Designing Incentive Mechanisms for Fair Participation in Federated Learning

1st Han Xu*, 2nd Priyadarsi Nanda† and 3rd Jie Liang‡
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
*0000-0002-3580-9403, †0000-0002-5748-155X and ‡0000-0001-7179-5208

***Abstract*—Federated learning offers a collaborative method for training machine learning models using distributed data sources. Given its reliance on diverse contributions from multiple participants, equitable representation and rewards are paramount to its success. Upholding fairness is crucial to encourage active participation in this collaborative process. This paper presents a comprehensive analysis of existing mechanisms designed to incentivise fair engagement in federated learning, structured around a taxonomy aligned with the evolving dynamics of the federated learning sources. The study offers insights into cultivating environments that prioritise fairness and broad participation while suggesting avenues for future research.**

***Index Terms*—Fairness, Federated learning, Incentive mechanism, Data valuation.**

I. Introduction

The widespread adoption of machine learning technology has made big data an invaluable asset in the information age. However, the existence of data silos often hinders the full utilisation of this resource. Collaborative approaches have emerged where multiple participants combine efforts to train machine learning models. While such collaboration offers significant benefits, centralising and merging data on a single server can compromise privacy and participants' rights.

Federated learning, heralded as a paradigm shift, enables collaborative model training without sharing raw data. Instead, it involves exchanging model updates to refine a global model iteratively [1]. The procedure of federated learning can be broken down into several steps, as shown in Fig. 1.

1) Objective Setting: Determine the machine learning problem to be solved, and identify the type of model to be co-trained.
2) Model Distribution: The initialised model parameters are distributed to all participating parties.
3) Local Training: Each participant trains the model using local data without sharing original data.
4) Model Aggregation: Participants send the updated model parameters to a central server for aggregation.
5) Model Update: The central server updates the model based on the received parameters and distributes the updated model parameters to all participants.
6) Iterative Training: Repeat the process of local training, model aggregation, and model update until a predefined stopping condition is met.

This structure allows participants to collaboratively improve a model without revealing their raw data. This methodology has led to significant advancements in handling sensitive data, finding applications in sectors like healthcare, finance, and the Internet of Things (IoT).

However, ensuring fairness among participants becomes paramount as federated learning gains traction. 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 [2]. The imbalance in contributions could deter participation and undermine the efficiency and viability of the federated learning ecosystem [3].

In this context, we note two distinct categories of participants: the dominant, who have abundant resources and typically favour a *contribution fairness* approach, and the vulnerable, who, having fewer resources, stress *equilibrium fairness*. Addressing the concerns and aspirations of both these groups is a formidable challenge in promoting federated fairness.

We embark on a thorough exploration of incentive and distribution mechanisms from the inception of federated learning. Within the *contribution fairness* framework, we discuss *marginal contribution*, *resource allocation*, and *reputation mechanisms*, shedding light on their influence on participant behaviour. Simultaneously, under the *equilibrium fairness* paradigm, we probe *fair participant selection* methods, *weight redistribution* strategies, and *personalization* approaches. Our overarching goal is to offer insights that foster a collaborative, fairness-centric federated learning environment.

Our contributions to this article include the following:

1) *Introducing a novel taxonomy for federated learning fairness.*
2) *Offering a synthesis of cutting-edge methodologies based on our taxonomy, underscoring their strengths and limitations.*
3) *Delving into unresolved challenges within federated fairness, suggesting avenues for future research.*

The structure of the remainder of this paper unfolds as follows: In Section II, we delve into the underlying motivations propelling the evolution of federated learning and introduce a refined taxonomy of federated learning fairness. In Sections III and IV, we systematically elucidate the state-of-the-art

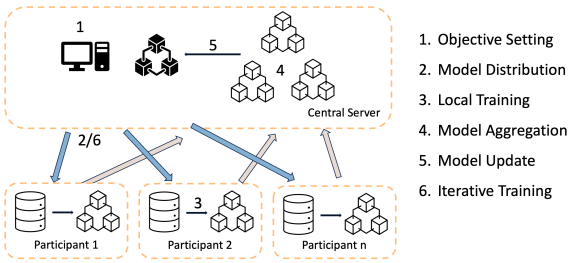

Fig. 1. Federated learning structure

advancements within the domains of *contribution fairness* and *equilibrium fairness*, respectively. These sections also critically assess the merits and pitfalls of extant literature in these realms. Section V encapsulates contemporary research's salient points, highlighting its commendable achievements and areas demanding further exploration while setting the stage for prospective trajectories in the quest for fairness within federated learning. Finally, Section VI offers a culminating reflection on the discourse presented, drawing together the paper's core insights and contributions.

II. Proposed Taxonomy for Federated Learning Fairness

In the introductory section, we highlighted the pivotal role of fair collaboration in federated learning. Grounded in this perspective and line with guiding principles outlined in the literature [2], we identify three primary motivations that drive collaboration in federated learning:

1) **Sustained Development and Participant Enthusiasm:** Sustainability is critical to the successful progression of federated learning. Ensuring active and consistent participation is crucial for the evolution and effectiveness of the learning process. The continuity in the willingness of participants to contribute resources is foundational to sustained model training.
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 better model performance to access to cutting-edge predictive models. Recognising and offering clear advantages to these participants can foster more active and consistent engagement.
3) **Ethical Considerations:** With federated learning pooling resources from many contributors, it becomes imperative to ensure nondiscrimination. Equity and social ethics dictate that all participants perceive the learning process as fair and that no resultant model unduly disadvantages any participant.

Two discernible perspectives emerge from these motivations: Vulnerable participants emphasise the importance of a global model that offers equivalent utility to all, adhering to the principle known as *equilibrium fairness*. In contrast, dominant participants focus on benefits commensurate with their contributions, advocating for what is termed *contribution fairness*.

A harmonious balance between technical viability, ethical standards, and participant preferences is essential to crafting effective incentives for federated learning. By accounting for these factors, we can promote an environment ripe for active collaboration in federated learning, formulating more precise and resilient models.

This discussion lays the foundation for our proposed taxonomy of fairness in federated learning collaborations, depicted in Fig. 2. Subsequent sections will delve into pertinent research aligned with this taxonomy, spotlighting embraced fairness concepts and their associated methodologies and constraints.

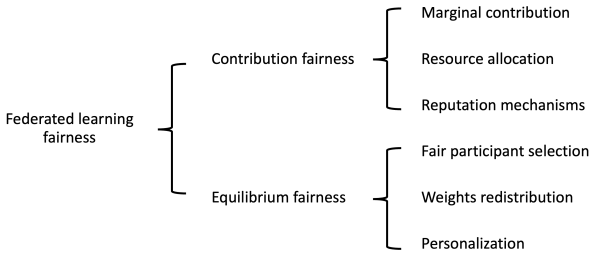

Fig. 2. Proposed Taxonomy for Federated Learning Fairness

III. 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 [4], 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*.

A. 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 [5] and have been extensively employed to evaluate each participant's contribution to a game.

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

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. [6]. 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 [7]. 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. [8] and a multi-dimensional contribution method based on stepwise computation by Nishio et al. [9].

B. 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 [10]. 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 [11]. 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 [12]. 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 [13]. 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 [14]. 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 [15]. 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 [16]. 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.

C. 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 [17]. 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 [18]. 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 [19]. 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 [20]. 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.

IV. 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.

A. 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 [21]. 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 [22].

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 [23]. 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) [24]. 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 [25]. 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 [26]. 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.

B. 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 [27]. 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 [28]. 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 [29]. 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 [30]. 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 [31].

Meanwhile, Zhao et al. presented an alternative to the q-FFL's loss amplification mechanism by proposing a direct weight redistribution methodology [32]. 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 [33].

However, all of the approaches above assume that participants are honest. If participants 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 [34].

C. 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 [35].

Data-based personalisation approaches are geared towards bridging the discrepancies in the distribution of client data, which often exhibit statistical heterogeneity in a federated learning context. In federated learning setups, it is frequently necessary to consider data-sharing strategies or acquire virtual datasets that comprehensively represent the overall data distribution. For instance, Zhao et al. introduced a data-sharing strategy involving the equitable allocation of a small portion of global data to individual clients based on categorical balance [36]. These studies' empirical findings illustrate that model accuracy can be substantially enhanced with minimal addition of data. On the contrary, Jeong et al. devised a federated augmentation technique (FAug) wherein generative adversarial network (GAN) models are trained within a federated learning server [37]. This method involves uploading data samples from minority groups to the server to train the GAN model. Subsequently, this trained GAN model is disseminated to each client to generate additional data, thus enriching their local datasets to create a uniformly distributed dataset.

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 FL 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 [38]. 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 FL framework that employs Transfer Learning (TL) and Knowledge Distillation (KD), allowing clients to develop autonomous models utilising their private data [39]. 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.

V. Critique and Future Directions

As shown in Fig. 3, on the historical development of the related technology, maintaining fairness has always been a pivotal issue in federated learning since the introduction of the federated learning framework. Contribution fairness has been widely applied in contribution evaluation and incentive mechanism research, while equilibrium fairness is extensively used in model optimisation and client selection. These new explorations propel the advancement of fair federated learning towards diversification of solutions and comprehensive scenario development.

Although there have been some exploratory studies in this domain, challenges persist: With the enlargement of the federated learning system scale, efficiently managing fairness amongst many participants (each possibly possessing unique data distributions) becomes arduous. As participants increase, the complexity of tracking and rewarding contributions multiplies. Real-time resource allocation and contribution tracking demand extensive computational resources and communication overhead, diverting attention from the primary task of model training. Ensuring data privacy while tracking contributions requires a delicate balance. Also, some participants might be incentivised to misreport or manipulate their local updates to skew the global model, especially when weights are reallocated based on contributions. Identifying and reducing such dishonest behaviours is crucial but presents its own set of challenges. Defining "fairness" can be subjective; although there are measures like the Gini coefficient and Jain's fairness index, there is no universally accepted standard for measuring fairness in federated learning. Different applications and environments may demand distinct fairness metrics, challenging establishing a universally accepted fairness criterion.

For future research, we suggest exploring the following crucial domains:

A. Providing a comprehensive definition for fairness federated learning

Different fairness metrics often lead to divergent evaluation outcomes when juxtaposing various federated learning frameworks for fairness. It suggests that fairness in federated learning is relative, depending on the model, data, and task requirements. A holistic, multi-dimensional algorithm fairness assessment and evaluation system is essential at this juncture. Such a system is necessary to effectively quantify inherent fairness risks in federated learning, thus hindering the guarantee of fairness in existing federated learning models. Therefore, a thorough definition of fairness needs further exploration. When crafting this definition, intertwining legal, regulatory, and social fairness principles is crucial to avoid narrowly focused technical solutions.

B. Encompassing more realistic application scenarios

An underlying assumption of ongoing research is that the model owner monopolises the formation and implementation of the incentive mechanism, and other participants can only choose to join or abstain. This phenomenon originates from traditional data collaboration models where the model owner or initiator, often a large corporation or organisation, has a technological and resource advantage. On the other hand, participants tend to be smaller entities or individual data

Fig. 3. Literature timeline

providers. Within such a structure, due to their technological and resource advantages, model owners usually dominate decision-making processes throughout federated learning, including the distribution of rewards and incentives. Still, this assumption needs revisiting in an open and competitive market, as it hinders the progression of federated learning. As federated learning continues to evolve, there is an urgent need to probe further and understand fair incentive mechanisms. In our recent work [40], we introduced a novel federated learning framework emphasising the pivotal role of data owners. This foundational effort serves as a precursor to our future direction—aiming to solidify data owners' standing within incentive frameworks and stimulate competition among multiple model owners. We believe that such endeavours can significantly propel the development of more balanced incentive mechanisms in federated learning.

C. Exploring interpretable fairness within the federated learning

Explainability has proven effective in addressing biases in machine learning and can enhance fairness in federated learning. A lack of accurate comprehension of fairness might deter participants. Setting explainable goals for joint fairness aims to elucidate how decisions impact each participant's interests. Existing explainability research mainly focuses on metrics such as accuracy assessments based on test sets, information-theoretic information gains, model similarities, and statistical properties measuring contribution value. However, every metric has pros and cons, depending on tailor-made joint solutions and value assessment constraints.

Moreover, quantifying contributions from rare data providers or ethically vulnerable participants is challenging. Complete transparency in fair explanations might reveal every participant's contributions, characteristics, and data patterns. For instance, if a participant has significantly contributed to a model in a specific scenario, this unique contribution could unveil that participant's distinct data attributes or behavioural patterns. Sometimes, this information could relate to an individual's sensitive traits or behaviours, potentially exposing their identity or other private details. Furthermore, if participants are competitors, complete transparency could enable some to exploit this information for an unfair competitive advantage. Therefore, achieving a balance in these conditions presents an intriguing research direction.

D. Ensuring robustness in fairness within federated learning

Current methods often presuppose a level of trust among participants in federated learning. However, in the real world, there may be participants with varying levels of malicious intent, some negligible and others more pronounced. These participants might manipulate gradients and transmit information to federated learning servers, severely compromising model performance. While many strategies currently address privacy attacks, there is a need for greater focus on data privacy protection, with inadequate attention given to its implications for fairness. Differential privacy techniques and secure multi-party computation might produce more effective fairness-centric learning algorithms. Integrating the latest cryptographic breakthroughs with existing fairness mechanisms represents a promising avenue for future research.

E. Adapting fairness models to dynamic scenario fluctuations

Existing definitions and metrics for fairness are primarily static. Current research prioritises fairness in federated learning without considering feedback or recent effects, overlooking these decisions' future implications for fairness. Recent studies suggest a temporary misalignment between contributions and returns. To achieve fairness, the waiting time participants endure for future returns from the final model needs consideration. Additionally, the sequence and hierarchical differences in participant involvement should be factored in. Forthcoming research must develop definitions of fairness and algorithms that incorporate the dynamism, rich feedback, and persistent resonances of decision systems. Fairness in federated learning is a dynamic game in which current research needs to be improved. Using adversarial settings and advancements in fairness enhancement to design dynamic fairness detection mechanisms presents a promising research path.

VI. Conclusion

As federated learning technology evolves, fairness stands out as a pivotal concern. The academic world has delved into numerous strategies to address this, but substantial challenges persist. From this paper's insights, we'll first design an evolution matrix for fairness comparisons and then address dynamic

fairness challenges in our next research phase. The complexity of fairness grows with the expansion and diversification of systems, and defining it requires navigating diverse fields like technology, law, and ethics. Future research should encompass realistic scenarios, explainable fairness, and adaptability. A collaborative approach among researchers, practitioners, and policymakers is essential for genuine fairness in federated learning. Despite existing challenges, federated learning offers vast opportunities, especially when prioritising fairness, ensuring a heightened standard of data collaboration in society.

REFERENCES

[1] 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.

[2] 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.

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

[4] L. K. Wang Y, Li GL, "Contribution evaluation for federated learning: A survey," *Journal of Software(in Chinese)*, vol. 34, no. 3, pp. 0–0, 2022.

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

[6] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the shapley value," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1167–1176.

[7] 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.

[8] 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.

[9] 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.

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

[11] L. Lyu, J. Yu, K. Nandakumar, Y. Li, X. Ma, J. Jin, H. Yu, and K. S. Ng, "Towards fair and privacy-preserving federated deep models," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 11, pp. 2524–2541, 2020.

[12] 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.

[13] 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.

[14] T. H. T. Le, N. H. Tran, Y. K. Tun, M. N. Nguyen, S. R. Pandey, Z. Han, and C. S. Hong, "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.

[15] 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.

[16] Y. Deng, F. Lyu, J. Ren, Y.-C. Chen, P. Yang, Y. Zhou, and Y. Zhang, "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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

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

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

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

[40] 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. 103521, 2023.