Benefits of the Federation? Analyzing the Impact of Fair Federated Learning at the Client Level

Luca Corbucci*
University of Pisa
KDD Lab
Pisa, Italy
luca.corbucci@phd.unipi.it

Xenia Heilmann*
Johannes Gutenberg University
Institute of Computer Science
Mainz, Germany
xenia.heilmann@uni-mainz.de

Mattia Cerrato
Johannes Gutenberg University
Institute of Computer Science
Mainz, Germany
mcerrato@uni-mainz.de

Abstract

Federated Learning (FL) enables collaborative model training while preserving participating clients' local data privacy. However, the diverse data distributions across different clients can exacerbate fairness issues, as biases inherent in client data may propagate across the federation. Although various approaches have been proposed to enhance fairness in FL, they typically focus on mitigating the bias of a single binary-sensitive attribute. This narrow focus often overlooks the complexity introduced by clients with conflicting or diverse fairness objectives. Such clients may contribute to the federation without experiencing any improvement in their own model's performance or fairness regarding their specific sensitive attributes. In this paper, we compare three approaches to mitigate model unfairness in scenarios where clients have differing and potentially conflicting fairness requirements. By analysing disparities across sensitive attributes and model performance, we investigate the conditions under which clients benefit from federation participation. Our findings emphasise the importance of aligning federation objectives with diverse client needs to enhance participation and equitable outcomes in FL settings.

CCS Concepts

• **Security and privacy → Social aspects of security and privacy**; • **Computing methodologies → Distributed computing methodologies**.

Keywords

Federated Learning, Fairness in Machine Learning, Bias, Algorithmic Fairness

ACM Reference Format:
Luca Corbucci, Xenia Heilmann, and Mattia Cerrato. 2025. Benefits of the Federation? Analyzing the Impact of Fair Federated Learning at the Client Level. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3715275.3732152

*Both authors contributed equally to this research.

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732152

1 Introduction

The widespread adoption of Machine Learning (ML) across various fields not only required the introduction of numerous Artificial Intelligence (AI) regulations [6, 10, 19, 24, 32] but also emphasised the focus on Responsible AI practices. In recent years, we witnessed a growing interest in developing explainable models [7], protecting users' privacy [31], and preventing unfair model behavior [28].

FL [27] was introduced to protect user privacy and enable collaborative ML model training. While FL avoids sharing raw data with a central training server, it remains vulnerable to privacy attacks [21, 23, 36] and is often combined with privacy-enhancing technologies such as Differential Privacy to mitigate these risks [29]. To prevent unfair model behaviour, numerous approaches were proposed to train fair FL models [33], mainly inspired by traditional centralised learning techniques. However, the training of fair models in FL settings presents unique challenges due to the heterogeneity of the participating clients and the inherent complexities of learning in a decentralised environment. Heterogeneity can include that clients from different regions hold data biased toward different groups, define fairness differently, and have varying motivations for participating in FL. They may also prioritise different sensitive attributes, i.e., features of an individual such as gender, race, or age, which are usually linked to biases and societal inequalities and require consideration to ensure fairness.

Existing approaches to training fair FL models often assume a single binary sensitive attribute that applies uniformly to all participants. Consequently, the fairness objectives addressed and evaluated may not align with the diverse fairness needs of all clients. Some clients might already achieve fairness for the chosen sensitive attribute, but could face significant unfairness for another, unaddressed attribute.

Recent research showed that bias can be propagated across the federation [8], resulting in models that are less fair for a part of the clients than if they had trained a model independently on their data. When clients notice that contributing their resources and data to a federation does not improve their locally trained model, their motivation to participate in further FL training decreases. From their perspectives, FL has failed to achieve the desired and expected outcome. This individual perspective could be changed to a more global one if these clients had access to all participants' data and could assess the success globally. For instance, in an FL setting with 20 clients where 15 reports *better* models than before, the 5 clients who see no improvements could view their contribution as successful if they knew it helped improve others' models. However, this is purely hypothetical since detailed statistics of each client are rarely shared across the federation or made publicly accessible.

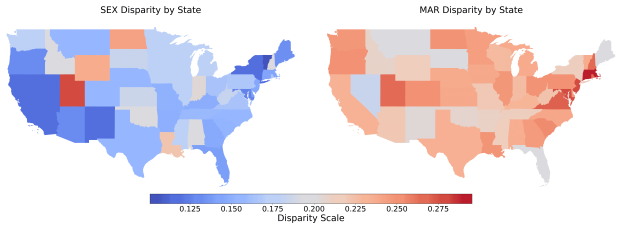

Figure 1: The fairness measured using Demographic Disparity on the ACS Income dataset. If we consider SEX and MAR as two separate sensitive attributes, we can observe that some states are more biased with respect to one attribute than to the other.

Therefore, clients taking part in FL expect an individual benefit. More specifically, they expect to train a *better* model by collaborating with others rather than training alone. Here, each client's definition of a "*better* model" can vary strongly among clients in an FL setting. This is the reason why, usually, FL objectives are defined before taking part in the federation to convince clients to join it. While these objectives have the potential to include a wide range of things, they are mostly focused on optimising accuracy and/or other metrics such as fairness. Specifically for fairness, in general, a fairness metric is incorporated into the FL objective for one or more sensitive attributes.

Ideally, defining these objectives can help to find clients willing to join the FL setting in the hope of benefiting from it. However, this is not always achieved as participation in federations can be compulsory for some or all clients for political, socioeconomic, or other reasons. When fairness objectives are involved, client heterogeneity presents two main challenges:

(1) Clients can have data that is unfair toward different values of the same sensitive attribute.
(2) Clients can possess data that is unfair to other sensitive attributes than the one the FL setting is focused on mitigating.

The first problem can appear when training a race-fair FL model on district hospitals, where clients may hold more data of one race depending on their location. This might be problematic in an FL scenario where clients could question the benefit of being part of the federation if the outcome is a model that is unfair toward their data. This scenario can happen if the majority of clients participating have data biased toward one sensitive group (e.g., black people) and therefore dominate the fairness objective, while only a minority of clients have data biased toward another group. In this setting, the competing interests of the different clients could even make the optimisation of fairness with respect to race impossible.

The second problem can occur if a fraction of clients holds data that is fair toward the sensitive attribute defined by the federation's objective but biased toward another attribute. This can emerge because of the data collection process, implemented laws, regulations, and economic or social factors. To show that this problem occurs, Figure 1 depicts how the Demographic Disparity (more details about this metric in Section 2.2) for the sensitive attributes SEX and MAR are distributed across states using the ACS Income dataset [13] (see Section 4.1).

In this paper, we focus on the second problem and simulate an FL training scenario where clients have conflicting fairness objectives to evaluate when and if joining a federation is more beneficial than local model training. Additionally, we compare the FL approach to a cluster-based method that groups clients according to their fairness objectives before executing the FL training to show how such an approach can better serve client needs compared to standard FL or purely local training.

Our key contributions are:

- A comprehensive comparison of three popular unfairness mitigation techniques for FL, evaluating their performance in scenarios with conflicting fairness objectives;
- A simulation of these scenarios using two popular tabular datasets to assess practical outcomes;
- A comparison of the standard FL approach with a cluster-based approach, to analyse if grouping clients with similar fairness preferences before training can lead to improved fairness outcomes for individual clients.

2 Background

2.1 Federated Learning

FL [27] is a collaborative learning approach introduced by Google in 2016, allowing K entities, usually called clients, to train a shared ML model without exposing their private training dataset D_k to external entities. Based on the number of clients K, FL can be categorised into two scenarios: cross-silo and cross-device [34]. In the cross-device scenario, K can increase to millions, clients possess a few samples, and their availability is limited to specific circumstances. In contrast, in the cross-silo scenario, K ranges from tens to hundreds, clients have more data available, and they are always available during training. The clients involved in federated training can be institutions, such as hospitals in a cross-silo context, and devices like smartphones or edge devices in a cross-device context. This paper focuses on a cross-silo scenario, where an organisation holds data about multiple individuals but would still like to treat them fairly with regard to some group fairness definitions. We note that, compared to a cross-device setting, there is a disconnect here between the client (i.e. silo, organisation) and the individuals represented in the data. This further motivates our critical investigation of the cross-silo setting, as it highlights the risk of a "trickle-down effect", where the potential harms of the organization's decisions about whether to join a federation will ultimately impact (groups of) individuals. Yet, our analysis can easily be extended to cross-device settings.

A second distinction for FL is between Horizontal and Vertical FL settings [34], which defines how data is distributed across clients. In Horizontal FL, clients hold datasets within the same feature space but with completely different samples. In Vertical FL, instead, clients share the same ID space (e.g., same users), but the feature space is different. For our work here, we assume a Horizontal FL scenario. The training of an FL model is usually orchestrated by a central server S that is responsible for the selection of a subset χ of available clients in each training round $r \in [0, R]$ as well as the aggregation of the models trained by them. At the initial round $r = 0$, the server selects a subset χ of clients and shares a model θ_r initialised with random weights with these selected clients. The

following training procedure depends on the chosen aggregation algorithm. With Federated-SGD (FedSGD) [27], clients receiving θ_r perform a single local training step. During this single step, each client computes the gradient $g_k = \nabla\mathcal{L}(\theta_r, b_i)$ of the model θ_r on the batch b_i from its local dataset D_k and shares g_k with the server. The server then aggregates the gradients and updates the global model $\theta_{r+1} \leftarrow \theta_r - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k$ where η is a fixed learning rate and $n = \sum_{k=1}^{K} n_k$. This aggregation algorithm entails high communication costs as gradients are exchanged after each local update. This problem can be solved with a more efficient algorithm called Federated Average (FedAvg) [27]. In this case, the clients perform a model update $\tilde{\theta}^k \leftarrow \theta^k - \eta\nabla\mathcal{L}(\theta^k, b_i)$ for E local epochs before sharing the update with the server. The server aggregates these final updates to compute the global model $\theta_{r+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \tilde{\theta}^k$. This aggregation reduces communication overhead while maintaining model performance, making FedAvg the preferred choice in many FL applications.

2.2 Group Fairness in Machine Learning

Group Fairness in ML models refers to their ability to predict an outcome without exhibiting a bias toward a specific demographic group. Researchers proposed several different metrics to quantify a model's fairness in recent years, including Equality of Opportunity [20], Equalized Odds [20], Predictive Parity [9], and Demographic Parity [16]. Which metric or metrics to choose is very specific to the application, as each metric offers a different interpretation and may conflict with others. In this paper, we rely on the Demographic Parity [16], a popular and widely used metric employed in the three approaches we compare.

Definition 2.1. Demographic Parity is a fairness metric based on the principle of Independence [3], which requires that the likelihood of a particular prediction outcome must not depend on the membership in a sensitive group. Formally, Demographic Parity can be expressed as:

$$\mathbb{P}(\hat{Y} = y \mid Z = z) = \mathbb{P}(\hat{Y} = y \mid Z \neq z) \tag{1}$$

where y is one of the possible targets predicted by the model and z is one of the possible values of the sensitive attribute. For example, the sensitive group could correspond to the gender of the individuals in the dataset. The closer the two probabilities are, the fairer the model will be. The approaches compared in this paper do not directly report the Demographic Parity of the different groups, instead, they evaluate the maximum difference between the Demographic Parity of the different sensitive groups. More specifically, they use the definition of Demographic Disparity.

Definition 2.2. Demographic Disparity $\Gamma(y, z)$ is the difference between the probability of predicting class y for samples with sensitive value z and the probability of predicting class y for samples with sensitive value different than z:

$$\Gamma(y, z) = \mathbb{P}(\hat{Y} = y \mid Z = z) - \mathbb{P}(\hat{Y} = y \mid Z \neq z) \tag{2}$$

The closer the Demographic Disparity is to 0, the less biased the model is in favour of one group over another. Notably, the ideal scenario for Demographic Parity occurs in a completely random model, where predictions are made uniformly at random. In such a case, the differences between the probabilities for different groups approach ~ 0, indicating no bias.

2.3 Fairness in Federated Learning

Fairness in FL can refer to many principles such as *client fairness*, *selection fairness*, *contribution fairness*, *individual fairness*, and *group fairness* [33]. Client Fairness refers to the principle that the performance of an FL model is distributed evenly across the participating clients. Selection Fairness focuses on choosing participants for FL rounds without bias, while Contribution Fairness ensures clients get rewarded proportionally to their contribution to the global model. While these fairness principles are specific to FL settings, individual and group fairness also exist in centralised learning settings. This work focuses specifically on **Group Fairness** in FL. Recent approaches have attempted to integrate group fairness into FL settings [1]. Here, many challenges arise due to data heterogeneity, restricted information about sensitive attributes, resource constraints, or client participation. Specifically, the limited information about sensitive data in each client's dataset leads to several concerns about fairness, such as how to integrate intersectional fairness into FL settings.

3 Related Work

Evaluating a client's benefit from participating in FL training is crucial for analysing model performances on individual client data and not only on a global test set [14, 35]. While performance metrics typically include accuracy, F1-score, and similar measures, to our knowledge, no prior work has analysed the benefits of FL participation from each client's fairness perspective. However, several relevant works inspired and informed the analysis conducted in our paper. Firstly, researchers proved that FL is highly sensitive to bias propagation when the sensitive attribute is included as an input feature. This finding was first highlighted in a preliminary study [17] and then explored in depth in [8]. In particular, in [8] the authors proved a correlation between the bias encoded in each client's data and the fairness benefit gained from joining FL training. Clients with an initially greater bias tend to obtain fairer models through FL, while clients with less initial bias often receive more biased models. This is due to bias propagation, where even a small subgroup of biased clients can influence the overall model fairness. Our work builds on the statement that FL participants can significantly influence the final model properties and that these properties are perceived differently by each client. However, while [8] relies on FedAvg as an FL algorithm, we focus on FL methods enhancing fairness for sensitive attributes. Also, we do not include any sensitive attributes in our feature input space and restrict our dataset to at most 20,000 data points per client, differently from the 3,000 considered in [8]. Furthermore, we focus on the individual's benefit of taking part in a federation with defined objectives.

Secondly, [33] highlighted that current research on fairness in FL lacks contributions related to intersectional fairness. Intersectional fairness describes forms of discrimination and societal effects happening when different features intersect with each other [12]. This analysis partially inspired our paper as we consider scenarios in which, in centralized learning, an intersectionally fair model

[1]We refer the reader to [33] for a broad Group Fairness in FL literature overview

would be a near solution. As an intersectional fair FL method does not exist yet [33], our work aims to provide insights into how FL settings respond to clients with multiple fairness objectives until research advances in this direction.

4 Experimental Setting

To simulate an FL setting where clients in the federation have conflicting fairness objectives, we preprocessed the ACS Income and ACS Employment [13] dataset. Furthermore, we evaluate three different fair FL methods: PUFFLE [11], Reweighing [1], and FedMinMax [30]. Details on preprocessing, our choice of sensitive attributes, the applied methods, and our local baseline models are described in this section.

4.1 Datasets

We conduct experiments on two popular tabular datasets: ACS Income and ACS Employment [13]. The American Community Survey (ACS) collected data for these two datasets across all 50 states of the USA and Puerto Rico between 2014 and 2018. The natural division of the dataset into 51 entities is particularly useful for FL research because it allows treating each entity as a separate client, exploiting the natural non-iid distribution of the samples. The ACS Income dataset contains 1,664,500 samples, the task involves predicting if an individual's income is above 50, 000 or not. The ACS Employment dataset contains 3,236,107 samples, the task involves predicting an individual's employment status.

4.2 Preprocessing

As explained in Section 1, this paper aims to analyse if and how the different clients benefit from being involved in a federation in terms of model utility and fairness. Therefore, we require a data distribution where different groups of clients exhibit distinct fairness concerns. Specifically, we want half of the clients to demonstrate a bias toward a particular sensitive attribute and aim to mitigate the model's unfairness with respect to this attribute. Meanwhile, the other half of the clients should show a bias toward a different sensitive attribute and seek to reduce model unfairness related to this one. To simulate this scenario, we preprocess the two datasets in the following way:

- We consider the same two sensitive attributes for both datasets: SEX, which indicates the gender of the different samples, and MAR, which indicates their marital status. In both datasets, SEX is already a binary attribute. Instead, MAR can have the following possible values: "Married", "Widowed", "Divorced", "Separated" and "Never married". To have datasets compatible with all applied unfairness mitigation approaches, we reduce MAR to a binary attribute with two values: "Married" and "Not Married". In particular, we map "Widowed", "Divorced", "Separated" and "Never married" to "Not married".
- As participants in our setting, we select 20 of the 51 clients available in the dataset to reduce the computational power required to perform our tests across the various methods and settings under consideration. For the same reason, we limit the maximum number of training samples for each client to 20,000.

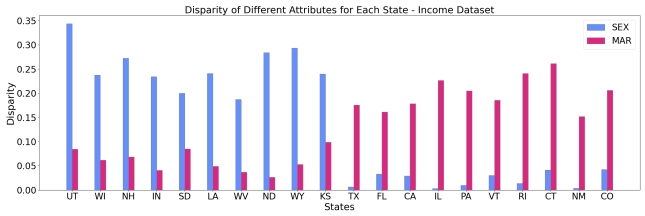

Figure 2: Demographic Parity computed on the ACS Income dataset of the 20 clients selected in the FL computation. The first 10 clients (in blue) are unfair w.r.t. the SEX sensitive attribute but low MAR disparity. The last 10 clients (in red) are unfair toward MAR with ~ 0 SEX disparity.

- A subgroup of the clients was already unfair toward the two sensitive attributes considered in this paper. For the other participants, we artificially made the clients unfair toward MAR and SEX or exacerbated the unfairness. To this end, we removed samples from 10 clients' datasets to increase bias toward SEX, and from the other 10 clients' datasets to increase bias toward MAR. We report the distribution of the Demographic Disparity computed on the training dataset of the 20 clients for the ACS Income dataset in Figure 2. A corresponding figure for ACS Employment is reported in Figure 10 in Appendix B.
- We exclude from the feature input space both sensitive attributes SEX and MAR, as well as another sensitive attribute called RAC1P which encodes the race.

We published the preprocessed dataset used throughout the experiments in our GitHub repository [2].

4.3 Fair FL Methods

4.3.1 FedMinMax. FedMinMax [30] is based on the minimax group fairness criterion [25], which aims to optimise a model by maximising the prediction performance of the worst-performing demographic group while avoiding unnecessary degradation in the performance of other demographic groups.

To archive this in a FL scenario, FedMinMax [30] solves the following optimization problem:

$$\min_{\theta \in \Theta} \max_{\mu \in \Delta^{|Z|-1}_{\geq \epsilon}} \sum_{z \in Z} \mu_z \hat{r}_z(\theta). \tag{3}$$

In Equation 3, the objective is in parallel minimised for optimising the model parameters θ and maximised to optimise the weighting coefficients μ. The summation is taken over the estimated empirical risk $\hat{r}_z$ for each value $z \in Z$ of the sensitive attribute. Since the clients only have access to their local finite datasets, rather than to the complete data distribution, the risk is estimated in a weighted way. Therefore, Equation 3 focuses on minimising the risk for the worst-performing demographic group. This approach can be extended to scenarios where not all clients participate in every round of training by estimating the risk based on only the available participants. Additionally, the method allows for different distributions of the demographic groups among the clients in the FL

[2] https://github.com/xheilmann/FairnessBenefitsFL

setting as well as different fairness metrics. To evaluate FedMinMax, we adapted the source code [3] for Demographic Parity.

4.3.2 PUFFLE. PUFFLE [11] is a recently proposed in-process method that employs Differential Privacy [15] and a Regularization approach [22] to protect clients' privacy while reducing model unfairness measured with Demographic Parity. The clients execute the unfairness mitigation process, which is based on regularization. When the server selects a client in an FL round, it starts the training of the model as in classic FL. However, given a batch of data, the clients not only compute the gradient with respect to the model output but also to the model's unfairness measured on the same batch. This allows the incorporation of an additional regularization term in the model update that mitigates the unfairness. The two gradients are summed and weighted using a hyperparameter λ indicating the importance of the model's utility and its unfairness. Choosing a $\lambda \simeq 1$ would lead to a perfectly fair model with accuracy close to 0.5. On the contrary, a $\lambda \simeq 0$ would optimise the model utility without caring about the model's unfairness. PUFFLE can be used with a tunable and fixed λ. In a classic FL scenario, which does not involve distribution shifts during the training, tunable and fixed λ guarantee similar results. Instead, when distribution shifts happen during the training, the tunable outperforms the fixed λ. Since we do not consider the problem of distribution shift in this paper, we use PUFFLE with fixed λ and without the differential privacy mechanism, which is outside the scope of this paper. To evaluate PUFFLE, we used the source code provided here [4].

4.3.3 Reweighing Approach. Reweighing is a preprocessing approach to reduce the unfairness of models trained with FL [1]. This approach involves the computation of weights that are assigned to the training dataset samples. The weights depend on the composition of the datasets. The authors proposed two possible solutions to compute the weights: Local and Global Reweighing. In the Global Reweighing approach, the clients involved in the computation calculate a set of statistics $C_k(z, y)$ regarding the local training dataset D_k and true labels Y, which they then share with the central server.

$$C_k(z, y) = |(\mathcal{X} \in D_k | Z = z) \wedge (Y = y)| \ \forall z \in Z, y \in Y \quad (4)$$

The server is responsible for the aggregation of all the statistics received by the clients and for the computation of the weights $W_k(z, y)$. At the end of this process, the server holds a weight for each possible pair (z, y). These weights are then applied during training to weigh the importance of the prediction mistakes made for the different groups (z, y).

$$W_k(z, y) = \frac{\sum_{k, y \in Y} C_k(z, y) * \sum_{k, z \in Z} C_k(z, y}{C_k(z, y) \sum_{k, z \in Z, y \in Y} C_k(z, y)} \quad (5)$$

The alternative solution proposed in the paper is Local Reweighing. In this case, the clients compute the weights directly on their training dataset without sharing any information with the server. This means that each client has a different set of weights. In this paper, we report the results obtained with Local Reweighing. We implemented Reweighing with the AI Fairness 360 library [4] based on the guidance provided in [1].

[3] FedMinMax GitHub repository: https://github.com/oscardilley/federated-fairness
[4] PUFFLE GitHub repository: https://github.com/lucacorbucci/PUFFLE

4.4 Simulating FL

We perform our experiments in a cross-silo FL scenario. To simulate this environment, each client divides its dataset into a train and a test set. We keep this split of the dataset fixed during all the experiments for all considered settings. During the hyperparameter tuning phase, the clients divide the training set into a train and a validation set to perform proper hyperparameter tuning. We simulate the FL scenario using the popular Flower Framework [5], selecting all 20 clients to train the model in each FL round. The metrics reported in the paper are computed by aggregating the results of the metrics computed by the participating clients on their test sets. For this evaluation, the model trained by the clients on their local training dataset and the best hyperparameters found in the hyperparameter tuning phase are used (see Appendix A for more details about the hyperparameter tuning).

4.5 Local Training

To evaluate if joining the federation and training a collective model with FL benefits the different clients, we trained 20 Logistic Regression models [26] (one for each client involved in the FL training) on each client's local data. To ensure consistency when comparing with the FL models, each client holds the same train set and test set split as in the FL settings. Additionally, we train one local model while mitigating bias. In this optimisation, we mitigate the unfairness of each client with respect to one sensitive attribute that shows the highest disparity in the baseline model. For instance, if we consider Figure 2, the first client, "UT" has a higher disparity for the SEX attribute compared to MAR, so the fair model is trained to mitigate the unfairness specifically for the SEX attribute. To mitigate the unfairness, we apply the Exponentiated Gradient Reduction method [2] implemented in the AI Fairness 360 library [4].

5 Empirical Analysis

To evaluate the benefits of participating in fair FL settings, we conduct an extensive experimental analysis with the two described datasets and three unfairness mitigation methods for FL (FedMinMax [30], PUFFLE [11], and Reweighing [1]). We compare these methods against local models and a clustered FL approach in which clients are grouped based on their fairness objectives. We measure the accuracy and group fairness, operationalised as demographic disparity, as in Equation 2. Specifically, the reported demographic disparity is $\max_{y \in \hat{Y}, z \in Z} \Gamma(y, z)$. This allows us to understand the benefit of being in a federation in terms of bias mitigation and accuracy.

The following research questions are the basis for our analysis:

- **Q1:** In which situations is it beneficial for clients to join a federation?
- **Q2:** Do clients benefit from participating in a federation that does not align with their individual objectives?
- **Q3:** How do clients perform in fairness-aligned clusters where all participants share the same objectives, compared to their performance in an FL scenario focused on a single objective?

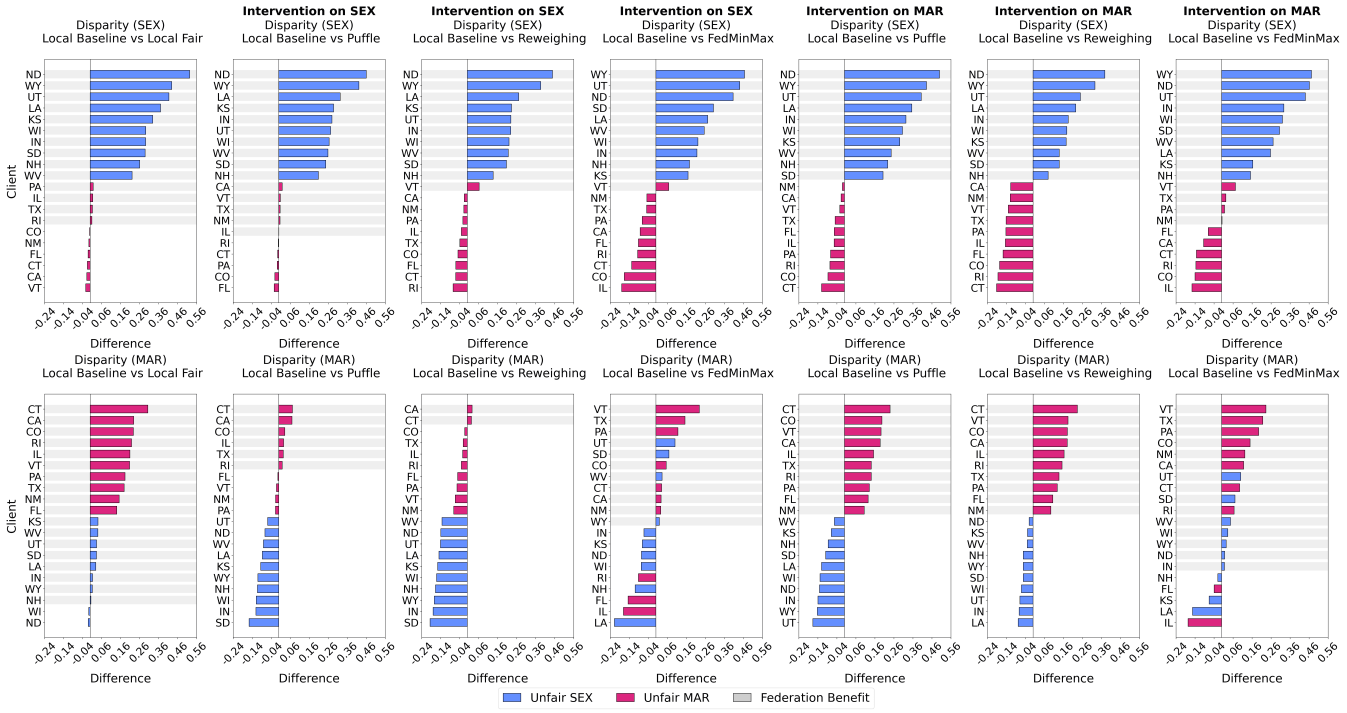

Figure 3: Income Dataset: Difference in disparities w.r.t. SEX sensitive attribute and MAR sensitive attribute between the baseline local model *without bias mitigation* (one model for each client) and four different fairness mitigation strategies (local fair model, PUFFLE, Reweighing and FedMinMax)

In this section, we analyse our findings reporting the results obtained using the ACS Income dataset [13], comparable results using the ACS Employment dataset are reported in Appendix C [5].

5.1 Benefits and Harms of the Federation

Our experiments reveal a consistent pattern regarding which groups or clients benefit from participating in fair FL. Figure 3 compares the local baseline models (trained without any fairness interventions on the local clients' training data) against four scenarios for the ACS income dataset: 1) A local fair baseline, 2) FL with PUFFLE, 3) FL with Reweighing, and 4) FL with FedMinMax. Clients showing improvements in these settings are highlighted in the figure with a grey shadow: for example, clients from "ND" to "RI" in the first plot of the first row of Figure 3 show a benefit in terms of disparity reduction when using the local model with unfairness mitigation compared to the baseline without any mitigation.

Notably, clients with data unfair toward the SEX attribute consistently show a reduction in disparity for SEX across all methods and all fairness interventions (first row of Figure 3). These clients, consistently ranked at the top of the plot, demonstrate the greatest benefits. A similar trend emerged for clients facing unfairness toward the MAR attribute. For this group, clients are generally ranked lower in terms of disparity reduction for MAR compared to those with SEX-related unfairness (second row in Figure 3). We also observe that, even compared to a baseline with no intervention, joining a bias-mitigating federation negatively impacts demographic parity if the chosen attribute for intervention is not aligned with the local-level requirements. In the algorithms employed in Figure 3, the between-group disparity often gets worse if a client had significant statistical disparity w.r.t. SEX but the federation intervened on MAR, or vice versa. Furthermore, it is important to note that any benefits of participating in a federation are primarily limited to disparity reduction. We observe only very few improvements in terms of accuracy, which provides additional evidence for a trade-off between fairness and accuracy [18, 28] (see Figure 11 in Appendix C). For the ACS employment dataset, we notice similar results for individual client rankings as well as client benefits. However, a larger proportion of clients benefit overall. Detailed results for this dataset are reported in Appendix C.1.

When comparing local fair models to fair FL settings, the consistency highlighted in the previous scenario emerges even more at the level of individual clients, as shown in Figure 4. The state "VT" shows a consistent benefit across all five methods in the first row. Additionally, "WV" benefits in four out of five methods and is ranked highly in the fifth. In the second row, states such as "UT", "SD", "VT", and "WY" show a consistent benefit across all methods. Interestingly, when using the FedMinMax method, the group "UT", "SD", "WY" not only benefits in terms of fairness across all fairness intervention settings but also from an accuracy improvement, as can be seen in Figure 5, which shows how the fairness-accuracy trade-off is contingent to the underlying data and setting. In general,

[5] Our code is available at https://github.com/xheilmann/FairnessBenefitsFL

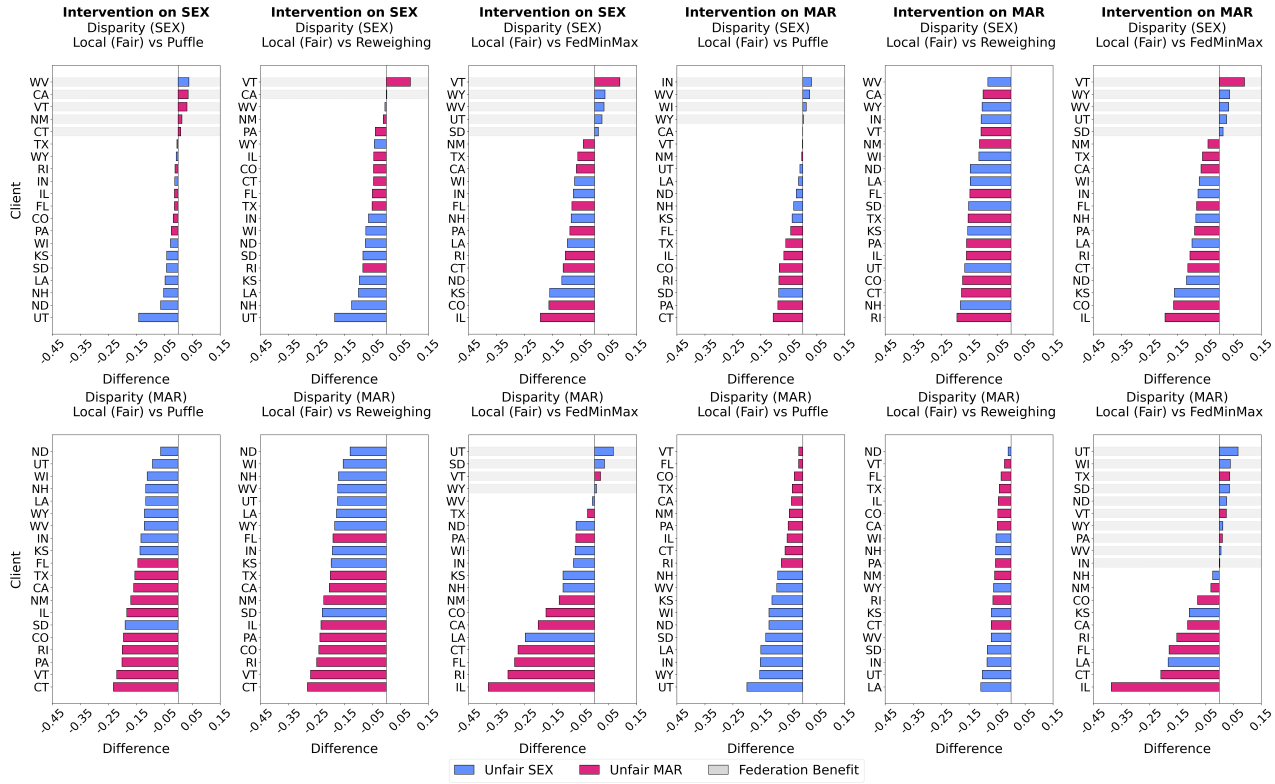

Figure 4: Income Dataset: Difference in disparities w.r.t. SEX sensitive attribute and MAR sensitive attribute between the local fair model (one model for each client) and three different fair FL methods (PUFFLE, Reweighing and FedMinMax).

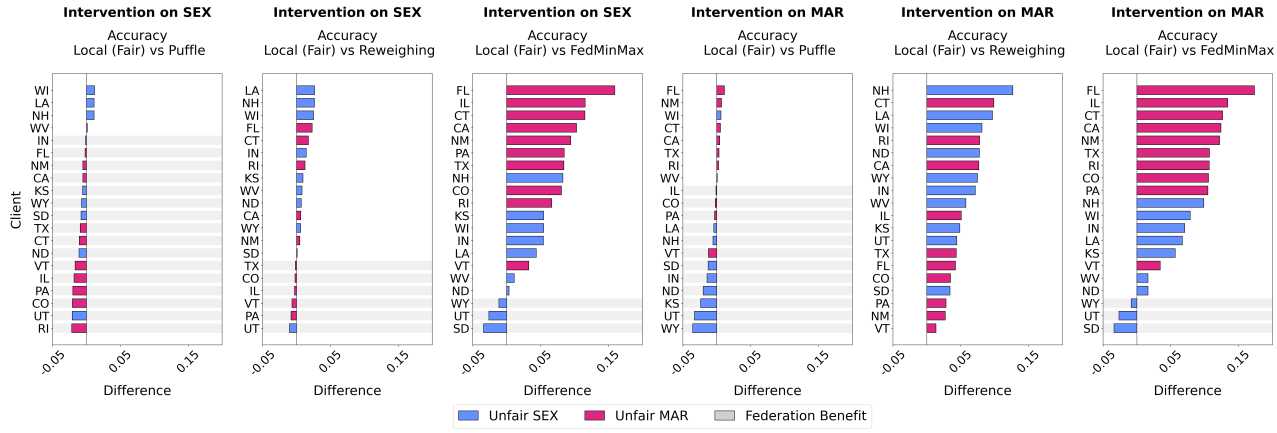

Figure 5: Income Dataset: Difference in accuracy between the local fair model and three different fair FL methods (PUFFLE, Reweighing and FedMinMax).

we observe that, for most clients, locally intervening on intersectional disparities provides a stronger bias mitigation than joining a federation which intervenes at a global level.

For the ACS employment dataset, we provide additional results in Appendix C.1. Taking these results into account, some of the states, such as "UT" or "VT" seem to benefit across most methods, interventions, and also datasets.

Overall, the disparity reduction achieved by participating in a fair FL setting is limited to a specific subset of clients in the federation when compared to a local fair baseline. Figure 5 shows that more clients benefit in terms of accuracy than in terms of disparity

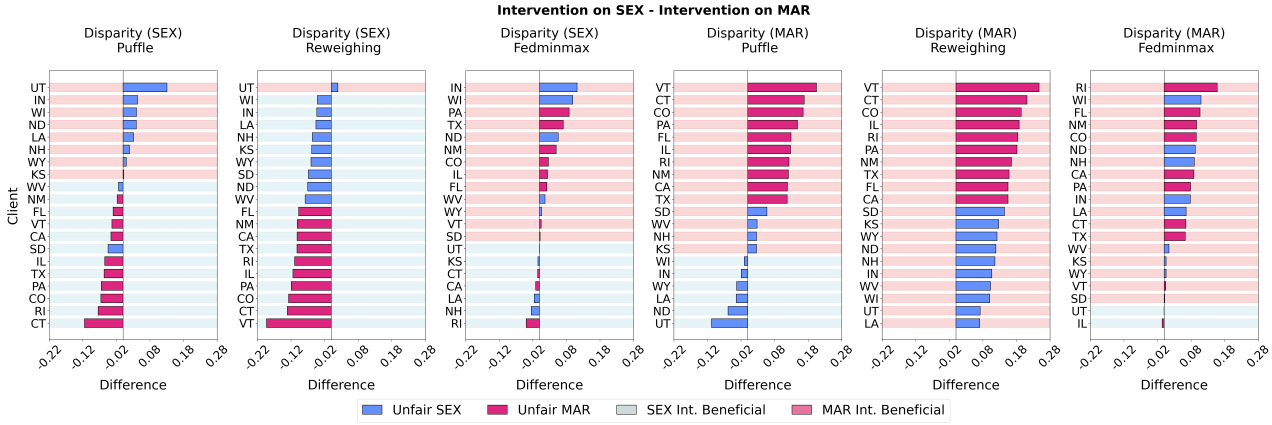

Figure 6: Income Dataset: Comparison of disparity benefits from taking part in FL settings with different fairness objectives across three methods (PUFFLE, Reweighing, and FedMinMax).

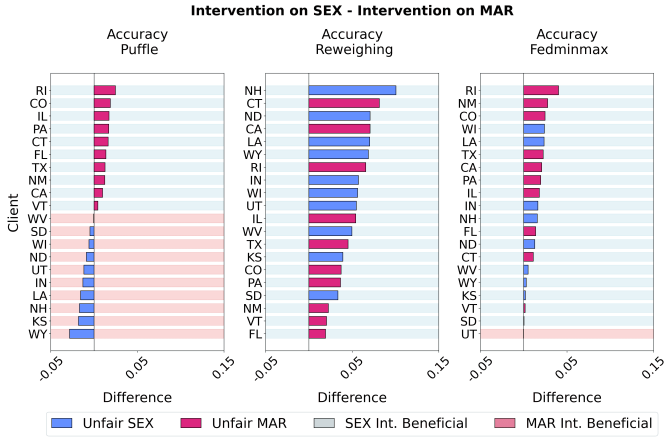

Figure 7: Income Dataset: Comparison of accuracy benefits from taking part in FL settings with different fairness objectives across three methods (PUFFLE, Reweighing and FedMinMax).

reduction. This aligns with findings in [8], where it was shown that participating in FL without fairness interventions typically provides stronger accuracy improvements than fairness gains as opposed to training on local data only. However, the small disparity magnitude in our experiments suggests the need to consider a margin of benefit rather than focusing solely on the absolute difference.

We believe this observation has important implications for federation formation. Clients who benefit from FL participation tend to consistently do so across different settings, making them more likely to join another setting. This is profitable for these clients as it enables them to compare multiple models from different FL settings and decide which one to apply in practice to their data. However, further research is needed to determine if this poses a threat to the other clients participating in FL settings, especially when the frequent participants hold large and very specific datasets. Conversely, the overall benefits of participating in fair FL in terms of disparity reduction may be less substantial than what institutions promote when recruiting federation participants. This discrepancy could lead to clients' frustration or, even worse, to the application of a model which is more biased than locally trained fair models. Therefore, we recommend that, when possible, clients taking part in fair FL settings compare locally-trained fair models against the fair FL model. Furthermore, we advocate for clear communication during the recruitment phase of a federation to ensure that client needs and expectations are taken into account.

5.2 Joining other Federations is Beneficial

In Figure 6 and Figure 7, we compare the results of the fair FL settings intervening on the SEX attribute with the settings intervening on the MAR attribute. We highlight the ones benefiting from SEX intervention in light blue and those benefiting from MAR intervention in light red. Our analysis shows that taking part in an FL setting with MAR intervention is more beneficial for disparity reduction, except for the Reweighing method for SEX disparity reduction. An interesting observation can be made for the PUFFLE method. Here,

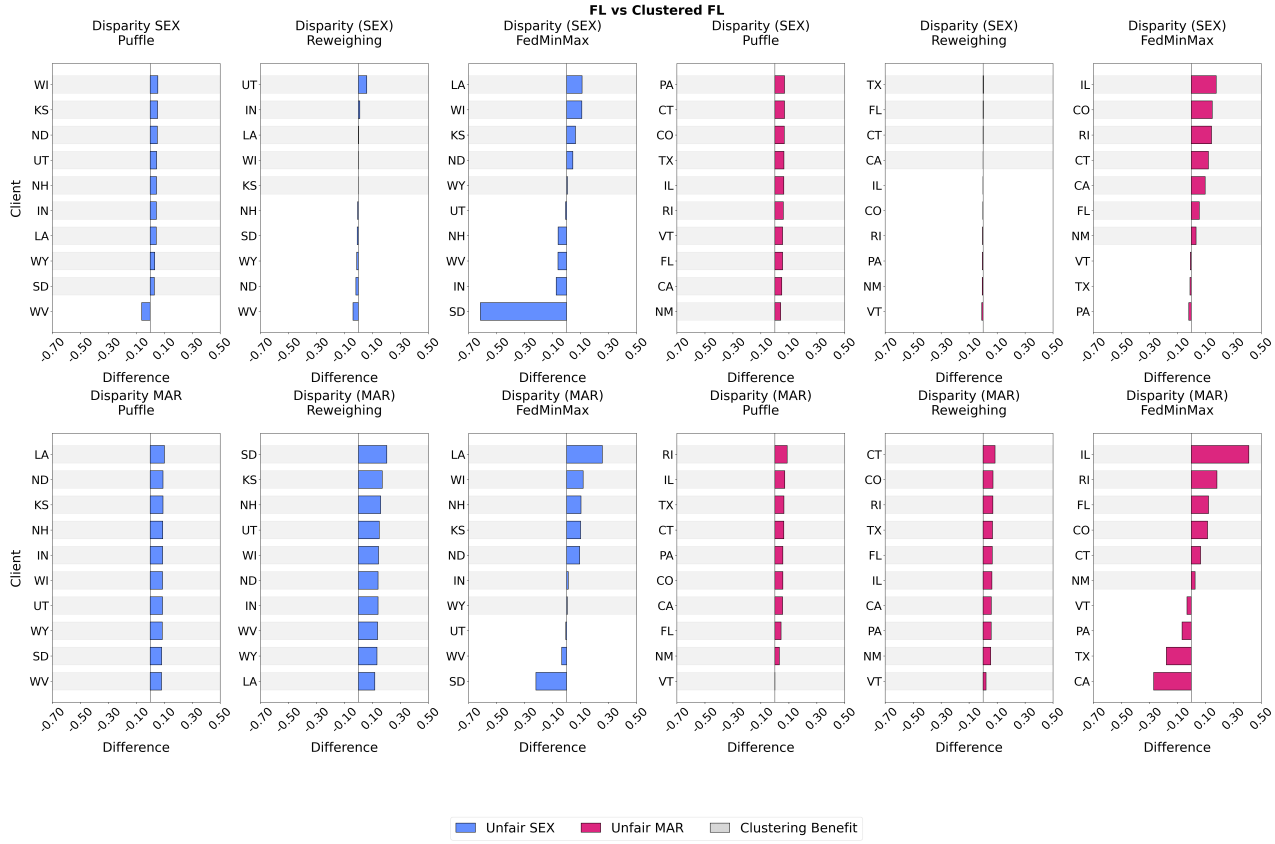

Figure 8: Income Dataset: Comparison of disparity benefits from taking part in clustered FL settings versus mixed FL settings across three methods (PUFFLE, Reweighing, and FedMinMax).

clients with SEX-related unfairness benefit more from participating in the FL setting intervening on MAR and vice versa for both accuracy and disparity.

These results are interesting when considering the initial disparity distribution across clients shown in Figure 2. Clients with MAR-related unfairness show lower MAR disparity than clients who are unfair toward SEX show for SEX. The same holds for the other group of clients. Therefore, we expect SEX intervention to result in a larger accuracy decrease than MAR intervention since it needs to mitigate higher disparities. Surprisingly, the opposite occurs, intervention on MAR proves to be more effective in reducing disparities for MAR and SEX but comes with higher accuracy degradation. This could be caused by the higher initial MAR disparities among clients with SEX-related unfairness, suggesting that intervening on MAR has more influence on their local learning process than vice versa.

We report corresponding results for Employment in Appendix C.2. They follow the same trend with an even clearer indication that more clients benefit from FL settings intervening on MAR. In conclusion, clients with SEX-related unfairness achieve better disparity reductions through participating in FL settings intervening on MAR. However, this is not the case for accuracy improvement. Here, clients with MAR-related unfairness benefit more from settings intervening on SEX. Altogether, this shows that there can be a benefit for clients to take part in federations that do not have the client's objectives as a goal. We believe that these outcomes can be related to the initial data distributions and the relationship between the two sensitive attributes SEX and MAR. However, further analysis is needed to examine this hypothesis.

5.3 Client Clustering Improves Fairness

In Figure 8, we compare the performances of the clustered federation. In this experiment, clients are divided into two clusters based on their fairness objectives. This results in training two separate models: one for the first cluster, reducing the unfairness towards SEX, and another for the second cluster, reducing unfairness towards MAR. The results show that clustering benefits the fairness of the individual clients. Specifically, clustering not only improves the disparity for the attribute that is intervened upon (SEX for the blue clients, MAR for the red clients) but is also beneficial for the sensitive attribute, which is not intervened upon.

In terms of accuracy, Figure 9 shows a slight degradation when using the clustering approach. This trade-off matches our expectations: improvements on the fairness metric may come at the cost of accuracy. Results for ACS Employment are shown in Appendix C.3.

Altogether, we find that clustering the clients in a federation with respect to their fairness objective can be an off-the-shelf solution

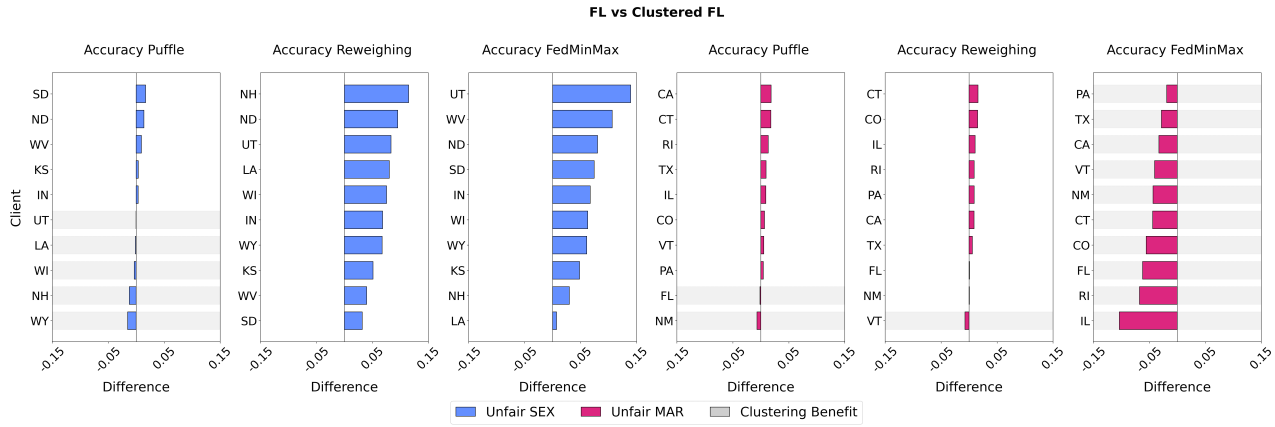

Figure 9: Income Dataset: Comparison of accuracy benefits from taking part in clustered FL settings versus mixed FL settings across three methods (PUFFLE, Reweighing, and FedMinMax).

to improve fairness. Here, the goal is to train an individual model incorporating the objectives of each cluster for each cluster. As clusters are created after the initial federation is found, this method heavily relies on clients being aware of the bias encoded in their data and therefore, their objectives. Also, to form these clusters, clients need to reveal highly sensitive information and statistics about their local data. Additionally, this methodology comes with higher costs for the central server as multiple models are trained and need to be aggregated. Hence, this highlights the need for custom strategies which preserve the privacy of each client's sensitive information and data, reduce computational cost and enhance the potential of fair FL settings when clients hold multiple different objectives.

6 Future Directions and Conclusion

This study evaluates the benefits for individual clients participating in fair FL settings with varying objectives. We focus on scenarios where clients hold data biased toward different sensitive attributes, reflecting real-world FL challenges. A key question that will be explored in further research is how federations where clients are biased toward the same sensitive attribute but different attribute values behave. Our work considers binary sensitive attributes, as they dominate current fair FL methods [33]. Further investigation is needed to understand the benefits of Fair FL methods when applied to non-binary sensitive attributes. Understanding whether the patterns observed in this paper hold in that scenario could significantly broaden the impact and applicability of these methods.

Our findings reveal that while individual clients exhibit consistent patterns in benefiting from fairness-aware federations across different metrics and methods, locally trained fair models often yield greater benefits. We recommend that locally trained fair models should always be considered as a baseline at the client level as well as to provide support on how to train these models to clients who are unfamiliar with fairness mitigation techniques. Additionally, we emphasise the importance of transparent communication during the recruitment phase of a federation. This ensures that client needs and expectations are considered, leading to outcomes in the FL setting that are perceived as more rewarding and equitable.

For the methods and datasets analysed, most clients benefit more from joining federations which intervene on the sensitive attribute `MAR` instead of `SEX`, likely because of data and bias distribution among clients. However, broader experimentation with additional datasets, distributions and a larger client pool is necessary to determine the relationship between clients' local data and benefits derived from joining a specific federation. Overall, we encourage clients to join FL settings with objectives that differ from their own to potentially benefit from these.

Finally, we highlight how clustering can be a simple yet effective solution for managing disparity distribution across various sensitive attributes in a federation. However, privacy concerns often limit the information sharing needed for cluster formation. This underscores the need for custom strategies to reach the full potential of fair FL settings while respecting privacy requirements.

Acknowledgments

This research was partially supported by the "TOPML: Trading Off Non-Functional Properties of Machine Learning" project funded by Carl Zeiss Foundation, grant number P2021-02-014. And, the European Commission under the NextGeneration EU programme – National Recovery and Resilience Plan (PNRR), under agreements: PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI". The European Union Horizon 2020 program under grant agreement No. 101120763 (TANGO). Views and opinions expressed are, however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. 2020. Mitigating bias in federated learning. https://arxiv.org/abs/2012.02447

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[3] Solon Barocas and Moritz Hardt. 2017. Fairness in machine learning.

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[5] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. 2020. Flower: A Friendly Federated Learning Research Framework. https://arxiv.org/abs/2007.14390

[6] Joseph R Biden. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

[7] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2023. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery* 37, 5 (2023), 1719–1778.

[8] Hongyan Chang and Reza Shokri. 2023. Bias Propagation in Federated Learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=V7CYzdruWdm

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] European Commission. 2019. Ethics guidelines for trustworthy AI.

[11] Luca Corbucci, Mikko A. Heikkilä, David Solans Noguero, Anna Monreale, and Nicolas Kourtellis. 2024. PUFFLE: Balancing Privacy, Utility, and Fairness in Federated Learning. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024) (Frontiers in Artificial Intelligence and Applications, Vol. 392)*, Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto José Bugarín Diz, Jose Maria Alonso-Moral, Senén Barro, and Fredrik Heintz (Eds.). IOS Press, 1639–1647. https://doi.org/10.3233/FAIA240671

[12] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. 23–51 pages.

[13] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 6478–6490. https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html

[14] Christoph Düsing and Philipp Cimiano. 2022. Towards predicting client benefit and contribution in federated learning from data imbalance. 23–29 pages.

[15] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 4052)*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.). Springer, 1–12. https://doi.org/10.1007/11787006_1

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, Shafi Goldwasser (Ed.). ACM, 214–226. https://doi.org/10.1145/2090236.2090255

[17] Michele Fontana, Francesca Naretto, Anna Monreale, and Fosca Giannotti. 2022. Monitoring Fairness in HOLDA. In *HHAI 2022: Augmenting Human Intellect - Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence, Amsterdam, The Netherlands, 13-17 June 2022*, Stefan Schlobach, María Pérez-Ortiz, and Myrthe Tielman (Eds.). Frontiers in Artificial Intelligence and Applications, Vol. 354. IOS Press, 246–248. https://doi.org/10.3233/FAIA220205

[18] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[19] Expert group on how AI principles should be implemented. 2023. AI Governance in Japan.

[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[21] Malhar S. Jere, Tyler Farnan, and Farinaz Koushanfar. 2021. A Taxonomy of Attacks on Federated Learning. *IEEE Security & Privacy* 19, 2 (2021), 20–28. https://doi.org/10.1109/MSEC.2020.3039941

[22] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*, Myra Spiliopoulou, Haixun Wang, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu (Eds.). IEEE Computer Society, 643–650. https://doi.org/10.1109/ICDMW.2011.83

[23] Lingjuan Lyu, Han Yu, and Qiang Yang. 2020. Threats to federated learning: A survey. https://arxiv.org/abs/2003.02133

[24] Tambiama Madiega. 2021. Artificial intelligence act.

[25] Natalia Martínez, Martín Bertrán, and Guillermo Sapiro. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 6755–6764. http://proceedings.mlr.press/v119/martinez20a.html

[26] P. McCullagh and J. A. Nelder. 1989. *Generalized Linear Models*. Chapman & Hall / CRC, London.

[27] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Xiaojin (Jerry) Zhu (Eds.). PMLR, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a.html

[28] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (2021), 35 pages. https://doi.org/10.1145/3457607

[29] Ahmed El Ouadrhiri and Ahmed Abdelhadi. 2022. Differential Privacy for Deep and Federated Learning: A Survey. *IEEE Access* 10 (2022), 22359–22380. https://doi.org/10.1109/ACCESS.2022.3151670

[30] Afroditi Papadaki, Natalia Martínez, Martín Bertrán, Guillermo Sapiro, and Miguel R. D. Rodrigues. 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*. ACM, 142–159. https://doi.org/10.1145/3531146.3533081

[31] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research* 77 (2023), 1113–1201.

[32] Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2021. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation.

[33] Teresa Salazar, Helder Araújo, Alberto Cano, and Pedro Henriques Abreu. 2024. A Survey on Group Fairness in Federated Learning: Challenges, Taxonomy of Solutions and Directions for Future Research. https://arxiv.org/abs/2410.03855

[34] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. https://arxiv.org/abs/1902.04885

[35] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. 2020. Salvaging federated learning by local adaptation. https://arxiv.org/abs/2002.04758

[36] Joshua Zhao, Saurabh Bagchi, Salman Avestimehr, Kevin Chan, Somali Chaterji, Dimitris Dimitriadis, Jiacheng Li, Ninghui Li, Arash Nourian, and Holger Roth. 2025. The Federation Strikes Back: A Survey of Federated Learning Privacy Attacks, Defenses, Applications, and Policy Landscape. *Comput. Surveys* 57, 9 (2025), 1–37.

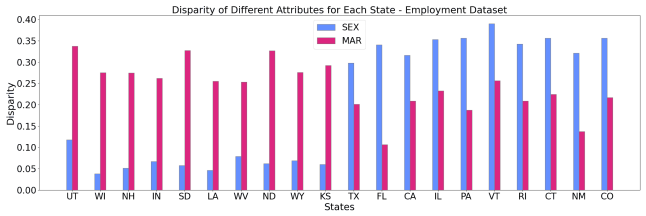

Figure 10: Demographic Parity computed on the ACS Employment dataset of the 20 clients selected in the FL computation.

A Hyperparameter Tuning

To perform hyperparameter tuning when using PUFFLE, we followed the authors' suggestions reported in the paper [11]. Therefore, we performed a Bayesian optimisation in order to minimise the model validation accuracy while keeping the model unfairness under a target $T = 0.05$. The parameters that we optimised are: Learning Rate, Batch Size, optimiser, number of local epochs, and the value of the λ used for the unfairness mitigation. We performed a similar hyperparameter tuning for Reweighing [1]. We searched for the hyperparameters able to maximise the model accuracy while staying under $T = 0.05$. The parameters that we optimised are: Learning Rate, Batch Size, Optimiser, and number of local epochs.

For FedMinMax [30], we tuned the hyperparameters with the code provided by the paper's authors (https://github.com/oscardilley/federated-fairness). Learning rate as well as adverse learning rate were optimised by applying a grid search over all combinations of $[0.002, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0]$. The final goal was to optimise the Equation 3 reported in Section 4.3. We applied the model which showed the best accuracy while having an aggregated disparity smaller than 0.1.

B Preprocessing

In Figure 10, we report the Demographic Parity of the 20 clients used to train the models when using the ACS Employment dataset. As we did with ACS Income, we artificially made the clients unfair toward MAR and SEX attributes to exacerbate the unfairness. As you can see in Figure 10, the first 10 clients of the ACS Employment dataset have a higher disparity for the MAR sensitive attribute than the SEX sensitive attribute. For the last group of 10 clients, instead you can see the opposite scenario with predominant unfairness toward the SEX attribute.

C Additional Results

In this Section, we provide detailed results on the ACS Employment dataset. As we observed similar results for the ACS income dataset, we will only provide a short addition to the statements provided in Section 5.

C.1 Benefits and Harms of the Federation

In Figure 12, we show the comparison of a local baseline without fairness interventions to a local fair baseline, PUFFLE, FedMinMax, and Reweighing for the ACS employment dataset. Clients highlighted in grey show an improved metric by participation in the specified settings. We observe that all unfair clients toward the attribute SEX benefit from reduced disparity scores with regard to SEX across six out of seven methods (first row in Figure 12). The same is observed for clients who hold data biased toward the MAR attribute. These clients are ranked consistently at higher positions if any benefit is observed. In terms of accuracy, Figure 13 shows no improvements, aligning with the well-known trade-off between fairness and accuracy.

For the comparison between local fair models to the fair FL settings for each client, we refer to Figure 14. Here, for clients unfair with regard to MAR we can observe a benefit for five out of six methods in reducing the disparity for the attribute SEX. Concerning the SEX attribute, the subgroup UT, WV, SD, WY, and WI benefit from improved disparity whenever improvements are observed. In the second row, clients from VT, RO, PA, and UT show fairness improvements. In Figure 15, we show that sometimes small accuracy benefits exist when taking part in fair FL settings opposed to learning a fair local model.

C.2 Joining other Federations is Beneficial

In Figure 16 and Figure 17, we show results from comparing the settings which intervene on the SEX attribute with the settings that intervene on the MAR attribute. Clients benefiting from taking part in an FL setting which intervenes on the sensitive attribute SEX are highlighted in light blue, and clients benefiting from an FL setting with intervention on MAR are highlighted in light red. Concerning disparity, in five out of six methods, clients benefit more when taking part in the settings which intervene on MAR. For accuracy, there is no trend across methods. In Figure 10, we show the disparities before training on the client level. Here, clients who are unfair toward SEX still have a high disparity for the attribute MAR. We believe that this initial data distribution is the reason why intervening on MAR has a stronger beneficial influence on reducing disparities for the blue group of clients. Reducing disparity for the MAR attribute is also a reasonable objective for the blue clients and seems to have the side effect that disparities for the attribute SEX are reduced as well.

C.3 Client Clustering improves Fairness

In Figure 18, we analyse the performance of a clustering-based federation, where clients are grouped according to their fairness objectives, in comparison to mixed federations. In this experiment, clients are divided into two clusters based on their fairness objectives. This results in training two separate models: one for the first cluster, reducing the unfairness towards SEX, and another for the second cluster, reducing unfairness towards MAR. The results demonstrate that clustering enhances fairness for clients in the cluster with the objective to enhance fairness for the attribute SEX (this is equal to intervening on the attribute SEX). For the cluster which trains a model intervening on MAR, we do not see an improvement of the disparity with respect to MAR.

Regarding accuracy, Figure 19 reveals a slight improvement when adopting the clustering strategy.

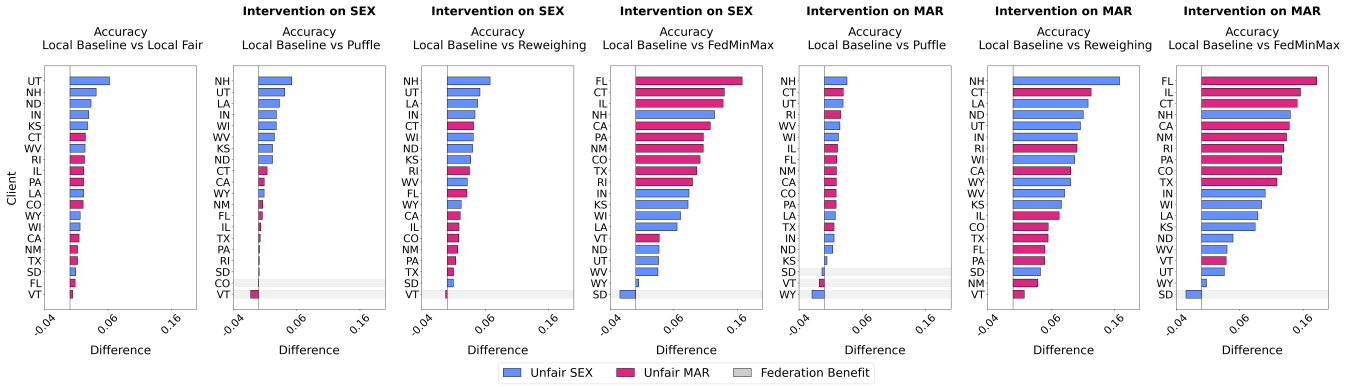

Figure 11: Income Dataset: Difference in accuracy between the baseline local models and four different fairness mitigation strategies (local fair model, PUFFLE, Reweighing and FedMinMax).

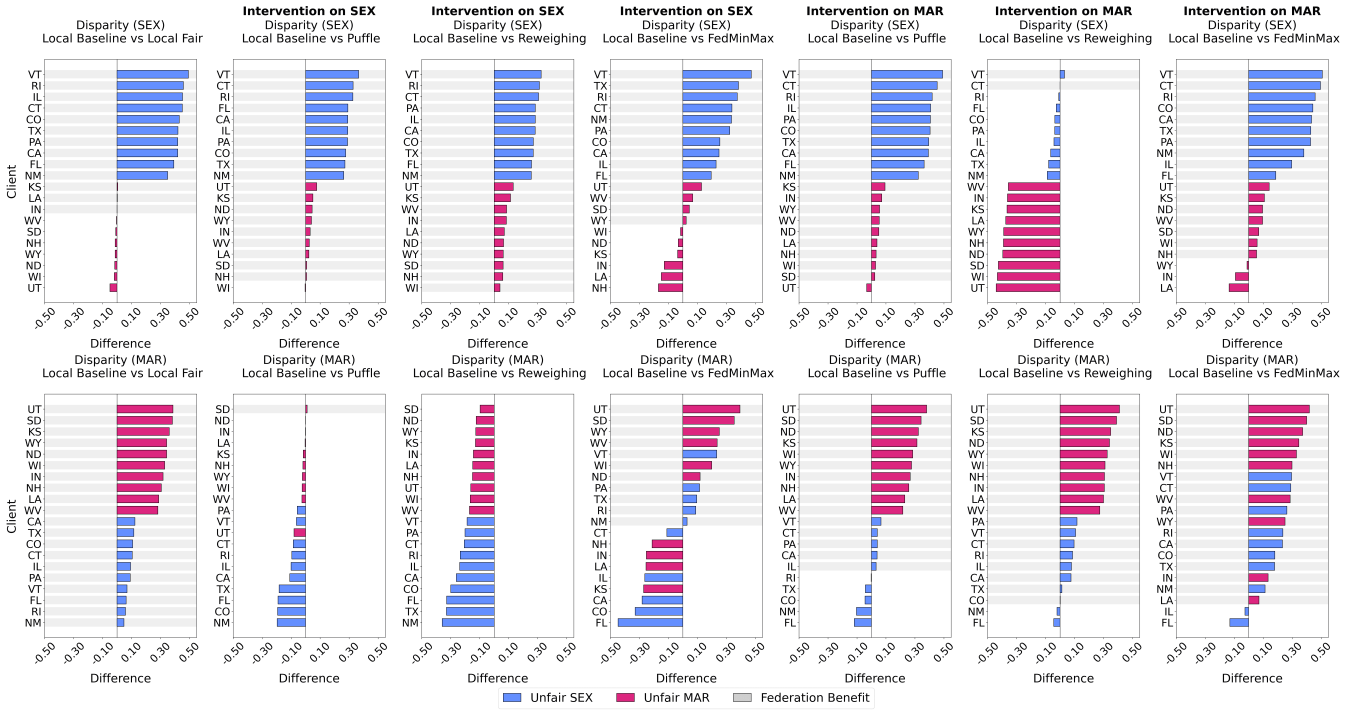

Figure 12: Employment Dataset: Difference in disparities w.r.t. SEX sensitive attribute and MAR sensitive attribute between the baseline local model (one model for each client) and four different fairness mitigation strategies (local fair model, PUFFLE, Reweighing and FedMinMax).

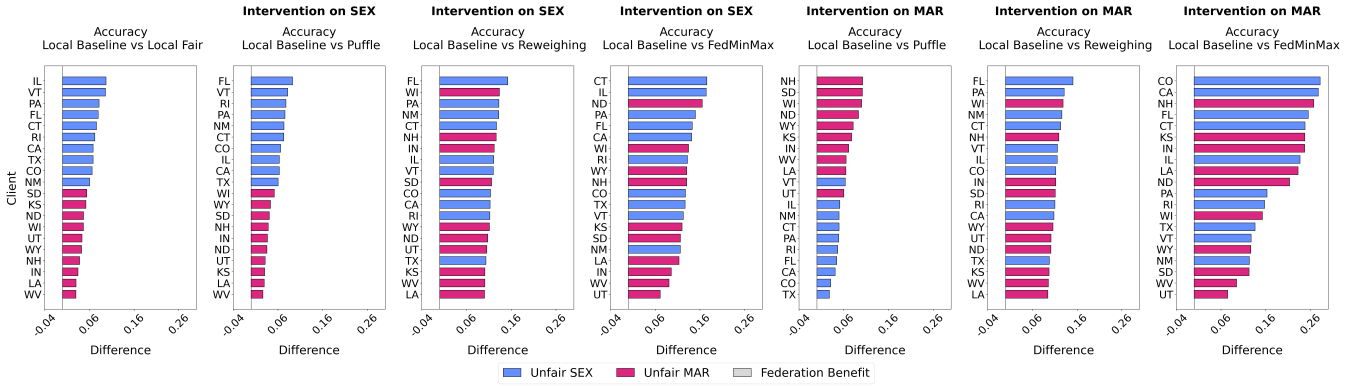

Figure 13: Employment Dataset: Difference in accuracy between the baseline local models and four different fairness mitigation strategies (local fair model, PUFFLE, Reweighing and FedMinMax).

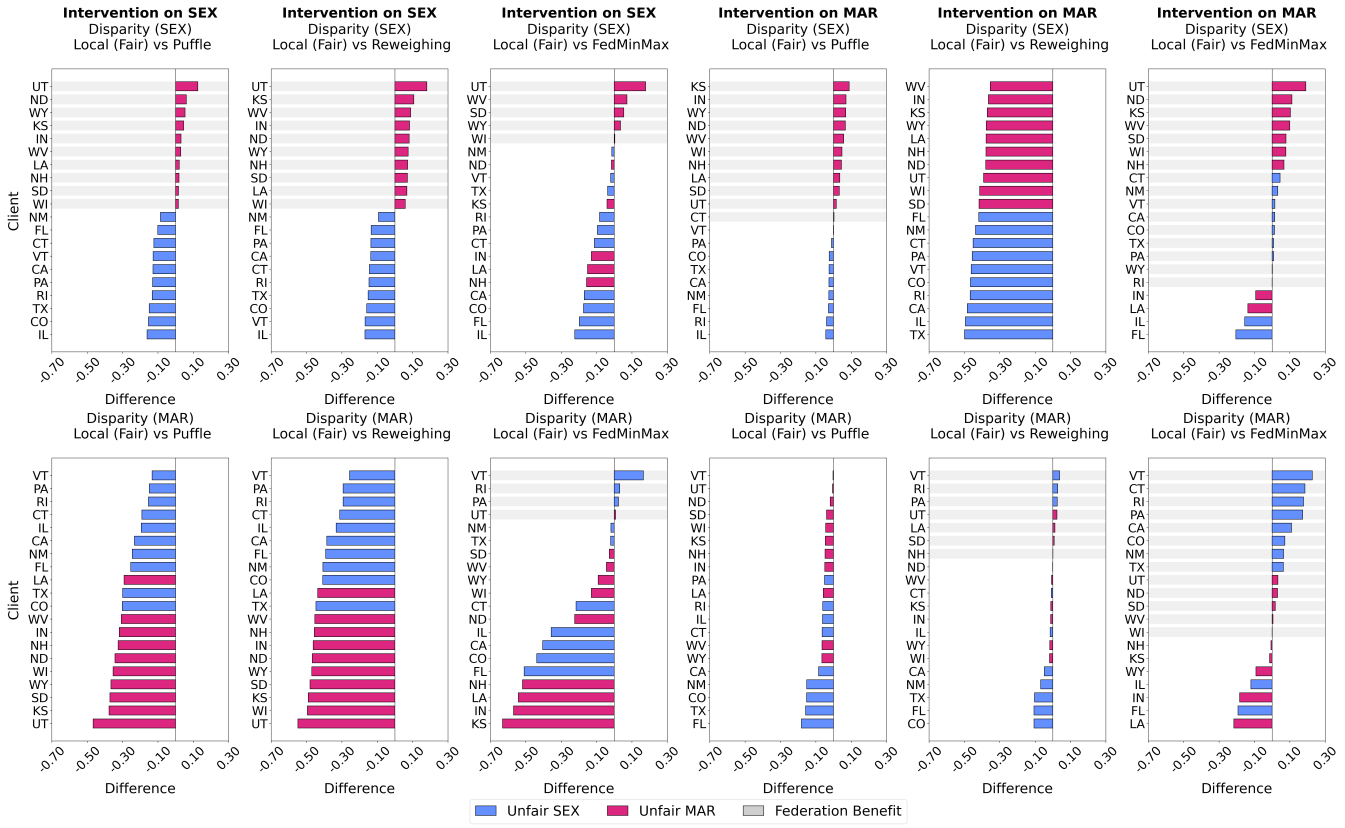

Figure 14: Employment Dataset: Difference in disparities w.r.t. SEX sensitive attribute and MAR sensitive attribute between the local fair model (one model for each client) and three different fair FL methods (PUFFLE, Reweighing and FedMinMax).

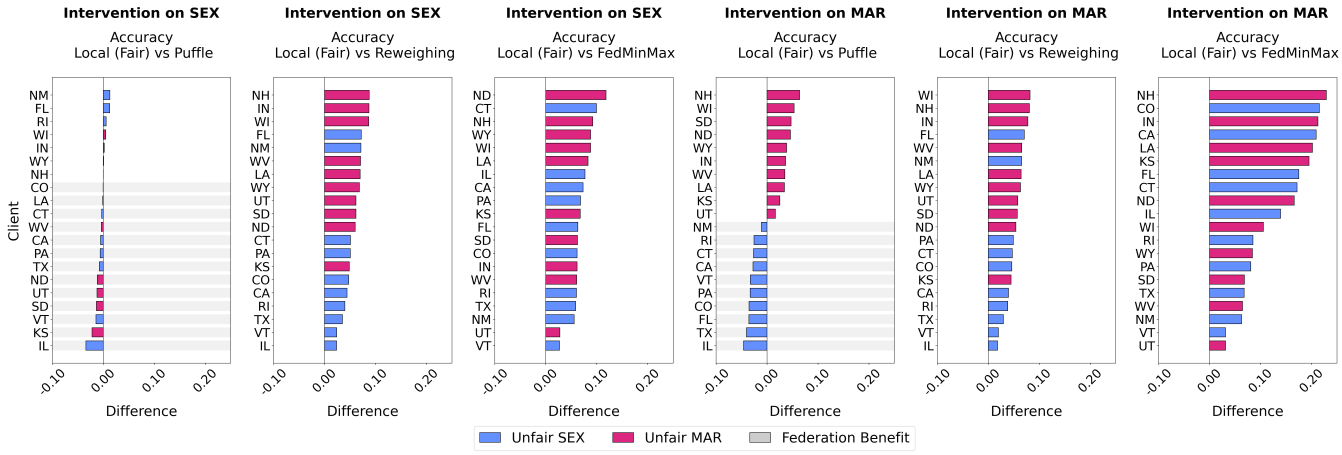

Figure 15: Employment Dataset: Difference in accuracy between the local fair model (one model for each client) and three different fair FL methods (PUFFLE, Reweighing and FedMinMax).

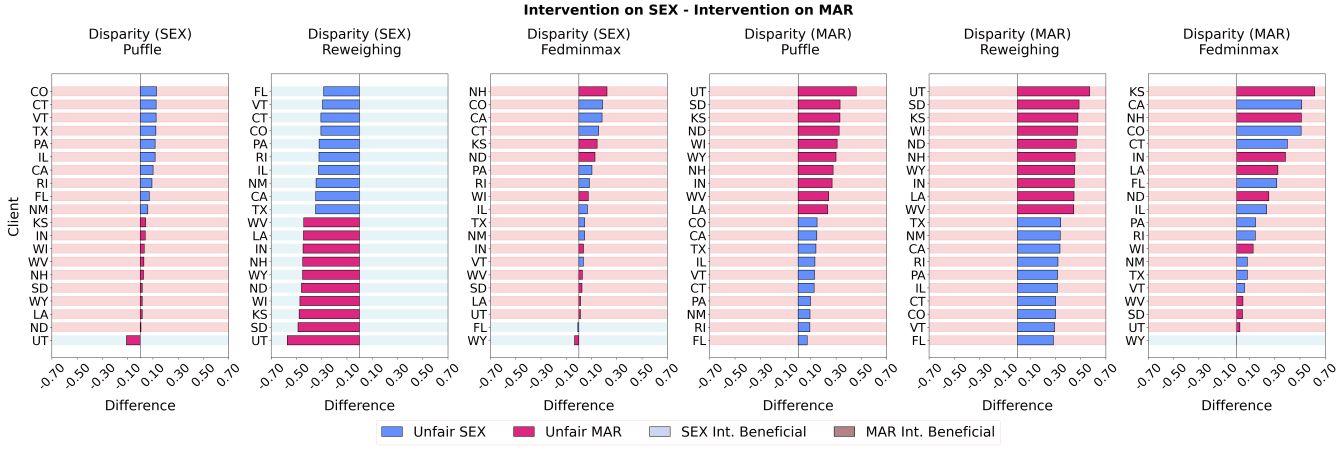

Figure 16: Employment Dataset: Comparison of disparity benefits from taking part in FL settings with different fairness objectives across three methods (PUFFLE, Reweighing and FedMinMax)

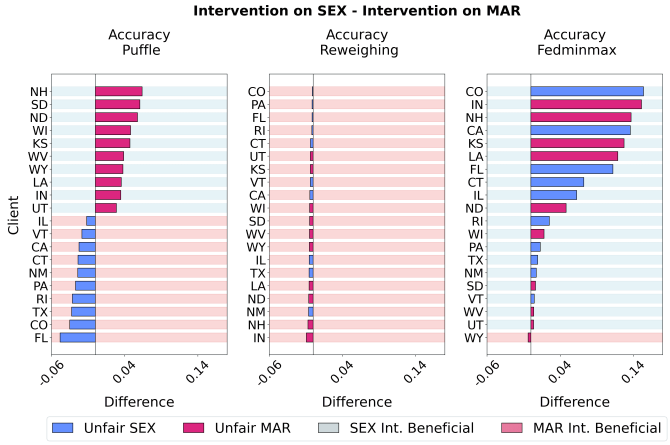

Figure 17: Employment Dataset: Comparison of accuracy benefits from taking part in FL settings with different fairness objectives across three methods (PUFFLE, Reweighing and FedMinMax)

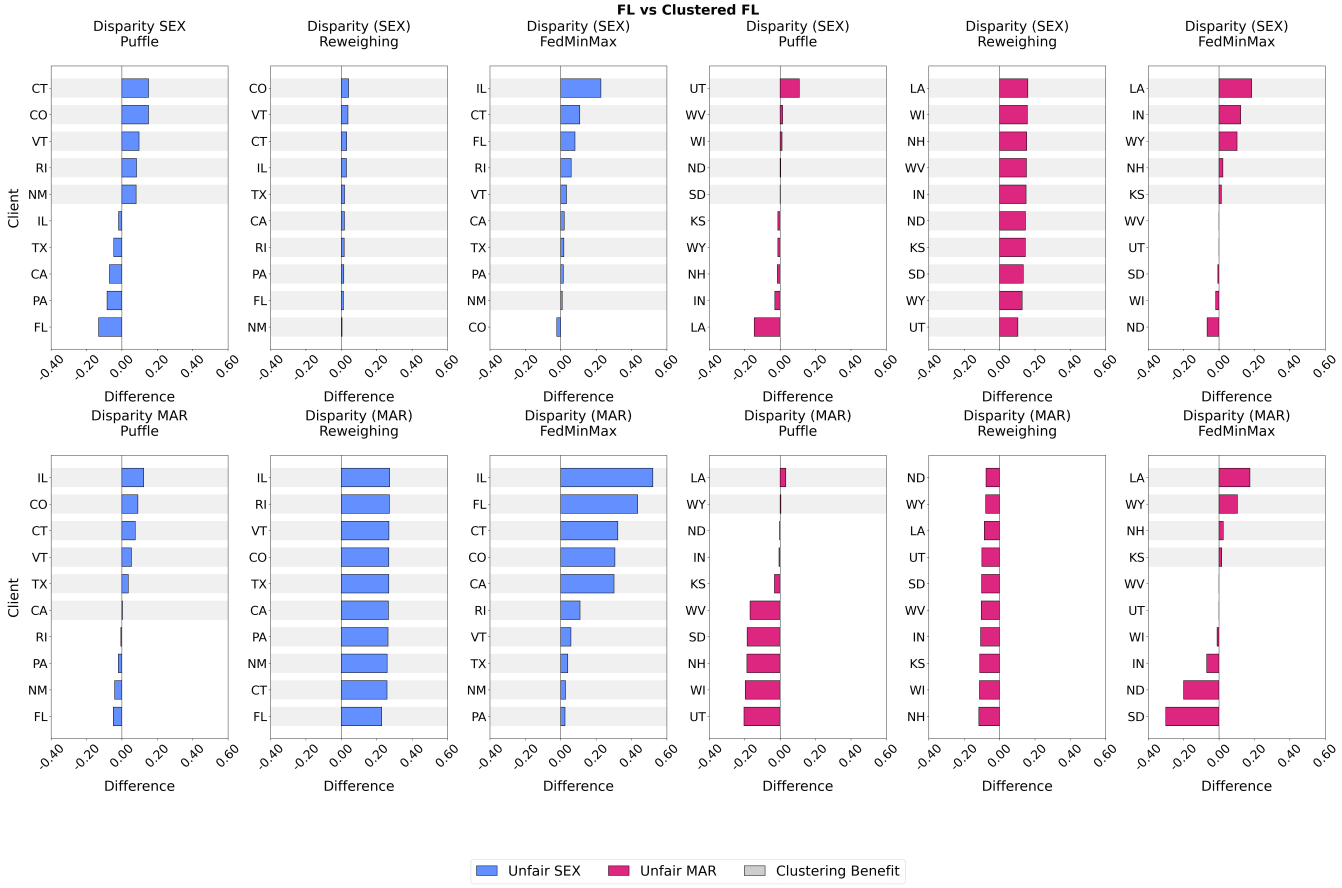

Figure 18: Employment Dataset: Comparison of disparity benefits from taking part in clustered FL settings versus mixed FL settings across three methods (PUFFLE, Reweighing and FedMinMax).

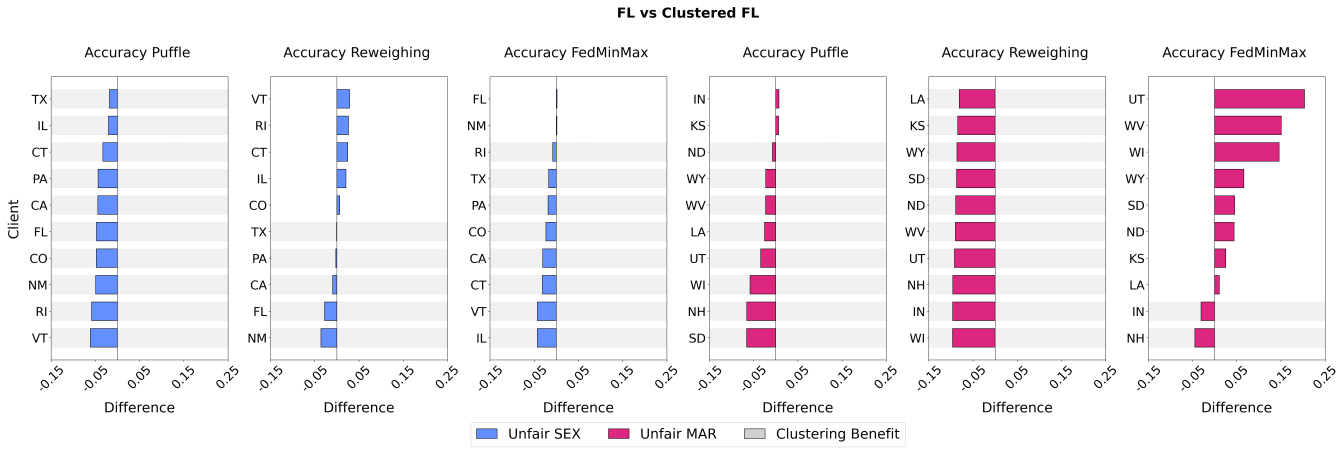

Figure 19: Employment Dataset: Comparison of accuracy benefits from taking part in clustered FL settings versus mixed FL settings across three methods (PUFFLE, Reweighing and FedMinMax).