

Fairness in Federated Learning: Fairness for Whom?

Afaf Taik¹, Khaoula Chehbouni^{2,3}, Golnoosh Farnadi^{2,3,4}

¹ Université de Sherbrooke,

² Mila - Quebec AI Institute,

³ McGill University,

⁴ Université de Montréal

afaf.taik@usherbrooke.ca, khaoula.chehbouni@mila.quebec, farnadig@mila.quebec

Abstract

Fairness in federated learning (FL) has emerged as a rapidly growing area of research, with numerous works proposing formal definitions and algorithmic interventions. Yet, despite this technical progress, fairness in FL is often defined and evaluated in ways that abstract away from the sociotechnical contexts in which these systems are deployed. In this paper, we argue that existing approaches tend to optimize narrow system-level metrics — such as performance parity or contribution-based rewards — while overlooking how harms arise throughout the FL lifecycle and how they impact diverse stakeholders. We support this claim through a critical analysis of the literature, based on a systematic annotation of papers for their fairness definitions, design decisions, evaluation practices, and motivating use cases. Our analysis reveals five recurring pitfalls: (1) fairness framed solely through the lens of server–client architecture; (2) a mismatch between simulations and motivating use-cases and contexts; (3) definitions that conflate protecting the system with protecting its users; (4) interventions that target isolated stages of the lifecycle while neglecting upstream and downstream effects; and (5) a lack of multi-stakeholder alignment where multiple fairness definitions can be relevant at once. Building on these insights, we propose a harm-centered framework that links fairness definitions to concrete risks and stakeholder vulnerabilities. We conclude with recommendations for more holistic, context-aware, and accountable fairness research in FL.

1 Introduction

Machine learning (ML) systems are increasingly deployed in high-stakes domains such as healthcare (Hadi et al. 2023) and hiring (Qin et al. 2018), where their decisions carry significant social implications. Ensuring fairness in such systems is critical, yet difficult. While fairness in centralized ML has been widely studied (Mitchell et al. 2021; Saxena et al. 2019), collaborative ML paradigms such as Federated Learning (FL) introduce new complexities to this endeavor (Konečný 2016; Kairouz et al. 2021). In FL, model training is distributed across multiple clients (e.g., smartphones or institutions), each with their own local data and constraints. This setup brings unique challenges, including heterogeneity in data, compute, and context, as well as lim-

ited observability into client distributions or downstream impact.

To address these challenges, new fairness definitions have emerged specifically for FL (Shi, Yu, and Leung 2024), including participation fairness, performance fairness, group fairness, and collaborative fairness. Several recent surveys have documented these proposals (Shi, Yu, and Leung 2024; Vucinich and Zhu 2023; Annareddy, Preston, and Fox 2023; Salazar et al. 2024; Ude et al. 2023; Rafi et al. 2024; Chen et al. 2024; Wang et al. 2024b; Balbierer et al. 2024). While this body of work provides valuable taxonomies and technical insight, it often treats fairness as a standalone optimization problem, isolated from the social context in which FL systems are deployed.

In this paper, we argue that much of the fairness in FL literature suffers from a critical abstraction error (Selbst et al. 2019). Fairness is often defined and operationalized in terms of narrow system-level metrics, such as minimizing performance variance across clients or allocating rewards based on contribution, without attention to the structural inequalities, institutional dynamics, or real-world harms that shape FL participation and impact. These abstractions risk obscuring who is protected, what harms are being addressed, and whose interests are prioritized.

To examine these concerns, we adopt a harm-centered lens and conduct a systematic annotation of 121 papers on fairness in FL. We map fairness definitions, intervention points, and evaluation practices to the FL lifecycle—from problem formulation to deployment—and identify recurring gaps in how fairness is conceptualized and operationalized. Our findings reveal five core pitfalls: (i) a narrow focus on server–client architecture; (ii) a disconnect between fairness definitions and simulated use-cases; (iii) the conflation of system interests’ protection with user protection; (iv) a limited intervention scope for mitigation techniques; and (v) a lack of multi-stakeholder alignment where many fairness definitions could be relevant.

We also highlight how fairness efforts in FL interact with other concerns such as privacy and robustness—sometimes introducing new harms, such as disproportionate privacy degradation for minorities (Ling et al. 2024), or the misclassification of underrepresented clients as adversaries (Touat and Bouchenak 2023).

Furthermore, to address these gaps, we propose a harm-

centered framework for guiding context-aware fair FL solutions. Our framework maps the FL lifecycle—spanning problem formulation, client selection, aggregation, evaluation, and deployment—against different sources of harm and stakeholder vulnerabilities. It encourages researchers to ask: Who is (or could be) harmed at this step? What kinds of errors or exclusions may occur? And how do these relate to existing fairness definitions? This approach offers a more grounded and context-aware lens for evaluating fairness claims and designing interventions.

Our contributions are twofold:

- We provide a critical, annotated review of the fairness in FL literature, uncovering patterns in definitions, motivations, and lifecycle coverage.
- We propose a harm-centered framework for fairness in FL, connecting technical decisions to stakeholder impact, and offering actionable recommendations for more accountable, context-aware research.

2 Background

Federated learning (FL) is a decentralized ML paradigm in which a global model is trained collaboratively across multiple participants (e.g., smartphones, hospitals, or institutions), without exchanging raw data. Clients train the model locally and share updates (e.g., gradients or model weights) with a central server, which aggregates them to produce an updated model. This process repeats over multiple communication rounds (Kairouz et al. 2021). Depending on the type of participants, we refer to the setting as cross-silo FL for institutional participants, and cross-device FL when clients are massively distributed IoT and mobile devices. This distinction helps move beyond the abstract client-server model by grouping use cases with similar resource constraints and aligning them with the types of stakeholders both operating the clients and affected by the resulting model—for example, mobile users both provide data and are affected by the model in cross-device FL, whereas in cross-silo FL, institutions (e.g., banks, hospitals) operate as the clients, but use their models to make decisions impacting their customers.

While FL is often introduced as a privacy-preserving alternative to centralized ML, its distributed nature introduces new technical and ethical challenges. In particular, fairness in FL must be reconsidered in light of structural asymmetries between clients. These asymmetries are not incidental; they are deeply ingrained into the realities of FL deployment.

2.1 Federated Learning Lifecycle

Although implementations of FL vary across settings and applications, most systems follow a common lifecycle that structures the learning process and coordination between clients and the central server. The following outline represents a basic FL lifecycle. While researchers have proposed many variations—including hierarchical (Abad et al. 2020) and clustered (El-Rifai et al. 2025) setups, and personalization training schemes (Tan et al. 2022)—these core steps remain the most widely adopted and studied in practice:

1. **Problem Formulation:** The task to be solved is selected, formalized, and translated into an ML problem.

This includes defining objectives, choosing input/output formats, loss functions, and defining the different constraints.

2. **Model Initialization:** An initial global model is defined—often using a standard architecture and optionally initialized with pretrained weights. This phase also includes setting key hyperparameters (e.g., learning rate, batch size), and establishing the stopping criterion (e.g., number of rounds or convergence threshold). Then, in each iteration, referred to as communication round, the following steps 3 to 5 are repeated until the stopping criterion is met.
3. **Client Selection:** During training, a subset of clients (e.g., institutions, mobile devices) is selected in each round based on varying criteria such as availability, resources, or data attributes. These choices affect which clients contribute to the model.
4. **Local Training:** Clients train the model locally on their private data. Key design decisions at this stage include number of local epochs, optimizer choices, and any personalization applied before sending updates.
5. **Model Aggregation:** The central server aggregates client updates (e.g., using weighted averaging). The aggregation strategy and timing (synchronous/asynchronous) influence which updates contribute and how they are valued.
6. **Evaluation:** The model is evaluated—either centrally, on a held-out dataset, or by clients locally. Evaluation choices affect which metrics are prioritized and accounted for.
7. **Deployment and Incentives:** The final model (or personalized variants) is deployed to clients or users. In some settings, reward mechanisms may be used to incentivize participation, often based on estimated contributions.

In addition to this cycle, further mechanisms are often added to enhance privacy and guarantee adversarial robustness.

2.2 Why Fairness in FL is Challenging

FL is fundamentally shaped by two structural properties of FL systems: *heterogeneity* and *scarcity*.

Heterogeneity refers to differences across clients along three critical dimensions (Molamohammadi et al. 2023):

- *Data heterogeneity:* Local datasets may differ in size, label distribution, feature representation, or demographic composition.
- *Resource heterogeneity:* Clients vary in computational capacity, network stability, and energy constraints.
- *Contextual heterogeneity:* Clients operate under diverse legal, institutional, or social conditions that shape what data can be collected or shared.

Scarcity, meanwhile, refers to limited availability of data and compute resources—especially among clients serving minority populations or operating under infrastructural constraints. Scarcity amplifies the risk that some clients will be

consistently underrepresented or excluded from the training process.

Together, heterogeneity and scarcity lead to two key limitations for FL: (i) *Partial participation*: Not all clients participate equally or consistently, leading to sampling biases in model updates; (ii) *Limited visibility*: The server typically lacks access to protected attributes, demographic breakdowns, or downstream outcomes—making it difficult to measure or enforce fairness during training.

2.3 How Fairness is Defined in FL Research

Across the literature, several fairness notions have been proposed for FL (Shi, Yu, and Leung 2024), often adapted from centralized ML or cooperative systems. These include:

- **Performance-centered fairness**: The dominant definition of performance fairness found in the FL literature (Li et al. 2020; Yan et al. 2024; Xu et al. 2022; Chu et al. 2024a; Cong et al. 2024) refers to achieving similar or comparable model performance across different clients or groups of clients. A model ω_1 is considered fairer than model ω_2 with respect to the participants if, for a metric of interest (e.g., accuracy, loss), it has a smaller variance across clients. This definition can also be applied when comparing multiple client groups. Two additional definitions were proposed in this category: 1) Rawlsian definition of fairness, where the objective is to optimize the performance of the worst performing clients or group of clients (Öksüz et al. 2024; Mohri, Sivek, and Suresh 2019); and 2) individual fairness, where similar clients should have similar performance (Mashhadi et al. 2022).
- **Group fairness-inspired definitions**: Group fairness in ML emphasizes that predictions should not disadvantage underrepresented or unprivileged groups based on sensitive attributes such as race or gender. Building on these principles, researchers have sought to adapt classical group fairness criteria—such as equal opportunity, equalized odds, or demographic parity—to the federated learning setting, applying them both to global aggregates and to individual client models (Shi, Yu, and Leung 2024; Salazar et al. 2024). While these definitions are applicable to cross-device settings, they are mostly studied in the context of cross-silo collaborations, where institutions develop models that need to be fair to the populations they serve.
- **Collaborative fairness**: Ensuring that clients receive rewards, utility, or model quality proportionally to their contributions. This definition of fairness is particularly relevant for cross-silo FL settings, where data owners often compete within the same market and face increased risks from malicious actors and free-riders. Yu et al. (2020) define three fairness criteria for a fair distribution of the rewards: (1) contribution fairness: the data owner’s payoff should be positively related to its contribution, (2) regret distribution fairness: the difference of regret among data owners should be minimized and (3) expectation fairness: the variability of data owners’ regret should be minimized. Here, regret refers to the difference between the reward already received by the data

owner and what they are supposed to receive.

- **Participation fairness**: Including under-represented and never-represented clients (Shi, Yu, and Leung 2024) in the training process. Client selection in FL (Cho, Wang, and Joshi 2022) is among the most studied steps of the FL lifecycle, as the subset selection can determine who gets to influence the model and how fast the FL model will converge. Partial participation and uneven representation of client groups might appear to be the easiest aspect to assess. Yet, given the complex and heterogeneous nature of the clients, in addition to the private nature of their data and other properties, it is challenging to determine which properties should guide client selection and how to assess the fairness of this process. The most straightforward approach is to define it at an individual level, where all clients are given an equal chance to contribute. This is often formulated as a long-term fairness constraint (Huang et al. 2021), which requires setting a minimum threshold on the number of times a client needs to be included in the FL process. Another option is to define client groups based on the general knowledge of the system, such as geographical location (Lee, Kim, and Joo 2024; Saputra et al. 2019) (e.g., applications that might be impacted by the weather or the culture), or language preferences (e.g., next word prediction) (Hard et al. 2018).

While these definitions offer technically tractable objectives, they are often introduced in isolation from questions of power, structural inequality, or downstream harm. In the following section, we examine how these fairness notions fall short when applied without regard to the use case context, stakeholders involvement and systemic constraints.

3 Method

To inform our position, we conducted a systematic literature review covering 121 papers that focus on bias and fairness in FL. We used DBLP to find papers with keywords (‘Fair*’+ ‘Federated learning’), (‘Bias’+ ‘Federated Learning’), which were published up until December 2024. As FL is multidisciplinary, we did not limit our search to ML venues, but also included papers from networking/ telecommunications and economics venues. We filtered out: (a) Comments, abstracts, thesis, tutorials, slides, and other reports which are not research papers, (b) Papers that use of term “bias” in other contexts, such as bias-variance trade-off, (c) duplicates, and duplicates under different titles, (d) papers in predatory venues, and those behind paywalls and inaccessible through the authors’ institutions.

Categorization: We manually categorize the papers depending on the fairness definition that was used, i.e., participation fairness, performance fairness, group fairness, and collaborative fairness. Note that many of these papers do not explicitly use this terminology, and that some overlap exists across the definitions. Further filtering was done for papers that explore the intersection with privacy and adversarial robustness, which we considered as a separate category when needed, although they might use the same fairness definitions.

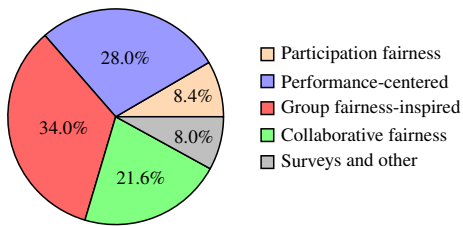


Figure 1: Fairness paradigms covered in our analysis.

Annotation: For each paper, we identified (1) the motivations of the paper and if it came with real-world examples, (2) the fairness definitions and the considered stakeholders, (3) the proposed interventions and which stages of the FL lifecycle they focus on; and (4) the datasets used in the evaluation.

4 From Definition to Deployment: Five Pitfalls in Federated Learning Fairness Research

Despite increasing attention to fairness in FL, many existing approaches operate with narrow or conflicting definitions, often optimized for system-level performance or developer convenience rather than equity for clients or end users. Drawing from our annotation of 121 papers, we identify five recurring pitfalls that hinder meaningful fair FL research and development.

4.1 Abstract System Formulation: Narrow Focus on Client-Server Architecture

One common observation is that fairness in FL is defined without explicit articulation of *who is being protected*, *what harms are being mitigated*, or *whose values are being encoded*. Definitions are often not related to stakeholder groups and contextual risks and harms. Several papers define fairness objectives without specifying whether they are optimizing for institutional clients or individual users, nor do they specify in which use-cases their approaches would be applicable.

FL is often described as a two-party protocol between a central server and distributed clients. This technical framing—rooted in system architecture—focuses on communication efficiency (Lim et al. 2020), model convergence (Lu et al. 2024), and privacy guarantees (Yin, Zhu, and Hu 2021) between these two roles. However, real-world FL deployments rarely involve just servers and clients. Instead, they are embedded in larger sociotechnical systems with multiple stakeholders, each with distinct roles, interests, and vulnerabilities (Chehbouni et al. 2025; Antunes et al. 2022).

For example, in cross-silo FL between healthcare institutions, the “client” may represent an entire hospital, but the downstream impact of the model affects patients, physicians, and administrators. Similarly, in financial applications, in addition to customers on whom decisions apply, regulators and auditors may also have a stake in how models are trained and evaluated. Developers, platform providers, and even the

designers of benchmark datasets exert influence over what is optimized, who is included, and what trade-offs are made.

This framing risks can lead to what (Selbst et al. 2019) call an “abstraction error”: treating fairness as a property of a technical subsystem, rather than of the broader sociotechnical system it serves. Justice is not an outcome of optimizing model convergence or performance variance in isolation; it requires attention to several aspects, especially power dynamics and harm. It requires understanding who is behind each abstract client, who bears the risk when performance degrades, and who gets excluded when design decisions are made. Treating clients as impersonal and theoretical entities misses the nested and the broader institutional, social, and economic factors that shape FL. This lack of clarity is not just a technical oversight. It reflects a deeper failure to treat fairness as a *sociotechnical* concept in this subset of the research community.

4.2 Abstract Evaluation: Synthetic Clients vs Real Harms

There is a significant mismatch between the concept that researchers aim to measure (fairness) and the measurement methods they use to evaluate it. This mismatch appears in two ways: the creation of synthetic clients focused solely on data heterogeneity, and the use of datasets unrelated to the motivating use-cases.

The majority of the annotated papers evaluate their framework using centralized ML datasets (e.g., MNIST, CIFAR-10, Adult, or COMPAS), by splitting them into synthetic “clients.” For instance, for group fairness papers, many studies cite real-world applications in finance, justice, and healthcare, but evaluation typically relies on centralized ML datasets like Adult ((Kohavi and Becker 1996)), COMPAS ((Angwin et al. 2016)), and CelebA ((Liu et al. 2015)). A few use the more realistic ACS dataset ((Ding et al. 2021)), which models demographic variation across U.S. states, and only one paper both motivates and evaluates using real-world federated data ((Chu et al. 2024b)). While useful for prototyping, this practice obscures the sociotechnical context in which fairness actually matters. This is particularly concerning given that federated learning introduces unique challenges not present in centralized settings (§ 2.2).

Only a small fraction of the literature uses datasets from real-world federated settings or evaluate across realistic data from silos or clients. In our analysis, we found that fewer than 10% of papers evaluated their fairness methods on domain-specific data.

For instance, in *performance fairness* papers, healthcare is the most commonly cited motivating use case, with the goal of reducing disparities across silos (e.g., hospitals) (Yan et al. 2024; Xu et al. 2022). Similar concerns apply to cross-device FL, especially in personalized applications like mobile phones (Hard et al. 2018), where QoS must be consistent across clients. Yet, most research is evaluated on ML benchmarks like MNIST and CIFAR10, simulating heterogeneity through class imbalance or rotated images. These datasets do not capture real-world distributional shifts, which are shaped by demographic, geographic, or temporal factors, nor do they capture all the domain-specific

challenges that come with the application. Instead, fairness is reduced to an oversimplified distributional problem, where the minority class serves as a proxy for real-world disadvantaged groups. Only five performance fairness papers evaluate on domain-specific datasets (Yan et al. 2024; Xia et al. 2024; Xu et al. 2022; Mashhadi et al. 2022; Selialia, Chandio, and Anwar 2022), mostly in healthcare.

Likewise, of the 26 papers on *collaborative fairness*, only five are grounded in real use cases (Maheswari et al. 2024; Albaseer et al. 2024; Xu, Nanda, and Liang 2024; Donahue and Kleinberg 2021; Lu et al. 2022), three offer motivating examples (Lyu, Xu, and Wang 2020; Yu et al. 2020; Liu et al. 2023), and just one (Albaseer et al. 2024) conducts experiments using data relevant to its stated application.

While these setups enable reproducibility and control, they abstract away the institutional and social contexts in which FL systems operate. As the technical interventions for fairness in FL are evaluated only over synthetic clients, it is easy to miss the real impact of these solutions.

Additionally, clients are modelled as interchangeable units (i.e., similar resources, contexts), as the focus is predominantly on data heterogeneity. The abstraction of equally resourced, equally motivated, or equally impacted, hides some of the inequalities fairness mechanisms are meant to address.

4.3 Protection for the System, Not the Vulnerable

Another trend we observed is the prioritization of fairness definitions that aim to protect the system from free riders and adversaries, rather than the people the system serves. *Collaborative fairness*, which represents almost a third of the annotated papers, allocates rewards based on contribution to global utility, often estimated via Shapley values. While this may seem reasonable in settings with symmetric power (e.g., corporate collaboration), in reality, it easily breaks down, as participation is shaped by unequal access to resources and heterogeneous data and contexts. In such settings, clients with noisy or underrepresented data would be penalized not because their data is intentionally low quality, but because it simply deviates from the majority’s data. These frameworks’ claimed goal is to provide structured methods for evaluating contributions and distributing rewards, and by doing so, they may inadvertently prioritize individual gains over collective outcomes. This individualistic perspective can conflict with the inherently cooperative nature of FL, where the primary objective is to collaboratively train a high-quality global model that benefits *all* participants.

This becomes obvious with how collaborative fairness papers assign degraded models or lower rewards to clients deemed “low contributors” (Lyu, Xu, and Wang 2020; Yi et al. 2024; Adamek and Darup 2024), without considering that those clients might face structural barriers to participation. This logic is particularly troubling in high-stakes domains like healthcare or finance—often cited as motivating use-cases for this fairness definition—where a poorly performing model may harm patients or customers rather than the institution itself. Ultimately, it is the patient or customer who pays the price.

Although free-riders present a potential risk, tackling such issues aligns more closely with adversarial robustness than algorithmic fairness. Collaborative fairness papers appear to prioritize adversarial robustness under the guise of protecting fairness for “honest” clients. Framing these mechanisms as fairness-preserving obscures the reality that they primarily aim to safeguard the system from misuse. While adversarial robustness frameworks often acknowledge the risk of false positives and incorporate mechanisms for correction (Touat and Bouchenak 2023), collaborative fairness schemes frequently cast exclusionary decisions, such as penalizing low-contribution clients, as inherently just. This framing leaves little room to question whether such clients are disadvantaged due to resource constraints rather than bad faith (Green 2021).

4.4 Disconnected interventions: One-Stage Solutions to Lifecycle Harms

Existing interventions in FL in general, and fairness in FL in particular, tend to concentrate on a limited subset of the development pipeline, most notably aggregation schemes (Qi et al. 2024) and client selection (Fu et al. 2023). However, other critical stages such as problem formulation, model initialization, and evaluation remain largely overlooked. This narrow focus creates blind spots: biases introduced in earlier stages may propagate unchecked, and the criteria used to assess fairness may fail to capture the outcomes that matter most. Our analysis reveals that few papers adopt a lifecycle perspective, which is essential to understanding how harms emerge and to designing effective and equitable interventions. Most interventions are localized on a few steps while others are overlooked, as illustrated in Fig 2.

Indeed, participation fairness papers focus on client selection (Cho, Wang, and Joshi 2022; Huang et al. 2021; Javaherian et al. 2024a) and asynchronous model aggregation (Gao et al. 2024; Wang et al. 2024a). While a large portion of the surveyed literature in group fairness in FL (Ezzeldin et al. 2021; Makhija et al. 2024; Roy, Sharma, and Salekin 2024; Yue, Nouiehed, and Kontar 2021; Meerza et al. 2024; Papadaki et al. 2022; Selialia, Chandio, and Anwar 2023; Abay et al. 2020; Li et al. 2024b) proposed techniques that combine adapting local training (e.g., using regularizers, adversarial debiasing) with another intervention at the server-level during model aggregation (e.g., reweighting). For this line of work, only one intervention was proposed for client selection (Zhang, Kou, and Wang 2020), and one at the personalization stage (Chu et al. 2024b). Meanwhile, to achieve performance fairness, the proposed solutions span a larger portion of the FL pipeline compared to group fairness, but the different steps have received varying attention. The interventions include model initialization (Chu et al. 2024a), client selection (Javaherian et al. 2024b), adjusting local training (Cong et al. 2024), model aggregation (Li et al. 2020), or personalization techniques (Xie et al. 2024) such as clustered FL (Chu et al. 2022). Despite these contributions, the disjointed nature of current interventions leaves key sources of harm unaddressed; without a holistic, lifecycle-oriented approach, fairness efforts risk being reactive, fragmented, and insufficiently adapted to the clients heterogeneity.

4.5 Disconnected Impact: No Single Fairness Definition is Enough

In many real-world FL scenarios, fairness cannot be reduced to a single definition. Each of the paradigms discussed (i.e., performance-centric, group fairness, and collaborative fairness) captures only one axis of concern. Treating them as mutually exclusive leads to incomplete or even counterproductive interventions. In our analysis, few papers considered multiple fairness at once (Amiri et al. 2022; Gálvez et al. 2021), however, these papers were mostly exploring the impact of privacy techniques on performance and group fairness, rather than their interplay. Instead, we believe a holistic evaluation for different fairness aspects should be adopted. FL is a multi-stakeholder system, thus, it is important to evaluate the harm to each stakeholder.

Consider a cross-silo FL deployment in the financial sector; one of the few recurring examples in the three different parts of the annotated literature. Multiple banks may collaborate to train a shared fraud detection model. A performance-centric definition might require that the global model perform well for all participating institutions. However, each institution may serve different customer bases, raising concerns about group fairness: will the model perform equally across demographic subgroups within each silo? At the same time, institutions have competing interests. Supposing that the institutions are equally powerful, a so-called collaborative fairness mechanism may be considered to ensure that no institution intentionally degrades the performance for others. In this example, the fairness concerns are not only relevant, they are entangled. Optimizing for just one could undermine the others. For instance, maximizing global performance may disproportionately benefit larger clients; enforcing group fairness locally may reduce global accuracy; prioritizing contribution-based rewards may punish institutions that serve minority populations.

This illustrates a broader point: fairness in FL is not a property of a single metric or stage in the pipeline. It is an emergent property of the full system, shaped by technical design choices, institutional power dynamics, and the social roles of clients and end users. Addressing fairness meaningfully requires acknowledging this multidimensionality/multi-stakeholder nature, and moving beyond one-size-fits-all definitions.

5 A Harm-Centered Framework for Fairness in Federated Learning

ML systems can lead to harms such as unequal performance across demographic groups, inequitable resource distribution, or the reinforcement of stereotypes (Shelby et al. 2023). FL inherits these risks from traditional ML while introducing new ones due to its decentralized structure and structural asymmetries across clients. To design fairer FL systems, we must shift from identifying isolated biases to understanding how harm emerges from developer decisions throughout the lifecycle, influenced by *heterogeneity* (in data, context, resources) and *scarcity* (of data and compute power).

Following our analysis in Section 4, we propose a framework for identifying potential harms and guiding fairness in-

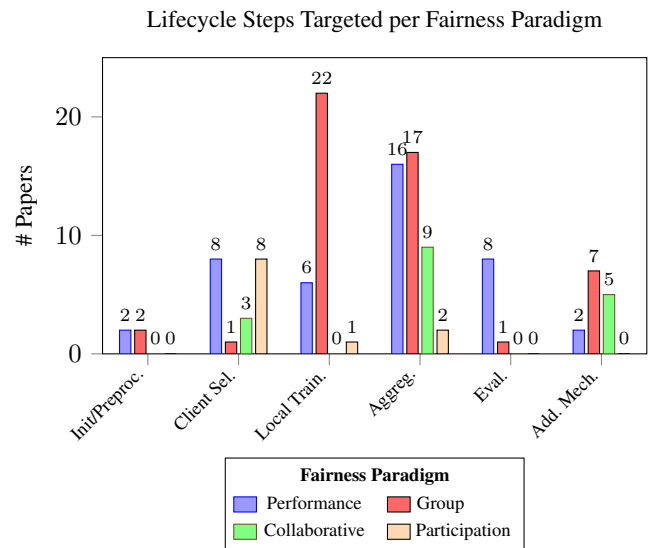


Figure 2: Lifecycle steps addressed in fairness papers, by fairness paradigm.

terventions throughout the FL pipeline. Rather than viewing bias as a fixed category, we examine how developer choices, constrained by heterogeneity across clients, may introduce harms such as quality of service (QoS), allocative, and representational harms. Below, we walk through each stage of the FL lifecycle and highlight where structural conditions and technical design interact to produce or amplify harm. We illustrate our framework in Figure 3.

5.1 Federated Learning Lifecycle Decisions and Harms

Developers have little control and visibility on data collection and local model training, in comparison with centralized settings. However, different decisions that are made throughout the FL lifecycle propagate existing biases and introduce new ones. Indeed, data generation and collection introduces 1) **historical bias**: Since data reflects the cultural, institutional, political, and socioeconomic context surrounding its extraction, it can encompass past or existing structural inequalities and injustices (e.g., COMPAS dataset (Angwin et al. 2016)); 2) **representation/ selection bias** where these biases arise when training data is not representative of the real-world population or use-case distribution (e.g., ImageNet data (Yang et al. 2020)); and 3) **measurement bias**: Such bias is introduced by how features or labels are collected, chosen, and used (e.g., faulty sensors, inadequate proxies (Kleinberg et al. 2018)). Such biases can be amplified during local training, introducing **learning bias**. In the following, we focus on the different steps where there is more control and visibility, as well as biases that are specifically tied to FL’s collaborative nature.

Problem Formulation Developers often retain full control over how problems are framed, including defining learning objectives, inputs and outputs, and success metrics (Passi and Barocas 2019; Suresh and Guttat 2021). In FL, contex-

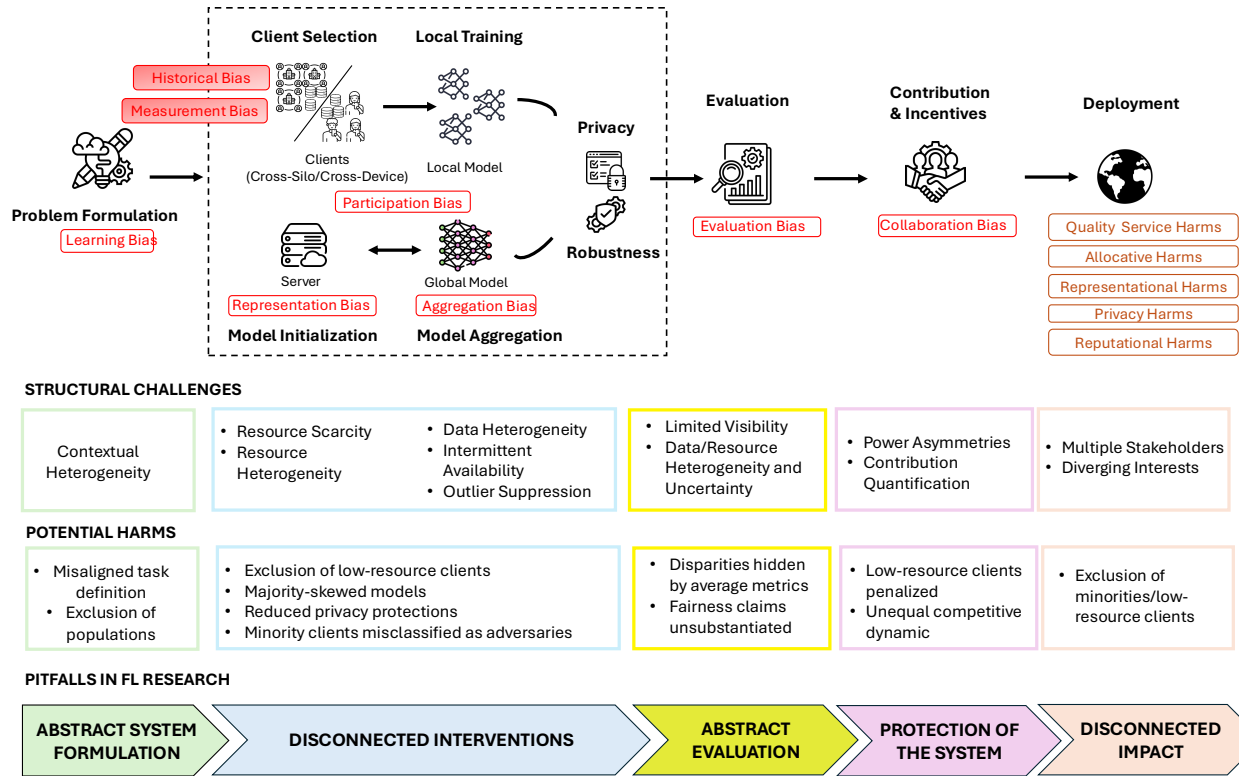


Figure 3: Overview of our harm-centered framework for fairness in FL. At each step of the FL lifecycle, we identify biases within stakeholders control, or outside of their control (red-filled boxes), structural challenges and potential harms that might emerge at each step.

tual heterogeneity across clients, such as differing institutional goals or regional constraints, complicates this step. When problem formulations assume a universal objective, they may be misaligned with local needs (Raji et al. 2022). For instance, health prediction tasks that ignore regional disease prevalence or local medical practices risk producing models that systematically underperform for some populations (Asiedu et al. 2024). Additionally, oversimplified fairness criteria may be selected for convenience rather than appropriateness (Simson, Fabris, and Kern 2024), leading to **learning bias** (Suresh and Guttag 2021). Thus, rather than just asking technical questions about the solution, moving away from the client-server abstraction is necessary. Considering *who this model is for?* *what goals are embedded in this task?* and *whose outcomes matter?* can positively impact the formulation to be more applicable.

Model Initialization Model choices, including size, architecture, and pretraining, are often made before the start of the training, and in tandem with problem formulation. However, these decisions can exclude clients with limited resources, especially in cross-device FL, where large mod-

els may not run on older or lower-powered devices (McMahan et al. 2017; Bonawitz 2019). This creates **participation bias**, where only well-resourced clients can meaningfully contribute to training. Similarly, using a pre-trained model may introduce **representation bias** if the pretraining dataset does not reflect the data distributions of FL clients (Wang et al. 2023; Chuang et al. 2023; Kaur and Jadhav 2023). As a result, it is important to consider the possibility that *the chosen model architecture/ pretrained model disadvantage clients with limited resources or underrepresented data.*

Client Selection While many aspects about data collection are not under the developers’ control, especially in cross-device FL, some of the biases and harms found in data collection are shifted to client selection.

In cross-device FL, the ideal scenario would include a full participation from all clients at each communication round. Nonetheless, clients are often offline, have low battery/ poor connectivity, or are unavailable (Tak and Cherkaoui 2021). Additionally, privacy concerns in FL due to memorization remain present (Thakkar et al. 2020). If clients participate too often, their data may be over-exposed during model up-

dates. This leads to the ideal to be purely random/ round robin selection. Moreover, the most adopted FL aggregation scheme is synchronous, meaning that if a client does not send their update before a set deadline, their update is discarded. Additionally, the communication bandwidth is a bottleneck for FL, hence only a few clients can participate in each round. Thus, optimizing the client selection step has become a key strategy in FL (Nishio and Yonetani 2019; Zhu et al. 2022; Cho, Wang, and Joshi 2022). The proposed schemes often try to maximize the number of collected updates in order to improve the performance of the global model. While these schemes might accelerate the convergence of the training on average, these schemes are often biased towards clients with more powerful devices and better connectivity. Slow clients are often labeled as stragglers and dropped out of the training round. On the opposite side of the spectrum, other strategies focus on data properties, where clients are selected in proportion to the size or quality of their data, meaning that clients with more data (or potentially more valuable data) are more likely to be selected (Goetz et al. 2019). However, data size is not a good proxy for the meaningfulness of a contribution due to redundancy (Tak and Cherkaoui 2021). Moreover, clients with larger datasets (e.g., big companies or institutions with more resources) may dominate the model, potentially drowning out smaller clients or those with less data but more diverse or unique insights.

Client selection strategies that overlook resource and data heterogeneity and their interplay are how **participation bias** could manifest in FL. Consequently, it is important to ask *Are some clients less likely to participate? Are selection rules exclusionary?*

Model Aggregation FL aggregation combines updates from clients to form a global model. There are two decisions that developers make for this stage: *When to aggregate?* and *How to weigh the updates?* For the first question, synchronous aggregation is the most commonly adopted scheme. It requires that model updates are averaged or combined at fixed intervals, with updates being sent after the fixed deadlines being ignored. This inevitably exacerbates **participation bias**, as clients with slower internet connections or lower computational power might not be able to update their models in time for synchronization (Gao et al. 2024). Over time, this would lead to a model that better represents the data from faster clients but fails to accurately capture the diversity and variability of the entire population of clients.

The second decision is choosing the updates weighting scheme. The first proposed and most adopted method (McMahan et al. 2017) requires weighting models based on the size of each client’s dataset. Such technique amplifies aggregation bias by prioritizing dominant clients and suppressing minority contributions. Over time, this leads to models that reflect majority distributions and perform poorly for smaller, less represented clients, leading to **aggregation bias**. Different aggregation schemes were proposed (Qi et al. 2024), tackling various challenges in FL, including robustness and asynchronous updates. This step is also a popular

target of group fairness in FL techniques.

As a decisive step, it is important to ask *1) How does the aggregation deadline affect which clients are able to participate, and which are excluded?, and 2) How does the aggregation weighting scheme influence the fairness outcome being targeted?*

Evaluation FL evaluations often rely on benchmark datasets or average client performance. However, due to limited visibility into local data, these metrics may obscure disparities (Lai et al. 2022). Similarly to centralized ML, this can be referred to as **evaluation bias**, where the methods or conditions used to assess a model’s performance do not accurately represent the real-world scenarios in which it will be deployed. A model that performs well on average may still harm specific clients, especially if fairness is not evaluated across subgroups. Additionally, relying solely on benchmarks and simulated environments can further distort conclusions. While these tools offer necessary step for developers to prototype, these remain controlled conditions with standardized datasets and predefined assumptions, whereas real-life conditions would differ significantly. Thus it is necessary to adopt detailed metrics (e.g., performance, client participation and drop-out, group-fairness metrics), and to continually monitor the various aspects of the FL process after deployment.

For this part we should ask *Which metrics are we using? On which data do we evaluate? For whom do these metrics matter? and how often do we need to assess the system’s performance?*

Contribution and Incentives Incentive mechanisms in FL aim to reward clients based on their contributions to the global model. However, measuring “contribution” remains an open and contested problem. Existing approaches often favour clients whose data aligns with dominant patterns or leads to measurable performance gains in global utility. Meanwhile, clients providing rare or minority data, which may be critical for generalization and group or performance fairness, are undervalued or excluded. This creates a feedback loop that reinforces inequality and marginalizes those already underrepresented, we refer to this issue as **collaboration bias**.

In FL, clients already bear real costs—computational, communicational, and sometimes organizational. In theory, participation should offer sufficient returns to justify these investments, whether through improved local model performance, access to the global model, or monetary compensation. While improving contribution evaluation is a meaningful technical objective, it also risks missing the larger question: *What purpose do additional incentives serve in the first place?* To move forward, future work must grapple with contextual questions: *Who is affected by the final model? What are the actual costs borne by clients? How would an incentive scheme shape participation and power within the federation?* Without answers to these questions, incentive design risks introducing new harms—through **collaborative bias** where incentive mechanisms introduce misattributed rewards and the punishment of honest but nonconforming clients.

Privacy and Robustness Mechanisms Privacy guarantees and robustness defences are essential in FL, but they are not neutral.

Privacy: As keeping the data local was proven not enough for privacy protection, a combination of FL with privacy preserving mechanisms such as differential privacy (DP) and secure aggregation have been adopted in the literature and practice through the use of DP-SGD for instance. However, balancing privacy and fairness presents a challenge. On one hand, DP degrades the performance of the models, and disproportionately affect underrepresented groups (Bagdasaryan, Poursaeed, and Shmatikov 2019; Ling et al. 2024). On the other hand, prioritizing fairness can heighten privacy risks due to revealing sensitive attributes and memorization (Pentyala et al. 2022). Given these tradeoffs and disparate impacts on privacy and performance, it is important to consider *what are the side effects of the privacy mechanisms?* and ask if *everyone is equally protected?*

Adversarial Robustness: Adversarial attacks in FL, such as model/data poisoning, free riding, and backdoor attacks, are often detected through anomaly detection or using similarity metrics, using techniques like reputation systems Lyu, Xu, and Wang (2020), performance evaluation, and statistical tests (e.g., clustering or outlier detection) (Sattler, Müller, and Samek 2020). However, due to data heterogeneity, falsely identifying a legitimate client as an adversary is a common risk with significant consequences (Touat and Bouchenak 2023; Ren et al. 2024). Misclassified clients may face unjust reputation damage, exclusion from future tasks, financial losses tied to missed incentives, and reduced opportunities for collaboration. These outcomes can erode trust, foster alienation, and ultimately discourage participation in FL. When designing robustness mechanisms, developers must ask: *What are the consequences of a false-positive detection—and is the cost of excluding a legitimate client greater than the risk posed by an undetected adversary?* It is thus necessary to ensure that these additional mechanisms *do not introduce additional harms*, while providing *equal protection* to different stakeholders.

5.2 Downstream Harms in Federated Learning

FL inherits many of the fairness concerns from centralized ML but introduces new layers of complexity due to its distributed nature and client heterogeneity. In this section, we examine different types of harms that can arise in FL, and we map each to the corresponding biases and stages in the FL lifecycle that contribute to them. Understanding how harms emerge and propagate is essential for developing targeted and context-aware fairness interventions.

In Section 5.1, we noted that several decisions lead to the exclusion of minority populations. Such harms in practice translate into QoS, allocative, and representational harms. While such harms are noted in centralized ML, they take different forms in FL settings. Additionally, additional mechanisms for privacy and adversarial robustness introduce disparities and possible damages. In particular, the biases in the FL lifecycle translate into the following harms:

- **Quality of Service (QoS) Harms.** These occur when the

FL model performs unequally across clients. A typical cause is *aggregation bias*, where global model updates disproportionately reflect data from dominant clients (e.g., those with more frequent participation or more standard data distributions). This harm is often introduced during model aggregation (Michieli and Ozay 2021), but can be seeded earlier through biased model initialization or improper problem formulation (e.g., assuming all clients have similar objectives (Suresh and Gutttag 2021)).

- **Allocative Harms.** These relate to unfair allocation of benefits or resources informed by FL models. They may result from *representation bias* or *learning bias* at the data and training stages, where certain groups are underrepresented or oversimplified. In financial applications, for instance, credit scoring systems built using FL can exclude certain regions or demographics if their data was sparsely sampled or downweighted (Ding et al. 2021). Incentive mechanisms that reward “high-contributing” clients also risk allocative harms when contributions are unfairly evaluated (Albaseer et al. 2024).
- **Representational Harms.** FL systems may reinforce stereotypes or marginalize specific groups when measurement or historical bias is present at the data collection step (Yang et al. 2020; Kleinberg et al. 2018). These harms are especially prominent in applications involving vision (Li et al. 2024a) language (Gallegos et al. 2024).
- **Privacy Harms.** While FL aims to preserve privacy by design, privacy-enhancing techniques like differential privacy can have disparate effects. For instance, clients with small datasets or outlier data distributions may suffer from higher performance degradation due to added noise (Bagdasaryan, Poursaeed, and Shmatikov 2019). These harms are introduced by protection mechanisms, but are shaped by earlier choices around data representation and client sampling.
- **Reputational Harms.** When robustness mechanisms such as anomaly detection or reputation scoring are deployed, clients with nonconforming updates (e.g., due to non-IID data) risk being misclassified as adversaries (Touat and Bouchenak 2023; Ren et al. 2024). This can result in exclusion from training, lost incentives, or damaged reputations, even when clients acted in good faith. These harms often emerge during aggregation or incentive distribution stages but stem from unaccounted-for data heterogeneity.

6 Takeaways and Path Forward

Our analysis of fairness in FL highlights a disconnect between how fairness is defined in research and how harms manifest in practice. If fairness is to be meaningful in FL, it must be grounded not only in optimization objectives, but in the structural conditions and institutional realities shaping the collaboration.

We propose a framework emphasizing that fairness in FL is not a single technical problem, but a distributed, sequential, multi-stakeholder and context-dependent challenge. Addressing fairness requires understanding where

structural inequalities shape developer decisions, and how those decisions affect different stakeholders in different ways. While we focused on the classic horizontal FL lifecycle, it is important to recognize that the pipeline is not fixed—additional steps such as personalization or clustering may be introduced, and design choices at each stage can evolve to better address fairness and contextual needs.

Following our analysis, We note some key takeaways for future work:

- 1. Fairness requires contextually grounded evaluation.** The datasets most commonly used in FL fairness research are inherited from centralized ML and fail to reflect the complexity of collaborative scenarios. Many rely on synthetic or oversimplified distributions, with minority classes used as proxies for structural disadvantage. These abstractions risk obscuring real harms and may result in cherry-picked evaluations (Lones 2021). Only a few datasets, such as folktables (Ding et al. 2021) and flamby (Ogier du Terrail et al. 2022), reflect real-world fairness/ FL concerns, yet they remain limited in scope. Advancing fairness in FL will require better datasets, more realistic simulations, and evaluation strategies that reflect real-world risks.
- 2. Fairness must account for real-world harms and stakeholder impact.** Understanding fairness in FL requires identifying who is affected, by which kinds of errors, and how these translate into tangible harms. This includes underrepresented clients misclassified as adversaries (Touat and Bouchenak 2023; Ren et al. 2024), institutions excluded due to limited data (Goetz et al. 2019), or regions harmed by biased global updates (Asiedu et al. 2024). Researchers must go beyond convergence metrics to consider harms like reduced quality of service, loss of trust, and perpetuation of inequity.
- 3. Lifecycle harms require lifecycle thinking.** Many fairness interventions in FL target isolated steps, with a focus on client selection and aggregation. However, harms often emerge from a chain of design choices across the FL lifecycle. For example, performance disparities may stem as much from problem formulation or model initialization as from aggregation. Tackling fairness effectively thus demands a holistic view: interventions should account for how earlier decisions constrain fairness downstream, and solutions must span multiple stages to achieve the fairness goals.
- 4. Fairness definitions are not interchangeable.** Performance, group, and collaborative fairness objectives may conflict, especially in heterogeneous systems. In some FL settings—such as financial or healthcare collaborations—multiple fairness notions may be simultaneously relevant. For instance, clients may seek equitable performance (performance fairness), while also serving diverse subpopulations (group fairness), and contributing to shared resources (collaborative fairness). Future work should investigate how these fairness objectives interact and how conflicts between them can be resolved.
- 5. Stakeholders must shape the design process.** Fairness cannot be defined externally to those it affects. Partic-

ipatory approaches (Birhane et al. 2022) that engage clients, domain experts, and impacted communities can improve problem formulation, risk assessment, and modeling decisions. As FL distributes control and responsibility across many actors, it also demands more inclusive governance mechanisms. Harms cannot be addressed without first identifying who could be harmed, which errors are most consequential, and how these manifest in specific contexts (Chehbouni et al. 2025; Shelby et al. 2023; Abercrombie et al. 2024).

7 Limitations

Our analysis is subject to several limitations. First, we focus primarily on horizontal FL, as it is the most widely studied variant in fairness literature; however, other forms such as vertical FL present different challenges that merit separate examination. Second, the coverage of our annotations may be constrained by the chosen library and search criteria, although we believe the sample is representative of the current research landscape. Finally, despite our effort to systematize the annotations, the process inevitably involves interpretation, which may introduce subjectivity or overlook subtle nuances.

Acknowledgements

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google award, CIFAR Catalyst Grant award, FRQNT and NSERC Discovery Grants program. We also express our gratitude to Compute Canada for their support in providing facilities for our evaluations.

References

- Abad, M. S. H.; Ozfatura, E.; Gunduz, D.; and Ercetin, O. 2020. Hierarchical federated learning across heterogeneous cellular networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8866–8870. IEEE.
- Abay, A.; Zhou, Y.; Baracaldo, N.; Rajamoni, S.; Chuba, E.; and Ludwig, H. 2020. Mitigating Bias in Federated Learning. *CoRR*, abs/2012.02447. ArXiv: 2012.02447.
- Abercrombie, G.; Benbouzid, D.; Giudici, P.; Golpayegani, D.; Hernandez, J.; Noro, P.; Pandit, H.; Paraschou, E.; Pownall, C.; Prajapati, J.; et al. 2024. A collaborative, Human-Centred taxonomy of AI, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*.
- Adamek, J.; and Darup, M. S. 2024. Privacy-preserving gradient-based fair federated learning. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2132–2138. ISSN: 2576-3555.
- Albaseer, A.; Abdi, N.; Abdallah, M.; Qaraq, M.; and Al-Kuwari, S. 2024. FedPot: A Quality-Aware Collaborative and Incentivized Honeypot-Based Detector for Smart Grid Networks. *IEEE Transactions on Network and Service Management*, 21(4): 4844–4860. Conference Name: IEEE Transactions on Network and Service Management.

- Amiri, S.; Belloum, A.; Nalisnick, E. T.; Klous, S.; and Gommans, L. 2022. On the impact of non-IID data on the performance and fairness of differentially private federated learning. In *52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN Workshops 2022, Baltimore, MD, USA, June 27-30, 2022*, 52–58. IEEE.
- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There’s software used across the country to predict future criminals. *And it’s biased against blacks*. *ProPublica*, 23: 77–91.
- Annapareddy, N.; Preston, J.; and Fox, J. 2023. Fairness and Privacy in Federated Learning and Their Implications in Healthcare. *CoRR*, abs/2308.07805. ArXiv: 2308.07805.
- Antunes, R. S.; André da Costa, C.; Küderle, A.; Yari, I. A.; and Eskofier, B. 2022. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4): 1–23.
- Asiedu, M. N.; Dieng, A.; Haykel, I.; Rostamzadeh, N.; Pfohl, S.; Nagpal, C.; Nagawa, M.; Oppong, A.; Koyejo, S.; and Heller, K. 2024. The Case for Globalizing Fairness: A Mixed Methods Study on Colonialism, AI, and Health in Africa. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–24. San Luis Potosi Mexico: ACM. ISBN 9798400712227.
- Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Balbierer, B.; Heinlein, L.; Zipperling, D.; and Kühl, N. 2024. A Multivocal Literature Review on Privacy and Fairness in Federated Learning. *CoRR*, abs/2408.08666. ArXiv: 2408.08666.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.
- Bonawitz, K. 2019. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*.
- Chehbouni, K.; Cock, M. D.; Caporossi, G.; Taik, A.; Rab-bany, R.; and Farnadi, G. 2025. Enhancing Privacy in the Early Detection of Sexual Predators Through Federated Learning and Differential Privacy. *arXiv:2501.12537*.
- Chen, H.; Zhu, T.; Zhang, T.; Zhou, W.; and Yu, P. S. 2024. Privacy and Fairness in Federated Learning: On the Perspective of Tradeoff. *ACM Comput. Surv.*, 56(2): 39:1–39:37.
- Cho, Y. J.; Wang, J.; and Joshi, G. 2022. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, 10351–10375. PMLR.
- Chu, W.; Xie, C.; Wang, B.; Li, L.; Yin, L.; Zhao, H.; and Li, B. 2022. FOCUS: Fairness via Agent-Awareness for Federated Learning on Heterogeneous Data. *CoRR*, abs/2207.10265. ArXiv: 2207.10265.
- Chu, Y.-W.; Han, D.-J.; Hosseinalipour, S.; and Brinton, C. G. 2024a. Rethinking the Starting Point: Enhancing Performance and Fairness of Federated Learning via Collaborative Pre-Training. *CoRR*, abs/2402.02225. ArXiv: 2402.02225.
- Chu, Y.-W.; Hosseinalipour, S.; Tenorio, E.; Cruz, L.; Douglas, K.; Lan, A. S.; and Brinton, C. G. 2024b. Multi-layer personalized federated learning for mitigating biases in student predictive analytics. *IEEE Transactions on Emerging Topics in Computing*.
- Chuang, C.-Y.; Jampani, V.; Li, Y.; Torralba, A.; and Jegelka, S. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.
- Cong, Y.; Qiu, J.; Zhang, K.; Fang, Z.; Gao, C.; Su, S.; and Tian, Z. 2024. Ada-FFL: Adaptive computing fairness federated learning. *CAAI Trans. Intell. Technol.*, 9(3): 573–584.
- Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P. S.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 6478–6490. Curran Associates, Inc.
- Donahue, K.; and Kleinberg, J. M. 2021. Models of fairness in federated learning. *CoRR*, abs/2112.00818. ArXiv: 2112.00818.
- El-Rifai, O.; Ali, M. B.; Megdiche, I.; Peninou, A.; and Teste, O. 2025. A Survey on Cluster-based Federated Learning. *arXiv preprint arXiv:2501.17512*.
- Ezzeldin, Y. H.; Yan, S.; He, C.; Ferrara, E.; and Avestimehr, S. 2021. FairFed: Enabling Group Fairness in Federated Learning. *CoRR*, abs/2110.00857. ArXiv: 2110.00857.
- Fu, L.; Zhang, H.; Gao, G.; Zhang, M.; and Liu, X. 2023. Client Selection in Federated Learning: Principles, Challenges, and Opportunities. *IEEE Internet of Things Journal*, 10(24): 21811–21819.
- Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3): 1097–1179.
- Gao, J.; Mavromatis, I.; Li, P.; Carnelli, P. E.; and Khan, A. 2024. Mitigating System Bias in Resource Constrained Asynchronous Federated Learning Systems. *CoRR*, abs/2401.13366. ArXiv: 2401.13366.
- Goetz, J.; Malik, K.; Bui, D.; Moon, S.; Liu, H.; and Kumar, A. 2019. Active Federated Learning. ArXiv:1909.12641 [cs].
- Green, B. 2021. The contestation of tech ethics: FA sociotechnical approach to technology ethics in practice. *Journal of Social Computing*, 2(3): 209–225.
- Gálvez, B. R.; Granqvist, F.; Dalen, R. C. v.; and Seigel, M. 2021. Enforcing fairness in private federated learning via the modified method of differential multipliers. *CoRR*, abs/2109.08604. ArXiv: 2109.08604.
- Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

- Hard, A.; Rao, K.; Mathews, R.; Ramaswamy, S.; Beaufays, F.; Augenstein, S.; Eichner, H.; Kiddon, C.; and Ramage, D. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.
- Huang, T.; Lin, W.; Wu, W.; He, L.; Li, K.; and Zomaya, A. Y. 2021. An Efficiency-boosting Client Selection Scheme for Federated Learning with Fairness Guarantee. *IEEE Transactions on Parallel and Distributed Systems*, 32(7): 1552–1564. ArXiv:2011.01783 [cs].
- Javaherian, S.; Panta, S.; Williams, S.; Islam, M. S.; and Chen, L. 2024a. FedFair³: Unlocking Threefold Fairness in Federated Learning. *CoRR*, abs/2401.16350. ArXiv: 2401.16350.
- Javaherian, S.; Panta, S.; Williams, S.; Islam, M. S.; and Chen, L. 2024b. FedFair³: Unlocking Threefold Fairness in Federated Learning. In *IEEE International Conference on Communications, ICC 2024, Denver, CO, USA, June 9-13, 2024*, 3622–3627. IEEE.
- Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.
- Kaur, I.; and Jadhav, A. J. 2023. A Comprehensive Study on Model Initialization Techniques Ensuring Efficient Federated Learning. ArXiv:2311.02100 [cs].
- Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2018. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, 22–27. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.
- Kohavi, R.; and Becker, B. 1996. Uci adult data set. *UCI Machine Learning Repository*, 5.
- Konečný, J. 2016. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv preprint arXiv:1610.05492*.
- Lai, F.; Dai, Y.; Singapuram, S.; Liu, J.; Zhu, X.; Madhyastha, H.; and Chowdhury, M. 2022. FedScale: Benchmarking model and system performance of federated learning at scale. In *International conference on machine learning*, 11814–11827. PMLR.
- Lee, M.; Kim, H.; and Joo, C. 2024. Geographical Node Clustering and Grouping to Guarantee Data IIDness in Federated Learning. *arXiv preprint arXiv:2410.15693*.
- Li, B.; Yao, Y.; Tan, J.; Gong, R.; Lu, J.; and Luo, Y. 2024a. Rectify representation bias in vision-language models for long-tailed recognition. *Neural Networks*, 172: 106134.
- Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2020. Fair Resource Allocation in Federated Learning. ArXiv:1905.10497 [cs].
- Li, Y.; Zhang, J.; Zhao, Y.; Chen, B.; and Yu, S. 2024b. Fairness-Aware Federated Learning Framework on Heterogeneous Data Distributions. In *IEEE International Conference on Communications, ICC 2024, Denver, CO, USA, June 9-13, 2024*, 728–733. IEEE.
- Lim, W. Y. B.; Luong, N. C.; Hoang, D. T.; Jiao, Y.; Liang, Y.-C.; Yang, Q.; Niyato, D.; and Miao, C. 2020. Federated learning in mobile edge networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(3): 2031–2063.
- Ling, X.; Fu, J.; Chen, Z.; Wang, K.; Li, H.; Cheng, T.; Xu, G.; and Li, Q. 2024. FedFDP: Federated Learning with Fairness and Differential Privacy. *CoRR*, abs/2402.16028. ArXiv: 2402.16028.
- Liu, L.; Kong, Y.; Li, G.; and Han, M. 2023. FairShare: An Incentive-Based Fairness-Aware Data Sharing Framework for Federated Learning. In Yang, H.; Liu, H.; Zou, J.; Yin, Z.; Liu, L.; Yang, G.; Ouyang, X.; and Wang, Z., eds., *Intelligent Robotics and Applications - 16th International Conference, ICIRA 2023, Hangzhou, China, July 5-7, 2023, Proceedings, Part II*, volume 14268 of *Lecture Notes in Computer Science*, 115–126. Springer.
- Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.
- Lones, M. A. 2021. How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*.
- Lu, J.; Liu, H.; Zhang, Z.; Wang, J.; Goudos, S. K.; and Wan, S. 2022. Toward Fairness-Aware Time-Sensitive Asynchronous Federated Learning for Critical Energy Infrastructure. *IEEE Trans. Ind. Informatics*, 18(5): 3462–3472.
- Lu, Z.; Pan, H.; Dai, Y.; Si, X.; and Zhang, Y. 2024. Federated learning with non-iid data: A survey. *IEEE Internet of Things Journal*.
- Lyu, L.; Xu, X.; and Wang, Q. 2020. Collaborative Fairness in Federated Learning. ArXiv:2008.12161 [cs].
- Maheswari, G. U.; Jeslin, J. G.; Rajasuguna, M.; and S, A. 2024. Transforming Healthcare with Federated Learning-based Artificial Intelligence: Concepts, Classifications, and Challenges. In *2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, 209–216.
- Makhija, D.; Han, X.; Ghosh, J.; and Kim, Y. 2024. Achieving Fairness Across Local and Global Models in Federated Learning. *CoRR*, abs/2406.17102. ArXiv: 2406.17102.
- Mashhadi, A.; Tabaraei, A.; Zhan, Y.; and Parizi, R. M. 2022. An Auditing Framework for Analyzing Fairness of Spatial-Temporal Federated Learning Applications. In *2022 IEEE World AI IoT Congress (AIoT)*, Seattle, WA, USA, June 6-9, 2022, 699–707. IEEE.
- McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.
- Meerza, S. I. A.; Liu, L.; Zhang, J.; and Liu, J. 2024. GLOCALFAIR: Jointly Improving Global and Local Group Fairness in Federated Learning. *CoRR*, abs/2401.03562. ArXiv: 2401.03562.
- Michieli, U.; and Ozay, M. 2021. Are All Users Treated Fairly in Federated Learning Systems? In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

- Workshops (CVPRW)*, 2318–2322. Nashville, TN, USA: IEEE. ISBN 978-1-6654-4899-4.
- Mitchell, S.; Potash, E.; Barocas, S.; D’Amour, A.; and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1): 141–163.
- Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic federated learning. In *International conference on machine learning*, 4615–4625. PMLR.
- Molamohammadi, M.; Taïk, A.; Le Roux, N.; and Farnadi, G. 2023. Unraveling the Interconnected Axes of Heterogeneity in Machine Learning for Democratic and Inclusive Advancements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO ’23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703812.
- Nishio, T.; and Yonetani, R. 2019. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications (ICC)*, 1–7. IEEE.
- Ogier du Terrail, J.; Ayed, S.-S.; Cyffers, E.; Grimberg, F.; He, C.; Loeb, R.; Mangold, P.; Marchand, T.; Marfoq, O.; Mushtaq, E.; et al. 2022. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, 35: 5315–5334.
- Papadaki, A.; Martínez, N.; Bertrán, M.; Sapiro, G.; and Rodrigues, M. R. D. 2022. Minimax Demographic Group Fairness in Federated Learning. In *FACCT ’22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, 142–159. ACM.
- Passi, S.; and Barocas, S. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. Atlanta GA USA: ACM. ISBN 978-1-4503-6125-5.
- Pentyala, S.; Neophytou, N.; Nascimento, A. C. A.; Cock, M. D.; and Farnadi, G. 2022. PrivFairFL: Privacy-Preserving Group Fairness in Federated Learning. *CoRR*, abs/2205.11584. ArXiv: 2205.11584.
- Qi, P.; Chiaro, D.; Guzzo, A.; Ianni, M.; Fortino, G.; and Piccialli, F. 2024. Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems*, 150: 272–293.
- Qin, C.; Zhu, H.; Xu, T.; Zhu, C.; Jiang, L.; Chen, E.; and Xiong, H. 2018. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 25–34.
- Rafi, T. H.; Noor, F. A.; Hussain, T.; and Chae, D.-K. 2024. Fairness and privacy preserving in federated learning: A survey. *Inf. Fusion*, 105: 102198.
- Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.
- Ren, P.; Qi, K.; Li, J.; Yan, T.; and Dai, Q. 2024. CosPer: An adaptive personalized approach for enhancing fairness and robustness of federated learning. *Inf. Sci.*, 675: 120760.
- Roy, S.; Sharma, H.; and Salekin, A. 2024. Fairness Without Demographics in Human-Centered Federated Learning. *CoRR*, abs/2404.19725. ArXiv: 2404.19725.
- Salazar, T.; Araújo, H.; Cano, A.; and Abreu, P. H. 2024. A Survey on Group Fairness in Federated Learning: Challenges, Taxonomy of Solutions and Directions for Future Research. *CoRR*, abs/2410.03855. ArXiv: 2410.03855.
- Saputra, Y. M.; Hoang, D. T.; Nguyen, D. N.; Dutkiewicz, E.; Mueck, M. D.; and Srikanteswara, S. 2019. Energy demand prediction with federated learning for electric vehicle networks. In *2019 IEEE global communications conference (GLOBECOM)*, 1–6. IEEE.
- Sattler, F.; Müller, K.-R.; and Samek, W. 2020. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8): 3710–3722.
- Saxena, N. A.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D. C.; and Liu, Y. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 99–106.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.
- Selialia, K.; Chandio, Y.; and Anwar, F. M. 2022. Federated Learning Biases in Heterogeneous Edge-Devices: A Case-Study. In Gummesson, J.; Lee, S. I.; Gao, J.; and Xing, G., eds., *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, SenSys 2022, Boston, Massachusetts, November 6-9, 2022*, 980–986. ACM.
- Selialia, K.; Chandio, Y.; and Anwar, F. M. 2023. Mitigating Group Bias in Federated Learning for Heterogeneous Devices. *CoRR*, abs/2309.07085. ArXiv: 2309.07085.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Ros-tamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.
- Shi, Y.; Yu, H.; and Leung, C. 2024. Towards Fairness-Aware Federated Learning. *IEEE Trans. Neural Networks Learn. Syst.*, 35(9): 11922–11938.
- Simson, J.; Fabris, A.; and Kern, C. 2024. Lazy data practices harm fairness research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 642–659.
- Suresh, H.; and Gutttag, J. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. – NY USA: ACM. ISBN 978-1-4503-8553-4.
- Tak, A.; and Cherkaoui, S. 2021. Federated Edge Learning: Design Issues and Challenges. *IEEE Network*, 35(2): 252–258.

- Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12): 9587–9603.
- Thakkar, O.; Ramaswamy, S.; Mathews, R.; and Beaufays, F. 2020. Understanding unintended memorization in federated learning. *arXiv preprint arXiv:2006.07490*.
- Touat, O.; and Bouchenak, S. 2023. Towards Robust and Bias-free Federated Learning. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, 49–55. Rome Italy: ACM. ISBN 979-8-4007-0084-2.
- Ude, B.; Odeyomi, O. T.; Roy, K.; and Yuan, X. 2023. A Survey on Bias Mitigation in Federated Learning. In *IEEE Symposium Series on Computational Intelligence, SSCI 2023, Mexico City, Mexico, December 5-8, 2023*, 1170–1175. IEEE.
- Vucinic, S.; and Zhu, Q. 2023. The Current State and Challenges of Fairness in Federated Learning. *IEEE Access*, 11: 80903–80914.
- Wang, C.; Huang, H.; Li, R.; Liu, J.; Cai, T.; and Zheng, Z. 2024a. Libra: A Fairness-Guaranteed Framework for Semi-Asynchronous Federated Learning. In *44th IEEE International Conference on Distributed Computing Systems, ICDCS 2024, Jersey City, NJ, USA, July 23-26, 2024*, 797–808. IEEE.
- Wang, L.; Zhu, T.; Zhou, W.; and Yu, P. S. 2024b. Linkage on Security, Privacy and Fairness in Federated Learning: New Balances and New Perspectives. *CoRR*, abs/2406.10884. ArXiv: 2406.10884.
- Wang, Y.; Tang, X.; Lu, Y.; and Liu, N. 2023. Research on the Fairness of Cold-start Recommender System Based on Federated Learning Framework. In *Proceedings of the 5th International Conference on Internet of Things, Automation and Artificial Intelligence, IoTAAI 2023, Nanchang, China, November 24-26, 2023*, 802–807. ACM.
- Xia, Y.; Ma, B.; Dou, Q.; and Xia, Y. 2024. Enhancing Federated Learning Performance Fairness via Collaboration Graph-Based Reinforcement Learning. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part X*, volume 15010 of *Lecture Notes in Computer Science*, 263–272. Springer.
- Xie, R.; Li, C.; Zhou, X.; and Dong, Z. 2024. Accelerating Communication-Efficient Federated Multi-Task Learning With Personalization and Fairness. *IEEE Trans. Parallel Distributed Syst.*, 35(11): 2239–2253.
- Xu, G.; Wu, Y.; Hu, J.; and Shi, Y. 2022. Achieving Fairness in Dermatological Disease Diagnosis through Automatic Weight Adjusting Federated Learning and Personalization. *CoRR*, abs/2208.11187. ArXiv: 2208.11187.
- Xu, H.; Nanda, P.; and Liang, J. 2024. Reciprocal Federated Learning Framework: Balancing incentives for model and data owners. *Future Generation Computer Systems*, 161: 146–161.
- Yan, Y.; Zhu, L.; Li, Y.; Xu, X.; Goh, R. S. M.; Liu, Y.; Khan, S. H.; and Feng, C.-M. 2024. A New Perspective to Boost Performance Fairness For Medical Federated Learning. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part X*, volume 15010 of *Lecture Notes in Computer Science*, 13–23. Springer.
- Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Rusakovsky, O. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 547–558.
- Yi, L.; Shi, X.; Wang, N.; Zhang, J.; Wang, G.; and Liu, X. 2024. FedPE: Adaptive Model Pruning-Expanding for Federated Learning on Mobile Devices. *IEEE Transactions on Mobile Computing*, 23(11): 10475–10493. Conference Name: IEEE Transactions on Mobile Computing.
- Yin, X.; Zhu, Y.; and Hu, J. 2021. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6): 1–36.
- Yu, H.; Liu, Z.; Liu, Y.; Chen, T.; Cong, M.; Weng, X.; Niyato, D.; and Yang, Q. 2020. A Fairness-aware Incentive Scheme for Federated Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20*, 393–399. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-7110-0.
- Yue, X.; Nouiehed, M.; and Kontar, R. A. 2021. GIFAIR-FL: An Approach for Group and Individual Fairness in Federated Learning. *CoRR*, abs/2108.02741. ArXiv: 2108.02741.
- Zhang, D. Y.; Kou, Z.; and Wang, D. 2020. FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models. In Wu, X.; Jermaine, C.; Xiong, L.; Hu, X.; Kotevska, O.; Lu, S.; Xu, W.; Aluru, S.; Zhai, C.; Al-Masri, E.; Chen, Z.; and Saltz, J., eds., *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, 1051–1060. IEEE.
- Zhu, H.; Zhou, Y.; Qian, H.; Shi, Y.; Chen, X.; and Yang, Y. 2022. Online client selection for asynchronous federated learning with fairness consideration. *IEEE Transactions on Wireless Communications*, 22(4): 2493–2506.
- Öksüz, H. Y.; Molinari, F.; Sprekeler, H.; and Raisch, J. 2024. Boosting Fairness and Robustness in Over-the-Air Federated Learning. *IEEE Control. Syst. Lett.*, 8: 682–687.