effort and promote fairness in federated learning scenarios.

Chapter 4 introduces an incentive framework for data owner-led federated learning scenarios, the Fair Clearing House (FCH) framework, based on the multi-stage incentive mechanism we introduced in Chapter 2. In particular, the data owner-led multi-stage incentive mechanism necessitates clearing updated models and rewards at the conclusion of each stage. However, the decentralised and transparent clearing of the blockchain provides a solution to this problem. To this end, we have constructed an efficient and secure two-party clearing protocol by combining cryptography and smart contracts for fair reward settlement. The empirical evaluation demonstrates that FCH outperforms traditional federated learning frameworks on various metrics, including participant effort optimisation and unethical behaviour mitigation.

However, the work presented in Chapters 3 and 4 does not fully capture the self-organising and dynamic aspects of multiple data owners in federated learning. The work presented in these two chapters still assumes the existence of a data owner or a group of data owners with aligned interests. It is a limitation, as the data and model fit of data owners can change dynamically as federated learning proceeds. This data and model mismatch can result in shifts in the returns to data owners, potentially leading to an inequitable distribution of returns among data owners. We propose a multi-stage data contribution buyback mechanism in Chapter 5 to address this. We introduce the Reciprocal Federated Learning Framework (RFLF), which further advances incentive design in federated learning. The framework supports a dynamic and reciprocal incentive structure that promotes fairness and accountability among data and model owners. The integration of blockchain technology ensures secure and transparent transactions, and empirical evaluations based on existing datasets demonstrate the effectiveness of the RFLF in facilitating collaboration, mitigating free-riding behaviours, and improving the overall effectiveness of federated learning.

6.2 Broader Implications

The research presented in this thesis has significant implications for the widespread adoption and democratization of federated learning across various domains. By

addressing the crucial challenges of incentive alignment, trust-building, and fairness, our work paves the way for a more collaborative and equitable approach to model development, especially within the dynamic landscape of the Industrial Internet of Things (IIoT) and beyond.

In the realm of smart agriculture, our contributions have the potential to revolutionize the way farmers and agricultural stakeholders participate in data-driven innovation. The collaborative nature of federated learning is crucial for addressing complex agricultural challenges such as pest detection, disease diagnosis, and resource optimization. However, the reluctance to share data due to concerns about fair compensation and potential misuse has often hindered progress. Our research directly addresses this barrier by introducing incentive mechanisms that reward farmers for their valuable data contributions, fostering trust and encouraging active participation in federated learning initiatives. Moreover, the incentive framework we designed has the potential to attract a wide range of stakeholders, including not only farmers but also machine learning specialists, researchers, government agencies, and capital investors.

Machine learning specialists and researchers, motivated by the prospect of financial rewards and access to valuable agricultural data, are incentivized to develop cutting-edge algorithms and applications. These innovations can lead to more efficient crop health monitoring systems, precision fertilizer and water management tools, and accurate yield prediction models, ultimately benefiting farmers and contributing to the overall sustainability of the agricultural sector. Government agencies and agricultural organizations can also leverage our frameworks to facilitate data sharing among farmers, leading to more informed policy decisions and the development of sustainable agricultural practices. Meanwhile, capital investors are attracted to the potential for high social and economic impact, injecting much-needed resources into the development and deployment of these innovative solutions.

Our research findings extend beyond agriculture, offering transformative potential for other IIoT domains, notably the Internet of Medical Things (IoMT). In healthcare, where data sharing is often hampered by stringent privacy regulations and concerns about patient data security, our incentive mechanisms can facilitate the

formation of secure and equitable collaborations. By ensuring fair compensation for data contributors, such as hospitals and research institutions, and incentivizing the development of high-quality models, our frameworks can catalyze the creation of innovative healthcare solutions, including personalized medicine, early disease detection, and improved treatment outcomes. Moreover, the transparent and trustworthy nature of our approach can attract a wider range of stakeholders to the healthcare sector. Machine learning experts and medical researchers, motivated by the opportunity to make a real-world impact and contribute to the advancement of medical knowledge, are drawn to federated learning projects facilitated by our frameworks. Capital investors, recognizing the immense potential for both financial returns and positive societal impact, are also more likely to invest in these collaborative ventures. This convergence of expertise, data, and resources can significantly accelerate the pace of medical innovation and improve patient care.

The impact of our research also extends to supply chain management, logistics, and other traditional industries where data-driven decision-making is crucial. By integrating federated learning with our incentive mechanisms, we enable businesses to leverage decentralized data sources while maintaining confidentiality. This collaborative approach fosters the development of predictive models for demand forecasting, inventory optimization, and route planning, leading to increased operational efficiency and a more equitable distribution of benefits among supply chain partners. The transparent and fair nature of our frameworks can also attract capital investors seeking to support innovative projects with the potential for high social and economic impact, further fueling the growth and development of the federated learning ecosystem.

By democratizing machine learning through innovative incentives, our research has the potential to transform the landscape of collaborative model development. The ability to harness the power of decentralized data while maintaining privacy and ensuring equitable outcomes is poised to revolutionize various fields. As federated learning becomes more accessible and inclusive, we anticipate a surge in data-driven innovation, with a broader range of stakeholders, including farmers, healthcare providers, researchers, investors, and machine learning specialists, actively participating in the development of solutions that address critical challenges and drive

progress across industries.

6.3 Future Works

While this thesis has made significant strides in addressing the challenges of incentive alignment and trust-building in federated learning, there are several avenues for further exploration and refinement.

- *Dealing with Malicious Participants:* A fundamental assumption of our current work is that all federated learning participants are semi-honest participants who will honestly adhere to the protocol but may attempt to collect unauthorised information, but real-world federated learning is not immune to the presence of malicious participants who may deliberately deviate from the protocol to gain an unfair advantage, or even sabotage the entire federated learning process. These malicious participants may attack federated learning through, for example, data poisoning (i.e., a malicious participant injects false or misleading data into the training set) or model poisoning (i.e., a malicious participant manipulates model updates to introduce backdoors or vulnerabilities)[93]. In addition, malicious participants may attempt to infer sensitive information about other participants' data through model inversion attacks or other privacy breaches [94]. In order to safeguard the integrity and reliability of federated learning systems, future research should delve deeper and introduce robust mechanisms capable of detecting and mitigating such malicious behaviour. This may involve employing reputation systems to track the historical behaviour of participants [27], anomaly detection algorithms to identify anomalous patterns in model updates or data contributions [95], or even sophisticated cryptographic techniques (e.g., Secure Multi-Party Computing) to validate the authenticity of shared information and the integrity without revealing its content [96].

- *Scalability and Complex Federated Learning Networks:* While our framework has demonstrated promising results in relatively controlled environments, their scalability to larger and more complex federated learning networks remains an open question. Future research should investigate strategies for efficient incen-

tive distribution, model aggregation, and conflict resolution in such large-scale collaborations. For instance, exploring decentralized consensus mechanisms [97], such as those used in blockchain networks, or sharding techniques [98] could help address the computational and communication bottlenecks that may arise as the number of participants increases. Additionally, investigating novel incentive allocation algorithms that consider the heterogeneous contributions of diverse participants [99], and the dynamic nature of their involvement [100], could further enhance the fairness and efficiency of large-scale federated learning systems.

- *Addressing Unlabeled Data in Federated Learning:* A significant limitation of our research is its primary focus on scenarios where data is labeled. While this reflects the current state of most federated learning approaches, it's important to acknowledge that real-world IIoT applications often involve a substantial amount of unlabeled or mislabeled data [101]. This poses challenges for the server to identify participants with suitable data for model training, and it may require addressing issues of scalability, heterogeneity, and privacy within federated learning systems. Future research could explore techniques for enabling devices to learn labels from each other or investigate semi-supervised learning-inspired methods to tackle the unlabeled data problem in federated learning [102]. Such advancements could significantly expand the applicability and effectiveness of our proposed frameworks in real-world scenarios where labeled data is scarce or costly to obtain.

- *Integrating Capital Investors and Expanding the Incentive Ecosystem:* While our research has expanded the incentive framework to accommodate data-owner-led federated learning scenarios, it is crucial to acknowledge that in real-world applications, especially those involving high-risk innovative projects, specialized capital investors like venture capitalists (VCs) or government funding agencies often play a primary role in providing financial resources [103]. This necessitates a deeper exploration of how to integrate these external investors into the federated learning ecosystem. Future research could investigate mechanisms to fairly distribute rewards among data owners, model owners, and capital investors, taking into account their respective contribu-

tions and risk tolerances [27]. Additionally, it would be valuable to explore novel governance structures that balance the interests of all stakeholders while ensuring transparency, accountability, and efficient decision-making in these complex collaborations [103].

Bibliography

[1] C. o. t. E. U. European Parliament, *Guide to the general data protection regulation*, 2018. [Online]. Available: `https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation`.

[2] *California consumer privacy act (ccpa)*, California Government Code § 1798.100 et seq., 2018. [Online]. Available: `https://oag.ca.gov/privacy/ccpa/regs`.

[3] 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.

[4] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, 2016.

[5] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, 2017, pp. 587–601.

[6] H. Xu, P. Nanda, and J. Liang, "Designing incentive mechanisms for fair participation in federated learning," in *2023 IEEE International Conference on High Performance Computing and Communications*, IEEE, 2023, pp. 357–373.

[7] H. Xu, P. Nanda, J. Liang, and X. He, "The force of compensation, a multi-stage incentive mechanism model for federated learning," in *International Conference on Network and System Security*, Springer, 2022, pp. 357–373.

[8] H. Xu, P. Nanda, J. Liang, and X. He, "Fch, an incentive framework for data-owner dominated federated learning," *Journal of Information Security and Applications*, vol. 76, p. 103 521, 2023.

[9] H. Xu, P. Nanda, and J. Liang, "Reciprocal federated learning framework: Balancing incentives for model and data owners," *Future Generation Computer Systems*, vol. 161, pp. 146–161, 2024.

[10] M. Asad, A. Moustafa, and C. Yu, "A critical evaluation of privacy and security threats in federated learning," *Sensors*, vol. 20, no. 24, p. 7182, 2020.

[11] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[12] R. H. Myers and R. H. Myers, *Classical and modern regression with applications*. Duxbury press Belmont, CA, 1990, vol. 2.

[13] 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*, PMLR, 2017, pp. 1273–1282.

[14] 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*, vol. 2, pp. 429–450, 2020.

[15] M. Asad, A. Moustafa, and T. Ito, "Fedopt: Towards communication efficiency and privacy preservation in federated learning," *Applied Sciences*, vol. 10, no. 8, p. 2864, 2020.

[16] M. Abadi *et al.*, "{Tensorflow}: A system for {large-scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[17] T. Ryffel *et al.*, "A generic framework for privacy preserving deep learning," *arXiv preprint arXiv:1811.04017*, 2018.

[18] C. He *et al.*, "Fedml: A research library and benchmark for federated machine learning," *arXiv preprint arXiv:2007.13518*, 2020.

[19] Y. Liu, T. Fan, T. Chen, Q. Xu, and Q. Yang, "Fate: An industrial grade platform for collaborative learning with data protection," *Journal of Machine Learning Research*, vol. 22, no. 226, pp. 1–6, 2021.

[20] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[21] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[22] R. Böhme, N. Christin, B. Edelman, and T. Moore, "Bitcoin: Economics, technology, and governance," *Journal of economic Perspectives*, vol. 29, no. 2, pp. 213–238, 2015.

[23] Y. Qu, M. P. Uddin, C. Gan, Y. Xiang, L. Gao, and J. Yearwood, "Blockchain-enabled federated learning: A survey," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–35, 2022.

[24] Z. Wang and Q. Hu, "Blockchain-based federated learning: A comprehensive survey," *arXiv preprint arXiv:2110.02182*, 2021.

[25] 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)*, IEEE, 2019, pp. 151–159.

[26] K. Toyoda and A. N. Zhang, "Mechanism design for an incentive-aware blockchain-enabled federated learning platform," in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 395–403.

[27] J. Kang, Z. Xiong, D. Niyato, H. Yu, Y.-C. Liang, and D. I. Kim, "Incentive design for efficient federated learning in mobile networks: A contract theory approach," in *2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS)*, IEEE, 2019, pp. 1–5.

[28] S. K. Lo *et al.*, "Toward trustworthy ai: Blockchain-based architecture design for accountability and fairness of federated learning systems," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3276–3284, 2022.

[29] T. Rückel, J. Sedlmeir, and P. Hofmann, "Fairness, integrity, and privacy in a scalable blockchain-based federated learning system," *Computer Networks*, vol. 202, p. 108 621, 2022.

[30] T.-T. Kuo, R. A. Gabriel, and L. Ohno-Machado, "Fair compute loads enabled by blockchain: Sharing models by alternating client and server roles,"

Journal of the American Medical Informatics Association, vol. 26, no. 5, pp. 392–403, 2019.

[31] J. Weng, J. Weng, J. Zhang, M. Li, Y. Zhang, and W. Luo, "Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2438–2455, 2019.

[32] L. Gao, L. Li, Y. Chen, C. Xu, and M. Xu, "Fgfl: A blockchain-based fair incentive governor for federated learning," *Journal of Parallel and Distributed Computing*, vol. 163, pp. 283–299, 2022.

[33] J. Kang *et al.*, "Scalable and communication-efficient decentralized federated edge learning with multi-blockchain framework," in *Blockchain and Trustworthy Systems: Second International Conference, Bloc-kSys 2020, Dali, China, August 6–7, 2020, Revised Selected Papers 2*, Springer, 2020, pp. 152–165.

[34] Y. Shi, H. Yu, and C. Leung, "Towards fairness-aware federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[35] L. K. Wang Y Li GL, "Contribution evaluation for federated learning: A survey," *Journal of Software(in Chinese)*, vol. 34, no. 3, pp. 1168–1192, 2022.

[36] L. S. Shapley *et al.*, *A value for n-person games*. Princeton University Press Princeton, 1953.

[37] R. Jia *et al.*, "Towards efficient data valuation based on the shapley value," in *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019, pp. 1167–1176.

[38] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *International conference on machine learning*, PMLR, 2019, pp. 2242–2251.

[39] T. Song, Y. Tong, and S. Wei, "Profit allocation for federated learning," in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 2577–2586.

[40] T. Nishio, R. Shinkuma, and N. B. Mandayam, "Estimation of individual device contributions for incentivizing federated learning," in *2020 IEEE Globecom Workshops (GC Wkshps*, IEEE, 2020, pp. 1–6.

[41] J. Zhang, C. Li, A. Robles-Kelly, and M. Kankanhalli, "Hierarchically fair federated learning," *arXiv preprint arXiv:2004.10386*, 2020.

[42] L. Lyu *et al.*, "Towards fair and privacy-preserving federated deep models," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 11, pp. 2524–2541, 2020.

[43] Y. Sarikaya and O. Ercetin, "Motivating workers in federated learning: A stackelberg game perspective," *IEEE Networking Letters*, vol. 2, no. 1, pp. 23–27, 2019.

[44] T. H. T. Le *et al.*, "An incentive mechanism for federated learning in wireless cellular networks: An auction approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 8, pp. 4874–4887, 2021.

[45] R. Zeng, S. Zhang, J. Wang, and X. Chu, "Fmore: An incentive scheme of multi-dimensional auction for federated learning in mec," in *2020 IEEE 40th international conference on distributed computing systems (ICDCS)*, IEEE, 2020, pp. 278–288.

[46] Y. Deng *et al.*, "Fair: Quality-aware federated learning with precise user incentive and model aggregation," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, IEEE, 2021, pp. 1–10.

[47] L. Lyu, X. Xu, Q. Wang, and H. Yu, "Collaborative fairness in federated learning," in *Federated Learning*, Springer, 2020, pp. 189–204.

[48] Y. Zhao, J. Zhao, L. Jiang, R. Tan, and D. Niyato, "Mobile edge computing, blockchain and reputation-based crowdsourcing iot federated learning: A secure, decentralized and privacy-preserving system," *arXiv preprint arXiv:1906.10893*, pp. 2327–4662, 2019.

[49] M. H. ur Rehman, K. Salah, E. Damiani, and D. Svetinovic, "Towards blockchain-based reputation-aware federated learning," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, IEEE, 2020, pp. 183–188.

[50] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 72–80, 2020.

[51] M. Yang, X. Wang, H. Zhu, H. Wang, and H. Qian, "Federated learning with class imbalance reduction," in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 2174–2178.

[52] T. Huang, W. Lin, W. Wu, L. He, K. Li, and A. Y. Zomaya, "An efficiency-boosting client selection scheme for federated learning with fairness guarantee," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1552–1564, 2020.

[53] 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)*, IEEE, 2019, pp. 1–7.

[54] P. Zhou, P. Fang, and P. Hui, "Loss tolerant federated learning," *arXiv preprint arXiv:2105.03591*, 2021.

[55] H. Wang, Z. Qu, S. Guo, X. Gao, R. Li, and B. Ye, "Intermittent pulling with local compensation for communication-efficient distributed learning," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 2, pp. 779–791, 2020.

[56] R. Hu and Y. Gong, "Trading data for learning: Incentive mechanism for on-device federated learning," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, IEEE, 2020, pp. 1–6.

[57] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*, PMLR, 2019, pp. 4615–4625.

[58] Z. Hu, K. Shaloudegi, G. Zhang, and Y. Yu, "Federated learning meets multi-objective optimization," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 4, pp. 2039–2051, 2022.

[59] S. Cui, W. Pan, J. Liang, C. Zhang, and F. Wang, "Addressing algorithmic disparity and performance inconsistency in federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 091–26 102, 2021.

[60] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[61] J. Tian, X. Lü, R. Zou, B. Zhao, and Y. Li, "A fair resource allocation scheme in federated learning," *Journal of Computer Research and Development*, vol. 59, no. 2022-06-1240, p. 1240, 2022.

[62] Z. Zhao and G. Joshi, "A dynamic reweighting strategy for fair federated learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8772–8776.

[63] T. Li, A. Beirami, M. Sanjabi, and V. Smith, "Tilted empirical risk minimization," *arXiv preprint arXiv:2007.01162*, 2020.

[64] M. H. ur 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*, vol. 17, no. 12, pp. 8485–8494, 2021.

[65] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[66] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[67] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *arXiv preprint arXiv:1811.11479*, 2018.

[68] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

[69] D. Bergemann and U. Hege, "Venture capital financing, moral hazard, and learning," *Journal of Banking & Finance*, vol. 22, no. 6-8, pp. 703–735, 1998.

[70] R. Elitzur and A. Gavious, "A multi-period game theoretic model of venture capitalists and entrepreneurs," *European Journal of Operational Research*, vol. 144, no. 2, pp. 440–453, 2003.

[71] L. Lipper *et al.*, "Climate-smart agriculture for food security," *Nature climate change*, vol. 4, no. 12, pp. 1068–1072, 2014.

[72] L. Da Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Transactions on industrial informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.

[73] J. Hopkins and P. Hawking, "Big data analytics and iot in logistics: A case study," *The International Journal of Logistics Management*, vol. 29, no. 2, pp. 575–591, 2018.

[74] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[75] J. Su *et al.*, "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring," *IEEE transactions on industrial informatics*, vol. 17, no. 3, pp. 2242–2249, 2020.

[76] Y. Zhu, J. Song, and F. Dong, "Applications of wireless sensor network in the agriculture environment monitoring," *Procedia Engineering*, vol. 16, pp. 608–614, 2011.

[77] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Computers and electronics in agriculture*, vol. 153, pp. 69–81, 2018.

[78] P. Boniecki, H. Piekarska-Boniecka, K. Świerczyński, K. Koszela, M. Zaborowicz, and J. Przybył, "Detection of the granary weevil based on x-ray images of damaged wheat kernels," *Journal of Stored Products Research*, vol. 56, pp. 38–42, 2014.

[79] M. Dworkin, *Sha-3 standard: Permutation-based hash and extendable-output functions*, en, 2015. DOI: https://doi.org/10.6028/NIST.FIPS.202.

[80] J. Camenisch, M. Drijvers, T. Gagliardoni, A. Lehmann, and G. Neven, "The wonderful world of global random oracles," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Springer, 2018, pp. 280–312.

[81] S. Dziembowski, L. Eckey, and S. Faust, "Fairswap: How to fairly exchange digital goods," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 967–984.

[82] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Conference on the theory and application of cryptographic techniques*, Springer, 1987, pp. 369–378.

[83] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, no. 2014, pp. 1–32, 2014.

[84] J. P. Bharadiya, "Artificial intelligence and the future of web 3.0: Opportunities and challenges ahead," *American Journal of Computer Science and Technology*, vol. 6, no. 2, pp. 91–96, 2023.

[85] 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.

[86] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," in *Proceedings of the 33rd international conference on Very large data bases*, 2007, pp. 315–326.

[87] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[88] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[89] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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

[91] S. Yang, F. Wu, S. Tang, X. Gao, B. Yang, and G. Chen, "On designing data quality-aware truth estimation and surplus sharing method for mobile crowdsensing," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 4, pp. 832–847, 2017.

[92] S. Gollapudi, K. Kollias, D. Panigrahi, and V. Pliatsika, "Profit sharing and efficiency in utility games," in *25th Annual European Symposium on Algorithms (ESA 2017)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[93] X. Zhang, C. Chen, Y. Xie, X. Chen, J. Zhang, and Y. Xiang, "A survey on privacy inference attacks and defenses in cloud-based deep neural network," *Computer Standards & Interfaces*, vol. 83, p. 103 672, 2023.

[94] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[95] Y. Liu *et al.*, "Trojaning attack on neural networks," in *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*, Internet Soc, 2018.

[96] 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*, 2020, pp. 1–9.

[97] D. C. Nguyen *et al.*, "Federated learning meets blockchain in edge computing: Opportunities and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 806–12 825, 2021.

[98] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," *Proceedings of machine learning and systems*, vol. 1, pp. 374–388, 2019.

[99] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–44, 2023.

[100] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[101] N. Alghanmi, R. Alotaibi, and S. M. Buhari, "Hlmcc: A hybrid learning anomaly detection model for unlabeled data in internet of things," *IEEE Access*, vol. 7, pp. 179 492–179 504, 2019.

[102] X. Lin *et al.*, "Federated learning with positive and unlabeled data," in *International Conference on Machine Learning*, PMLR, 2022, pp. 13 344–13 355.

[103] S. Si, J. Hall, R. Suddaby, D. Ahlstrom, and J. Wei, *Technology, entrepreneurship, innovation and social change in digital economics*, 2023.