Bias in Federated Learning: Factors, Effects, Mitigations, and Open Issues

Mourad Benmalek, Abdessamed Seddiki

IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol., No., Month, Year, pp. **-**
Journal homepage: http://iieta.org/journals/isi

Bias in Federated Learning: Factors, Effects, Mitigations, and Open Issues

Mourad Benmalek[1*], Abdessamed Seddiki[2]

[1] Computer Engineering Department, College of Engineering, Al Yamamah University, Riyadh, 11512, Saudi Arabia
[2] Ecole nationale Supérieure d'Informatique, BP 68M, Oued-Smar, Algiers 16309, Algeria

Corresponding Author Email: m_benmalek@yu.edu.sa

https://doi.org/10.18280/isi.xxxxxx

Received: 21 September 2024
Revised: 19 October 2024
Accepted: 9 December 2024
Available online:

Keywords:
artificial intelligence, machine learning, federated learning, bias, fairness

ABSTRACT

Federated learning (FL) enables collaborative model training from decentralized data while preserving privacy. However, biases manifest due to sample selection, population drift, locally biased data, societal issues, algorithmic assumptions, and representation choices. These biases accumulate in FL models, causing unfairness. Tailored detection and mitigation methods are needed. This paper analyzes sources of bias unique to FL, their effects, and specialized mitigation strategies like robust aggregation, cryptographic protocols, and algorithmic debiasing. We categorize techniques and discuss open challenges around miscoordination, privacy constraints, decentralized evaluation, data poisoning attacks, systems heterogeneity, incentive misalignments, personalization tradeoffs, emerging governance needs, and participation. As FL expands into critical domains, ensuring equitable access without ingrained biases is imperative. This study provides a conceptual foundation for future research on developing accurate, robust and fair FL through tailored technical solutions and participatory approaches attuned to the decentralized environment. It aims to motivate further work toward trustworthy and inclusive FL.

1. INTRODUCTION

Machine Learning (ML) has become pervasive across domains, powering services from image recognition to personalized recommendations. The success of ML critically depends on access to massive datasets that fuel model development. However, aggregation of large centralized datasets poses significant privacy concerns and security risks [1]. Federated learning (FL) is a distributed collaborative learning paradigm introduced to address this limitation.

In FL, data remains decentralized on the client devices like mobile phones or hospital servers. A shared global model is trained collectively without direct access to raw private data [2]. The model training process involves multiple rounds of communication between the clients and a central aggregation server [3]. In each round, the server first broadcasts the current state of the global model to a subset of available clients. Each client then performs local computation on this model using their private local dataset to generate model parameter updates. Only these update vectors are shared with the server which are aggregated to improve the global model. After several rounds of this federated training loop, the model converges to an optimal solution trained on the collective data [4, 5].

FL provides multiple advantages compared to ML:

- Enhanced privacy and confidentiality as raw data remains decentralized.
- Reduced security risks from single point of failure.
- Compliance with regulations on control of user data.
- Mitigates data silos problem across different entities.
- Ability to leverage scattered, non-Independent and Identically Distributed (non-IID) data located on endpoints.

This has enabled deployment of FL across diverse applications like next word prediction on smartphones [6], disease detection in healthcare [7], fraud detection in finance [8] and more. Leading technology companies like Google, Apple, Meta, Microsoft, Uber and IBM have incorporated FL into products and services [2]. The decentralized nature of FL also makes it suitable for emerging paradigms like edge computing which push intelligence to the network edge [9].

However, FL introduces statistical and systems challenges compared to centralized ML [5]:

- Non-IID, unbalanced and sparse local client data.
- Systems heterogeneity in compute capabilities of clients.
- Threat of inference attacks and privacy leaks.
- Communication overhead and convergence issues.
- Misaligned incentives between competitive clients.

These unique constraints impact the collaborative learning and can introduce biases that lead to discrimination and unfairness issues [10]. Recent studies have empirically shown that ignoring the heterogeneous decentralized distributions in FL can lead to suboptimal accuracy and algorithmic harms against certain subgroups [11, 12].

There are growing societal concerns around potential biases encoded in AI systems [13]. Bias refers to any systematic error which can lead to unfairness, discrimination and suboptimal model performance on certain subgroups [14]. Biases manifest

due to historical prejudices, representation imbalance a systemic inequity ingrained in the data get propagated into the algorithms, leading to discriminatory predictions and decisions [15]. Some sources of bias specific to FL include:

- **Sample selection bias**: The local datasets available to different FL clients may not represent the true population distribution. This can introduce sampling bias against minorities who are underrepresented. Location biases are common as the data may overrepresent certain geographic regions based on app usage.
- **Population distribution shift**: The demographics and data distributions often drift over time and vary across clients. Models trained on past biased data may not generalize fairly to new user populations [16].
- **Biased local data**: The local data itself can contain systemic biases and discrimination against protected groups which get encoded into the models [11]. For instance, differential judgement and labeling of minority groups.
- **Algorithmic biases**: The objective functions and model architectures may implicitly embed assumptions that disadvantage certain subgroups. Optimizing accuracy often comes at the cost of fairness [17].
- **Systemic societal biases**: Long-standing structural inequalities along gender, racial and socioeconomic divides manifest as background statistical biases [14].

These biases accumulate during training and get propagated to the global model. Ignoring fairness can exacerbate harms against already marginalized communities. This underscores the critical need for tackling biases in FL to prevent exclusions and ensure equitable access to the technology benefits. However, mitigating biases is significantly more complex in federated settings compared to centralized ML due to constraints around visibility into local client distributions, coordination between distrusting entities, systems heterogeneity, and misaligned incentives [2]. Care must be taken to balance accuracy, fairness and privacy tradeoffs given the decentralized nature of FL [14]. Therefore, tailored solutions are required that account for the unique challenges.

In this paper, we provide a comprehensive analysis of the various sources of bias that can manifest in FL systems and examine mitigation strategies tailored to the decentralized environment. We categorize techniques based on principles such as robust aggregation algorithms, cryptographic protocols, algorithmic and data debiasing, personalized modeling, and participative evaluation. We also discuss key open challenges around miscoordination, privacy constraints, security, systems heterogeneity, incentives, governance, and participation that remain to be addressed to realize trustworthy and fair FL.

The key contributions of this work are:

- A taxonomy of bias sources unique to FL including sample selection bias, population drift, locally biased data, societal biases, algorithmic biases, and representational biases.
- Categorization of bias mitigation techniques tailored for FL based on principles like robust aggregation algorithms, cryptographic protocols, data and algorithmic debiasing, personalized modeling, and participative evaluation.
- Discussion of open research challenges around miscoordination, privacy, systems heterogeneity, incentives, governance, and participation that remain to be tackled.

The rest of this paper is organized as follows: Section 2 provides background on FL, contrasting it with centralized ML and discussing architectures, algorithms, applications, and challenges. Section 3 analyzes sources of bias in FL and their effects. Section 4 examines bias mitigation techniques in FL, categorizing them into data-based, algorithmic, and hybrid approaches. Section 5 examines key open research issues and future directions around developing fair and accountable FL systems. Finally, Section 6 concludes with a summary and outlook on bias mitigation for trustworthy FL.

2. BACKGROUND ON FEDERATED LEARNING

As mentioned above, FL enables training ML models collaboratively without directly sharing private raw data from participants. It involves coordinating decentralized clients like mobile devices or hospitals to build shared models while keeping data localized [2]. In this section, we provide background on FL and contrast it with centralized approaches. We also discuss architectures, algorithms, applications, and key challenges.

2.1 Centralized vs federated learning

In traditional centralized ML, data from various sources is aggregated to a single centralized server or data warehouse for model training [2]. This allows applying powerful computational resources and statistical techniques optimized for independent and identically distributed data [18]. However, centralizing raw data from diverse sources raises significant privacy and security concerns. Sharing personal data like healthcare records or financial information to a centralized pool also risks violating regulations and user trust [19].

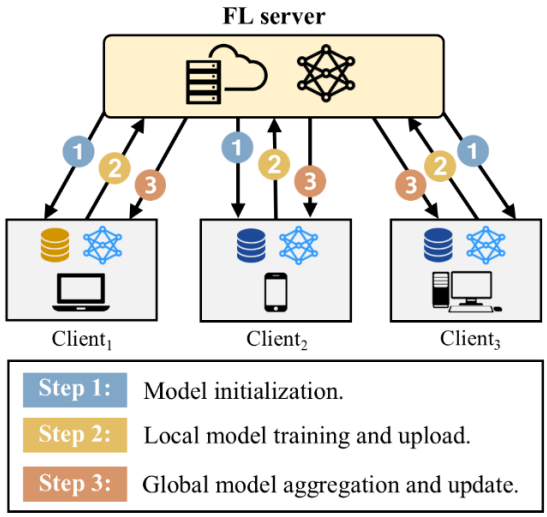

Figure 1. Federated Learning Architecture [5]

With increasing focus on data privacy and confidentiality, FL has emerged as an alternative distributed approach that enables collaborative training of ML models without directly pooling private raw data [2]. As depicted in Figure 1, FL coordinates many decentralized edge devices or organizations to build shared models, while keeping sensitive data localized on the devices. In FL, individual clients train models locally

using their own private data subsets. The clients could be hospitals, banks, smartphones, wearables, vehicles, etc. Rather than sharing this local private data, the clients transmit only model parameter updates to a central aggregation server [4, 20]. The server averages these updates from several clients to build an enhanced global model. This global model is then shared back with the clients for further improvement in the next round. Over several rounds of this federated training loop, the ML model converges to an optimal state, learning from the collective data at all the clients without directly accessing any raw private data. Differential privacy techniques may be used to anonymize the model updates [21]. Data remains decentralized throughout, enhancing privacy and compliance.

Table 1. Comparison of centralized ML and FL

Parameter	Centralized ML	FL
Data Storage	Centralized data warehouse	Decentralized on client devices
Data Privacy	Low, single point of failure	High, data remains localized
Data Variety	Typically, IID	Non-IID, unbalanced
Scalability	Limited by centralized compute	Highly scalable with distributed clients
Model Training	On centralized servers	Distributed across clients
Communication	Low overhead	High overhead for coordination
Incentives	Single objective	Potentially misaligned incentives
Personalization	Global model for all users	Scope for personalized models
Trust	High, single trusted entity	Varying trust across clients
Robustness	Vulnerable to data poisoning	Robust against single point failure
Regulations	Difficult to comply	Better compliance

As shown in Table 1, some key differences between centralized and FL include [2, 3]:

- FL enhances privacy as raw data stays localized on devices while centralized ML aggregates data to servers.
- Statistical assumptions differ, with centralized ML relying on abundant IID. data while FL handles decentralized non-IID. data.
- Communication overhead is higher in FL to coordinate clients and transmit model updates.
- There is a lack of coordination between competitive clients with misaligned incentives in FL.
- FL enables learning from data siloed across organizations by coordinating privately without direct data sharing.
- Hardware is homogeneous in centralized ML whereas FL must account for heterogeneous systems like smartphones, servers, IoT devices.
- Centralized learning converges faster with IID. data while FL can be slower and unstable due to systems heterogeneity and statistical variations.

While FL enhances privacy and decentralized participation, the heterogeneous and sparse data as well as systems diversity introduce new challenges compared to centralized settings [3]. Common issues include:

2.1.1 Statistical challenges

The decentralized data in FL often exhibits properties like non-IID distributions, unbalanced quantities, sparsity, and concept drift over time. Specific statistical issues include:

- **Non-IID data:** The data distributions across clients can vary significantly based on their location, demographics, usage patterns and other factors.
- **Unbalanced data:** The quantity of local data available at each client may differ, with some having relatively small datasets.
- **Sparse local data:** Each client typically has limited local data compared to the population level.
- **Concept drift:** The data distributions can shift dynamically over time rather than being static.

2.1.2 Systems challenges

There is typically significant diversity in the capabilities of client devices participating in FL. Heterogeneity in systems resources leads to challenges including:

- **Systems heterogeneity:** The capabilities like compute, storage, network capacity can vary greatly across different hardware clients like mobiles, edge devices, servers.
- **Limited communication**: Bandwidth constraints limiting coordination especially for remote clients.
- **Client availability:** Not all clients may always be online to participate in each training round.
- **Client reliability:** Some devices may drop out halfway due to technical glitches or lost connectivity.
- **Constraints on local computation:** Power or connectivity limitations restricting local model training.

2.2 Applications and benefits

The decentralized and privacy-preserving nature of FL has led to adoption across diverse domains, both in industry and academia [22-27]. As shown in Figure 2, real-world applications of FL include:

- **Next word prediction:** Google deployed FL in Gboard mobile keyboard to improve word suggestions without accessing typed data [28].
- **Fraud detection:** Banks can jointly build models to detect fraudulent transactions while keeping client data decentralized [8, 29].
- **Disease prediction:** Hospitals can collaboratively improve models for risk prognosis without sharing private health records [30, 31].
- **Intrusion detection:** Network providers can coordinate edge devices like routers and mobiles to identify threats without direct data access [32-34].
- **Personalized recommendation:** Retailers can collectively train models to provide individualized product recommendations which maintains user privacy [20, 35].
- **Traffic flow optimization:** Automotive companies can coordinate vehicles to predict congestion without tracking individual cars [36, 37].
- **User profiling:** Firms can derive insights from user preferences across applications to improve services while respecting privacy [38, 39].
- **Smart agriculture:** In precision agriculture, FL can be employed for crop yield prediction and pest control by aggregating data from various farms without disclosing specific farm details [40, 41].

Some key potential benefits of FL include [2]:

- Privacy and confidentiality as raw data remains decentralized rather than getting pooled. This provides strong safeguards for sensitive data like healthcare records and financial information.
- Compliance with regulations that limit sharing of raw private data across borders or economic sectors. Federated models rely only on aggregated model updates.
- Mitigates the problem of isolated data silos spread across organizations by enabling collaborative learning without direct data exchange. This unlocks previously trapped data.
- Robustness against single points of failure since compromising any one device does not reveal full raw data. Attacks require coordinated access to many federated nodes.
- Inclusive utilization of widely distributed data on heterogeneous nodes like mobiles and edge devices. Allows leveraging data where it already resides.
- Scalability due to distributed compute across nodes. Training tasks get divided across available clients.
- Personalization by allowing individual nodes to build models customized for their local population from on-device data.

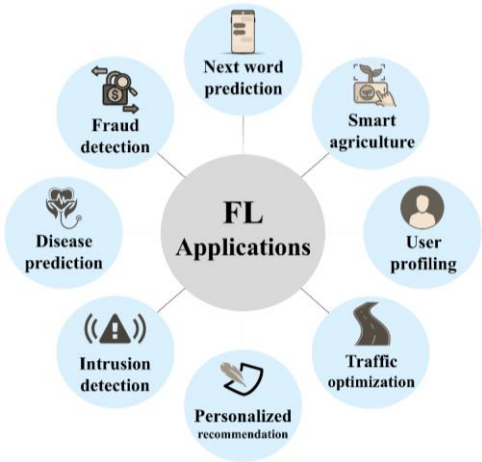

Figure 2. Applications of Federated Learning

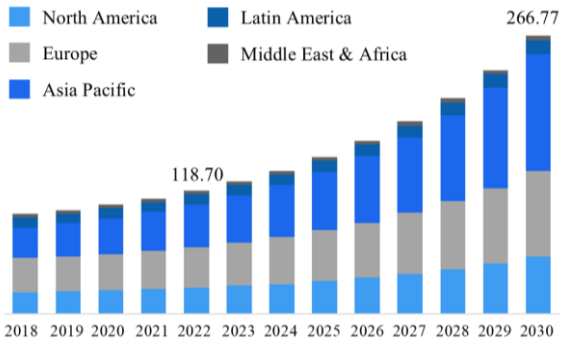

Figure 3. Growing Adoption of Federated Learning

The adoption of FL has rapidly increased over the past few years as illustrated in Figure 3. The global FL market size reached nearly $118.70 million in 2022. The market is projected to grow at a Compound Annual Growth Rate (CAGR) of 10.7% between 2023 and 2030 to reach a value of around $266.77 million by 2030, according to a new study by *Polaris Market Research* [42].

Technology companies at the forefront of deploying FL include Google, Apple, Samsung, Huawei, IBM, Microsoft, Intel, and Nvidia [3, 43-45]. Various open source frameworks have been developed like TensorFlow Federated and PySyft to support wider adoption [44, 46]. Academic research on FL is also accelerating with new innovations in algorithms, system design and applications [47-51]. However, there are still challenges and open problems related to systems heterogeneity, statistical variations, communication overhead, privacy, security, and incentive alignments that must be addressed to fully realize the potential benefits of FL [4, 47-51]. As solutions emerge to the unique constraints of decentralized orchestration, FL is poised to see massive growth as an enabler for collaborative intelligence while preserving confidentiality.

2.3 Architectures

Based on network topology, FL systems can be categorized into centralized and fully decentralized architectures [52]:

2.3.1 Centralized federated learning (Cross-device)

As shown in Figure 1, even though FL is typically considered as a decentralized approach, a centralized server is required to collect clients' model updates and aggregate them to the global model. However, in contrast to traditional centralized ML, the raw private data stays on device in FL. Only model updates are communicated with the server. For example, Google uses centralized FL in Gboard mobile keyboard to train next-word prediction models from user typing data, without directly accessing sensitive text. The global model is optimized by aggregating updates from millions of mobiles while keeping data localized on device.

2.3.2 Fully decentralized federated learning (Cross-silo)

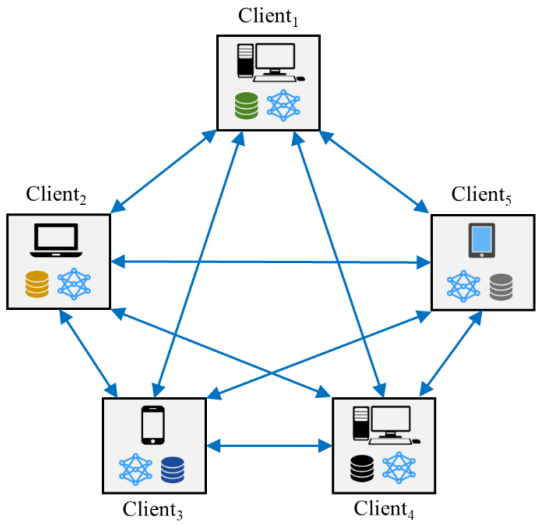

Figure 4. Fully Decentralized Federated Learning

As shown in Figure 4, fully decentralized FL eliminates the central aggregation server [53]. Clients communicate with each other in a Peer-to-Peer (P2P) manner to improve their local models. Key steps include [10]:

- Clients discover neighborhood peers based on proximity or other metrics.
- Model updates are exchanged over P2P links.
- Updates from peers are aggregated into local models using decentralized algorithms.

- Useful updates propagate across the network through transient links.
- Individual models converge to shared states through continual P2P exchanges.

Fully decentralized approaches have the advantage of not relying on any trusted central entity. However, they introduce challenges related to discovery, incentive alignment, and convergence guarantees [53]. Hybrid architectures that balance centralized and peer-based control may provide optimal solutions.

2.4 Aggregation algorithms

A variety of aggregation algorithms have been developed to enable robust and efficient FL that accounts for statistical and systems heterogeneity, while preserving privacy. These aggregation algorithms customize the training process for decentralized environments. Key algorithms adapted for federated settings include:

2.4.1 Federated averaging (FedAvg)

This aggregates client updates on the central FL server by taking a weighted average. The weight assigned to each client is determined based on factors like the size and quality of its local dataset. This gives higher priority to updates from clients with more representative data. FedAvg is employed in a big number of solutions like Google's federated keyboard predictions [54].

2.4.2 Federated stochastic gradient descent (FedSGD)

This applies distributed stochastic gradient descent, where gradients are computed locally on each client using their data and then averaged to optimize the global model. FedSGD is well-suited for non-IID data distributions prevalent in federated settings. It has been applied in domains like patient monitoring [31, 55].

2.4.3 FedProx

To address the challenges of heterogeneity in FL environments, Li et al. [55] proposed FedProx. As stated by the authors, "*FedProx algorithm can be viewed as a generalization and re-parametrization of FedAvg*". They proposed to add a proximal term to the local subproblem that helps to effectively limit the impact of variable local updates, and thus improve the stability of the method. Moreover, they proved that FedProx achieves better convergence and stability compared to FedAvg in heterogeneous FL environments.

2.4.4 Secure aggregation

This employs cryptographic protocols like differential privacy [56], multi-party computation [57] and homomorphic encryption [58, 59] during model aggregation to preserve privacy of the updates. This prevents inference attacks while aggregating updates from untrusted clients.

2.5 Privacy and incentive considerations

Although raw private data remains decentralized in FL, additional precautions are necessary to prevent inference attacks and preserve privacy [2, 60]. Participants may try to reconstruct sensitive attributes about data at other clients from the model updates. Common privacy risks include:

- **Membership inference:** Determining if a sample was part of a client's training set.
- **Attribute inference:** Predicting sensitive attributes like illness status.
- **Model inversion:** Reconstructing parts of the training data.
- **Generative modelling:** Synthesizing realistic proxy data.

Differential privacy techniques can be applied to perturb model updates before sharing to minimize risks of sensitive leakage [56, 61]. Noise is carefully calibrated and added to updates to prevent precise reconstruction while preserving utility. Secure multiparty computation protocols like homomorphic encryption and secret sharing can also enhance privacy during aggregation [57-59]. Moreover, access control mechanisms restricting visibility of updates from other participants based on trust can also improve privacy [62]. For example, a hospital may only share updates within consortiums of trusted healthcare institutions rather than with all clients. Fine-grained access policies, data sandboxing and hardware-based trusted execution environments are also being explored [63].

Furthermore, there is also a lack of coordination between clients who may be competing entities or have misaligned incentives, unlike the centralized setting. Individual users and organizations may act strategically to try to influence the model towards their local objectives rather than global accuracy [64, 65]. For example, a client may selectively contribute only biased updates that exclude certain demographics. Malicious clients can launch data poisoning attacks by submitting intentionally corrupted updates to compromise model integrity and performance [66]. Carefully addressing these emerging considerations around adversarial threats, incentive misalignments, and mechanisms for privacy-preserving coordination between untrusting participants remains an active research challenge for FL [66].

3. BIAS IN FEDERATED LEARNING

As mentioned in Section 1, bias refers to any systematic error in the data or algorithms that can lead to unfairness, discrimination, or suboptimal performance on certain subgroups [14, 67]. Biases can manifest in FL systems due to historical prejudices ingrained in the data, representation imbalance across decentralized datasets, systemic inequities encoded in the algorithms, as well as feedback loops that exacerbate minor statistical variations [68]. As shown in Figure 5, sources of bias that can arise in FL include sample selection biases, population distribution shift, biased local data, systemic societal biases propagated through data, algorithmic biases from objectives and assumptions, as well as social biases reflecting cultural prejudices [13]. These biases accumulate during the federated training process and get imprinted into the global model, leading to issues around fairness, accountability and exclusion of underrepresented groups. Therefore, technical solutions tailored to the unique characteristics of FL are necessary to mitigate biases and ensure equitable access to the benefits of this technology. In the following subsections, we analyze the sources of biases that can manifest in FL systems, how they may impact the models, and what are the potential mitigations. Table 2 provides a comparative summary of the different categories of bias that can manifest in FL, along with their associated factors, detrimental effects, and potential mitigation strategies.

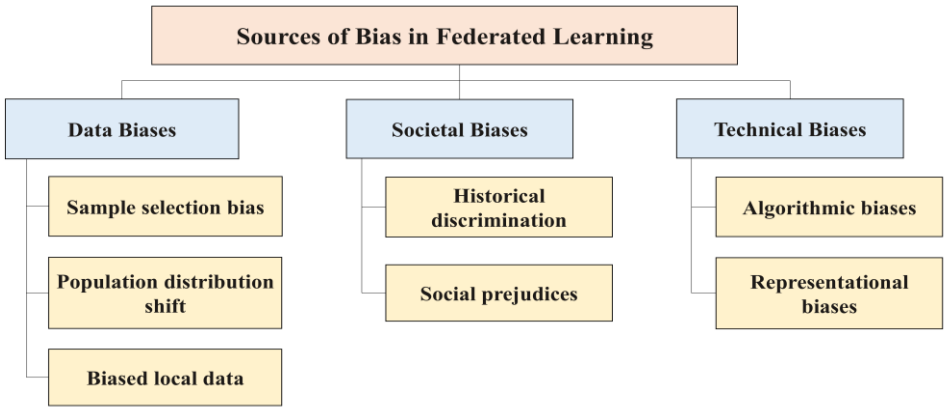

Figure 5. Taxonomy of Biases in Federated Learning

Table 2. Summary of Biases in Federated Learning

Bias Type	Factors	Effects	Potential Mitigations
Sample Selection Bias	Device and connectivity biases; user demographic biases; adversarial manipulation.	Model unfairness; bias amplification; adversarial manipulation; lack of visibility into representativeness.	Careful client selection; robust aggregation; statistical bias detection; differential privacy.
Population Distribution Shift	User demographic shifts; user behavior shifts; app/device version shifts; adversarial drift injection.	Overall performance degradation; subgroup performance issues.	Local drift detection; personalized FL; robust optimization; continuous analytics; synthetic data; hybrid cloud FL; adversarial adaptation.
Biased Local Data	Biased data collection; measurement biases; omitted variables; proxy encoding.	Data poisoning attacks; model evasion; difficulty auditing.	Client-side auditing; local debiasing; secure aggregation for bias; reputation systems; incentives.
Systemic Biases in Data	Historical discrimination; measurement biases; proxy discrimination; anchor biases; social stereotypes; unexamined assumptions.	Underrepresentation; measurement & feature bias; objective bias; miscalibration.	Decentralized bias auditing; data augmentation; re-weighting; debiasing algorithms; inclusive data collection; regulations.
Social Biases	Historical discrimination; representation imbalances; social stereotypes; anchor biases; implicit associations; feedback loops.	Underrepresentation; feature biases; poor generalization; stereotyping; exclusion; denial of opportunities; abusive targeting; loss of autonomy.	Diverse clients; bias auditing; data augmentation; nutrition labels; algorithmic fairness; subgroup validation.
Algorithmic Biases	Biased objectives; regularization; model assumptions; complex models; aggregated algorithmic biases.	Suboptimal performance; historical bias perpetuation; opportunity denial; privacy violations; deflected accountability.	Subgroup validation; fairness regularization; flexible model selection; multimodal modeling; counterfactual evaluation; modular transparency.
Representational Biases	Biased data formatting; problem framing; global model design; evaluation metrics.	Poor personalization; limited accessibility; skewed optimization; privacy violations; entrenched biases; lack of recourse.	Inclusive design; personalized models; subgroup validation; bias metrics, nutrition labels.

3.1 Sample selection bias

Sample selection bias can arise in FL due to differences in participation and selection of clients for training rounds [69]. The decentralized datasets in FL are determined by which users, devices or organizations participate as clients in the model training process. However, the client population may not accurately represent the true underlying population distribution [70]. Certain subgroups may end up being overrepresented or underrepresented among the clients based on factors like geographic location, demographics, device types, etc. [69]. For example, patients from larger hospitals may dominate in healthcare FL while smaller clinics are underrepresented [7]. This can lead to biased, non-representative data distributions among the decentralized client datasets. Minority demographic groups and less dominant data patterns may get excluded or underrepresented. Overrepresented groups will have an outsized influence on the model compared to underrepresented groups. As a consequence, the global model may not sufficiently capture diverse perspectives and vulnerabilities, potentially resulting in discrimination against certain minorities excluded from the training process. Furthermore, the model may also not generalize well to underrepresented segments of the population [71]. Without corrective techniques, this sample selection bias can get imprinted into the global model during federated training [72].

For instance, in Google's implementation of FL for Gboard, the mobile keyboard app, sample selection bias was observed due to uneven participation of users. Users with high-end devices and reliable internet connections were more likely to participate in training rounds, leading to overrepresentation of certain demographic groups. This resulted in the language model performing better for these groups while underperforming for underrepresented demographics. The bias manifested as less accurate next-word predictions for users from underrepresented groups, affecting user experience and potentially widening the digital divide.

3.1.1 Factors

There are several factors that can skew the sampling of clients in FL:

- **Device and Connectivity Biases:** Clients participate voluntarily based on device capabilities and connectivity. Lower-resourced devices may drop out due to power or bandwidth limitations [3]. Regions with poor connectivity provide limited client samples. This leads to biased geographic and demographic representations.
- **User Demographic Biases:** Due to variability in technology adoption, the decentralized samples tend to overrepresent certain population segments more proficient with technology, while underrepresenting minorities [21]. Age, income, education and cultural gaps can skew the client population.
- **Adversarial Manipulation:** Malicious entities can deliberately poison or manipulate client sampling through targeted device infections or Sybil attacks simulating fake clients [5]. This allows injecting biases by influencing which devices participate in FL.

3.1.2 Effects

These sampling biases lead to some groups being overrepresented while minorities are underrepresented among the clients. This has several detrimental effects [2]:

- Model fairness can degrade for underrepresented or excluded groups. Predictions tend to favor majority demographics.
- Biases against minorities and protected groups are amplified without diversity in sampling.
- Groups not represented in the client sample can be negatively affected by model decisions.
- There is a lack of visibility into client sample representativeness at the central server.

3.1.3 Potential mitigations

Addressing sample selection bias remains an active area of research in FL. A range of techniques have been proposed to detect, limit and correct for sampling biases in the decentralized client population:

- Careful client selection and recruitment strategies can help improve coverage of the underlying population distribution. The server can selectively recruit new clients to improve diversity on dimensions like geographic location, device types and demographics [73]. Incentive mechanisms through rewards or service benefits can also encourage broader user participation [74, 75]. However, this requires additional coordination overhead.
- Robust aggregation algorithms help limit the influence of any manipulated or biased local updates on the global model [16]. Techniques like trimmed mean ignore extreme outlier updates during aggregation [76]. The Federated Averaging algorithm assigns weights to clients based on characteristics like data size and reliability. This reduces the impact of falsified model contributions.
- Statistical bias detection techniques can diagnose sampling issues before launching each training round [77]. Exploratory analysis of client characteristics helps identify underrepresented groups. Proxy metrics can be derived to estimate representation biases. This enables corrective actions like targeted recruitment.
- Differential privacy mechanisms add noise to updates before aggregation to cloak client identities [78]. This limits reconstructing sensitive attributes that could enable manipulating client samples. Cryptographic secure aggregation also prevents leaking client properties [1].

While these approaches help mitigate sample selection bias, fully addressing the root causes requires expanding decentralized participation. Better incentives, user engagement and representativeness metrics can enhance diversity over time.

3.2 Population distribution shift

The independent datasets distributed across clients in FL are not static over time. There can be significant drift in the underlying data distributions as user populations and behaviors evolve [79, 80]. For example, demographics like age groups, language preferences, cultural affiliations, etc. can change across geographic regions that are represented by different clients [1].

New trends and emerging use cases lead to shifts in usage patterns and data characteristics. The interests and needs of users may also vary over time. In healthcare, new patient groups and disorders can arise while incidence of existing diseases may decline [7]. Such changes can lead to the problem of concept drift, where models trained on past client data distributions do not generalize well to new emerging distributions [81, 82]. Historical biases can get entrenched into the models without accounting for shifting populations and trends over time. As an illustration, consider banks training fraud detection models using FL across their branches. If the model was trained only on historical data, it may not catch new fraud patterns arising at a faster rate in certain newer regions [83]. Without explicit retraining or adaptation, model performance can degrade rapidly.

3.2.1 Factors

Several factors can cause shifts in the decentralized distributions that client devices see:

- **User Demographic Shifts:** Populations inherently change as new user groups emerge and old cohorts decline. For example, generational differences lead to varying app usage patterns [84]. Geographic migration also causes distribution drifts.
- **User Behavior Shifts:** Interests, habits and needs of users evolve over time. Social trends lead to changing how apps and devices are utilized. For instance, growth of short-form video apps like TikTok [85].
- **App and Device Version Shifts:** As apps get updated with new features, the resulting data patterns change. New device models alter data types, apps, and usage. Software and hardware evolution leads to drift [86].
- **Adversarial Drift Injection:** Attackers can manipulate the local data or model updates to intentionally cause distribution shifts that degrade model performance over time [87]. They may target particular subpopulations.

3.2.2 Effects

Unchecked, such drift can lead to models becoming stale and inaccurate on new data distributions, even if they were robust when first deployed [88]. Two key effects include:

- **Overall Performance Degradation:** The global model starts making increasingly erroneous predictions as its assumptions become outdated [89]. Metrics like accuracy drop without realizing data has changed.
- **Subgroup Performance Issues:** Model quality deteriorates rapidly on segments of users exhibiting the most distribution drift without being detected [90]. Minority groups often suffer the most.

3.2.3 Potential mitigations

As user behaviors, environments, and systems evolve in FL, adaptive solutions are necessary to detect and respond to concept drift across decentralized devices [91-93]:

- Local drift detection techniques can analyze data distributions before aggregation to identify shifts from historical baselines [94]. Devices flag significant deviations to the central server prompting retraining. However, drift invisible to individual clients can accumulate globally.
- Personalized FL customizes models to adapt to changing local distributions [95]. This allows specializing in emerging user cohorts and data types. However, personalization risks fragmenting the global model.
- Robust optimization methods like distributionally robust optimization assume worst-case distributions to maintain performance despite drift [96]. Model rigor comes at the cost of lower accuracy on stable distributions.
- Continuous aggregated analytics helps detect distribution shifts globally by monitoring metrics like decreasing accuracy, increasing loss, and higher variance in updates [97]. Retraining can then recalibrate models.
- Simulating anticipated drifts via synthetic data allows proactively adapting models [98]. But this requires resources to generate future data profiles. And unexpected shifts may still occur.
- Hybrid cloud-federated architectures utilize centralized cloud resources to rapidly retrain models when unmanageable drift is detected [99]. But this partially compromises privacy.
- Adversarial domain adaptation explicitly trains models to adapt to different distributions, enhancing generalization [94]. But assumptions of shiftable domains are required.
- Fully tackling decentralized drift requires coordination frameworks to align client objectives, server oversight to diagnose drift, and configurable robustness against uncertainty [16].

3.3 Biased local data

In FL, biases can originate from the local datasets held by the clients themselves even before training begins [100]. The data collection and annotation process may be skewed against certain protected groups leading to underrepresentation or measurement bias.

For instance, a study by Kaissis et al. [101] demonstrated that FL models for medical imaging inherited biases from local datasets. Hospitals with more advanced imaging equipment contributed higher-quality data, while those with older equipment supplied lower-quality images. This disparity led to a model that performed better on data similar to that from hospitals with advanced equipment, disadvantaging patients from under-resourced hospitals.

3.3.1 Factors

There are several factors that can introduce bias into the decentralized data:

- **Biased Data Collection:** The data collection process may systemically underrepresent certain population groups. For example, clinical trials often focus on majority demographics while excluding minorities [102].
- **Measurement Biases:** Systemic issues in how data is measured and annotated can lead to bias. Examples include label skew, stereotypes in human labeling, and distortion in self-reported data [103].
- **Omitted Variables:** Relevant factors needed to make fair decisions may be missing from the data. Sensitive attributes like race, age or gender might be excluded.
- **Proxy Encoding:** Even if sensitive attributes are excluded, other features may encode similar information leading to indirect bias [104]. Location or income data could correlate to race for instance.

The local datasets may also disproportionately represent and further amplify existing societal biases [104]. Discrimination faced by marginalized communities propagates into the data. Seemingly objective data can perpetuate systemic inequities. If such issues in the decentralized data are not addressed, the biases will naturally get propagated to the global model during federated training and aggregation [11].

3.3.2 Effects

If left unaddressed, these biases in local data can get propagated to the global model during federated training. Biased data also makes models vulnerable to deliberate manipulation:

- **Data Poisoning Attacks:** Bad actors can inject poisoned local data with backdoors or malevolent biases that taint the global model [105]. This is harder to detect in decentralized models.
- **Model Evasion:** Biased data can be used to generate adversarial examples to evade detection by models like fraud classifiers [106].
- **Difficult to Audit:** Due to privacy requirements, auditing local data quality directly is difficult in FL, allowing bad data to go undetected.

Biased local data can thus lead to loss of fairness, performance issues, and security vulnerabilities. But mitigating decentralized data bias raises challenges around coordination, privacy, and incentive alignment between competitive clients [2].

3.3.3 Potential mitigations

Addressing biases in the decentralized datasets of FL clients poses unique challenges around preserving privacy and securely coordinating among untrusting parties. A range of techniques have been proposed to detect, limit, and correct for biases in local client data:

- **Client-Side Data Auditing:** Adapting centralized dataset auditing tools for decentralized execution allows clients to analyze their local data for biases and coverage issues before training models [107]. Privacy-preserving metrics quantify label skew, profiling gaps, stereotypes in annotations, proxy discrimination

through related variables, and other data quality issues without requiring to share actual data [108]. Aggregated reports allow assessment of biases across the federated network to prioritize improvements.

- **Local Data Debiasing:** Similar to centralized debiasing, privacy-preserving algorithms can pre-process local datasets to remove biases while preserving utility [109]. Removal or anonymization of sensitive variables, reweighting, resampling, and generative data augmentation changes balance representations before training. Debiasing also makes models more robust to data poisoning attacks.
- **Secure Aggregation for Bias:** Using secure multiparty computation and differential privacy, aggregation queries can be run across decentralized clients to detect outliers indicative of manipulated or intentionally poisoned local data [78, 110]. Noise addition during aggregation provides privacy for such diagnostics. Data trusts which hold local data encrypted can also run validation.
- **Reputation Systems:** Central servers can track metrics like variance between updates from a client, prediction errors on new data, anomalies in parameters, etc. to build reputation scores for reliable clients over time [105]. Biased or manipulated data contributions then get lower weightage during aggregation.
- **Incentive Mechanisms:** Proper incentives that reward good behavior can help improve data quality [111-113]. For example, well-behaved clients who provide useful updates are paid or provided higher model quality. Penalties can deter detectable manipulations and data poisoning.
- **Robust Secure Aggregation:** Carefully designed robust federated algorithms ignore or down-weight extreme parameter updates and give higher priority to trusted clients [76, 114]. This limits the impact of intentionally poisoned data contributions, but may overlook valid outliers.

While these help in mitigating biased local data, approaches to expand diversity and inclusion in data collection are also needed to address systemic root causes. Careful alignment of incentives, coordination and education can enhance data quality over time.

3.4 Systemic biases in data

The data used to train ML models often reflects and amplifies systemic societal biases that have persisted historically against certain groups [14]. Even datasets that appear neutral on the surface can propagate unfair social prejudices if not carefully examined [115]. For instance, healthcare data has been shown to contain inherent biases that disadvantage ethnic minorities. Clinical studies disproportionately focus on majority populations, excluding historically marginalized groups like racial minorities and poorer socioeconomic segments [116]. Predictive models trained on such data perpetuate the sampling and coverage biases, leading to inequitable healthcare access and outcomes for minorities. Further, face recognition systems have higher error rates for darker skin tones due to unbalanced training data [117]. Furthermore, finance data also reflects systemic biases in lending practices and income disparities across demographic factors like gender and race. Models built on such data can deny opportunities to minorities by repeating historical discrimination [118].

3.4.1 Factors

There are several factors that allow systemic biases to manifest in data:

- **Historical Discrimination:** Past discrimination faced by groups like racial minorities due to unethical laws, policies and practices over decades gets captured in data documenting those eras [119, 120]. Datasets thus bake in historical harms.
- **Measurement Biases:** The practices and tools used for data collection may systemically underrepresent certain population segments leading to sampling bias [121]. Surveys that fail to sample minorities, sensors unavailable in low-income regions, human annotation from majority demographics, etc. cause systemic data collection biases.
- **Proxy Discrimination:** Even when sensitive attributes like race, gender or age are explicitly excluded, other superficial attributes may correlate to them enabling indirect systemic biases [122]. Geography, income, language, profession and other factors can act as proxies.
- **Anchor Biases:** When labeling training data or defining categories, majority priorities and contexts are unconsciously anchored on leading to biases against minorities with different needs [123].
- **Social Stereotypes:** Human cognitive biases and prevailing societal prejudices unconsciously get imprinted into data [124, 125]. Cultural stereotypes around race, gender, age, ethnicity, etc. distort human annotation of training data as well as user-generated data.
- **Unexamined Assumptions:** Design choices that ignore minority interests perpetuate the status quo. Questionnaire design, feature selection, problem formalization and other assumptions disadvantage minorities [126-128].

3.4.2 Effects

When applied to data reflecting systemic biases, machine learning models inherit these issues which then get amplified due to feedback loops:

- **Underrepresentation:** Minority groups most impacted by systemic biases have insufficient data leading to outcomes optimized for the majority groups [129]. Their interests are excluded.
- **Measurement and Feature Bias:** Systemically distorted data provides inaccurate ground truth for modeling minority behavior leading to skewed models, and sensitive attributes may be proxied [14]. Moreover, Models latch onto features correlated with identity rather than meaningful drivers leading to proxy discrimination [130].
- **Objective Bias:** Model objectives like accuracy optimize for the majority groups at the cost of minorities by assuming balanced data [131].
- **Miscalibration:** Even when overall metrics seem robust, performance on minority groups suffers due to systemic underrepresentation and measurement biases [129].

While FL keeps the sensitive raw data decentralized, the models can still ingest systemic biases present in the local datasets. This can lead to scenarios where the AI systems

exclude, misrepresent or disproportionately target groups suffering from structural marginalization [132]. Unless countermeasures are taken, the aggregated models will reflect the accumulated societal biases. Techniques to audit datasets and algorithms as well as incentivize equitable engagement are necessary to mitigate harm from historically ingrained biases [133]. The compounding effects of minor data imbalances also need to be considered.

3.4.3 Potential mitigations

Addressing systemic biases that have accumulated in data requires a multifaceted approach:

- **Decentralized Bias Auditing:** Tools adapted for decentralized execution can allow clients to audit local data and coordinate audits on aggregated data to quantify bias and identify impacted groups [108]. Privacy-preserving metrics assess representation imbalances, distortion, profiling harms.
- **Dataset Nutrition Labels:** Attaching documentation detailing known gaps, assumptions, reporting subgroups, etc. provides transparency into limitations to address in system design. Data statements enable informed usage.
- **Dataset Augmentation:** Synthetically generating additional data samples from minority groups can improve coverage [134]. However, measuring impact on bias needs care.
- **Re-weighting Data:** Assigning higher weights to minority samples during model training can rebalance contributions [135]. But directly optimizing on biased data has risks.
- **Debiasing Algorithms:** Algorithms that pre-process data to remove proxy variables, reweight groups, and normalize representations can be applied locally before federated training [136, 137].
- **Inclusive Data Collection:** Expanding diversity in data collection practices to cover beyond majority demographics, languages, contexts, etc. provides more representative data.
- **Regulations and Standards:** Requiring standardized reporting, impact assessments, risk management and other oversight measures allows governing use of systemically biased data [138].

The unique constraints of FL require adapting anti-bias approaches to the decentralized environment. Key technical interventions include improving local data quality, algorithmic debiasing, and rigorous subgroup validation. However, technical steps alone are insufficient to address systemic societal issues. Participative auditing, representation in governance, ensuring accountability, and reforming unjust structures are imperative [139].

3.5 Social biases

ML systems do not operate in isolation, but rather reflect prevailing societal attitudes and ingrained human prejudices [140]. Data generated by humans naturally captures embedded cultural stereotypes and unconscious biases around factors like race, gender, age, ethnicity, etc. [141]. When datasets exhibiting social biases are used to train AI systems, the models inherit and amplify these biases. Seemingly neutral factors can encode demographic attributes leading to proxy discrimination [142].

3.5.1 Factors

There are various complex sociotechnical factors that allow social biases to become ingrained in data and algorithms [143]:

- **Historical Discrimination:** Legacies of unethical laws, policies, practices, and social norms against minority groups in the past propagate biases. Discriminatory practices get captured in historical datasets documenting those eras [119, 120]. Models trained on such data perpetuate historical prejudices.
- **Representation Imbalances:** Due to gaps in access, awareness, and technical skills, certain demographic segments end up underrepresented in datasets used to train AI systems. Without diversity, systems optimize for majority groups [129].
- **Social Stereotypes:** Human cognitive biases and prevailing cultural stereotypes unconsciously get imprinted into data generated by people. Annotators apply societal prejudices when labeling training data, which gets propagated into models [124].
- **Anchor Biases:** When defining problems, choosing categories, and labeling data, majority priorities, contexts and needs are anchored onto. Minority interests are overlooked, leading to biases against them.
- **Implicit Associations:** Models pick up on correlations in data that reflect social stereotypes and societal associations, even if not present in the variables directly used for training [144]. It get imprinted into models.
- **Feedback Loops:** Due to self-reinforcing cycles, minor biases accumulate and get amplified over time. Biased models produce skewed outputs which further distort data used for retraining [145].

3.5.2 Effect

Unchecked social biases perpetuated through data and algorithms lead to the following discriminatory impacts [143]:

- **Underrepresentation:** Minority groups left out of training data receive lower quality outcomes as models never learn to optimize for them. Their interests are excluded.
- **Feature Biases:** Models rely on factors correlating with sensitive attributes like race or gender rather than meaningful drivers of decisions. This leads to proxy discrimination.
- **Poor Generalization:** Performance disparities arise for minority dialects, cultural contexts, age groups, etc. that differ from majority training data.
- **Stereotyping:** Recommendations align with and reinforce stereotypes around race, gender, age or ethnicity leading to pigeonholing.
- **Exclusion:** Requirements like video interviews, majority cultural settings, etc. disadvantage subgroups with less access to technology and dominant cultures.
- **Denial of Opportunities:** Biased models result in fewer opportunities in areas like credit, employment, housing, etc. for impacted groups.
- **Abusive Targeting:** Models explicitly target minorities for predatory practices like aggressive policing, high interest loans, or addiction promotion.
- **Loss of Autonomy:** Biased systems override minority self-determination by encoding majority priorities and perspectives.

Social biases in FL lead to disproportionate errors, exclusion or problematic recommendations against protected

groups facing structural inequities [146-148]. Without concerted efforts, federated models will encode accumulated societal prejudices leading to discriminatory impacts.

3.5.3 Potential mitigations

Addressing social bias remains an active area of research in FL requiring a multifaceted approach:

- **Diverse Clients:** Promoting participation from a diverse and representative range of users, devices, and organizations helps create more varied decentralized data to train on.
- **Bias Auditing:** Adapting bias quantification tools to run locally allows clients to audit their data. Secure aggregation enables collective auditing to identify systemic issues [108, 149].
- **Data Augmentation:** Generative models can be used to synthesize new data samples from underrepresented groups, locally improving data diversity [150].
- **Dataset Nutrition Labels:** Attaching documentation about known data biases improves transparency for those aggregating models [151].
- **Algorithmic Fairness:** Regularization terms and constraints are incorporated into the global model optimization to encourage fairness across groups [152].
- **Subgroup Validation:** Maintaining segmented test sets for evaluating model performance on minority groups allows detecting disparities [153].

3.6 Algorithmic biases

In addition to data issues, biases can also arise from the model architectures, objective functions, and assumptions made during the ML pipeline [14]. Choices that seem neutral can unintentionally introduce algorithmic harms against certain groups. For instance, commonly used performance metrics like accuracy implicitly assume class balance and can optimize for the majority groups, disadvantaging minorities [124]. Maximizing accuracy leads models to disproportionately focus on improving predictions for well-represented groups.

For instance, an implementation of FL for music recommendation revealed algorithmic bias due to the optimization objective favoring majority user preferences [154]. The model prioritized genres favored by the dominant user groups in the training data, underrepresenting music genres preferred by minority users.

3.6.1 Factors

There are various ways that bias can inadvertently become encoded into the algorithms and models themselves:

- Model objectives like accuracy and Root Mean Square Error (RMSE) optimize for overall aggregate performance but ignore effects on subgroups. Maximizing accuracy leads systems to focus on improving predictions for well-represented majority groups at the cost of minorities [131]. Classifier performance metrics assume balanced data and proportional subgroups.
- Many regularization techniques to prevent overfitting like parameter norm penalties are optimized for centralized settings with abundant IID data. They do not sufficiently account for the decentralized unbalanced distributions often present in FL [54]. This degrades model robustness and fairness on underrepresented groups.
- The choice of model family, architectures and hyperparameters reflects inherent assumptions that may not universally hold across diverse subgroups. Different subgroups may require different model forms to capture nuanced data relationships.
- Deep learning models can latch onto spurious correlations during training that reflect historical biases and discrimination rather than meaningful drivers of decisions [155]. The black box nature of complex models makes auditing these latent biases difficult.
- During federated training, the process of aggregating algorithmic biases from various decentralized clients can compound harms. Differing optimization constraints and incentives between competitive clients further complicate the effects [5].

3.6.2 Effect

Unchecked algorithmic biases can result in the following discriminatory impacts on minority groups:

- Suboptimal performance due to poor generalization on data patterns not fitting simplified modeling assumptions based on majority trends [129]. Tailored model selection is needed to address subgroup needs.
- Perpetuating historical harms by replicating ingrained patterns of discrimination that correlate superficially in the data but have no true explanatory relationship [156].
- Denying opportunities by optimizing only for the overall metric improvements which enables sacrificing minority subgroups if it benefits the majority aggregate.
- Violating privacy and self-determination by enabling inference of sensitive attributes from model outputs even if not an explicit input [157]. Individuals lose control over use of their personal information.
- Avoiding accountability by providing ambiguous opaque systems that deflect responsibility for unfair outcomes by attributing them to algorithmic determinism [158].

Unless algorithmic biases are mitigated through thoughtful selection of performance metrics, model forms and training objectives, FL risks exacerbating discrimination through its algorithms.

3.6.3 Potential mitigations

While data biases entering models can be addressed through preprocessing and augmentation techniques, biases can also arise from the model development process itself. From unfair performance metrics to poor generalizability on minority data patterns, a range of technical choices can inadvertently introduce algorithmic harms:

- **Subgroup Validation:** Maintain segmented test sets for each demographic group to evaluate model performance across subgroups and detect disparities [153, 159].
- **Fairness Regularization:** Incorporate constraints or terms in the global model optimization that penalize discrimination and encourage fairness across groups [160, 161].
- **Flexible Model Selection:** Use adaptive model selection and hyperparameter tuning tailored to optimize performance for individual clients to improve personalization [95].

- **Multimodal Modeling:** Ensemble approaches combining diverse model families tuned on representations learned from subset data can improve robustness.
- **Counterfactual Evaluation:** Assess models by systematically simulating conditions with biases removed to quantify impact on subgroups [162].
- **Modular Transparency:** Adopt interpretable, modular and auditable model architectures to enable better oversight [163].

While technical interventions help, addressing algorithmic biases requires examining how problems are formulated, performance is measured, and who is centered in development. Wider community participation in designing and auditing algorithms can surface harmful assumptions [164].

3.7 Representational biases

Representational biases refer to issues that arise from how data is structured, problems formalized, and models designed in ways that marginalize certain populations [115, 158]. Choices that may seem neutral can inadvertently encode assumptions that disadvantage minority groups. For example, the way data is formatted often normalizes attributes common in majority demographics while minorities end up represented as edge cases or exceptions [159].

For instance, FL models trained on text data from globally distributed users may underrepresent low-resource languages [165, 166]. Users typing in less common languages contribute less to the model updates, leading to poorer language processing capabilities for those languages.

3.7.1 Factors

There are various ways that representational biases can manifest in the FL pipeline:

- **Client Data Formatting:** The way local decentralized data is preprocessed, formatted and encoded often normalizes attributes common in majority demographics while minorities are edge cases [14].
- **Problem Framing:** Objectives set by aggregators often ignore needs of underrepresented groups. Narrow assumptions result in poor personalization [151].
- **Global Model Design:** Choices in model family, architecture, hyperparameters may be suboptimal for tailoring to diverse clients' data patterns [115].
- **Evaluation Metrics:** Global validation datasets and metrics like overall accuracy insufficient to assess subgroup impacts [167].

Together these representational choices by central aggregators can center majority groups while marginalizing minorities in the federated environment. This leads to embedding biases that disadvantage subgroups among the decentralized clients.

3.7.2 Effect

Representational biases manifest in FL models through the following effects:

- **Poor Personalization:** Global models fail to specialize well for minority clients with atypical data distributions, dialects, and use cases [168-170].
- **Limited Accessibility:** Systems lack localization, inclusivity, and accessibility limiting adoption by diverse decentralized groups [171].
- **Skewed Optimization:** Aggregate metrics like overall accuracy lead to models optimized for majority clients at the cost of minorities [16].
- **Privacy Violations:** Centralized coordinators could infer sensitive attributes about local client data from model updates [78].
- **Entrenched Biases:** Global models perpetuate systemic representational biases without reforms to inclusive coordination [172].
- **Lack of Recourse:** Opaque federated systems deflect accountability for unfair outcomes by attributing them to algorithmic determinism [165].

Decentralized participatory approaches are necessary to ensure representations do not exclude or disadvantage minority clients.

3.7.3 Potential mitigations

Addressing representational biases remains an active area of research in FL requiring technical interventions combined with inclusive participative design and decentralized governance approaches:

- **Inclusive Design:** Engage wider community participation in problem formulation, model prototyping and testing to surface excluded perspectives.
- **Personalized Models:** Adapt model selection and hyperparameters for each client to improve personalization and prevent one-fits-all effects [95].
- **Subgroup Validation:** Use segmented test sets and metrics to ensure model performance is evaluated across diverse clients and environments.
- **Bias Metrics:** Define quantifiable metrics to assess model fairness and inclusion issues during development [108].
- **Nutrition Labels:** Require documentation of design choices, assumptions, limitations to inform aggregation and oversight.

While technical interventions help, addressing root causes requires examining who is centered in FL design. Participative, decentralized, peer-based approaches can help reform exclusionary assumptions and structures. Community voices should guide problem formulation, not just passive data contributors. Standards preventing extractive, unethical data practices are also necessary [173]. Dual technical and social responses attuned to marginalized groups can accelerate progress.

4. BIAS MITIGATION TECHNIQUES IN FEDERATED LEARNING

In this section, we provide a detailed examination of specific bias mitigation techniques in FL, highlighting concrete examples and comparing their effectiveness. We categorize these techniques into: (1) data-based, (2) algorithmic, and (3) hybrid approaches and discuss their implementations and outcomes in real-world scenarios. Table 3 summarizes and compares these techniques.

4.1 Data-based mitigation techniques

Data-based techniques focus on manipulating the training data to reduce biases. These methods are implemented at the client level, where the data resides:

Table 3. Comparison of Bias Mitigation Techniques in Federated Learning

Technique	Category	Advantages	Limitations
q-FFL [16]	Data-Based / Algorithmic	Improves fairness across clients; simple implementation.	May slow convergence; requires client loss information.
FAug [150]	Data-Based	Enhances data diversity; no raw data sharing.	Requires client coordination; may not address all biases.
Local Adversarial Debiasing [174]	Data-Based	Reduces bias at source; preserves privacy.	Requires sensitive attribute labels; potential utility loss.
AFL [152]	Algorithmic	Robust to data heterogeneity; improves worst-case performance.	May reduce overall accuracy; conservative optimization.
GKT [175]	Algorithmic	Preserves group characteristics; improves group fairness.	Increased communication and computation due to clustering.
SCAFFOLD [176]	Algorithmic	Corrects client drift; improves convergence.	Additional communication overhead for control variates.
FedHealth [91]	Hybrid	Benefits clients with limited data; improves personalization.	Requires feature alignment; privacy concerns.
MOCHA [177]	Hybrid	Adapts to client-specific distributions; reduces biases.	Computationally intensive; complex optimization.

4.1.1 Reweighting and resampling strategies

These type of techniques aim to address sample selection bias and class imbalances by adjusting the importance of data samples or altering the sampling probability. In this direction, a solution, called q-Fair Federated Learning (q-FFL), proposed by Li et al. [16] introduces a fairness-aware objective function that adjusts weights based on the inverse of the client's loss. This approach emphasizes underperforming clients or minority groups by allocating them more weight during aggregation. q-FFL has been shown to improve fairness across clients in terms of model performance disparities. However, it may slow down overall convergence and requires careful calibration of the fairness parameter q. Additionally, it depends on clients sharing their loss values, which may raise privacy concerns.

4.1.2 Federated data augmentation

Data augmentation techniques enhance the diversity of training samples by generating synthetic data, reducing biases due to limited or skewed local datasets. Federated Augmentation (FAug) introduced by Jeong et al. [150] allows clients to share data augmentation strategies instead of actual data. By agreeing on common augmentation policies, clients can simulate a more balanced and diverse dataset locally. FAug improves model generalization and reduces biases arising from data heterogeneity. However, it requires coordination among clients to agree on augmentation policies, which may not be feasible in all federated settings.

4.1.3 Client-side data debiasing

Clients perform local data preprocessing and debiasing to mitigate biases inherent in their datasets. In this direction, Local Adversarial Debiasing proposed by Du et al. [174] involves training a debiasing model adversarially to remove sensitive attribute information from the representations learned locally. This approach reduces biases related to sensitive attributes (e.g., gender, race) at the source and preserves privacy since debiasing is performed locally. However, it requires clients to have access to sensitive attribute labels, which may not always be available or permissible due to privacy regulations.

4.2 Algorithmic mitigation techniques

Algorithmic techniques modify the FL process to incorporate fairness directly into model training and aggregation.

4.2.1 Fair federated learning algorithms

These algorithms adjust the training objective to consider fairness across clients or groups. For example, Agnostic Federated Learning (AFL) proposed by Mohri et al. [152] optimizes for the worst-case weighted combination of client losses. By focusing on the minimax optimization problem, AFL aims to improve fairness and robustness across heterogeneous client data distributions. AFL has demonstrated improved fairness metrics and robustness to non-IID data. However, it may lead to a reduction in overall model accuracy due to its conservative optimization approach that prioritizes the worst-performing clients.

4.2.2 Fair averaging and aggregation

Aggregation methods can be designed to account for fairness during the model update phase. For instance, Group Knowledge Transfer (GKT) introduced by Wang et al. [175] clusters clients into groups based on data distributions and aggregates models within each group before combining them globally. This preserves group-specific characteristics and mitigates biases due to group differences. GKT improves fairness by ensuring that group-specific information is not lost during aggregation. The method may increase communication overhead and computational complexity due to the need for clustering and multiple aggregations.

4.2.3 Adaptive optimization techniques

These methods adjust the learning rates or update rules to account for data heterogeneity. SCAFFOLD proposed by Karimireddy et al. [176] uses control variates to correct for client drift resulting from heterogeneity in data distributions. It introduces a variance reduction technique to better align local updates with the global objective. SCAFFOLD achieves faster convergence and reduces the biases caused by client drift. The trade-off includes additional communication overhead due to the need to transmit control variates between clients and the server.

4.3 Hybrid mitigation techniques

Hybrid techniques combine data-based and algorithmic approaches to leverage the strengths of both.

4.3.1 Federated transfer learning

It enables clients with limited data to benefit from models trained on larger datasets from other clients. For instance, FedHealth proposed by Chen et al. [91] is a federated transfer learning framework designed for wearable healthcare data. It leverages knowledge from rich datasets at some clients to improve the personalized models at others. FedHealth improves model accuracy and fairness for clients with scarce data and reduces biases due to data scarcity. Challenges include ensuring feature space alignment across clients and managing privacy concerns.

4.3.2 Multi-Task learning approaches

Multi-task learning allows clients to learn personalized models while sharing representations. For example, MOCHA proposed by Smith et al. [177] is a federated multi-task learning framework that models each client's task separately but jointly learns shared representations. MOCHA reduces biases by accommodating client-specific data distributions and enhances personalization. It requires solving complex optimization problems and may have higher computational demands.

5. CHALLENGES AND OPEN ISSUES

While progress has been made in developing bias mitigating solutions tailored to FL, significant open questions and challenges remain to be addressed. The unique constraints arising from decentralization, systems heterogeneity, competing incentives, and privacy considerations pose difficulties in directly applying centralized techniques. Novel advancements are needed across areas ranging from coordinated evaluation to privacy-preserving auditing and incentive mechanisms for promoting voluntary adoption of debiasing techniques. The emerging field of fair and accountable FL continues to be an active research domain, with impacts on developing trustworthy and inclusive AI systems. In this section, we explore critical unresolved challenges and promising research directions to fully harness FL's potential for addressing societal needs and enhancing public welfare. These challenges span technical, ethical, and practical dimensions that must be examined to ensure FL can effectively serve humanitarian causes.

5.1 Lack of coordination

A core assumption in many bias mitigation techniques is the ability to coordinate across the full dataset or training process. However, FL involves decentralized clients that are often distrusting entities with misaligned incentives and competing objectives. This poses challenges for bias mitigation compared to traditional centralized training where full coordination can be enforced. For example, techniques like reweighting underrepresented groups or oversampling minorities require a global view of the overall data distribution to ensure proper balancing. But transparency into other clients' local data distributions to calculate appropriate weights may violate privacy expectations and business interests. Introducing fake simulated data also needs coordination to prevent overlapping samples.

Furthermore, distributed validation of models on segmented test sets representing diverse groups is important to assess biases consistently. But this requires collective coordination in defining evaluation methodology and sharing results. Adversarial attacks exploiting lack of coordination are also harder to detect without global visibility across clients. Moreover, mechanisms for limited coordination could help balance these tensions. For example, secure distributed analytics and differential privacy may provide aggregated insights into bias without exposing raw client data. Further, economic incentives and reputation systems could encourage coordination behaviors aligned with mitigating bias. Furthermore, transfer learning can propagate useful patterns across models without sharing actual data [178].

However, fully decentralized bias mitigation without any coordination remains an open research challenge. Options like trusted industry-specific authorities to govern coordination, norms-based self-organization, and voluntary coordination around social responsibilities may help in specific contexts. But ensuring mitigation at scale without centralized oversight remains an open issue.

5.2 Privacy constraints

While transparency and auditing can help identify biases, strict privacy protections in FL pose challenges. Directly analyzing raw decentralized data to quantify bias metrics would provide the most insight but violates client privacy expectations [179]. Furthermore, differential privacy and secure aggregation techniques allow some validation and analytics in a privacy-preserving manner. But these introduce noise that limits utility [180]. The level of noise required to fully mask membership or attribute inference may render validation metrics too distorted to draw fair conclusions.

Moreover, transparency reports on aggregate demographics and bias metrics can help coordinate mitigation efforts. But this reveals sensitive information about clients so is often avoided. Detailed model cards on performance across groups require profiling users. Therefore, novel privacy-preserving auditing techniques are needed that enable unbiased evaluation without compromising confidential data. Secure multi-party computation shows promise to run distributed diagnostics. Models can also be trained to predict biases without exposing data [181].

However, partially relaxing privacy, such as within industry consortiums, can enable transparency for bias mitigation. But this risks exclusion if marginalized groups lie outside such consortiums. Limited, accountable disclosure may help balance risks [94]. Thus, fully decentralized approaches without transparency remain challenging. Social coordination through norms and voluntary self-disclosure could raise accountability. But ensuring comprehensive, privacy-preserving auditing at scale remains an open research area.

5.3 Evaluating decentralized Non-IID data

A key challenge in FL is evaluating models for bias on decentralized non-IID client data. Overall performance metrics like accuracy, calculated centrally after aggregating model updates, can miss disparate impacts on underrepresented groups that exhibit distribution skew [14]. Drilling down to assess model behavior across heterogeneous local datasets is important to quantify fairness. However, this requires sharing samples from sensitive client data, which violates privacy expectations.

While centralized evaluation on representative test sets is common in ML, this poses difficulties in federated settings.

Clients are often unwilling to share local data for pooled validation due to confidentiality concerns. Coordinating clients to create standardized test sets that cover diverse subgroups is also arduous. Differences in evaluation methodology, metrics, and labeling schemas across organizations further complicate consistent assessment [182]. Therefore, new decentralized protocols are required that can validate model performance and fairness on heterogeneous local data in a privacy-preserving manner. Secure multiparty computation techniques enable aggregating subgroup metrics across clients without exposing raw data. Differentially private mechanisms allow discerning bias while limiting attribute disclosure. Emerging techniques like federated meta-learning assess models by exchanging trainable parameters instead of actual data samples.

However, practical adoption of such emerging secure evaluation techniques faces barriers. Compliance from competitive clients is difficult to ensure without oversight [183]. For instance, peer auditing raises additional tensions around proprietary model comparisons. Relative benchmarking against fluctuating baselines provides limited insight into absolute model biases. Thus, beyond technical solutions, progress requires establishing community standards, participative governance and incentives promoting accountability [184].

5.4 Adversarial attacks and model poisoning

The decentralized nature of FL makes models susceptible to adversarial attacks and intentional data poisoning aimed at compromising fairness [185]. Malicious clients could inject poisoned updates with embedded biases against certain subgroups that taint the global model. Since the origin of contributed updates is obscured, it becomes challenging to detect such manipulated contributions encoding malevolent biases. Discriminatory attacks that would be evident in centralized settings can remain invisible in FL without full visibility into heterogeneous client behaviors [186].

While techniques like robust aggregation can ignore or downweight outlier updates, advanced attacks craft poisoned updates that appear as normal local changes to bypass defenses. Carefully coordinated injection of poisoning data across multiple colluding devices can further mask bias detection, which is harder without global oversight. Simulating centralized retraining on all raw data can diagnose poisoning but leaks private data. Developing robust anomaly detection tailored to federated environments remains an open area needing novel privacy-preserving inspection of contributions across distrusting parties to identify manipulation of biases. Monitoring metrics indicating emerging skew against protected groups can help [187]. Cryptographic reward systems that incentivize fair updates and flag suspicious behaviors could enhance resilience. Addressing these emerging threats is critical for securing FL against adversarially encoded biases.

5.5 Communication and computation constraints

Many bias detection and mitigation techniques designed for centralized environments involve additional communication and computational overhead, which poses challenges in adopting them for FL. Complex multi-round algorithms or hosting large validation datasets may be infeasible given the limitations of decentralized devices and connectivity constraints. For example, bandwidth-heavy gradient exchange between clients for consensus-based federated averaging results in high latency [188]. Similarly, continuously monitoring detailed metrics and rerunning expensive model diagnostics like full factorization for bias can be prohibitive [1].

Therefore, developing efficient and optimized solutions is crucial to enable practical bias mitigation under federated constraints. Approaches tailored to minimize communication such as exchanging only model updates rather than full parameters can assist adoption. Further, strategies like secure hierarchical aggregation avoids all-to-all exchanges. Furthermore, predictive modeling using proxy data can analyze biases without heavy diagnostics. Moreover, sparse collective matrices reduce intersection computations. Parallelizing operations across clients hastens convergence. Thus, achieving efficacy and robustness against biases within the cost limitations of decentralized environments remains an active research goal needing novel frugal approaches.

5.6 Personalization Vs. fragmentation for bias mitigation

There exists a tradeoff between customizing FL models to mitigate local biases versus maintaining a synchronized global model. Personalization allows adapting to address distinct biases arising from localized data distributions and needs [95]. However, excessive flexibility can fragment the global model with heterogeneous variations that compound bias [189].

On one hand, personalized federated algorithms tailored to each client's constraints and subgroups enable improved fairness on specific populations [91]. Specialization captures nuanced correlations behind local biases invisible in one-size-fits-all models [190]. However, unchecked personalization can result in clients diverging into isolated localized models that overfit unique biases and fail to generalize fairly.

Carefully regulating model customization is therefore necessary to balance bias mitigation gains from specialization with robustness arising from global coordination. Selective relaxation of certain parameters for local adaptation while coordinating on a shared core model offers one path. Regularization terms that penalize divergence away from global fairness solutions incentivize alignment. Peer-based consensus algorithms propagating useful personalized variations globally provide another approach. Further research is needed to develop adaptive, secure frameworks enabling responsible flexibility balanced with coordination for bias mitigation.

5.7 Incentives for voluntary bias mitigation

In federated settings without centralized control, competitive clients lack inherent incentives to employ voluntary bias mitigation practices that may compromise their individual utilization goals. Enterprises may resist techniques like restricting sensitive attributes or balanced sampling that reduce biases but lower accuracy on their local objectives. Uncoordinated self-interests can perpetuate inequities even if harms are collectively suboptimal [112].

To encourage voluntary adoption of debiasing techniques, novel incentive mechanisms tuned to the decentralized nature of FL are needed. Crypto-economic approaches based on token rewards for good behavior show promise to incentivize fairness [191]. For example, clients can be compensated for adopting updates that enhance equity while temporarily

reducing private metrics. Gamification through leaderboards celebrating top contributors tackling bias creates motivational incentives. However, malicious actors may still manipulate such mechanisms if robustness is insufficient [192]. Careful incentive alignment balanced with security remains an open research challenge. Beyond technical solutions, legislation and policies may also be required to mandate responsible bias mitigation practices.

5.8 Emerging best practices for bias mitigation

As FL expands, best practices around ethical data sharing, representation, model interpretability, and algorithmic accountability tailored to decentralized environments are still evolving. While technical interventions help, holistic responses spanning governance, transparency, and participative design are imperative for mitigating bias.

Establishing norms against excluding subgroups, requiring localized testing, instituting peer audits, attaching model fact sheets, and enacting regulations around algorithmic harms can steer progress. Community participation in shaping problem formulation, metrics, and standards prevents narrowly technocratic solutions. Translating technical insights about model behaviors and uncertainties using interactive visualizations improves transparency.

However, operationalizing responsible bias mitigation practices faces barriers around coordination, privacy, and misaligned incentives in federated ecosystems. Developing minimal disclosures, distributed audits, and incentives balancing rigor with confidentiality offers paths forward. Focusing bias mitigation on enhancing equity instead of just detecting deficiencies reorients efforts towards social impact. Ultimately, the path towards trustworthy and inclusive FL necessitates interdisciplinary perspectives attuned to marginalized communities.

5.9 Governance, transparency and participation

Beyond specific technical solutions, addressing biases in FL raises broader societal questions around governance, transparency, and participation that remain open issues. Mechanisms for oversight and accountability are unclear in decentralized ecosystems involving distrusting parties. Standards and policies specifically regulating bias and representation in federated contexts are still emerging across sectors [193].

Transparency around bias mitigation practices, performance differences, and accountability is limited but critical for detecting exclusion errors. However, this conflicts with privacy expectations in federated ecosystems. Developing minimal, participative transparency frameworks balancing accountability with confidentiality is an open challenge. Enabling impacted communities to shape problem formulation, audit biases, and steer solutions centrally recognizes their self-determination.

Realizing the benefits of FL for social good requires inclusive governance and participative design. Co-creating decentralized architectures integrating peer oversight, contestations, and transparency with marginalized groups' interests centered can enhance legitimacy and representation. Beyond technical bias mitigation, broader questions around reforming unjust social structures enabling comprehensive participation remain imperative [194].

5.10 Scalability concerns

As FL systems expand to include a large number of clients, implementing bias mitigation strategies introduces scalability challenges. Techniques such as robust aggregation, personalized modeling, and cryptographic protocols may incur significant computational and communication overhead when applied at scale. The diversity in client devices and network conditions can further complicate the efficient deployment of these strategies in larger networks. Addressing scalability is crucial to ensure that bias mitigation remains practical and effective as FL systems grow, enabling widespread adoption without compromising performance or fairness.

6. CONCLUSIONS

In this paper, we provided a comprehensive analysis of the various sources of bias that can manifest in FL systems and examined tailored mitigation strategies. The decentralized and privacy-preserving nature of FL poses unique constraints compared to centralized ML when addressing biases. Care must be taken to balance fairness, accuracy, and confidentiality. We discussed how sample selection biases, population distribution shifts, locally biased data, systemic societal biases, algorithmic biases, and representational biases can accumulate in federated models, leading to discrimination and exclusion.

Our key contributions include developing a taxonomy of bias sources unique to FL, categorizing bias mitigation techniques tailored for FL, and discussing open research challenges. We categorized mitigation strategies into data-based, algorithmic, and hybrid approaches, highlighting specific methods such as reweighting and resampling strategies, federated data augmentation, fair federated learning algorithms, and privacy-preserving fairness regularization techniques. By analyzing these techniques, we provided insights into their practical implementations, advantages, and limitations in real-world scenarios.

Additionally, we summarized key open challenges and future directions around miscoordination, privacy constraints, decentralized evaluation, data poisoning attacks, systems heterogeneity, incentive misalignments, personalization trade-offs, emerging governance needs, participation, and scalability concerns that remain to be addressed. Addressing these challenges is crucial for developing fair and trustworthy FL systems.

This study aims to provide a conceptual foundation and starting point for future research focused on developing trustworthy and fair FL systems. As FL expands into critical domains like healthcare, finance, and mobility, ensuring equitable and inclusive access free from embedded biases will be imperative. Both technical solutions tailored to the decentralized environment, as well as participatory approaches, are needed to unlock the full potential of this emerging technology for social good. By combining advances in bias mitigation techniques with inclusive governance and community participation, we can work towards federated learning systems that are not only accurate and efficient but also fair and equitable for all stakeholders.

REFERENCES

[1] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, B., Van Overveldt, T., Petrou, D., Ramage, D., Roselander, J. (2019). Towards federated learning at scale: System design. In Proceedings of Machine Learning and Systems, 1: 374-388. https://doi.org/10.48550/arXiv.1902.01046

[2] Yang, Q., Liu, Y., Chen, T.J., Tong, Y.X. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2): 1-19. https://doi.org/10.1145/3298981

[3] Lim, W.Y.B., Luong, N.C., Hoang, D.T., Jiao, Y.T., Liang, Y.C., Yang, Q., Niyato, D., Miao, C.Y. (2020). Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(3): 2031-2063. https://doi.org/10.1109/COMST.2020.2986024

[4] Li, T., Sahu, A.K., Talwalkar, A., Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3): 50-60. https://doi.org/10.1109/MSP.2020.2975749

[5] Benmalek, M., Benrekia, M.A., Challal, Y. (2022). Security of federated learning: Attacks, defensive mechanisms, and challenges. Revue des Sciences et Technologies de l'Information-Série RIA: Revue d'Intelligence Artificielle, 36(1): 49-59. https://doi.org/10.18280/ria.360106

[6] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., Ramage, D. (2018). Federated learning for mobile keyboard prediction. arXiv Preprint arXiv: 1811.03604. https://doi.org/10.48550/arXiv.1811.03604

[7] Rieke, N., Hancox, J., Li, W.Q., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D.G., Baust, M., Cardoso, M.J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1): 119. https://doi.org/10.1038/s41746-020-00323-1

[8] Yang, W.S., Zhang, Y.H., Ye, K.J., Li, L., Xu, C.Z. (2019). Ffd: A federated learning based method for credit card fraud detection. In Big Data-BigData 2019: 8th International Congress, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, pp. 18-32. https://doi.org/10.1007/978-3-030-23551-2_2

[9] Brecko, A., Kajati, E., Koziorek, J., Zolotova, I. (2022). Federated learning for edge computing: A survey. Applied Sciences, 12(18): 9124. https://doi.org/10.3390/app12189124

[10] Savazzi, S., Nicoli, M., Rampa, V. (2020). Federated learning with cooperating devices: A consensus approach for massive IoT networks. IEEE Internet of Things Journal, 7(5): 4641-4654. https://doi.org/10.1109/JIOT.2020.2964162

[11] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V. (2020). How to backdoor federated learning. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy, 108: 2938-2948. https://doi.org/10.48550/arXiv.1807.00459

[12] Dinh, C.T., Tran, N., Nguyen, J. (2020). Personalized federated learning with moreau envelopes. Advances in Neural Information Processing Systems, 33: 21394-21405. https://doi.org/10.48550/arXiv.2006.08848

[13] Barocas, S., Crawford, K., Shapiro, A., Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning. In 9th Annual Conference of the Special Interest Group for Computing, Information and Society, pp. 1.

[14] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6): 1-35. https://doi.org/10.1145/3457607

[15] Friedman, B., Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3): 330-347. https://doi.org/10.1145/230538.230561

[16] Li, T., Sanjabi, M., Beirami, A., Smith, V. (2019). Fair resource allocation in federated learning. arXiv Preprint arXiv: 1905.10497. https://doi.org/10.48550/arXiv.1905.10497

[17] Zhang, B.H., Lemoine, B., Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335-340. https://doi.org/10.1145/3278721.3278779

[18] Lipton, Z.C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3): 31-57. https://doi.org/10.1145/3236386.3241340

[19] Voigt, P., Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing, 10(3152676): 10-5555. https://doi.org/10.1007/978-3-319-57959-7

[20] Konečný, J., McMahan, H.B., Ramage, D., Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. arXiv Preprint arXiv: 1610.02527. https://doi.org/10.48550/arXiv.1610.02527

[21] Kulkarni, V., Kulkarni, M., Pant, A. (2020). Survey of personalization techniques for federated learning. In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, UK, pp. 794-797. https://doi.org/10.1109/WorldS450073.2020.9210355

[22] Li, L., Fan, Y.X., Tse, M., Lin, K.Y. (2020). A review of applications in federated learning. Computers & Industrial Engineering, 149: 106854. https://doi.org/10.1016/j.cie.2020.106854

[23] Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N., Jararweh, Y. (2022). Federated learning review: Fundamentals, enabling technologies, and future applications. Information Processing & Management, 59(6): 103061. https://doi.org/10.1016/j.ipm.2022.103061

[24] Zheng, Z.H., Zhou, Y.Z., Sun, Y.L., Wang, Z., Liu, B.Y., Li, K.Q. (2022). Applications of federated learning in smart cities: Recent advances, taxonomy, and open challenges. Connection Science, 34(1): 1-28. https://doi.org/10.1080/09540091.2021.1936455

[25] Rahman, K.M.J., Ahmed, F., Akhter, N., Hasan, M., Amin, R., Aziz, K.E., Islam, A.K.M.M., Mukta, M.S.H., Islam, A.K.M.N. (2021). Challenges, applications and design aspects of federated learning: A survey. IEEE

Access, 9: 124682-124700. https://doi.org/10.1109/ACCESS.2021.3111118
[26] Shaheen, M., Farooq, M.S., Umer, T., Kim, B.S. (2022). Applications of federated learning; taxonomy, challenges, and research trends. Electronics, 11(4): 670. https://doi.org/10.3390/electronics11040670
[27] Zhang, T., Gao, L., He, C.Y., Zhang, M., Krishnamachari, B., Avestimehr, A.S. (2022). Federated learning for the internet of things: Applications, challenges, and opportunities. IEEE Internet of Things Magazine, 5(1): 24-29. https://doi.org/10.1109/IOTM.004.2100182
[28] Stremmel, J., Singh, A. (2021). Pretraining federated text models for next word prediction. In Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), 2: 477-488. https://doi.org/10.1007/978-3-030-73103-8_34
[29] Long, G.D., Tan, Y., Jiang, J., Zhang, C.Q. (2020). Federated learning for open banking. In Federated Learning: Privacy and Incentive, pp. 240-254. https://doi.org/10.1007/978-3-030-63076-8_17
[30] Kumar, Y., Singla, R. (2021). Federated learning systems for healthcare: Perspective and recent progress. Federated Learning Systems: Towards Next-Generation AI, 965: 141-156. https://doi.org/10.1007/978-3-030-70604-3_6
[31] Patel, V.A., Bhattacharya, P., Tanwar, S., Gupta, R., Sharma, G., Bokoro, P.N., Sharma, R. (2022). Adoption of federated learning for healthcare informatics: Emerging applications and future directions. IEEE Access, 10: 90792-90826. https://doi.org/10.1109/ACCESS.2022.3201876
[32] Khan, L.U., Saad, W., Han, Z., Hossain, E., Hong, C.S. (2021). Federated learning for internet of things: Recent advances, taxonomy, and open challenges. IEEE Communications Surveys & Tutorials, 22(3): 1759-1799. https://doi.org/10.1109/COMST.2021.3090430
[33] Rahman, S.A., Tout, H., Talhi, C., Mourad, A. (2020). Internet of things intrusion detection: Centralized, on-device, or federated learning? IEEE Network, 34(6): 310-317. https://doi.org/10.1109/MNET.011.2000286
[34] dos Santos, R.R., Viegas, E.K., Santin, A.O., Tedeschi, P. (2023). Federated learning for reliable model updates in network-based intrusion detection. Computers & Security, 133: 103413. https://doi.org/10.1016/j.cose.2023.103413
[35] Ammad-Ud-Din, M., Ivannikova, E., Khan, S.A., Oyomno, W., Fu, Q., Tan, K.E., Flanagan, A. (2019). Federated collaborative filtering for privacy-preserving personalized recommendation system. arXiv Preprint arXiv: 1901.09888. https://doi.org/10.48550/arXiv.1901.09888
[36] Otoum, S., Al Ridhawi, I., Mouftah, H.T. (2020). Blockchain-supported federated learning for trustworthy vehicular networks. In GLOBECOM 2020-2020 IEEE Global Communications Conference, Taipei, Taiwan, pp. 1-6. https://doi.org/10.1109/GLOBECOM42002.2020.9322159
[37] Qi, K.Q., Liu, T.T., Yang, C.Y. (2020). Federated learning based proactive handover in millimeter-wave vehicular networks. In 2020 15th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 1: 401-406. https://doi.org/10.1109/ICSP48669.2020.9320974
[38] Huang, Q.S., Jiang, W.Q., Shi, J., Wu, C.Y., Wang, D., Han, Z. (2023). Federated shift-invariant dictionary learning enabled distributed user profiling. IEEE Transactions on Power Systems, 39(2): 4164-4178. https://doi.org/10.1109/TPWRS.2023.3296976
[39] Wu, C.H., Wu, F.Z., Lyu, L.Y., Huang, Y.F., Xie, X. (2022). Communication-efficient federated learning via knowledge distillation. Nature Communications, 13(1): 2032. https://doi.org/10.1038/s41467-022-29763-x
[40] Kumar, P., Gupta, G.P., Tripathi, R. (2021). PEFL: Deep privacy-encoding-based federated learning framework for smart agriculture. IEEE Micro, 42(1): 33-40. https://doi.org/10.1109/MM.2021.3112476
[41] Mehta, S., Kukreja, V., Gupta, A. (2023). Transforming agriculture: Federated learning cnns for wheat disease severity assessment. In 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 792-797. https://doi.org/10.1109/ICCES57224.2023.10192885
[42] Polaris Market Research. (2022). Federated Learning Market Share, Size, Trends, Industry Analysis Report, By Application; By Industry Vertical; By Region; Segment Forecast, 2022-2030. https://www.polarismarketresearch.com/industry-analysis/federated-learning-market.
[43] El Ouadrhiri, A., Abdelhadi, A. (2022). Differential privacy for deep and federated learning: A survey. IEEE Access, 10: 22359-22380. https://doi.org/10.1109/ACCESS.2022.3151670
[44] Kholod, I., Yanaki, E., Fomichev, D., Shalugin, E., Novikova, E., Filippov, E., Nordlund, M. (2020). Open-source federated learning frameworks for IoT: A comparative review and analysis. Sensors, 21(1): 167. https://doi.org/10.3390/s21010167
[45] Lo, S.K., Lu, Q., Paik, H.Y., Zhu, L. (2021). FLRA: A reference architecture for federated learning systems. In European Conference on Software Architecture, pp. 83-98. https://doi.org/10.1007/978-3-030-86044-8_6
[46] Ziller, A., Trask, A., Lopardo, A., Szymkow, B., Wagner, B., Bluemke, E., Nounahon, J.M., Passerat-Palmbach, J., Prakash, K., Rose, N., Ryffel, T., Reza, Z.N., Kaissis, G. (2021). PySyft: A library for easy federated learning. Federated Learning Systems: Towards Next-Generation AI, 965: 111-139. https://doi.org/10.1007/978-3-030-70604-3_5
[47] Wen, J., Zhang, Z.X., Lan, Y., Cui, Z.H., Cai, J.H., Zhang, W.S. (2023). A survey on federated learning: Challenges and applications. International Journal of Machine Learning and Cybernetics, 14(2): 513-535. https://doi.org/10.1007/s13042-022-01647-y
[48] Ghimire, B., Rawat, D.B. (2022). Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things. IEEE Internet of Things Journal, 9(11): 8229-8249. https://doi.org/10.1109/JIOT.2022.3150363
[49] Ramu, S.P., Boopalan, P., Pham, Q.V., Maddikunta, P.K.R., Huynh-The, T., Alazab, M., Nguyen, T.T., Gadekallu, T.R. (2022). Federated learning enabled digital twins for smart cities: Concepts, recent advances, and future directions. Sustainable Cities and Society, 79: 103663. https://doi.org/10.1016/j.scs.2021.103663

[50] Liu, B.Y., Lv, N.Y., Guo, Y.C., Li, Y.W. (2024). Recent advances on federated learning: A Systematic survey. Neurocomputing, 597: 128019. https://doi.org/10.1016/j.neucom.2024.128019
[51] Nasri, S.A.E. M., Ullah, I., Madden, M.G. (2023). Compression scenarios for federated learning in smart manufacturing. Procedia Computer Science, 217: 436-445. https://doi.org/10.1016/j.procs.2022.12.239
[52] Coelho, K.K., Nogueira, M., Vieira, A.B., Silva, E.F., Nacif, J.A.M. (2023). A survey on federated learning for security and privacy in healthcare applications. Computer Communications, 207: 113-127. https://doi.org/10.1016/j.comcom.2023.05.012
[53] Mothukuri, V., Parizi, R.M., Pouriyeh, S., Huang, Y., Dehghantanha, A., Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115: 619-640. https://doi.org/10.1016/j.future.2020.10.007
[54] Lalitha, A., Shekhar, S., Javidi, T., Koushanfar, F. (2018). Fully decentralized federated learning. In Third Workshop on Bayesian Deep Learning (NeurIPS), 2. http://bayesiandeeplearning.org/2018/papers/140.pdf.
[55] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V. (2020). Federated optimization in heterogeneous networks. In Proceedings of Machine Learning and Systems, 2: 429-450. https://doi.org/10.48550/arXiv.1812.06127
[56] Ma, Y.Z., Zhu, X.J., Hsu, J. (2019). Data poisoning against differentially-private learners: Attacks and defenses. arXiv Preprint arXiv: 1903.09860. https://doi.org/10.48550/arXiv.1903.09860
[57] Geyer, R.C., Klein, T., Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv Preprint arXiv: 1712.07557. https://doi.org/10.48550/arXiv.1712.07557
[58] Xu, G.W., Li, H.W., Liu, S., Yang, K., Lin, X.D. (2019). VerifyNet: Secure and verifiable federated learning. IEEE Transactions on Information Forensics and Security, 15: 911-926. https://doi.org/10.1109/TIFS.2019.2929409
[59] Phong, L.T., Aono, Y., Hayashi, T., Wang, L., Moriai, S. (2018). Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 13(5): 1333-1345. https://doi.org/10.1109/TIFS.2017.2787987
[60] Hao, M., Li, H.W., Xu, G.W., Liu, S., Yang, H.M. (2019). Towards efficient and privacy-preserving federated deep learning. In ICC 2019-2019 IEEE International Conference on Communications (ICC), pp. 1-6. https://doi.org/10.1109/ICC.2019.8761267
[61] Zhang, J.P., Zhu, H., Wang, F.W., Zhao, J.Q., Xu, Q., Li, H. (2022). Security and privacy threats to federated learning: Issues, methods, and challenges. Security and Communication Networks, 2022(1): 2886795. https://doi.org/10.1155/2022/2886795
[62] Kalra, S., Wen, J., Cresswell, J.C., Volkovs, M., Tizhoosh, H.R. (2023). Decentralized federated learning through proxy model sharing. Nature Communications, 14(1): 2899. https://doi.org/10.1038/s41467-023-38569-4
[63] Shojafar, M., Mukherjee, M., Piuri, V., Abawajy, J. (2021). Guest editorial: Security and privacy of federated learning solutions for industrial IoT applications. IEEE Transactions on Industrial Informatics, 18(5): 3519-3521. https://doi.org/10.1109/TII.2021.3128972
[64] Pokhrel, S.R., Choi, J. (2020). Federated learning with blockchain for autonomous vehicles: Analysis and design challenges. IEEE Transactions on Communications, 68(8): 4734-4746. https://doi.org/10.1109/TCOMM.2020.2990686
[65] Yin, D., Chen, Y.D., Kannan, R., Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In International Conference on Machine Learning, pp. 5650-5659. https://proceedings.mlr.press/v80/yin18a
[66] Nguyen, T., Thai, M.T. (2023). Preserving privacy and security in federated learning. IEEE/ACM Transactions on Networking, 32(1): 833-843. https://doi.org/10.1109/TNET.2023.3302016
[67] Rodríguez-Barroso, N., Jiménez-López, D., Luzón, M. V., Herrera, F., Martínez-Cámara, E. (2023). Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. Information Fusion, 90: 148-173. https://doi.org/10.1016/j.inffus.2022.09.011
[68] Touat, O., Bouchenak, S. (2023). Towards robust and bias-free federated learning. In Proceedings of the 3rd Workshop on Machine Learning and Systems, pp. 49-55. https://doi.org/10.1145/3578356.3592576
[69] Gosselin, R., Vieu, L., Loukil, F., Benoit, A. (2022). Privacy and security in federated learning: A survey. Applied Sciences, 12(19): 9901. https://doi.org/10.3390/app12199901
[70] Zhang, M.Z., Yin, R.P., Yang, Z., Wang, Y.P., Li, K. (2023). Advances and challenges of multi-task learning method in recommender system: A survey. arXiv Preprint arXiv: 2305.13843. https://doi.org/10.48550/arXiv.2305.13843
[71] Ferraguig, L., Djebrouni, Y., Bouchenak, S., Marangozova, V. (2021). Survey of bias mitigation in federated learning. In Conférence Francophone d'informatique en Parallélisme, Architecture et Système, Lyon, France. https://hal.science/hal-03343288/
[72] Valadi, V., Qiu, X., De Gusmão, P.P.B., Lane, N.D., Alibeigi, M. (2023). FedVal: Different good or different bad in federated learning. In 32nd USENIX Security Symposium (USENIX Security 23), pp. 6365-6380. https://doi.org/10.48550/arXiv.2306.04040
[73] Zhang, D. Y., Kou, Z.Y., Wang, D. (2020). Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In 2020 IEEE International Conference on Big Data (Big Data) Atlanta, GA, USA, pp. 1051-1060. https://doi.org/10.1109/BigData50022.2020.9378043
[74] Huang, Y.X., Bert, C., Fischer, S., Schmidt, M., Dörfler, A., Maier, A., Fietkau, R., Putz, F. (2022). Continual learning for peer-to-peer federated learning: A study on automated brain metastasis identification. arXiv Preprint arXiv: 2204.13591. https://doi.org/10.48550/arXiv.2204.13591
[75] Zeng, R.F., Zeng, C., Wang, X.W., Li, B., Chu, X.W. (2022). Incentive mechanisms in federated learning and a game-theoretical approach. IEEE Network, 36(6): 229-235. https://doi.org/10.1109/MNET.112.2100706
[76] Zeng, R.F., Zeng, C., Wang, X.W., Li, B., Chu, X.W. (2021). A comprehensive survey of incentive mechanism for federated learning. arXiv Preprint arXiv: 2106.15406. https://doi.org/10.48550/arXiv.2106.15406

[77] Pillutla, K., Kakade, S.M., Harchaoui, Z. (2022). Robust aggregation for federated learning. IEEE Transactions on Signal Processing, 70: 1142-1154. https://doi.org/10.1109/TSP.2022.3153135

[78] Zheng, P., Zhu, Y., Hu, Y.L., Zhang, Z.M., Schmeink, A. (2023). Federated learning in heterogeneous networks with unreliable communication. IEEE Transactions on Wireless Communications, 23(4): 3823-3838. https://doi.org/10.1109/TWC.2023.3311824

[79] Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., Zhou, Y. (2019). A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, pp. 1-11. https://doi.org/10.1145/3338501.3357370

[80] Reisizadeh, A., Farnia, F., Pedarsani, R., Jadbabaie, A. (2020). Robust federated learning: The case of affine distribution shifts. Advances in Neural Information Processing Systems, 33: 21554-21565. https://doi.org/10.48550/arXiv.2006.08907

[81] Ding, J., Tramel, E., Sahu, A.K., Wu, S., Avestimehr, S., Zhang, T. (2022). Federated learning challenges and opportunities: An outlook. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, pp. 8752-8756. https://doi.org/10.1109/ICASSP43922.2022.9746925

[82] Casado, F.E., Lema, D., Iglesias, R., Regueiro, C.V., Barro, S. (2021). Concept drift detection and adaptation for robotics and mobile devices in federated and continual settings. In Advances in Physical Agents II: Proceedings of the 21st International Workshop of Physical Agents (WAF 2020), November 19-20, 2020, Alcalá de Henares, Madrid, Spain, pp. 79-93. https://doi.org/10.1007/978-3-030-62579-5_6

[83] Manias, D.M., Shaer, I., Yang, L., Shami, A. (2021). Concept drift detection in federated networked systems. In 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, pp. 1-6. https://doi.org/10.1109/GLOBECOM46510.2021.9685083

[84] Jin, Y.L., Liu, Y., Chen, K., Yang, Q. (2023). Federated learning without full labels: A survey. arXiv Preprint arXiv: 2303.14453. https://doi.org/10.48550/arXiv.2303.14453

[85] Liu, Z.W., Huang, L.P., Fan, C., Mostafavi, A. (2023). FairMobi-Net: A fairness-aware deep learning model for urban mobility flow generation. arXiv Preprint arXiv: 2307.11214. https://doi.org/10.48550/arXiv.2307.11214

[86] Li, Q.B., Wen, Z.Y., He, B.S. (2020). Practical federated gradient boosting decision trees. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(4): 4642-4649. https://doi.org/10.1609/aaai.v34i04.5895

[87] Cao, D., Chang, S., Lin, Z.J., Liu, G.H., Sun, D.H. (2019). Understanding distributed poisoning attack in federated learning. In 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), Tianjin, China, pp. 233-239. https://doi.org/10.1109/ICPADS47876.2019.00042

[88] Jothimurugesan, E., Hsieh, K., Wang, J.Y., Joshi, G., Gibbons, P.B. (2023). Federated learning under distributed concept drift. In International Conference on Artificial Intelligence and Statistics, pp. 5834-5853. https://doi.org/10.48550/arXiv.2206.00799

[89] Harth, N., Anagnostopoulos, C., Voegel, H.J., Kolomvatsos, K. (2022). Local & federated learning at the network edge for efficient predictive analytics. Future Generation Computer Systems, 134: 107-122. https://doi.org/10.1016/j.future.2022.03.030

[90] Sery, T., Shlezinger, N., Cohen, K., Eldar, Y.C. (2021). Over-the-air federated learning from heterogeneous data. IEEE Transactions on Signal Processing, 69: 3796-3811. https://doi.org/10.1109/TSP.2021.3090323

[91] Chen, Y.Q., Qin, X., Wang, J.D., Yu, C.H., Gao, W. (2020). Fedhealth: A federated transfer learning framework for wearable healthcare. IEEE Intelligent Systems, 35(4): 83-93. https://doi.org/10.1109/MIS.2020.2988604

[92] Duan, M.M., Liu, D., Ji, X.Y., Wu, Y., Liang, L., Chen, X.Z., Tan, Y.J., Ren, A. (2021). Flexible clustered federated learning for client-level data distribution shift. IEEE Transactions on Parallel and Distributed Systems, 33(11): 2661-2674. https://doi.org/10.1109/TPDS.2021.3134263

[93] Rafiee Sevyeri, L. (2022). Tackling distribution shift-Detection and mitigation (Doctoral dissertation, Concordia University). https://spectrum.library.concordia.ca/id/eprint/991835/

[94] Ghosh, A., Chung, J.C., Yin, D., Ramchandran, K. (2022). An efficient framework for clustered federated learning. IEEE Transactions on Information Theory, 68(12): 8076-8091. https://doi.org/10.1109/TIT.2022.3192506

[95] Fallah, A., Mokhtari, A., Ozdaglar, A. (2020). Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. Advances in Neural Information Processing Systems, 33: 3557-3568. https://dl.acm.org/doi/pdf/10.5555/3495724.3496024.

[96] Sattler, F., Wiedemann, S., Müller, K.R., Samek, W. (2019). Robust and communication-Efficient federated learning from non-Iid data. IEEE Transactions on Neural Networks and Learning Systems, 31(9): 3400-3413. https://doi.org/10.1109/TNNLS.2019.2944481

[97] Peng, X., Huang, Z., Zhu, Y., Saenko, K. (2019). Federated adversarial domain adaptation. arXiv Preprint arXiv: 1911.02054. https://doi.org/10.48550/arXiv.1911.02054

[98] Duan, M., Liu, D., Chen, X., Liu, R., Tan, Y., Liang, L. (2021). Self-balancing federated learning with global imbalanced data in mobile systems. IEEE Transactions on Parallel and Distributed Systems, 32(1): 59-71. https://doi.org/10.1109/TPDS.2020.3009406

[99] Nishio, T., Yonetani, R. (2019). Client selection for federated learning with heterogeneous resources in mobile edge. In ICC 2019-2019 IEEE International Conference on Communications (ICC), Shanghai, China, pp. 1-7. https://doi.org/10.1109/ICC.2019.8761315

[100] Abay, A., Zhou, Y., Baracaldo, N., Rajamoni, S., Chuba, E., Ludwig, H. (2020). Mitigating bias in federated learning. arXiv Preprint arXiv: 2012.02447. https://doi.org/10.48550/arXiv.2012.02447

[101] Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6): 305-311. https://doi.org/10.1038/s42256-020-0186-1

[102] Tyagi, S., Rajput, I.S., Pandey, R. (2023). Federated

learning: Applications, Security hazards and Defense measures. In 2023 International Conference on Device Intelligence, Computing and Communication Technologies (DICCT), Dehradun, India, pp. 477-482. https://doi.org/10.1109/DICCT56244.2023.10110075

[103] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11): 665-673. https://doi.org/10.1038/s42256-020-00257-z

[104] Olteanu, A., Castillo, C., Boy, J., Varshney, K. (2018). The effect of extremist violence on hateful speech online. Proceedings of the International AAAI Conference on Web and Social Media, 12(1): 221-230. https://doi.org/10.1609/icwsm.v12i1.15040

[105] Fung, C., Yoon, C.J., Beschastnikh, I. (2018). Mitigating sybils in federated learning poisoning. arXiv Preprint arXiv: 1808.04866. https://doi.org/10.48550/arXiv.1808.04866

[106] Juuti, M., Szyller, S., Marchal, S., Asokan, N. (2019). PRADA: Protecting against DNN model stealing attacks. In 2019 IEEE European Symposium on Security and Privacy (EuroS&P), Stockholm, Sweden, pp. 512-527. https://doi.org/10.1109/EuroSP.2019.00044

[107] Schelter, S., He, Y., Khilnani, J., Stoyanovich, J. (2020). FairPrep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. Proceedings of the 23nd International Conference on Extending Database Technology, EDBT. https://doi.org/10.5441/002/edbt.2020.41

[108] Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G., Thorne, B. (2017). Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv Preprint arXiv: 1711.10677. https://doi.org/10.48550/arXiv.1711.10677

[109] Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R. (2018). Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. IEEE Journal of Selected Topics in Signal Processing, 12(5): 1106-1119. https://doi.org/10.1109/JSTSP.2018.2865887

[110] Lian, X., Zhang, C., Zhang, H., Hsieh, C.J., Zhang, W., Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, pp. 5336-5346. http://dl.acm.org/doi/abs/10.5555/3295222.3295285.

[111] Martinez, I., Francis, S., Hafid, A.S. (2019). Record and reward federated learning contributions with blockchain. In 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Guilin, China, pp. 50-57. https://doi.org/10.1109/CyberC.2019.00018

[112] Zhan, Y., Zhang, J., Hong, Z., Wu, L., Li, P., Guo, S. (2021). A survey of incentive mechanism design for federated learning. IEEE Transactions on Emerging Topics in Computing, 10(2): 1035-1044. https://doi.org/10.1109/TETC.2021.3063517

[113] Li, S., Ngai, E.C., Voigt, T. (2023). Byzantine-robust aggregation in federated learning empowered industrial IoT. IEEE Transactions on Industrial Informatics, 19(2): 1165-1175. https://doi.org/10.1109/TII.2021.3128164

[114] Liu, W., Xu, X., Li, D., Qi, L., Dai, F., Dou, W., Ni, Q. (2023). Privacy preservation for federated learning with robust aggregation in edge computing. IEEE Internet of Things Journal, 10(8): 7343-7355. https://doi.org/10.1109/JIOT.2022.3229122

[115] Olteanu, A., Castillo, C., Díaz, F., Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2: 13. https://doi.org/10.3389/fdata.2019.00013

[116] Vyas, D.A., Eisenstein, L.G., Jones, D.S. (2020). Hidden in plain sight-Reconsidering the use of race correction in clinical algorithms. New England Journal of Medicine, 383(9): 874-882. https://doi.org/10.1056/NEJMms2004740

[117] Raji, I.D., Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, pp. 429-435. https://doi.org/10.1145/3306618.3314244

[118] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1): 5-47. https://doi.org/10.1111/jofi.13090

[119] Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Peixoto, R.M., Guimarães, G.A.S., Cruz, G.O.R., Araujo, M.M., Santos, L.L., Cruz, M.A.S., Oliveira, E.L.S., Winkler, I., Nascimento, E.G.S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1): 15. https://doi.org/10.3390/bdcc7010015

[120] Tan, S., Taeihagh, A., Baxter, K. (2022). The risks of machine learning systems. arXiv Preprint arXiv: 2204.09852. https://doi.org/10.48550/arXiv.2204.09852

[121] Suresh, H., Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, NY, USA, pp. 1-9. https://doi.org/10.1145/3465416.3483305

[122] Liang, P.P., Liu, T., Ziyin, L., Allen, N.B., Auerbach, R.P., Brent, D., Salakhutdinov, R., Morency, L.P. (2020). Think locally, act globally: Federated learning with local and global representations. arXiv Preprint arXiv: 2001.01523. https://doi.org/10.48550/arXiv.2001.01523

[123] Bender, E.M., Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6: 587-604. https://doi.org/10.1162/tacl_a_00041

[124] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of The 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, pp. 2979-2989. https://doi.org/10.18653/v1/D17-1323

[125] Solow-Niederman, A. (2019). Administering artificial intelligence. Southern California Law Review, 93(4): 633-696. http://southerncalifornialawreview.com/wp-content/uploads/2020/09/SolowNiederman_website.pdf.

[126] Levendowski, A. (2018). How copyright law can fix artificial intelligence's implicit bias problem. Washington Law Review, 93(2): 579-630. http://digitalcommons.law.uw.edu/wlr/vol93/iss2/2/

[127] Cooper, A.F., Abrams, E., Na, N. (2021). Emergent unfairness in algorithmic fairness-accuracy trade-off research. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, USA, pp. 46-54. https://doi.org/10.1145/3461702.3462519

[128] Santiago, T. (2019). AI bias: How does AI influence the executive function of business leaders? Muma Business Review, 3(16): 181-192. https://doi.org/10.28945/4380

[129] Buolamwini, J., Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR, 81: 77-91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[130] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv Preprint arXiv: 1708.00524. https://doi.org/10.48550/arXiv.1708.00524

[131] Djebrouni, Y., Benarba, N., Touat, O., De Rosa, P., Bouchenak, S., Bonifati, A., Felber, P., Marangozova, V., Schiavoni, V. (2024). Bias mitigation in federated learning for edge computing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7(4): 1-35. https://doi.org/10.1145/3631455

[132] Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., Anderljung, M., Krawczuk, I., Krueger, D., Lebensold, J., Maharaj, T. Zilberman, N. (2021). Filling gaps in trustworthy development of AI. Science, 374(6573): 1327-1329. https://doi.org/10.1126/science.abi7176

[133] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12): 86-92. https://doi.org/10.1145/3458723

[134] Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V. (2020). Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems, 33: 6256-6268. http://papers.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf.

[135] Wu, H., Wang, P. (2021). Fast-convergent federated learning with adaptive weighting. IEEE Transactions on Cognitive Communications and Networking, 7(4): 1078-1088. https://doi.org/10.1109/TCCN.2021.3084406

[136] McNamara, D., Ong, C.S., Williamson, R.C. (2017). Provably fair representations. arXiv Preprint arXiv: 1710.04394. https://doi.org/10.48550/arXiv.1710.04394

[137] Jovanović, N., Balunovic, M., Dimitrov, D.I., Vechev, M. (2022). Fare: Provably fair representation learning. In NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research. http://neurips.cc/virtual/2022/58646

[138] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, pp. 33-44. https://doi.org/10.1145/3351095.3372873

[139] Chen, I., Johansson, F.D., Sontag, D. (2018). Why is my classifier discriminatory? Advances in Neural Information Processing Systems, 31. http://proceedings.neurips.cc/paper_files/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf.

[140] Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-14. https://doi.org/10.1145/3313831.3376445

[141] Siddique, S., Haque, M.A., George, R., Gupta, K.D., Gupta, D., Faruk, M.J.H. (2023). Survey on machine learning biases and mitigation techniques. Digital, 4(1): 1-68. https://doi.org/10.3390/digital4010001

[142] Zhuang, W., Chen, C., Lyu, L. (2023). When foundation model meets federated learning: Motivations, challenges, and future directions. arXiv Preprint arXiv: 2306.15546. https://doi.org/10.48550/arXiv.2306.15546

[143] Chen, H., Zhu, T., Zhang, T., Zhou, W., Yu, P.S. (2023). Privacy and fairness in federated learning: On the perspective of tradeoff. ACM Computing Surveys, 56(2): 1-37. https://doi.org/10.1145/3606017

[144] Caliskan, A., Bryson, J.J., Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334): 183-186. https://doi.org/10.1126/science.aal4230

[145] Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR, 81: 160-171. http://proceedings.mlr.press/v81/ensign18a.html.

[146] Kleinberg, J., Mullainathan, S., Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv Preprint arXiv: 1609.05807. https://doi.org/10.48550/arXiv.1609.05807

[147] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications, 10(1): 1096. https://doi.org/10.1038/s41467-019-08987-4

[148] Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain, pp. 53-59. https://doi.org/10.18653/v1/W17-1606

[149] Guendouzi, B.S., Ouchani, S., Assaad, H.E., Zaher, M.E. (2023). A systematic review of federated learning: Challenges, aggregation methods, and development tools. Journal of Network and Computer Applications, 220: 103714. https://doi.org/10.1016/j.jnca.2023.103714

[150] Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.L. (2018). Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv Preprint arXiv: 1811.11479. https://doi.org/10.48550/arXiv.1811.11479

[151] Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K. (2020). The Dataset nutrition label: A

framework to drive higher data quality standards. Data Protection and Privacy, 12: 1-25.

[152] Mohri, M., Sivek, G., Suresh, A.T. (2019). Agnostic federated learning. Proceedings of the 36th International Conference on Machine Learning. PMLR, 97: 4615-4625. http://proceedings.mlr.press/v97/mohri19a.html.

[153] Abdulrahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Talhi, C., Guizani, M. (2021). A Survey on federated learning: The journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal, 8(7): 5476-5497. https://doi.org/10.1109/JIOT.2020.3030072

[154] Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, pp. 1171-1180. https://doi.org/10.1145/3038912.3052660

[155] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31. http://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

[156] Mouzannar, H., Ohannessian, M.I., Srebro, N. (2019). From fair decision making to social equality. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, pp. 359-368. https://doi.org/10.1145/3287560.3287599

[157] Fredrikson, M., Jha, S., Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, Colorado, USA, pp. 1322-1333. https://doi.org/10.1145/2810103.2813677

[158] Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, pp. 59-68. https://doi.org/10.1145/3287560.3287598

[159] Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12): 866-872. https://doi.org/10.7326/M18-1990

[160] Cotter, A., Jiang, H., Gupta, M., Wang, S., Narayan, T., You, S., Sridharan, K. (2019). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. Journal of Machine Learning Research, 20(172): 1-59. http://jmlr.org/papers/v20/18-616.html.

[161] Duchi, J.C., Jordan, M.I., Wainwright, M.J. (2013). Local privacy and statistical minimax rates. In 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, Berkeley, CA, USA, pp. 429-438. https://doi.org/10.1109/FOCS.2013.53

[162] Karimi, A.H., Barthe, G., Balle, B., Valera, I. (2020). Model-agnostic counterfactual explanations for consequential decisions. Proceedings of the International Conference on Artificial Intelligence and Statistics. PMLR, 108: 895-905. http://proceedings.mlr.press/v108/karimi20a.

[163] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5): 206-215. https://doi.org/10.1038/s42256-019-0048-x

[164] Iacob, A., Sani, L., Marino, B., Aleksandrov, P., Shen, W.F., Lane, N.D. (2024). Worldwide federated training of language models. arXiv Preprint arXiv: 2405.14446. https://doi.org/10.48550/arXiv.2405.14446

[165] Katyal, S.K. (2019). Private accountability in the age of artificial intelligence. UCLA Law Review, 66(1): 54-141. http://escholarship.org/uc/item/18n256z1.

[166] Hutchinson, B., Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, pp. 49-58. https://doi.org/10.1145/3287560.3287600

[167] Bergman, A.S., Hendricks, L.A., Rauh, M., Wu, B., Agnew, W., Kunesch, M., Duan, I., Gabriel, I., Isaac, W. (2023). Representation in AI evaluations. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, pp. 519-533. https://doi.org/10.1145/3593013.3594019

[168] Jeong, W., Yoon, J., Yang, E., Hwang, S.J. (2020). Federated semi-supervised learning with inter-client consistency & disjoint learning. arXiv Preprint arXiv: 2006.12097. https://doi.org/10.48550/arXiv.2006.12097

[169] Zhang, Z., Yang, Y., Yao, Z., Yan, Y., Gonzalez, J.E., Ramchandran, K., Mahoney, M.W. (2021). Improving semi-supervised federated learning by reducing the gradient diversity of models. In 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, pp. 1214-1225. https://doi.org/10.1109/BigData52589.2021.9671693

[170] Fan, C., Hu, J., Huang, J. (2022). Private semi-supervised federated learning. In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2009-2015. http://www.ijcai.org/proceedings/2022/0279.pdf.

[171] Sheikhi, S., Kostakos, P. (2023). DDoS attack detection using unsupervised federated learning for 5G networks and beyond. In 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Gothenburg, Sweden, pp. 442-447. https://doi.org/10.1109/EuCNC/6GSummit58263.2023.10188245

[172] Lee, M.K., Kusbit, D., Kahng, A., Kim, J.T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., Procaccia, A.D. (2019). WeBuildAI: Participatory framework for algorithmic governance. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW): 1-35. https://doi.org/10.1145/3359283

[173] Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J., Schwartz, O. (2018). AI now report 2018. New York: AI Now Institute at New York University. http://www.stc.org/roundtable/wp-content/uploads/sites/34/2019/06/AI_Now_2018_Report.pdf.

[174] Du, M., Jia, R., Song, D. (2019). Robust anomaly detection and backdoor attack detection via differential privacy. arXiv Preprint arXiv: 1911.07116. https://doi.org/10.48550/arXiv.1911.07116

[175] Wang, S., Tuor, T., Salonidis, T., Leung, K.K., Makaya, C., He, T., Chan, K. (2019). Adaptive Federated Learning in Resource Constrained Edge Computing Systems. IEEE Journal on Selected Areas in Communications, 37(6), 1205-1221. https://doi.org/10.1109/JSAC.2019.2904348

[176] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T. (2020). Scaffold: Stochastic controlled averaging for federated learning. Proceedings of the 37th International Conference on Machine Learning. PMLR, 119: 5132-5143. http://proceedings.mlr.press/v119/karimireddy20a.html.

[177] Smith, V., Chiang, C.K., Sanjabi, M., Talwalkar, A.S. (2017). Federated multi-task learning. Advances in Neural Information Processing Systems, 30. Curran Associates, Inc.

[178] Chu, L., Wang, L., Dong, Y., Pei, J., Zhou, Z., Zhang, Y. (2021). FedFair: Training fair models in cross-silo federated learning. arXiv Preprint arXiv: 2109.05662. https://doi.org/10.48550/arXiv.2109.05662

[179] Saha, S., Ahmad, T. (2021). Federated transfer learning: Concept and applications. Intelligenza Artificiale, 15(1): 35-44. https://doi.org/10.3233/IA-200075

[180] Gupta, O., Raskar, R. (2018). Distributed learning of deep neural network over multiple agents. Journal of Network and Computer Applications, 116: 1-8. https://doi.org/10.1016/j.jnca.2018.05.003

[181] Jayaraman, B., Wang, L., Evans, D., Gu, Q. (2018). Distributed learning without distress: Privacy-preserving empirical risk minimization. Advances in Neural Information Processing Systems, 31. http://proceedings.neurips.cc/paper_files/paper/2018/file/7221e5c8ec6b08ef6d3f9ff3ce6eb1d1-Paper.pdf.

[182] Song, J., Kalluri, P., Grover, A., Zhao, S., Ermon, S. (2019). Learning controllable fair representations. Proceedings of The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 89: 2164-2173. http://proceedings.mlr.press/v89/song19a.

[183] Passi, S., Barocas, S. (2019). Problem formulation and fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, pp. 39-48. https://doi.org/10.1145/3287560.3287567

[184] European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union, L119: 1-88. http://dvbi.ru/Portals/0/DOCUMENTS_SHARE/RISK_MANAGEMENT/EBA/GDPR_eng_rus.pdf.

[185] Veale, M., Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2). https://doi.org/10.1177/2053951717743530

[186] Bhagoji, A.N., Chakraborty, S., Mittal, P., Calo, S. (2019). Analyzing federated learning through an adversarial lens. Proceedings of the 36th International Conference on Machine Learning. PMLR, 97: 634-643. http://proceedings.mlr.press/v97/bhagoji19a.html.

[187] Yaacoub, J.P.A., Noura, H.N., Salman, O. (2023). Security of federated learning with IoT systems: Issues, limitations, challenges, and solutions. Internet of Things and Cyber-Physical Systems, 3: 155-179. https://doi.org/10.1016/j.iotcps.2023.04.001

[188] Nguyen, T.D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., Sadeghi, A.R. (2019). DÏoT: A federated self-learning anomaly detection system for IoT. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, pp. 756-767. https://doi.org/10.1109/ICDCS.2019.00080

[189] Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., Pedarsani, R. (2020). FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics. PMLR, 108: 2021-2031. http://proceedings.mlr.press/v108/reisizadeh20a.

[190] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V. (2018). Federated learning with non-iid data. arXiv Preprint arXiv: 1806.00582. https://doi.org/10.48550/arXiv.1806.00582

[191] Du, Z., Wu, C., Yoshinaga, T., Yau, K.L.A., Ji, Y., Li, J. (2020). Federated learning for vehicular internet of things: Recent advances and open issues. IEEE Open Journal of the Computer Society, 1: 45-61. https://doi.org/10.1109/OJCS.2020.2992630

[192] Kim, H., Park, J., Bennis, M., Kim, S.L. (2020). Blockchained on-device federated learning. IEEE Communications Letters, 24(6): 1279-1283. https://doi.org/10.1109/LCOMM.2019.2921755

[193] Raji, I.D., Yang, J. (2019). About ML: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. arXiv Preprint arXiv: 1912.06166. https://doi.org/10.48550/arXiv.1912.06166

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