FeDa4Fair: Client-Level Federated Datasets for Fairness Evaluation

XENIA HEILMANN, Institute of Computer Science, Johannes Gutenberg University, Germany
LUCA CORBUCCI*, Fondazione Bruno Kessler, Italy
MATTIA CERRATO, Institute of Computer Science, Johannes Gutenberg University, Germany
ANNA MONREALE, University of Pisa, Italy

Federated Learning (FL) enables collaborative training while preserving privacy, yet it introduces a critical challenge: the "illusion of fairness". A global model, usually evaluated on the server, appears fair on average while keeping persistent unfairness at the client level. Current fairness-enhancing FL solutions often fall short, as they typically mitigate biases for a single, usually binary, sensitive attribute, while ignoring two realistic and conflicting scenarios: Attribute-Bias (where clients are unfair toward different sensitive attributes) and Value-Bias (where clients exhibit conflicting biases toward different values of the same attribute). To support more robust and reproducible fairness research in FL, we introduce FeDa4Fair, the first benchmarking framework designed to stress-test FL fairness methods under these heterogeneous conditions. Our contributions are three-fold: **(1)** We introduce FeDa4Fair, a library designed to create datasets tailored to evaluating fair FL methods under heterogeneous client bias; **(2)** we release a benchmark suite generated by the FeDa4Fair library to standardize the evaluation of fair FL methods; **(3)** we provide ready-to-use functions for evaluating fairness outcomes for these datasets.

CCS Concepts: • **Computing methodologies** → **Distributed algorithms**; • **General and reference** → **Evaluation**.

ACM Reference Format:
Xenia Heilmann, Luca Corbucci, Mattia Cerrato, and Anna Monreale. 2026. FeDa4Fair: Client-Level Federated Datasets for Fairness Evaluation. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 40 pages. https://doi.org/10.1145/3805689.3812291

1 Introduction

With the increasing use of Machine Learning (ML) in critical economic and societal sectors, the demand for its responsible use is growing. This shift has led to the introduction of several Artificial Intelligence (AI) regulations [6, 11, 31, 39] and the emergence of new research fields such as explainability [7], fairness [9], and user privacy [28].

To mitigate privacy risks in decentralized settings, Federated Learning (FL) [32] has emerged as a standard solution enabling collaborative model training without requiring users, commonly called clients, to share raw data. However, a significant challenge in FL is the inevitable presence of non-independent and identically distributed (non-i.i.d.) data. While substantial progress has been made in addressing the utility degradation caused by non-i.i.d. conditions [23, 30, 36, 48], efforts to mitigate the resulting unfairness remain limited. Existing fair FL approaches [1, 2, 12, 18, 35, 37, 40, 42] typically reduce unfairness for underrepresented groups using group-level metrics such as Demographic Disparity or Equalized Odds Difference [3], and in doing so rely on a critical

*Work was partially done while at the University of Pisa.

Authors' Contact Information: Xenia Heilmann, Institute of Computer Science, Johannes Gutenberg University, Mainz, Germany, xenia.heilmann@uni-mainz.de; Luca Corbucci, Fondazione Bruno Kessler, Trento, Italy, lcorbucci@fbk.eu; Mattia Cerrato, Institute of Computer Science, Johannes Gutenberg University, Mainz, Germany, mcerrato@uni-mainz.de; Anna Monreale, University of Pisa, Pisa, Italy, anna.monreale@unipi.it.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812291

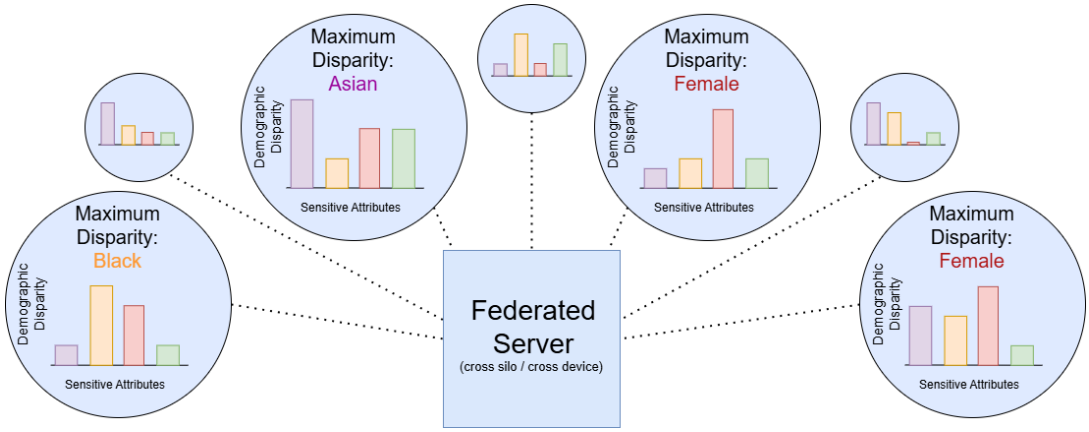


Fig. 1. A pictorial representation of the FL scenarios tackled by FeDa4Fair. Four simulated clients exhibit varying levels of maximal bias, depicted as Demographic Disparity values. Two clients are unfair toward different values of the same sensitive attribute (Value-Bias: one toward `African-American`, one toward `Asian` within Race). Across all four clients, they are unfair toward different sensitive attributes altogether (Attribute-Bias: two toward Race, two toward Sex).

simplifying assumption: that bias is uniform across the federation, both in terms of sensitive attributes and affected subpopulations.

In practice, this assumption rarely holds. Real-world federated settings are inherently heterogeneous. Clients operate within distinct legal, cultural, and socioeconomic contexts that induce conflicting biases. For instance, one hospital (client) may exhibit data bias against a specific ethnic group due to local demographics, whereas another may exhibit bias against a gender group due to differences in data collection pipelines. Under such heterogeneity, evaluating a global model on the server using a single fairness metric can create an illusion of fairness: the model may appear fair on average while remaining substantially unfair for individual clients or intersectional groups [13]. This has real consequences: Clients who are harmed by the global model have little incentive to further participate in the federation, bias propagates across clients [10], and the resulting global model can be both less fair and less accurate than models trained locally [13, 44], defeating the very purpose of federated collaboration.

Despite growing recognition of these issues [13, 22, 44], the field lacks the infrastructure to study them systematically. In this work, we argue that a more granular client-level evaluation framework is necessary to address the current limitations. We identify two specific heterogenous bias scenarios that current benchmarks fail to capture, together covering the primary axes along which client bias diverges in practice: **(I)** Value-Bias, where clients are biased toward conflicting *values* of the same sensitive attribute (e.g., one client is unfair toward `African-American`, while another toward `Asian` within the attribute Race); and **(II)** Attribute-Bias, where clients exhibit bias toward entirely different sensitive *attributes* (e.g., the two already mentioned clients are unfair toward Race while the two other clients are unfair toward Sex). We visualize these scenarios in Figure 1 with four clients and highlight their highest Demographic Disparity. To illustrate why Value-Bias fairness is particularly challenging, consider a federated facial recognition model trained across multiple smartphones. Since users typically store photos of themselves, each client's local dataset is naturally skewed toward their own demographic. A client whose data is biased toward `Asian` faces expects the global model to perform well on that subgroup; a client skewed toward `African-American` faces has the same expectation for theirs. These objectives are inherently in conflict. Standard fair FL methods typically address unfairness only where it is most

pronounced globally. Concretely, they implement fairness constraints for the demographic group (the attribute value) exhibiting maximal unfairness at the federation level. This means Value-Bias fairness constraints are often left unsatisfied. Worse, even when server-side evaluation reports that fairness constraints are met overall, the model may remain unfair for clients whose data is concentrated on specific minority values. Consequently, conflicting client interests can make it difficult, if not impossible, to achieve meaningful fairness in the federated model, even when federation-level mitigation is applied [13].

Attribute-Bias raises analogous but distinct challenges. When clients' data are biased with respect to entirely different attributes (`Sex` for some, `Race` for others), standard mitigation strategies that target a single attribute are misaligned with the problem. Federation-level interventions become ineffective or even harmful to participants whose fairness concerns are not represented in the global objective.

Empirically validating these failure modes requires real-world federated datasets in which both scenarios naturally occur across clients. However, such datasets are exceedingly rare. What is needed instead is a controlled simulation environment that can systematically stress-test fair FL solutions under worst-case bias heterogeneity before real-world deployment. Recent work by Taik et al. [44] and Corbucci et al. [13] explicitly calls for such datasets, highlighting their importance for evaluating FL models in settings where clients may hold diverse or even conflicting fairness objectives. To the best of our knowledge, however, no such resource currently exists.

We address this gap with "Client-Level **Fe**derated **Da**tasets **for Fair**ness Evaluation" (**FeDa4Fair**), a library and benchmarking suite designed to systematically stress-test fair FL solutions under heterogeneous and potentially conflicting bias conditions. While prior work has primarily identified these failure modes theoretically or through limited empirical studies [13, 44], FeDa4Fair provides the infrastructure needed for controlled, systematic, and reproducible experimentation. In doing so, it enables researchers to move beyond federation-averaged fairness metrics and evaluate mitigation strategies under the complex and heterogeneous conditions that real-world deployments are likely to face. Our contributions are as follows:

- **Federated dataset generation library.** We introduce FeDa4Fair[1], a library and dashboard for generating datasets with customizable, heterogeneous bias distributions (e.g., Attribute-Bias or Value-Bias). By providing programmatic mechanisms to simulate client-level biases, it streamlines the evaluation of fair FL in complex scenarios. To support reproducibility, FeDa4Fair includes automated datasheet generation that documents dataset parameters and injected bias configurations.
- **Benchmark suite for heterogeneous fairness evaluation.** We release a benchmarking suite spanning 8 tabular (`ACS Income`, `Dutch Census`) and 4 image (`CelebA`) datasets, each configured to instantiate both Value-Bias and Attribute-Bias conditions, reflecting realistic bias disparities across modalities and application contexts.
- **Empirical evaluation of fairness mitigation under heterogeneity.** We provide a comparative analysis of client-level fairness evaluation against standard federated aggregation (FedAvg) as well as pre-process and in-process fair-FL solutions under Value-Bias and Attribute-Bias conditions. Our results highlight the limitations of current "one-size-fits-all" mitigation strategies and underscore the need for federated fairness techniques that account for bias heterogeneity.

2 Background and Related Work

2.1 Federated Learning

FL [32] is a distributed training framework that enables K clients to train a shared ML model without exposing their private datasets $\mathcal{D}_k$. A central server S orchestrates training by selecting a subset χ of available clients for $r \in [0, R]$ rounds and aggregating their model updates. At round $r = 0$, the server shares a model θ_0 with random weights. The training procedure depends on the chosen aggregation method. A frequently used approach

[1] https://github.com/xheilmann/FeDa4Fair

is Federated Averaging (FedAvg) [32]. In this case, clients update θ_r locally for E epochs before sending updates to S, which aggregates them as $\theta_{r+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \theta_r^k$, where θ_r^k are the parameters updated locally by client k, n_k the dataset size of client k and $n = \sum_{k=1}^{K} n_k$. FedAvg minimizes communication overhead while preserving performance, making it widely applied in FL. Based on the number and the availability of the clients involved in the training, we can distinguish between cross-silo and cross-device FL. In the cross-silo scenario, there are typically tens to hundreds of clients, such as hospitals or companies, that are always available during the training and possess large volumes of data. On the contrary, the cross-device scenario involves a larger number of clients, each holding a small number of data samples. In this case, the clients are only available under specific circumstances, e.g., a client could be a smartphone and could be available only while charging. Several frameworks are available to simulate model training using FL [38]. One of the most well-known frameworks is Flower [5], which we adopted for all the FL experiments presented in this paper. Additionally, the FeDa4Fair library we introduce is fully compatible with Flower, ensuring a seamless integration for researchers and practitioners.

2.2 Fairness in Machine Learning

Fairness in ML refers to the principle that models' predictions should not systematically disadvantage individuals or groups based on sensitive attributes such as `Sex` or `Race`. To quantify how much an ML model is biased, multiple fairness metrics are available in the literature [9, 33].
In this paper, we measure fairness using Demographic Disparity (DD) and Equalized Odds Difference (EOD) [21]. The former builds upon Demographic Parity [3], which requires that the likelihood of a particular prediction outcome must not depend on the membership of 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. DD then evaluates the maximum difference between the Demographic Parity of the different sensitive groups:

$$\max_{y \in Y} \max_{z \in Z} \mathbb{P}(\hat{Y} = y \mid Z = z) - \mathbb{P}(\hat{Y} = y \mid Z \neq z). \tag{2}$$

Equalized Odds instead demands equality of both the true positive and false positive rates across groups. Thus, Equalized Odds Difference is defined as:

$$\max_{y \in Y} \max_{z \in Z} \mathbb{P}(\hat{Y} = y \mid Y = y, Z = z) - \mathbb{P}(\hat{Y} = y \mid Y = y, Z \neq z). \tag{3}$$

For both EOD and DD, lower values indicate fairer models. Additionally, intersectional fairness captures forms of discrimination and societal effects that emerge from the intersection of multiple sensitive features [14]. For instance, `young African-American` people may experience bias not because of `Age` or `Race` alone, but from their intersection. While each group may not be disadvantaged independently, their intersection can be. Addressing intersectional fairness is challenging, as these subgroups are often small and difficult to identify within the data.

2.3 Fairness in Federated Learning

In the FL context, fairness can be interpreted in multiple ways. Early work primarily focused on *Accuracy parity*, aiming to achieve consistent model performance across clients [17, 26, 27, 34, 47, 49], then on mitigating *Group Fairness* [15], which seeks to mitigate disparities across demographic groups. In this paper, we focus on *Group Fairness* mitigation. In the following, we discuss a selection of methods proposed to address group fairness and group bias in FL; we refer the reader to [4, 41] for more comprehensive reviews of the field.

Several families of approaches have been proposed to achieve group fairness in FL. A first family of solutions consists of making the model fair using a *pre-processing* approach, i.e., computing importance weights for different

groups based on their data distributions [1, 42]. We include the former [1] in the benchmarking phase of this paper. Similarly, Pentyala et al. [37] introduced a pre-processing and a post-processing algorithm. Specifically, the post-processing algorithm debiases the predicted outcomes by finding optimal classification thresholds for each group. Alternatively, unfairness in FL can be mitigated using *in-processing* methods. These methods [12, 40] introduce regularization terms during training to steer the model toward fairer outcomes. In the benchmarks of this paper, we adopted the former [12] to evaluate how it behaves on our datasets. A third family of approaches formulates federated training as a *multi-objective optimization* problem [2, 18], balancing multiple desiderata during learning. Typically, objectives include utility maximization, unfairness reduction, privacy preservation, and satisfaction of system-level constraints such as communication efficiency. However, exploring such multi-objective trade-offs is beyond the scope of this work. In all these methods, fairness is typically framed as a global objective, prioritizing bias reduction in the aggregated model [41, 46]. However, a critical limitation of current approaches is the assumption of a uniform fairness objective, i.e., that all clients wish to mitigate bias against the same sensitive attribute (e.g., `Race`) in the same direction (e.g., `African-American`). Recent studies suggest this assumption is invalid in real-world scenarios, where clients exhibit heterogeneous fairness needs due to varying local demographics and regulations [13, 44]. For instance, one client may prioritize gender fairness while another prioritizes racial fairness. These conflicting objectives can exacerbate bias propagation if ignored [10]. Standard mitigation strategies often fail in these settings, as optimizing for a global average can inadvertently harm clients with minority fairness objectives [13], making FL participation worthless or even harmful.

2.4 Federated Learning Benchmarks and Datasets

Despite the growing popularity of FL research, the field lacks standardized benchmarking practices. This is particularly evident in two areas: the absence of a gold standard in terms of datasets that should be experimented on[2] and the inconsistent methodologies for data partitioning used to simulate realistic FL scenarios [20].

LEAF [8] was the first attempt to establish a benchmarking dataset for FL. However, this benchmark only contains a dataset originally designed for the centralized context, which was then adapted to work for FL. Therefore, it does not effectively capture the different client-level distributions that could be present in an inherently federated dataset. To address this limitation, researchers have proposed various approaches for simulating non-i.i.d. data distributions across clients [20]. The use of the Dirichlet distribution is a popular partitioning method to simulate a non-i.i.d. split. Solutions like NIID-Bench [25] proposed the first benchmarking suite to compare different approaches and evaluate the impact on the model quality. More recently, FedArtML [19] proposed a similar solution to simulate and evaluate how data heterogeneity impacts the quality of the FL model.

A similar problem appears when fairness is taken into account during the FL process. Most existing datasets used in fairness literature were originally designed for centralized contexts and lack the natural client-level distributions needed for federated environments [41]. Beyond the issue of data heterogeneity, these datasets also fail to capture the diverse bias settings that can emerge across clients in real-world FL deployments. Specifically, to the best of our knowledge, no consensus exists on benchmark datasets or libraries designed to create data distributions with controlled bias properties such as Attribute-Bias or Value-Bias. Most of the bias patterns explored in the literature assume simplified settings that do not reflect the complex and conflicting bias distribution observed in practice. Real-world federated systems involve clients with different types and degrees of bias, yet such scenarios remain largely unexplored. As also highlighted by Taik et al. [44], the existence of datasets capable of simulating these conditions would benefit the evaluation of fair FL solutions, especially when dealing with clients with varying fairness preferences and objectives. A standardized benchmarking framework capable of

[2]For a summary of federated learning datasets in ML/AI conference publications, see https://flower.ai/blog/2024-12-02-federated-datasets-in-research/

reproducing such bias configurations would support both the development of new bias mitigation techniques and the rigorous evaluation of existing approaches in FL.

3 FeDa4Fair

Meaningful comparison of fair FL methods requires a shift in the evaluation pipeline: from evaluating single global models on the server in simplified client settings to individual-level evaluation in diverse, bias-heterogeneous client settings. To facilitate this and move beyond current fair FL evaluations, we introduce "Client-Level **Fe**derated **Da**tasets **for Fair**ness Evaluation" (**FeDa4Fair**), a library designed to create datasets for this purpose and help researchers to investigate fairness across a wide range of FL scenarios. Unlike existing evaluation pipelines for FL models' fairness evaluation, FeDa4Fair allows researchers to design complex scenarios in which clients have differing objectives and interests, such as different sensitive attributes on which to intervene. The library is designed with four core principles: modality agnosticism, configurable bias injection, fairness evaluation, and accessibility.

3.1 Modality-Agnostic Framework

FeDa4Fair is built on top of the Flower framework [5] and integrates seamlessly with the Hugging Face Hub [24]. This allows FeDa4Fair to ingest any dataset hosted on the Hub, spanning tabular, textual, and image data, and transform it into a partitioned federated benchmark[3] with optional bias injection. In addition to the open-ended opportunities afforded by FeDa4Fair, we release twelve benchmark datasets[4]. These are based on three well-known datasets, spanning two modalities:

- **Tabular:** We use `ACS Income` [16] (derived from US Census data) and the `Dutch Census` dataset [45]. The `ACS Income` dataset offers natural geographic partitioning (by U.S. State) to simulate realistic cross-silo settings. The `Dutch Census` dataset is similar to the previous one but provides a non-US demographic perspective.
- **Image:** We use the `CelebA` dataset [29], a large-scale face attributes dataset. We incorporate it to demonstrate the library's capability to handle tasks where bias is hidden in more complex feature correlations, rather than in tabular columns.

3.2 Configurable Bias Injection

The core contribution of FeDa4Fair is its ability to inject biases to simulate complex federated scenarios with fairness conflicts between clients. This ability allows practitioners to move beyond simple settings and define three distinct fairness scenarios:

- **Homogeneous Bias (Baseline):** The standard scenario found in current literature. The dataset is split into K clients, each biased toward the same value of the same sensitive attribute (e.g., all clients are biased against the value `Female` of the attribute `Sex`). This is the classic scenario considered until now in the literature.
- **Value-Bias:** A scenario where clients possess different levels of biases toward the values of the sensitive attribute they disadvantage. For example, the federation is split such that $K/2$ clients are more strongly biased against the sensitive attribute `Race` with value `African-American` and $K/2$ toward `Asian`.
- **Attribute-Bias:** The most complex scenario, where clients differ in the sensitive attributes toward which they are unfair. For instance, half of the clients are unfair toward the sensitive attribute `Sex` while the other half exhibit bias toward `Race`.

[3]The only prerequisite is the presence of a sensitive attribute (e.g., `Sex`, `Race`) to enable the library's bias injection mechanisms.
[4]https://huggingface.co/datasets/lucacorbucci/FeDa4Fair

To amplify or introduce these biases, FeDa4Fair adapts techniques from data robustness literature [43], specifically, sample dropping to simulate under-representation and label flipping to simulate mislabeling, both applied to data instances with negative labels. While traditionally used as "adversarial attacks", we repurpose these mechanisms to model structural inequities in our controlled environment, allowing researchers to systematically simulate specific demographic disparities and historical labeling errors rather than adversarial noise. To ensure high customizability, the injection process offers granular control over bias injection. Concretely, practitioners indicate the magnitude of bias injection and the location of its injections. In this regard, FeDa4Fair allows the definition of distinct bias profiles for clients individually or for groups of clients. In the first case, the practitioner specifies the sample dropping probability and label flipping probability per client individually. In the case of groups of clients, instead, the sample dropping or label flipping probabilities for each client are drawn from a Gaussian Distribution. Therefore, the practitioner has to specify the mean μ and variance σ of this distribution. This allows us to model varying degrees of severity across the federation.

For the location where bias should be injected, practitioners can specify for which attributes and for which attribute values the manipulation should be applied. Concretely, they select the sensitive attributes which should be considered (e.g., `Sex`, `Race`), and the attribute values to which the manipulation should apply (e.g., `Female`, `African-American`). Furthermore, more granular control is possible by defining an additional attribute and value, e.g., only flip labels for `Female` instances within the `African-American` group, to enable intersectional bias manipulation. Label flipping and sample dropping can be applied independently or combined: for instance, flipping of negative labels for the `African-American` subgroup while dropping at a different probability across datapoints with negative labels for the `Asian` subgroup.

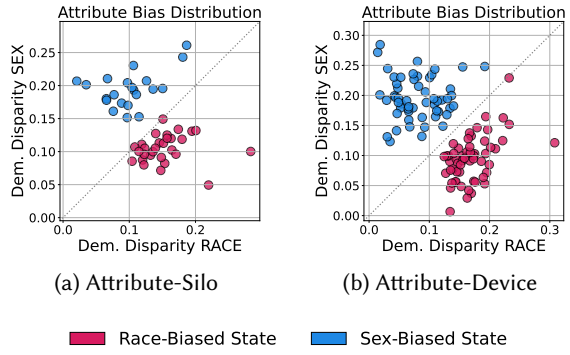
Formally, let $\mathcal{D}_k = \{(x_i, y_i, z_i^1, \ldots, z_i^m)\}_{i=1}^{n_k}$ represent the local dataset of client k, where $y_i \in Y$ represents the ground-truth label, $z_i^1 \in Z^1, \ldots, z_i^m \in Z^m$ values of the sensitive attributes and n_k the dataset size of client k. Practitioners then specify the bias injection policy based on a dropping probability P_{drop} and a flipping probability P_{flip}. They also decide on the sensitive attributes Z^j and attribute values z^j for which the bias should be injected. Finally, P_{drop} and P_{flip} are applied to the subset $\tilde{\mathcal{D}}_k = \{(x_i, y_i, z_i^1, \ldots, z_i^m) | z_i^j = z^j, y_i = y^-\}_{i=1}^{n_k}$ with y^- denoting the negative label.

Finally, the library is compatible with both naturally partitioned and centralized datasets. Here, centralized datasets refer to datasets that are originally designed for centralized ML experiments (e.g., `Dutch Census`, `CelebA`). For these, the library splits the data and optionally injects bias. Naturally partitioned datasets, instead, are the ones that are grouped into partitions directly during the data collection process (e.g., `ACS Income`). In this case, the library supports the natural partitioning into clients while optionally injecting additional biases, or can subdivide large clients into smaller sub-partitions to better simulate a cross-device scenario.

3.3 Fairness Evaluation

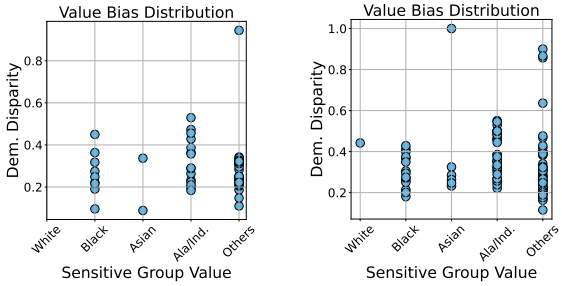
FeDa4Fair provides built-in evaluation tools to evaluate client-level bias distribution. FeDa4Fair supports both individual and intersectional fairness evaluation, allowing bias assessment based on single or combinations of attributes. The library integrates Demographic Disparity (DD) and Equalized Odds Difference (EOD) as metrics for fairness measurement, calculated at three levels of granularity to uncover different types of heterogeneity:

(1) **Attribute Level:** Reports the maximum unfairness for a specific sensitive attribute (e.g., "How unfair are the clients toward the sensitive attribute `Race`?").
(2) **Value Level:** Reports the unfairness of the clients toward specific values of a specific sensitive attribute (e.g., "How unfair are the clients toward the value `Asian` of the sensitive attribute `Race`?").
(3) **Attribute-Value Level:** A matrix of fairness metrics for all possible attribute-value combinations (e.g., "How unfair are the clients toward all different values of the sensitive attribute `Race`?").



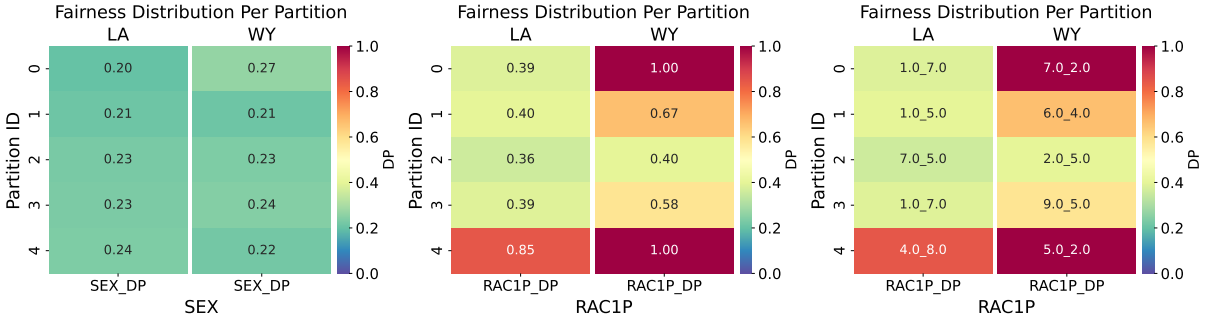
(a) Attribute-Silo (b) Attribute-Device

Fig. 2. Attribute-Bias measured with DD on the local XGBoost models for attribute benchmark datasets using ACS Income.



(a) Value-Silo (b) Value-Device

Fig. 3. Value-Bias measured with DD on the local XGBoost models for value benchmark datasets using ACS Income.



(a) Sex Attribute-Bias. (b) Race Attribute-Bias. (c) Race Value-Bias.

Fig. 4. Attribute-Bias and Value-Bias measured with DD on the true labels and partitioning data from "LA" and "WY" of ACS Income. These plots can be generated for any dataset created with FeDa4Fair.

Figure 2 illustrates examples of Attribute-Bias in datasets. Each dot represents a local XGBoost model trained on a dataset created via FeDa4Fair using ACS Income [16]. The plots highlight clients biased toward Race (in red) and Sex (in blue). Similarly, Figure 3 shows the Value-Bias in datasets, specifically how different clients exhibit varying degrees of maximum DD across the values of the Race attribute. Each dot represents the DD for a specific sensitive group value in a model trained locally on a client.

Similarly, in Figure 4, we present a more detailed analysis of fairness distributions for two states, now also partitioned into 5 clients each. Comparing Figures 4a and 4b, we observe distinct bias patterns across the Sex and Race attributes. Notably, Sex consistently shows lower bias than Race across the clients. Thus, we can easily identify the dominant bias toward specific attributes for each client. As shown in Figure 4c, the highest level of DD spans a broad range of attribute values. For instance, in the state of "LA", the maximum DD for Race occurs between $z = 4$ and $z = 8$ where 4 and 8 are two of the possible values for the sensitive attribute Race.

3.4 Accessibility and Reproducibility

To lower the barrier for interdisciplinary researchers and auditors to apply our framework, FeDa4Fair includes a web-based Interactive Dashboard [5]. This tool provides an interface to:

(1) Load datasets from the Hugging Face Hub or local repositories.
(2) Inspect the distribution of sensitive attributes across clients via interactive charts before training begins, ensuring the injected biases match the desired research scenario.
(3) Export the generated dataset to use it in simulations and benchmarks.

This workflow allows domain experts without deep programming expertise to design fairness benchmarks while ensuring reproducibility. We provide screenshots of the dashboard and details in Appendix A.

Recognizing recent concerns regarding the availability and reliability of FL research datasets [44], FeDa4Fair is designed such that all parameter choices to create the dataset are straightforward to disclose. To further support this goal, we have implemented a datasheet generation feature that semi-automatically provides documentation of the parameters employed to create any dataset, inheriting broader information about FeDa4Fair from a fixed template. With this, we aim for reproducibility and to ensure robust validation of fair FL research.

4 Benchmark Datasets

Besides releasing the FeDa4Fair library to build datasets for FL fairness evaluation, we also propose a new benchmark suite to standardize the evaluation of fair FL methods. These datasets enable a comprehensive evaluation of fair FL methods under diverse bias conditions. While the library supports generating any variation, we release pre-configured benchmarks for three datasets (`ACS Income`, `Dutch Census`, and `CelebA`) spanning two modalities (image and tabular). For each dataset, we consider both the cross-silo and cross-device FL scenario under the Value-Bias and Attribute-Bias setting. Thus, in total, we release twelve datasets grouped into four categories:

(I) **attribute-silo** datasets: Attribute-Bias dataset for the cross-silo setting;
(II) **value-silo** datasets: Value-Bias dataset for the cross-silo setting;
(III) **attribute-device** datasets: Attribute-Bias dataset for the cross-device setting;
(IV) **value-device** datasets: Value-Bias dataset for the cross-device setting.

4.1 `ACS Income`

The core of our benchmark relies on the 2018 `ACS Income` dataset [16]. We provide four configurations of this dataset to test different bias-conflicting scenarios. All configurations are pre-partitioned to simulate the FL clients.

In the case of a **attribute-silo** setting, we leverage the natural distribution of data across the 50 U.S. states and Puerto Rico, focusing on the `Race` and `Sex` attributes. To align with binary-focused fair FL methods [1, 12, 35], we binarize `Race` into `White` and `Others`. The bias of the different clients was exacerbated to achieve a scenario in which all clients satisfy $DD > 0.09$. To meet this target, we iteratively increase the sample drop rate of the more biased attribute between `Race` and `Sex`. Figure 2a illustrates the resulting bias distributions after these modifications. In our evaluation, 22 states exhibit a higher DD value for `Sex`, while 29 states show a higher DD value for `Race` across both models. Further details, including the list of affected states, EOD statistics, and applied data modifications, are provided in Table 1 and Table 3 in Appendix B.

For the **attribute-device** scenario, instead, we derived the dataset from the attribute-silo dataset by splitting each state into six subsets. We then sample from these subsets to create datasets satisfying our bias constraints ($DD > 0.09$ for each client). As a result, the modifications applied to these datasets are directly inherited from the corresponding parent state dataset in the attribute-silo setting (Table 1). Across these subsets, we observe that

[5]The dashboard is available here: https://feda4fair-submission.streamlit.app/

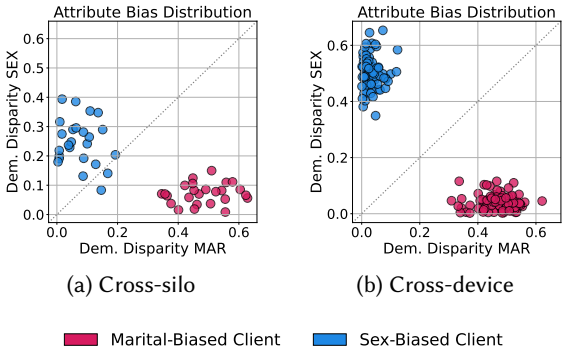


Fig. 5. Attribute-Bias measured with DD on the local XGBoost models for attribute benchmark datasets using Dutch Census.

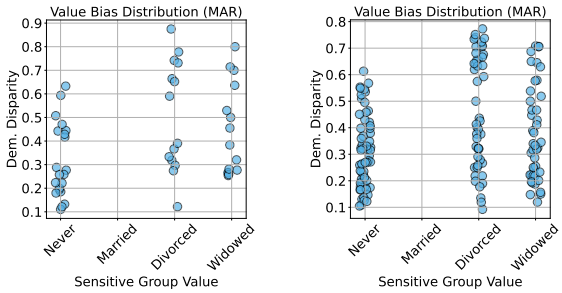


Fig. 6. Value-Bias measured with DD on the local XGBoost models for value benchmark datasets using Dutch Census.

55 states exhibit stronger bias toward Sex and 56 toward Race. These patterns are visualized in Figure 2b and summarized in Table 5 in Appendix B.

As with the attribute-silo dataset, the **value-silo** dataset is built on the natural distribution of the ACS Income dataset across 50 U.S. states and Puerto Rico. However, in this case, to better reflect real-world patterns and to induce different distributions of Value-Bias, we avoided binarizing the sensitive attribute. This choice is motivated by the observation that bias is often present in only a subset of the attributes' values. Moreover, introducing bias into groups that are historically not affected would be inappropriate and undesirable. Instead, we considered multiple classes for the Race attribute (White, African-American, Asian, Alaska Native/American Indian, Others) as a basis for analysis. Within these, we identify the attribute values exhibiting the highest DD values and apply data point dropping to amplify existing biases where $DD > 0.09$ was not satisfied (see Table 2 in Appendix B). The resulting Value-Bias distribution can be found in Figure 3a. In our analysis, we observe the most biased group is African-American in 9 states, Asian in 2 states, Alaska/Indian in 16 states, and Others in 24 states (see Table 4 in Appendix B).

We derived the **value-device** dataset by partitioning the value-silo dataset into four subsets per state and sampling from those that satisfy our bias constraints. The resulting client-level bias distribution includes 1 state with predominant White bias, 15 states with African-American bias, 6 with Asian bias, 31 with Alaska Native/American Indian bias, and 47 with Others bias, as shown in Figure 3b. Further details can be found in Table 2 and 6 in Appendix B.

4.2 Dutch Census

We release a second benchmark based on the Dutch Census [45] dataset. Unlike ACS Income, this dataset is originally centralized; therefore, we apply a partitioning to generate the cross-silo scenarios with 50 Clients and the cross-device scenarios with 150 clients. Here, for the **attribute-silo** and **attribute-device** datasets, we create a conflict between the two sensitive attributes Sex and Mar (marital status). The latter, in this case, we binarized into Married and Not Married. To create both datasets, we apply the library feature that allows us to define groups of users to inject bias into. In the former case, we created two groups of clients; the first one shows higher bias toward the sensitive attribute Sex while it is less biased toward Mar; the latter group has the opposite

configuration. The bias distribution of these datasets is visualized in Figure 5, which shows the DD distribution of clients that exhibit higher bias toward Sex in blue and clients that exhibit higher bias toward Mar in red.

For the **value-silo** and **value-device** datasets, we considered Mar as a sensitive attribute. As for ACS Income, we do not binarize the sensitive attribute in these scenarios. Thus, we have four possible values: Never Married, Married, Divorced, and Widowed. Figures 6a and 6b show the Value-Bias distribution of the different clients. Here, we can see that no clients are unfair toward the value Married. For the value-silo dataset (Figure 6a), clients are distributed among the other three values, while in the value-device scenario, most of the clients are unfair for Never Married and for Divorced. All details about the parameter configuration for these datasets are reported in Appendix B.

4.3 CelebA

Lastly, to show the compatibility of FeDa4Fair with image datasets, we also include in our benchmarking suite the four configurations derived from the CelebA dataset. Similar to Dutch Census, we have 50 clients in the cross-silo settings and 150 clients in the cross-device settings. CelebA, alongside images, also offers a tabular dataset containing information about the people in the images. For instance, for each image we know both the Sex and the Hair Color, two sensitive attributes that we build the bias injection procedure on. The configurations to obtain these datasets are group-based, as for the Dutch Census dataset; we report them in Appendix B Table 7.

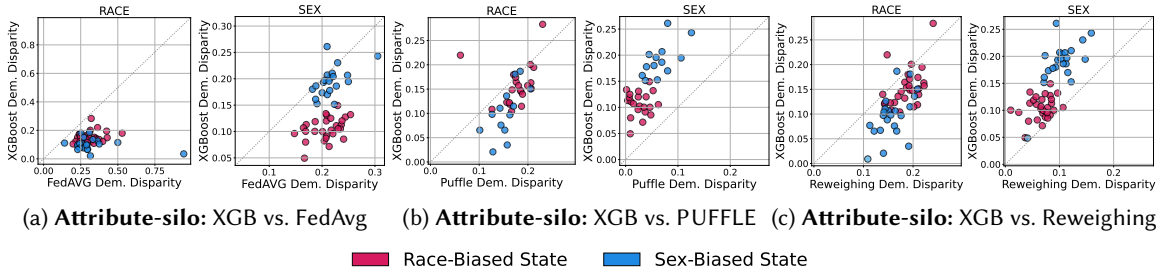
In particular, for the **attribute-silo** and **attribute-device** datasets, about half of the clients have a higher bias for the Sex sensitive attribute than for Hair Color; the other clients instead have the opposite setting. Due to space constraints, we report these results in Figure 31 in Appendix C.3. In Figure 32 in Appendix C.3, instead, we report the distribution of the maximum Demographic Disparity in the case of the **value-silo** and **value-device** datasets. In this case, we created a scenario in which the maximum Demographic Disparity measured on clients is, in about half of the cases, higher for Black Hair and in the other cases higher for Blond Hair.

5 Experimental Analysis

We report here an analysis of the results of the FL simulations executed with the four ACS Income datasets released in our benchmarking suite. Due to space constraints, we report the Dutch Census experiments in Appendix C.2. Unlike the previous two datasets, for CelebA, computational constraints prevented us from running extensive experiments across all dataset splits; we elaborate on this in Appendix C.3.

5.1 Setup

To contextualize our benchmark datasets, we stress-tested three representative FL paradigms. Our goal is not only to empirically show whether current Fair FL methods can handle the sociotechnical heterogeneity introduced in our scenarios, but also to show challenging FL settings, leaving open research questions for the future. We compare locally trained client models (Logistic Regression and XGBoost) with three FL training strategies: a Baseline FL model and two Fair FL methods. For the baseline model, we apply the FedAvg algorithm [32], simulating the FL training with Flower [5], as discussed in Section 2.1. For a first Fair-FL benchmark, we use PUFFLE [12] on top of the FedAvg algorithm to reduce model unfairness measured with DD. Specifically for PUFFLE, during training, selected clients compute the gradient and model fairness on a given batch. The computed fairness metric is then incorporated as an additional regularization term in the model update to mitigate unfairness by summing and weighting using a hyperparameter λ indicating the importance of the model's utility and its fairness. Here, a λ closer to 1 prioritizes fairness. In our experiments, we treat λ as a hyperparameter. More details about the hyperparameters and hardware underlying the experiments are reported in Appendix D, E.



(a) **Attribute-silo:** XGB vs. FedAvg (b) **Attribute-silo:** XGB vs. PUFFLE (c) **Attribute-silo:** XGB vs. Reweighing

Fig. 7. ACS Income: Individual, per attribute DD measured on the local XGBoost models versus the global FedAvg model, the global PUFFLE model, and the global Reweighing model. Values above the diagonal indicate that the DD of the local model is higher than that of the global model, and the global model reduces bias for this client and attribute. Values below the diagonal indicate that the DD of the local model is lower than that of the global model; the global model increases bias for this client and attribute.

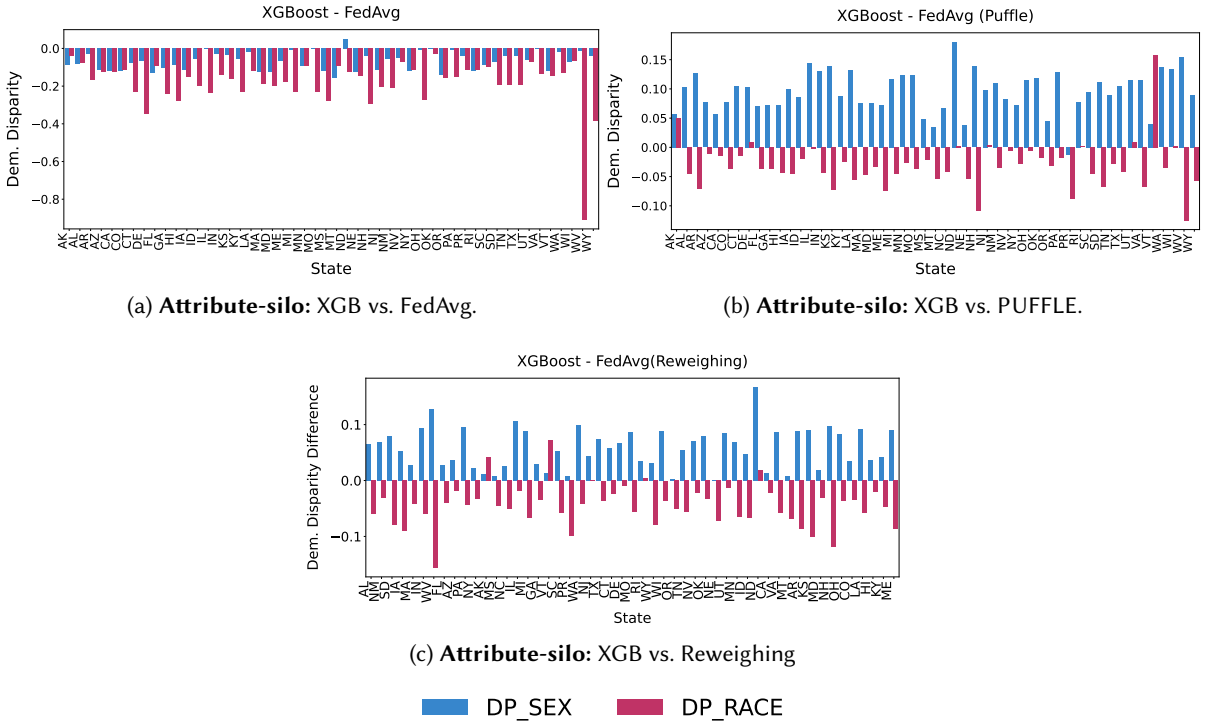
Lastly, we adopt a pre-processing approach for fairness mitigation, Reweighing [1], where each client computes weights based on its local demographic distribution. These weights are applied during the loss computation to penalize errors on underrepresented groups.

In the case of ACS Income, we set for both PUFFLE and Reweighing the strategy to mitigate the unfairness for the Sex sensitive attribute. This choice allows us to evaluate two distinct scenarios: In the attribute-silo datasets, where Sex is one of the sensitive attributes, we assess how mitigation affects the disparity for both Sex-biased clients and differentially biased clients. In contrast, in the value-silo and value-device datasets, we show how mitigating unfairness for Sex influences Value-Bias shifts across the clients. For the experiments run using the PUFFLE method, a target value for the DD is needed to run the experiments. Specifically, we set the target DD=0.05, which represents the DD value that the global model should maximally exhibit toward Sex; however, there is no formal guarantee that this holds for each individual client. A similar approach is not included in the Reweighing method, which performs the simulation while reducing the unfairness as much as possible.

5.2 Evaluation

We assess our methodology based on two principles: quantifying fairness at the individual client level and analysing shifts in bias distribution before and after the application of fair FL methods. Our analysis distinguishes between two FL scenarios:

- Cross-Silo: In this scenario, typically, clients possess sufficient data to perform a local train/test split. Therefore, we train individual client-specific models and compare their performance to that of the global FL models. This allows us to determine whether participation in FL mitigates bias for specific demographic groups or individuals, or conversely, if the existing bias is propagated or even exacerbated.
- Cross-Device: this case involves clients with limited local data, which is insufficient for training robust local models. Here, FL represents the only feasible way to obtain a usable model. In these scenarios, a common practice is to partition clients into a train and a test group. However, a challenge arises: local fairness metrics can only be computed on true labels or external model predictions. A potential solution is to compare the fairness metrics on test clients using the FL model against those obtained from a set of external models trained on similar datasets, possibly provided by the FL orchestrator. However, we leave this for future discussion and evaluate fairness on 30 test clients, which all hold enough data to train a local model.



(a) **Attribute-silo:** XGB vs. FedAvg.

(b) **Attribute-silo:** XGB vs. PUFFLE.

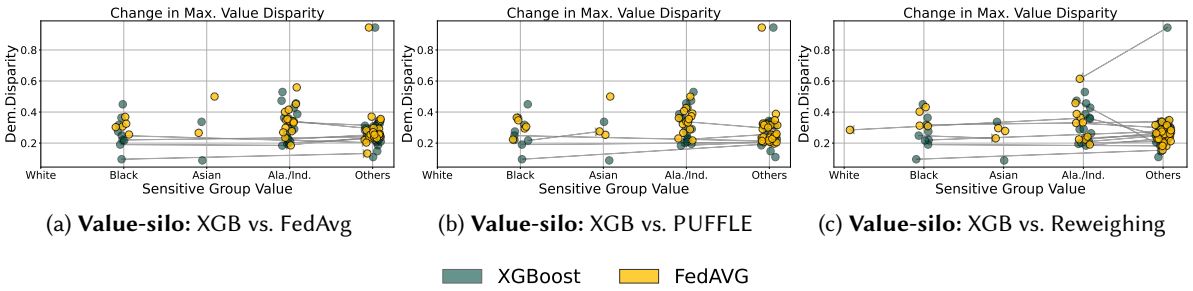
(c) **Attribute-silo:** XGB vs. Reweighing

Fig. 8. `ACS Income`: Individual, per attribute, bias differences in DD measured on the local XGBoost model minus the global FedAvg model, the global PUFFLE model, and the global Reweighing model. Values above 0 indicate that the DD of the local model is higher than that of the global model, and the global model reduces bias for this client and attribute. Values below 0 indicate that the DD of the local model is lower than that of the global model; the global model increases bias for this client and attribute.

Attribute-Bias benchmarks: The results for the **attribute-silo** dataset illustrated in Figure 7 reveal a trade-off. In these plots, the diagonal line represents a parity line: points below the diagonal line indicate an increase in DD, i.e., an increase in unfairness after FL training compared to the local models. On the contrary, dots above the diagonal suggest greater unfairness with the local XGBoost model. As shown in Figure 7a, training the baseline with the standard FedAvg drives the majority of clients below this line for both `Race` and `Sex`, increasing the DD unfairness and confirming that FL methods without fairness constraints propagate bias across the federation, as previous research has shown [10].

When PUFFLE is applied, it effectively enforces fairness constraints on the `Sex` attribute, resulting in a disparity reduction for `Sex` across all clients. This is visualized in Figure 7b on the right, where all clients are above the diagonal line. However, this comes at the cost of increased `Race` disparity compared to the local XGBoost model as depicted in Figure 7b on the left. Thus, clients with maximal bias towards the `Race` attribute profit less, in terms of bias reduction on their most biased attribute, from participating in the FL training when using PUFFLE.

For Reweighing, a similar trend as for PUFFLE is visible in Figure 7c. The method effectively enforces fairness constraints on the `Sex` attribute, resulting in a disparity reduction for `Sex` across all clients. This is visualized



(a) **Value-silo:** XGB vs. FedAvg (b) **Value-silo:** XGB vs. PUFFLE (c) **Value-silo:** XGB vs. Reweighing

Fig. 9. `ACS Income`: Individual, per value DD measured on the local XGBoost models versus the global FedAvg model, the global PUFFLE model, and the global Reweighing model. The lines indicate how the value changes from local to FL, showing a movement towards underrepresented values when measuring on the global FL models.

in Figure 7b. Again, on the right, all clients are above the diagonal line, indicating that the DD values for `Sex` decrease. However, with regard to `Race` disparity, Reweighing increases this value compared to the local XGBoost models. For both Puffle and Reweighing, however, the overall `Race` disparity is lower than for FedAvg, where the majority of DD values are above 0.25.

This trend is also visible from the **client-level** evaluation reported in Figure 8, where we compare the local XGBoost models with the FL models. While for FedAvg the unfairness increases upon the local models, PUFFLE and Reweighing decrease the unfairness with regard to the `Sex` attribute. Only for some, such as "AK", "VT", PUFFLE and Reweighing also decrease the unfairness for `Race`, highlighting the limitations of single-attribute mitigation in heterogeneous settings.

This trend remains consistent when using Logistic Regression as the local model (Figure 14, 20 in Appendix C.1.1). Additional results with the **attribute-device** dataset are detailed in Appendix C.1.1.

Value-Bias benchmarks. For the **value-silo** dataset, Figure 9 shows both how DD changes when moving from local training to FL and how the distribution of the attribute value associated with the maximum DD shifts. A clear pattern emerges: underrepresented groups, specifically, `Alaska Native/American Indian` and `Others`, are disproportionately harmed by participation in FL. In particular, clusters corresponding to the `Others` category increase density after FL training, indicating that this group is more frequently associated with the highest DD values. This effect becomes more pronounced under PUFFLE and even more with Reweighing, where an increasing number of clients report `Others` as the attribute value with the maximum DD. Despite these shifts, the overall `Race` disparity shows only a marginal improvement trend across all three settings. A more detailed individual-level analysis is provided in Figure 21 in Appendix C.1.2. We observe a consistent trend when using Logistic Regression as the local model, as shown in Figure 19 and Figure 22 in Appendix C.1.2. Additional results with the **value-device** dataset are detailed in Appendix C.1.2.

Overall, we observe consistent results across both cross-silo and cross-device settings for each type of bias-heterogeneous client configuration. However, the specific form of bias heterogeneity, namely, whether it arises from different attributes or from different attribute values, leads to different outcomes. This observation highlights the importance of evaluating fair FL approaches across a diverse range of bias scenarios to assess their robustness and ensure equitable performance.

5.3 Key Takeaways

Our benchmarking datasets reveal several limitations in current Fair FL paradigms:

- **Fairness effects are client-specific:** While global fairness metrics may show an improvement, client-level analysis may reveal substantial heterogeneity: FL participation can mitigate, preserve, or exacerbate bias depending on each client's original data distribution before potential manipulation.
- **Single-attribute mitigation has limitations:** Current Fair FL methods successfully reduce disparity for the targeted sensitive attribute but can increase unfairness for other attributes, particularly in bias-heterogeneous federations where clients have conflicting biases (Figure 7 and 8).
- **Value-Bias exposes limitations of current Fair FL solutions:** Existing approaches typically rely on minimizing the maximum DD for a given sensitive attribute. These methods ignore which specific attribute value is responsible for the highest disparity. In Value-Bias heterogeneous settings, this leads to patterns where underrepresented groups are negatively affected by participation in FL training despite the low unfairness of the global aggregated model (Figure 9).
- **Biases in heterogeneous settings:** A shift is required in both the design and evaluation of Fair FL methods. Developing solutions that can handle multiple Attribute-Bias and Value-Bias heterogeneity remains an open challenge, highlighting the need for future work on more advanced Fair FL solutions.

6 Conclusion

Current fairness evaluation in FL is usually based on the assumption of a uniform bias distribution across the clients. This often creates an "illusion of fairness" at the global level while ignoring the complex, heterogeneous biases that exist at the individual client's level. Previous work [10, 13] and our analysis performed on the benchmarking datasets created with the FeDa4Fair library revealed that current state-of-the-art solutions are insufficient in realistic scenarios involving Value-Bias and Attribute-Bias.

To address this limitation, we introduced FeDa4Fair, a library designed to create datasets for FL experimentation with heterogeneous client bias. By providing a reproducible and extensible framework, FeDa4Fair is a crucial first step towards enabling more robust and systematic evaluation of fair FL methodologies. To support this goal, we also published twelve benchmarking datasets spanning tabular and image data, cross-silo and cross-device scenarios, as well as Attribute-Bias and Value-Bias settings.

Thanks to its extensibility, FeDa4Fair is designed to support the continual expansion and diversification of the fairness evaluation landscape in FL. This facilitates two critical directions for future research: first, the inclusion of even more complex real-world scenarios and second, the investigation of fairness in cross-device settings where clients possess only limited data.

Acknowledgment

XH and MC were supported by the "TOPML: Trading Off Non-Functional Properties of Machine Learning" project funded by Carl Zeiss Foundation, grant number P2021-02-014. AM and LC were supported by the Italian Project Fondo Italiano per la Scienza FIS00001966 "MIMOSA" and by the European Union's Horizon Europe research and innovation program "TANGO" under G.A. 101120763. 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.

Generative AI usage statement

LLMs were used to aid non-native speakers with grammar and word corrections. During the implementation of the code needed for this submission, we used LLMs for autocompletion of code and to fix bugs.

Ethical Considerations Statement

FeDa4Fair allows for manipulating datasets such that they reflect more complex bias patterns to facilitate fair FL method evaluation and to improve bias mitigation development. However, we stress that manipulated data is, of course, no longer representative of the state-level distributions and demographics in the U.S. and Puerto Rico (for datasets based on ACS) or the Netherlands (for datasets based on Dutch). We strongly advise against employing data manipulated with FeDa4Fair for anything but the testing and development of federated bias mitigation strategies.

References

[1] Annie Abay et al. 2020. Mitigating Bias in Federated Learning. arXiv:2012.02447 [cs.LG]

[2] Maryam Badar et al. 2024. TrustFed: Navigating Trade-offs Between Performance, Fairness, and Privacy in Federated Learning. In *ECAI*, Vol. 392. 2370–2377.

[3] Barocas et al. 2017. Fairness in machine learning. *NeurIPS tutorial* 1, 2 (2017).

[4] Nawel Benarba and Sara Bouchenak. 2025. Bias in federated learning: A comprehensive survey. *Comput. Surveys* 57, 11 (2025), 1–36.

[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] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. 2019. LEAF: A Benchmark for Federated Settings. arXiv:1812.01097 [cs.LG] https://arxiv.org/abs/1812.01097

[9] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7, Article 166 (April 2024), 38 pages. https://doi.org/10.1145/3616865

[10] Hongyan Chang and Reza Shokri. 2023. Bias Propagation in Federated Learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023.*

[11] European Commission. 2019. *Ethics guidelines for trustworthy AI.* Publications Office. https://doi.org/doi/10.2759/346720

[12] 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

[13] Luca Corbucci, Xenia Heilmann, and Mattia Cerrato. 2025. Benefits of the Federation? Analyzing the Impact of Fair Federated Learning at the Client Level. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '25).* 2232–2248. https://doi.org/10.1145/3715275.3732152

[14] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories.* Routledge, 23–51.

[15] Dwork Cynthia et al. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*. 13 pages.
[16] 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
[17] Kate Donahue et al. 2021. Models of fairness in federated learning.
[18] Borja Rodríguez Gálvez et al. 2021. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS*.
[19] Daniel Mauricio Jimenez Gutierrez, Aris Anagnostopoulos, Ioannis Chatzigiannakis, and Andrea Vitaletti. 2024. Fedartml: A tool to facilitate the generation of non-iid datasets in a controlled way to support federated learning research. *IEEE Access* 12 (2024), 81004–81016.
[20] Daniel Mauricio Jimenez Gutierrez, David Solans, Mikko Heikkila, Andrea Vitaletti, Nicolas Kourtellis, Aris Anagnostopoulos, and Ioannis Chatzigiannakis. 2024. Non-IID data in Federated Learning: A Survey with Taxonomy, Metrics, Methods, Frameworks and Future Directions. arXiv:2411.12377 [cs.LG] https://arxiv.org/abs/2411.12377
[21] 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
[22] Xenia Heilmann, Luca Corbucci, and Mattia Cerrato. 2025. A Benchmark for Client-level Fairness in Federated Learning. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (Proceedings of Machine Learning Research, Vol. 294)*. PMLR.
[23] Mehdi Karami and Amin Karami. 2025. Harmony in federated learning: A comprehensive review of techniques to tackle heterogeneity and non-IID data. *Cluster Computing* 28, 9 (2025), 570.
[24] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 175–184. arXiv:2109.02846 [cs.CL] https://aclanthology.org/2021.emnlp-demo.21
[25] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. 2022. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 965–978.
[26] Tian Li et al. 2020. Fair Resource Allocation in Federated Learning. In *ICLR*.
[27] Tian Li et al. 2021. Ditto: Fair and Robust Federated Learning Through Personalization. In *ICML*.
[28] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* 54, 2, Article 31 (March 2021), 36 pages. https://doi.org/10.1145/3436755
[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*. IEEE Computer Society, 3730–3738. https://doi.org/10.1109/ICCV.2015.425
[30] Zili Lu, Heng Pan, Yueyue Dai, Xueming Si, and Yan Zhang. 2024. Federated Learning With Non-IID Data: A Survey. *IEEE Internet of Things Journal* 11, 11 (2024), 19188–19209. https://doi.org/10.1109/JIOT.2024.3376548
[31] Tambiama Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).
[32] 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
[33] 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
[34] Mehryar Mohri et al. 2019. Agnostic Federated Learning. In *ICML*, Vol. 97. 4615–4625.
[35] 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
[36] Jiaming Pei, Wenxuan Liu, Jinhai Li, Lukun Wang, and Chao Liu. 2024. A Review of Federated Learning Methods in Heterogeneous Scenarios. *IEEE Transactions on Consumer Electronics* 70, 3 (2024), 5983–5999. https://doi.org/10.1109/TCE.2024.3385440
[37] Pentyala et al. 2022. Privfairfl: Privacy-preserving group fairness in federated learning. *ArXiv prep.* abs/2205.11584 (2022).
[38] Walter Riviera, Ilaria Boscolo Galazzo, and Gloria Menegaz. 2023. FeLebrities: A User-Centric Assessment of Federated Learning Frameworks. *IEEE Access* 11 (2023), 96865–96878. https://doi.org/10.1109/ACCESS.2023.3312579

[39] Huw Roberts et al. 2021. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & society* (2021).
[40] Yves Rychener, Daniel Kuhn, and Yifan Hu. 2025. Global Group Fairness in Federated Learning via Function Tracking. arXiv:2503.15163 [cs.LG] https://arxiv.org/abs/2503.15163
[41] Teresa Salazar, Helder Araújo, Alberto Cano, and Pedro Henriques Abreu. 2026. A survey on group fairness in federated learning: Challenges, taxonomy of solutions and directions for future research. *Artificial Intelligence Review* 59, 2 (2026), 81.
[42] Khotso Selialia et al. 2024. Mitigating Group Bias in Federated Learning for Heterogeneous Devices. In *ACM FAccT*. 1043–1054.
[43] David Solans, Battista Biggio, and Carlos Castillo. 2021. Poisoning Attacks on Algorithmic Fairness. In *Machine Learning and Knowledge Discovery in Databases*, Frank Hutter, Kristian Kersting, Jefrey Lijffijt, and Isabel Valera (Eds.). Springer International Publishing, Cham, 162–177.
[44] Afaf Taik, Khaoula Chehbouni, and Golnoosh Farnadi. 2025. Fairness in Federated Learning: Fairness for Whom?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2456–2469.
[45] Paul Van der Laan. 2001. The 2001 census in the Netherlands: Integration of registers and surveys. In *Conference at the Cathie Marsh Centre*.
[46] Sean Vucinich and Qiang Zhu. 2023. The Current State and Challenges of Fairness in Federated Learning. *IEEE Access* 11 (2023), 80903–80914. https://doi.org/10.1109/ACCESS.2023.3295412
[47] Wang et al. 2021. Federated learning with fair averaging. *ArXiv prep.* abs/2104.14937 (2021).
[48] Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. 2023. Heterogeneous Federated Learning: State-of-the-art and Research Challenges. *ACM Comput. Surv.* 56, 3, Article 79 (Oct. 2023), 44 pages. https://doi.org/10.1145/3625558
[49] Zhiyuan Zhao et al. 2022. A dynamic reweighting strategy for fair federated learning. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE.

A FeDa4Fair Dashboard

FeDa4Fair has been developed with high accessibility in mind. This is why we included a dashboard developed using Streamlit [6] that allows practitioners to easily access all the features offered by FeDa4Fair without the need to write Python code [7].

As you can see in Figure 10, the dashboard allows practitioners to select a dataset (either from one of the locally supported datasets or one from the Hugging Face Hub) and split it, simulating the assignment of each partition to a separate FL client. It is also possible to define configurations to inject bias for certain groups of clients or for all participants. Moreover, our dashboard includes the ability to plot the unfairness of the different clients, measuring it with both Demographic Disparity and Equalized Odds Difference. As shown in Figure 11, this is useful for an easy visualization of the configurations defined for the bias injection.

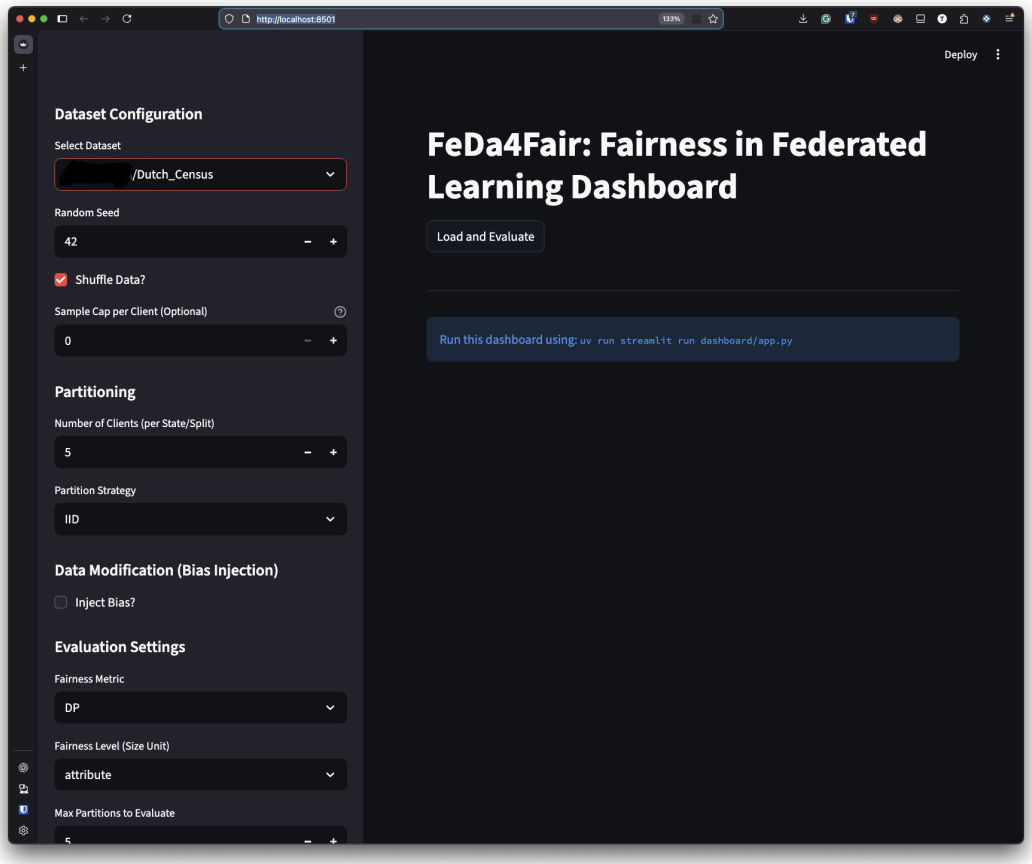


Fig. 10. Our dashboard allows practitioners to select a dataset that can be split into different clients and to define modifications to inject biases.

[6] Streamlit https://github.com/streamlit/streamlit
[7] The dashboard is available here: https://feda4fair-submission.streamlit.app/

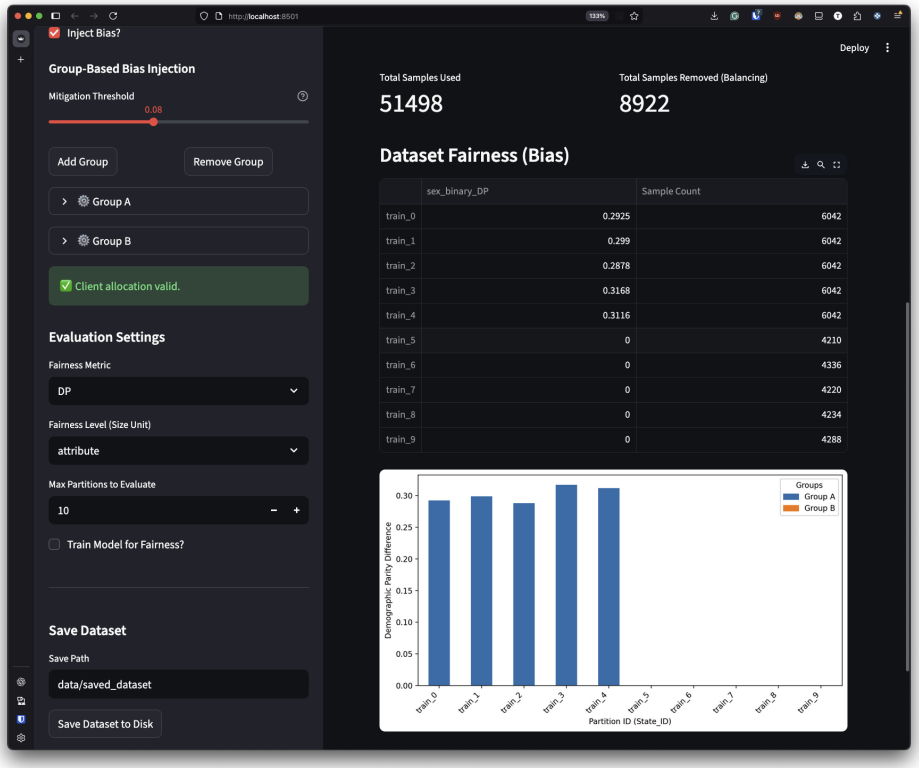


Fig. 11. Our dashboard includes the ability to plot the distribution of the fairness metrics of different dataset splits.

B Reproducibility of the Benchmark Suite

Alongside the FeDa4Fair library, we publish twelve datasets covering different bias scenarios to explore the behavior of fair FL methods. As discussed in Section 4, we exploited three different datasets in the benchmark suite. These datasets can be divided into four modalities:

(I) **Attribute-silo** datasets: Attribute-Bias dataset for the cross-silo setting;
(II) **Value-silo** datasets: Value-Bias dataset for the cross-silo setting;
(III) **Attribute-device** datasets: Attribute-Bias dataset for the cross-device setting;
(IV) **Value-device** datasets: Value-Bias dataset for the cross-device setting.

For the `ACS Income` dataset, we provide the concrete modifications applied for bias exacerbation in Table 1 for datasets (I) and (III) and in Table 2 for datasets (II) and (IV). Furthermore, for each of these four datasets, we list the counts and states sorted by where they take on the maximal bias in Table 3, 4, 5, and 6 respectively. Additionally, in Figures 12 and 13, we report the Demographic Disparity distribution for each of the four datasets on Logistic Regression models.

For the `Dutch Census` dataset, we split the clients into two groups (see Section 4.2), and we defined the bias injection settings for each of these groups. In this case, instead of setting a P_{drop} and P_{flip} for each client, we set the parameters μ and σ of the Gaussian Distribution for each group we define. Then, from this distribution, we draw the P_{drop} and P_{flip} for each client k of the group. This allows practitioners to easily set up the dataset

creation, especially when the dataset is split among hundreds of clients. All configurations utilize for P_{drop} the configuration $\mu = 0.5, \sigma = 0.05$ and for P_{flip} we set $\mu = 0.4, \sigma = 0.02$. Specifically, in the Attribute-Bias experiment, we considered the sensitive attributes `Sex` and `Mar` and fixed μ and σ for both of them. In the Value-Bias experiment, instead, we considered the `Mar` sensitive attribute with its values `Never Married`, `Married`, `Divorced`, and `Widowed`, fixing μ and σ for these groups.

For the `CelebA` dataset, we used the same approach we adopted with `Dutch Census`. We report in Table 7 the bias injection configurations used for the Attribute-Bias and Value-Bias datasets. Specifically, we set the parameters of P_{drop} to $\mu = 0.3, \sigma = 0.05$ and for P_{flip} we set $\sigma = 0.02$ while μ varies (see Table 7).

C Additional experiments

In line with Section 5, we provide additional experiments in settings we did not include in the main paper. In Section C.1 we present the results with the `ACS Income` dataset. Specifically, we present results obtained with all cross-device datasets, as well as with Logistic Regression as an additional local model. In Section C.2 we present the additional results obtained with the `Dutch Census` dataset. Lastly, in Section C.3 we present the results obtained with the `CelebA` dataset.

C.1 Experiments with `ACS Income`

C.1.1 Additional Attribute-Bias benchmark experiments. Group-level results for the attribute-device dataset are shown in Figures 15 and 16. Each point represents a dataset, plotting its Demographic Disparity (DD) value with local versus the three FL models (FedAVG, PUFFLE, Reweighing). Points below the diagonal indicate increased DD of the global models, while those above the diagonal show higher DD for the local model. For PUFFLE and Reweighing, fairness was constrained for the `Sex` attribute. Figures 15a and 16a demonstrate that FedAvg training often propagates bias, with DD increasing for both `Race` and `Sex` across most datasets. In contrast, Figures 15b and 16b reveal that PUFFLE's fairness constraints on `Sex` lead to reductions in `Sex` disparity across all clients. While some clients experience an increase in `Race` disparity with PUFFLE, it's unclear whether there exists a specific client group that benefits from improvements across both attributes. PUFFLE's performance on the `Race` attribute is comparable to FedAvg training without fairness constraints. For Reweighing, Figures 15c and 16c reveal that Reweighing successfully leads to reductions in `Sex` disparity across all clients. While for a few clients, `Race` disparity increases, the majority experience reduced `Race` disparity.

C.1.2 Additional Value-Bias benchmark experiments. For the value-device dataset, Figures 17 and 18 show results obtained when comparing FedAvg and PUFFLE models, respectively, with local XGBoost and local Logistic regression models. In particular, the plots highlight changes in Demographic Disparity (DD) and how the distribution of maximal DD values shifts. The trend indicates that underrepresented groups (here, values `Alaska Native/American Indian` and `Others`) experience a greater disadvantage from FL models. Their clusters grow, particularly for value `Others`, while other clusters, like `Black`, shrink. However, overall `Race` disparity appears to improve slightly in the visualization.

C.2 Experiments with `Dutch Census`

As explained in Section 5, besides the results obtained with the benchmarking dataset `ACS Income`, we also performed all experiments on `Dutch Census`. The settings considered for these experiments are the same; we run experiments on all four `Dutch Census` datasets.

C.2.1 Additional Attribute-Bias benchmark experiments. In Figures 23, 24, 25, and 26, we report the results for the attribute-silo and attribute-device datasets, comparing the Demographic Disparity (DD) of local models (XGBoost and Logistic Regression) against global models trained with FedAvg, PUFFLE, and Reweighing. Similar

to our previous analysis, the diagonal line acts as a parity line: points below it indicate that the global model increased unfairness compared to the local model, while points above indicate a reduction of bias. Across both cross-silo and cross-device settings, training with standard FedAvg pushes the clients across the parity line, except the blue clients when evaluating the DD on the `Mar` attribute. When applying mitigation strategies, we notice different behaviors with PUFFLE and Reweighing. PUFFLE overall reduces the DD value for most clients compared to local models when measured for the `Sex` attribute. Considering the `Mar` attribute, a similar trend can be seen for the attribute-silo datasets in Figures 23, 24. Conversely, for the cross-device datasets, Figures 25, and 26 show a different behavior. Here, the two groups of clients differ; PUFFLE improves the DD value for `Mar` biased clients, whereas the `Sex`-biased clients are below the diagonal, indicating their bias towards the `Mar` attribute was exacerbated by the global model. For Reweighing, instead, bias reduction is less pronounced, as most clients lie close to the diagonal and overall similar findings to the FedAvg baseline. These results validate our previous findings: single-attribute mitigation strategies struggle in heterogeneous federations and can inadvertently increase unfairness for clients with conflicting fairness objectives.

C.2.2 Additional Value-Bias benchmark experiments. In Figures 27 and 28 we show the results for the `Dutch Census` value-silo datasets. These figures show both how DD changes when moving from local training to FL and how the distribution of the attribute value associated with the maximum DD shifts. Firstly, we see that overall DD values are in a tighter range for the FL models. Furthermore, a shift towards `Divorced` can be noticed for the FL models, which is more pronounced for PUFFLE and Reweighing. Again, `Divorced` is the most underrepresented group in our dataset (see Figure 6a).

Figure 29, and 30 show the results for the `Dutch Census` value-device datasets. Here, we also see that the spread of DD is smaller for FL methods. However, only for FedAvg and Reweighting a shift towards `Divorced` is visible, while for PUFFLE, a clear shift towards the value `Never` can be observed.

C.3 Experiments with CelebA

Lastly, to show the compatibility of FeDa4Fair with image datasets, we also include in our benchmarking suite the four configurations obtained with `CelebA` as the underlying dataset.

Due to the higher computation needed to run the Celeba experiments, we did not perform the training of the local models for each of the clients. Therefore, the results that we report are analyses made directly on the federated dataset.

For the **attribute-silo** and **attribute-device** datasets, half of the clients have a higher bias for the `Sex` sensitive attribute than for `Hair Color`; the other half of the clients instead have the opposite setting. This is visualized in Figure 31; the clients on the left, in blue, have a higher bias for `Sex`; however, their bias toward `Hair Color` is close to 0.

In Figure 32, instead, we report the distribution of the maximum Demographic Disparity in the case of the **value-silo** and **value-device** datasets. In this case, we created a scenario in which the maximum Demographic Disparity measured on the clients is, in about half of the cases, the one corresponding to `Black Hair` and in the other half `Blond Hair`.

D Hyperparameter Tuning

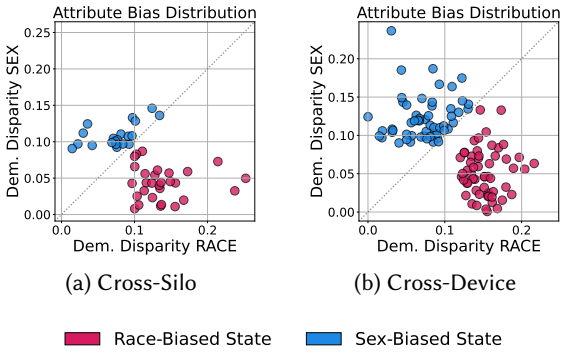
To perform hyperparameter tuning when training the FedAvg baseline, we performed a Bayesian optimization to maximize the model validation accuracy. The parameters that we optimized are: learning rate, batch size, optimizer, and number of local epochs. When training the FL model with PUFFLE[8], we followed the suggestions reported by Corbucci et al. [12]. Therefore, we performed a Bayesian optimization to minimize the model validation

[8]PUFFLE GitHub repository: https://github.com/lucacorbucci/PUFFLE

accuracy while keeping the model unfairness under $T = 0.05$. In this case, we optimized learning rate, batch size, optimizer, number of local epochs, and the value of the λ used for the unfairness mitigation. We trained the FL models for `ACS Income` and `Dutch Census` for 10 FL rounds, while we increased this value to 40 in the case of `CelebA`. This was needed because learning an image FL model is harder than learning an FL model on tabular data.

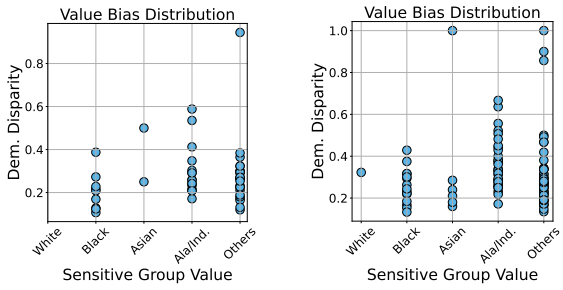
E Hardware

For most of the features offered by FeDa4Fair, a single CPU is sufficient, and no GPU is necessary. The memory requirements are in line with what is available on most commercial laptops. We believe that most resource-constrained laboratories will be able to create data with FeDa4Fair. Model fitting on the tabular datasets is similarly cheap. Moreover, to avoid the need to create datasets, we also provide 12 datasets to download. In the direction of reducing the computational needs and testing the fairness-aware FL in resource-constrained scenarios, it is possible to reduce the size of these fixed datasets by only subsampling a set of client datasets. We provide extensive statistics on each client dataset so that users may make well-founded decisions on which datasets to sample. Additionally, if restrictions on computation exist and practitioners want to create their dataset, it is not necessary to download the full `ACS Income` datasets, but only, e.g., preselected states. For all model-fitting experiments presented in this paper involving tabular datasets, we used a Mac Mini with an Apple M4 Processor and 24GB of RAM. In the experiments with images, instead, we used a NVIDIA DGX H100 with 224 CPU Intel(R) Xeon(R) Platinum 8480CL, 8 GPUs Nvidia H100 (80GB), and 2TB of RAM.



(a) Cross-Silo (b) Cross-Device

Fig. 12. Attribute-Bias measured with DD on the local Logistic Regression models for attribute benchmark datasets using `ACS Income`.



(a) Cross-Silo (b) Cross-Device

Fig. 13. Value-Bias measured with DD on the local Logistic Regression models for value benchmark datasets using `ACS Income`.

Table 1. `ACS Income` attribute-silo, `ACS Income` attribute-device dataset: Applied bias injections.

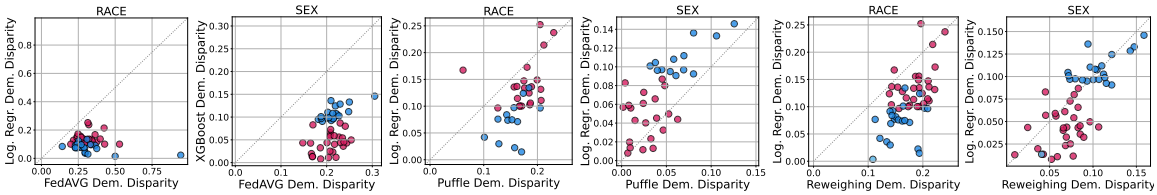
State	P_{drop}	Attribute	Value
WY	0.1	SEX	2
WI	0.1	RAC1P	2
ND	0.1	SEX	2
CA	0.1	RAC1P	2
MT	0.1	RAC1P	2
LA	0.1	SEX	2
KY	0.1	RAC1P	2
ME	0.1	RAC1P	2
AL	0.2	RAC1P	2
IN	0.2	SEX	2
MS	0.2	RAC1P	2
GA	0.2	RAC1P	2
VT	0.2	RAC1P	2
IL	0.3	SEX	2
WA	0.3	SEX	2
NH	0.3	SEX	2
PA	0.4	SEX	2
WV	0.4	SEX	2
AR	0.4	SEX	2
KS	0.4	SEX	2
OR	0.4	RAC1P	2
TX	0.4	SEX	2
DE	0.4	RAC1P	2
OK	0.4	SEX	2
ID	0.4	SEX	2
MI	0.5	SEX	2
VA	0.5	SEX	2
TN	0.5	SEX	2
OH	0.5	SEX	2
MO	0.6	SEX	2
PR	0.6	SEX	2

Table 2. ACS Income value-silo, ACS Income value-device dataset: Applied bias injections.

State	P_{drop}	Attribute	Value
AZ	0.1	RAC1P	5
OH	0.1	RAC1P	4
AR	0.2	RAC1P	4
MN	0.2	RAC1P	5
OR	0.2	RAC1P	2
WV	0.2	RAC1P	5
DE	0.3	RAC1P	4
LA	0.3	RAC1P	5
NE	0.3	RAC1P	4
AK	0.5	RAC1P	4
MS	0.5	RAC1P	4
PR	0.6	RAC1P	4

Table 3. **ACS Income attribute-silo dataset**: Counts and states for which the maximum Demographic Disparity/Equalized Odds Difference across all evaluated models is reached for the stated sensitive attribute. States where bias is distributed the same for both Demographic Disparity/Equalized Odds Difference are marked in bold.

Sensitive att.	Demographic Disparity		Equalized Odds Difference	
	Count	States	Count	States
Sex	21	AR, ID, IL, IN, KS, LA, MI, MO, ND, NH, OH, OK, PA, PR, SD, TN, TX, UT, VA, WA, WV, WY	0	
Race	29	**AK**, **AL**, **AZ**, CA, CO, CT, DE, FL, GA, **HI**, IA, KY, MA, MD, ME, **MN**, **MS**, MT, NC, **NE**, NJ, **NM**, **NV**, **NY**, OR, **RI**, **SC**, **VT**, **WI**	17	**AK**, **AL**, **AZ**, **HI**, **MN**, **MS**, ND, **NE**, **NM**, **NV**, **NY**, OK, **RI**, **SC**, **VT**, **WI**, WY

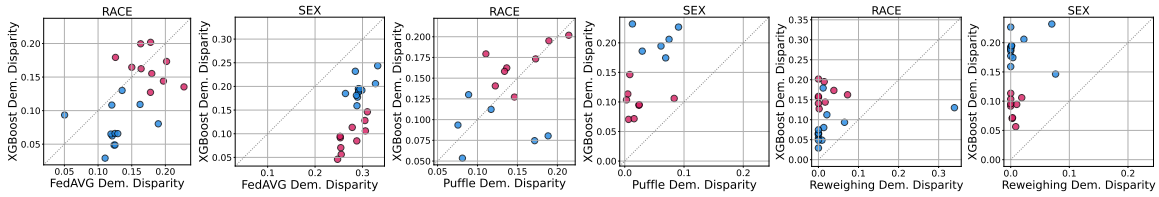


(a) **Attribute-silo:** LR vs. FedAvg (b) **Attribute-silo:** LR vs. PUFFLE. (c) **Attribute-silo:** LR vs. Reweighing.

Fig. 14. Attribute-Bias toward Race and Sex measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the attribute-silo ACS Income dataset.

Table 4. **`ACS Income` value-silo dataset**: Counts and states for which the maximum Demographic Disparity/Equalized Odds Difference across all evaluated models is reached for Race and the stated value. States where bias is distributed the same for both Demographic Disparity/Equalized Odds Difference are marked in bold.

Value (Race)	Demographic Disparity		Equalized Odds Difference	
	Count	States	Count	States
1	0		0	
2	9	FL, **ME**, MN, **MT**, ND, **NH**, OK, **SD**, **VT**	5	**ME**, **MT**, **NH**, **SD**, **VT**
3	2	PR, **WY**	2	ND, **WY**
4	16	AK, **AL**, **CO**, **CT**, **DE**, **GA**, **HI**, ID, **KY**, **MS**, **NE**, **NM**, **OH**, **OR**, **PA**, UT	17	**AL**, **CO**, **CT**, **DE**, FL, **GA**, **HI**, **KY**, **MS**, **NE**, NJ, **NM**, NY, **OH**, **OR**, **PA**, SC
5	24	**AR**, AZ, **CA**, IA, **IL**, **IN**, KS, **LA**, MA, MD, **MI**, MO, **NC**, NJ, **NV**, NY, **RI**, SC, TN, **TX**, VA, WA, **WI**, **WV**	13	**AR**, **CA**, ID, **IL**, **IN**, **LA**, **MI**, **NC**, **NV**, **RI**, **TX**, **WI**, **WV**

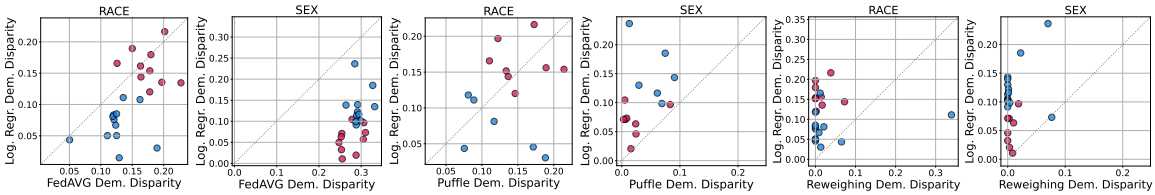


(a) **Attribute-device:** XGB vs. FedAvg (b) **Attribute-device:** XGB vs. PUFFLE(c) **Attribute-device:** XGB vs. Reweighing

Fig. 15. Attribute-Bias toward Race and Sex measured with Demographic Disparity on the local XGBoost models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the attribute-device `ACS Income` dataset.

Table 5. **`ACS Income` attribute-device dataset**: Counts and states for which the maximum Demographic Disparity/Equalized Odds Difference across all evaluated models is reached for the stated sensitive attribute. States where bias is distributed the same for both Demographic Disparity/Equalized Odds Difference are marked in bold.

Sensitive att.	Demographic Disparity		Equalized Odds Difference	
	Count	States	Count	States
Sex	55	ID_0, IL_2, IL_3, IL_4, IN_0, IN_2, IN_3, **KS_2**, KS_4, LA_4, MI_0, MI_1, MI_2, MI_3, MI_4, MO_1, MO_4, ND_1, **ND_4**, NE_5, NH_0, NH_1, **NH_4**, NH_5, OH_0, OH_1, OH_2, OH_3, OH_4, OK_0, OK_2, OK_3, PA_0, PA_1, PA_2, PA_3, PA_5, TN_2, TN_4, TN_5, TX_0, TX_1, TX_3, TX_4, TX_5, UT_0, UT_4, UT_5, VA_0, VA_1, VA_3, VA_4, WA_3, WA_5, WV_0	3	**KS_2**, **ND_4**, **NH_4**,
Race	56	**AL_2**, AL_3, AL_4, **AZ_0**, **AZ_5**, CA_1, **CO_0**, **CO_3**, **CT_2**, CT_5, FL_5, **HI_0**, **IA_0**, **IA_1**, **ID_2**, LA_2, MA_0, **MA_4**, MA_3, **ME_1**, **MN_0**, **MN_3**, **MN_4**, MN_5, **MS_4**, **MS_5**, NC_1, **NE_3**, **NE_4**, NJ_1, NJ_3, NJ_4, **NM_1**, **NM_5**, NV_0, **NV_1**, NV_3, NY_0, **NY_1**, NY_2, **NY_3**, NY_4, **NY_5**, OR_1, **OR_3**, **OR_4**, **RI_2**, **SC_0**, SC_1, SC_2, SC_4, **SD_3**, **UT_1**, WI_2, **WI_3**, **WV_5**	39	**AL_2**, **AZ_0**, **AZ_5**, **CO_0**, **CO_3**, **CT_2**, **HI_0**, **IA_0**, **IA_1**, **ID_2**, **MA_4**, **ME_1**, MI_2, **MN_0**, **MN_3**, **MN_4**, **MS_4**, **MS_5**, **NE_3**, **NE_4**, NH_0, NH_5, **NM_1**, **NM_5**, **NV_1**, **NY_1**, **NY_3**, **NY_5**, OK_3, **OR_3**, **OR_4**, **PA_3**, **RI_2**, **SC_0**, **SD_3**, TN_2, UT_5, **WI_3**, **WV_5**



(a) **Attribute-device:** LR vs. FedAvg (b) **Attribute-device:** LR vs. PUFFLE (c) **Attribute-device:** LR vs. Reweighing

Race-Biased State Sex-Biased State

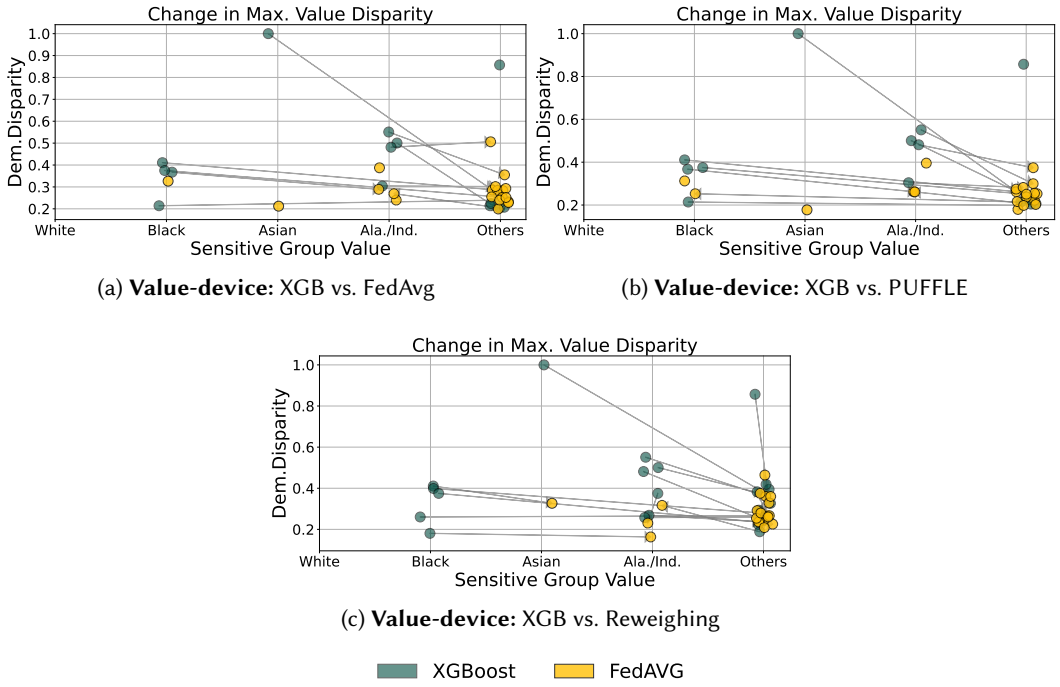
Fig. 16. Attribute-Bias toward Race and Sex measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the attribute-device `ACS Income` dataset.

Table 6. **`ACS Income` value-device dataset**: Counts and states for which the maximum Demographic Disparity/Equalized Odds Difference across all evaluated models is reached for race and the stated value. States where bias is distributed the same for both Demographic Disparity/Equalized Odds Difference are marked in bold.

Value (Race)	Demographic Disparity		Equalized Odds Difference	
	Count	States	Count	States
1	1	WV_2	0	
2	15	**ID_1**, **ME_2**, MI_1, MN_0, MN_1, **MN_2**, **MT_2**, **ND_0**, **ND_1**, **NH_2**, **SD_1**, **SD_2**, **VT_0**, WI_0, **WY_0**	17	HI_2, ID_0, **ID_1**, KY_0, ME_1, **ME_2**, **MN_2**, **MT_2**, **ND_0**, **ND_1**, NE_0, **NH_2**, PA_0, **SD_1**, **SD_2**, **VT_0**, **WY_0**
3	6	AK_0, **AR_2**, **MT_0**, MT_1, NE_1, **VT_1**	4	**AR_2**, **MT_0**, NE_2, **VT_1**
4	31	**AK_1**, AK_2, **CO_1**, **CT_1**, **CT_2**, **DE_0**, **DE_1**, **GA_1**, **GA_2**, **HI_0**, **HI_1**, **ID_2**, **IL_0**, **IL_1**, **KS_0**, **KY_2**, LA_1, MA_0, ME_1, **MI_2**, **MO_0**, **MO_2**, **MS_0**, **MS_1**, NE_0, NJ_2, NM_0, **NY_0**, **RI_0**, **SC_2**, **TN_2**	35	**AK_1**, AZ_2, **CO_1**, CO_2, **CT_1**, **CT_2**, **DE_0**, **DE_1**, GA_0, **GA_1**, **GA_2**, **HI_0**, **HI_1**, **ID_2**, **IL_0**, **IL_1**, IN_2, **KS_0**, **KY_2**, **MI_2**, **MO_0**, **MO_2**, **MS_0**, **MS_1**, NC_1, **NY_0**, NY_1, OH_0, **RI_0**, **SC_2**, TN_0, **TN_2**, VA_2, WI_0, WV_0
5	47	AL_0, AL_2, AR_0, **AR_1**, AZ_0, **AZ_1**, AZ_2, **CA_1**, **CA_2**, CO_2, **DE_2**, FL_0, GA_0, HI_2, ID_0, **IN_0**, IN_2, KS_2, KY_0, **MD_0**, **MD_1**, **MD_2**, MO_1, NC_0, NC_1, NC_2, NE_2, NJ_1, NV_0, **NV_2**, NY_1, NY_2, OH_0, PA_0, PA_2, RI_2, **SC_1**, TN_0, TX_0, **TX_1**, **TX_2**, VA_2, WA_1, WA_2, **WI_1**, **WI_2**, WV_0	17	**AR_1**, **AZ_1**, **CA_1**, **CA_2**, **DE_2**, **IN_0**, MA_0, **MD_0**, **MD_1**, **MD_2**, **NV_2**, **SC_1**, **TX_1**, **TX_2**, **WI_1**, **WI_2**, WV_2

Table 7. **`CelebA`:** Parameter settings for unfairness injection. All configurations set the group parameters of P_{drop} to $\mu = 0.3, \sigma = 0.05$ and for P_{flip} we set $\sigma = 0.02$.

Dataset Scenario	**Bias Configuration**	**Primary P_{flip} (μ)**	**Secondary P_{flip} (μ)**
Attribute-device/attribute-silo	`Sex` vs. `Hair Color`	0.28 (`Sex`)	0.60 (`Hair Color`)
Value-device/value-silo	`Black Hair` vs. `Blond Hair`	0.35 (`Black Hair`)	0.35 (`Blond Hair`)

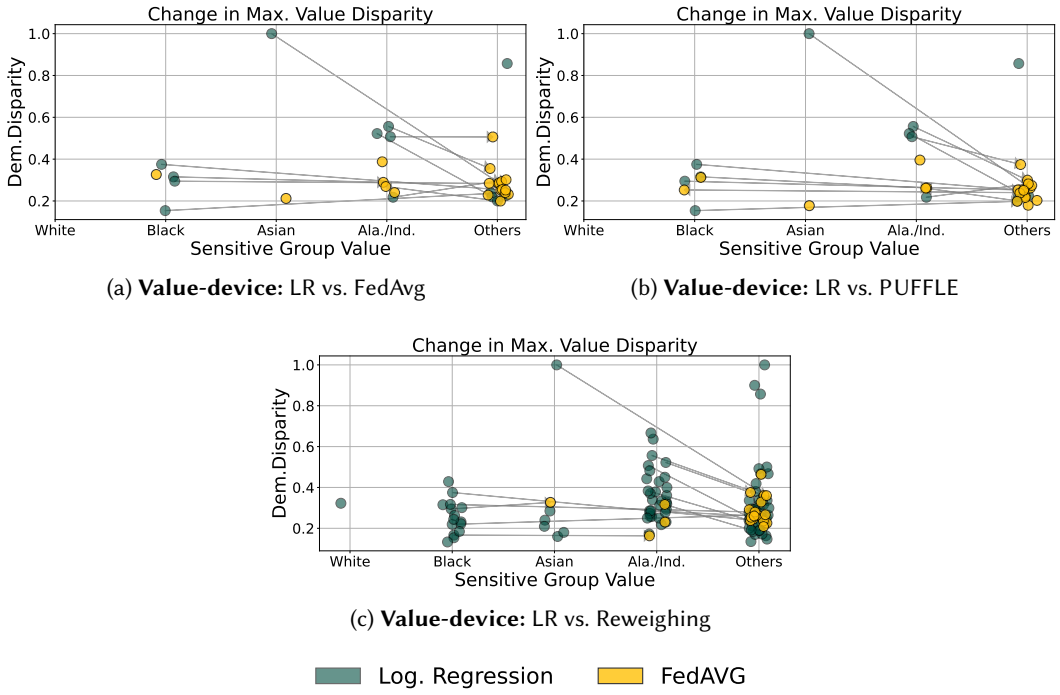


(a) **Value-device:** XGB vs. FedAvg

(b) **Value-device:** XGB vs. PUFFLE

(c) **Value-device:** XGB vs. Reweighing

Fig. 17. Value-Bias toward Race as well as value changes measured with Demographic Disparity on the local XGBoost models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-device ACS Income datasets.

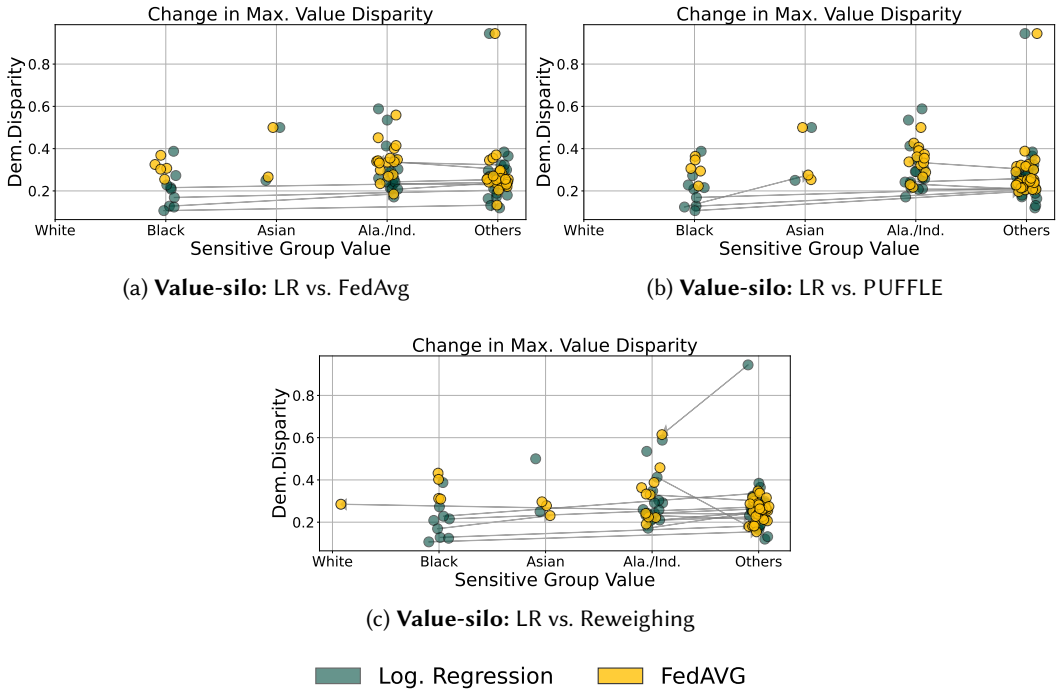


(a) **Value-device:** LR vs. FedAvg

(b) **Value-device:** LR vs. PUFFLE

(c) **Value-device:** LR vs. Reweighing

Fig. 18. Value-Bias toward Race as well as value changes measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-device ACS Income dataset.

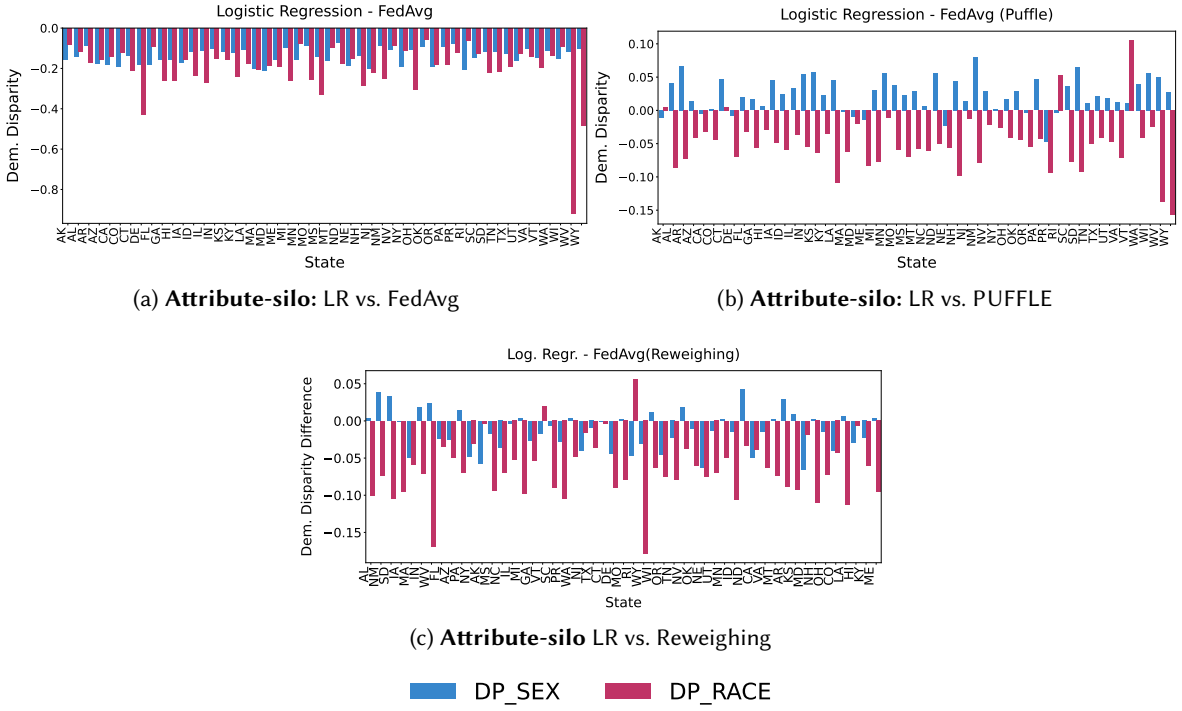


(a) **Value-silo:** LR vs. FedAvg

(b) **Value-silo:** LR vs. PUFFLE

(c) **Value-silo:** LR vs. Reweighing

Fig. 19. Value-Bias toward Race as well as value changes measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-silo ACS Income datasets.

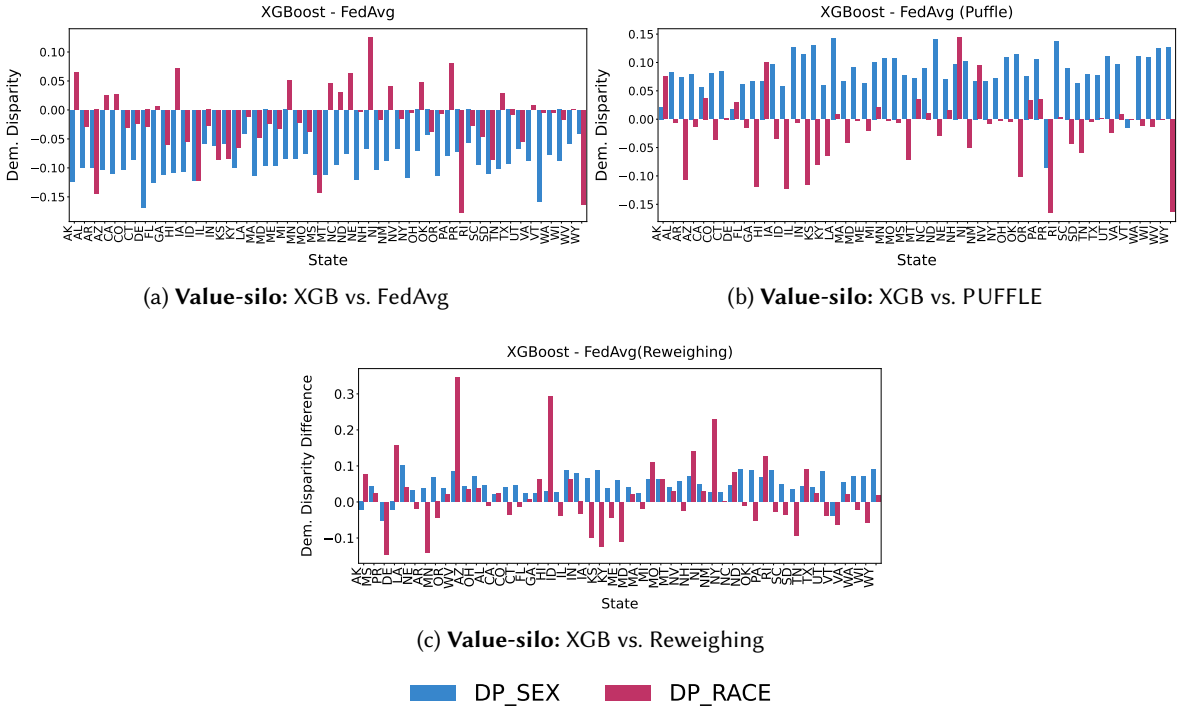


(a) **Attribute-silo:** LR vs. FedAvg

(b) **Attribute-silo:** LR vs. PUFFLE

(c) **Attribute-silo** LR vs. Reweighing

Fig. 20. ACS Income: Individual, per attribute, bias differences in Demographic Disparity between the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model.

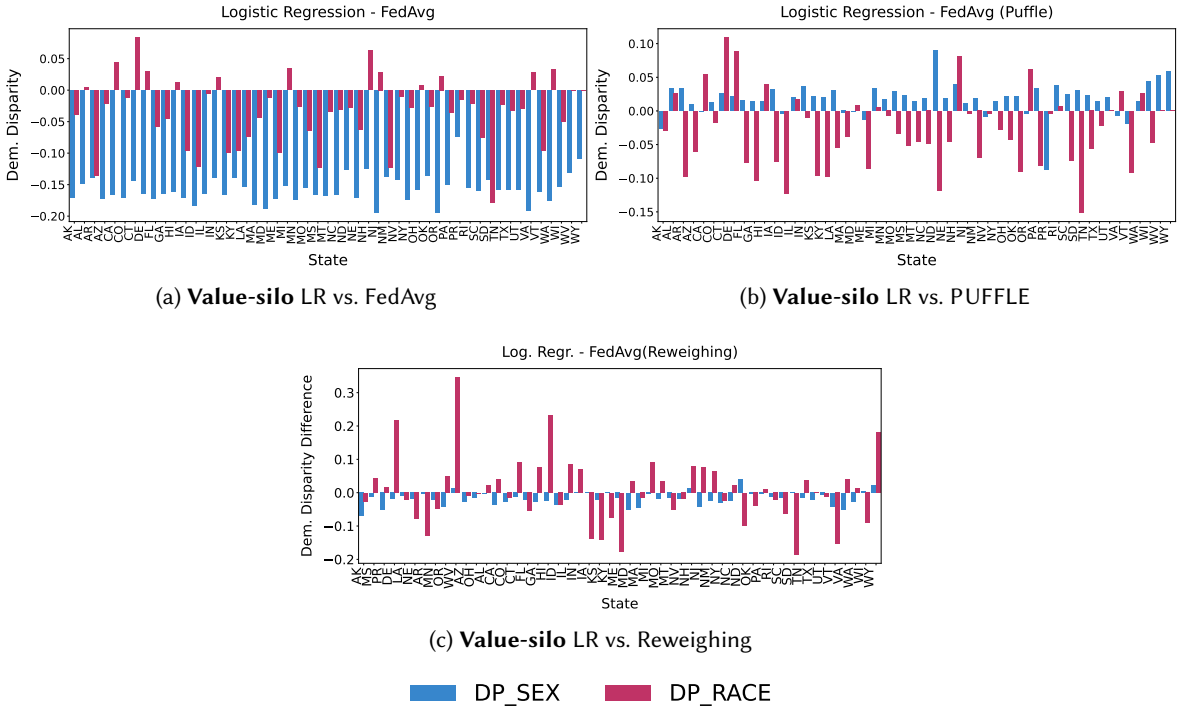


(a) **Value-silo:** XGB vs. FedAvg

(b) **Value-silo:** XGB vs. PUFFLE

(c) **Value-silo:** XGB vs. Reweighing

Fig. 21. Individual, per attribute, bias differences in Demographic Disparity between the local XGBoost models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-silo dataset.

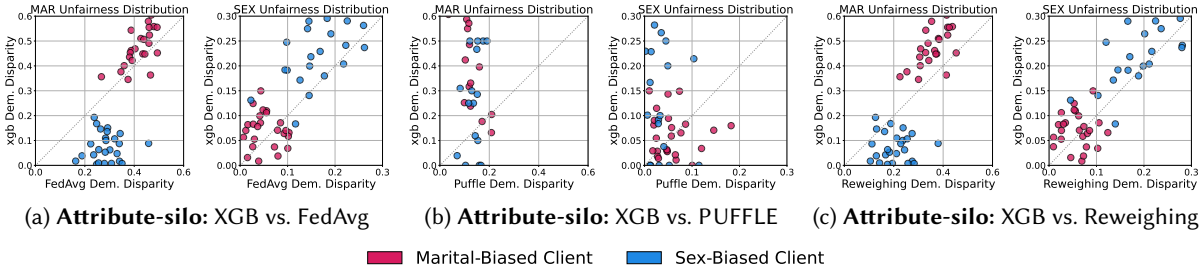


(a) **Value-silo** LR vs. FedAvg

(b) **Value-silo** LR vs. PUFFLE

(c) **Value-silo** LR vs. Reweighing

Fig. 22. Individual, per attribute, bias differences in Demographic Disparity between the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-silo dataset.

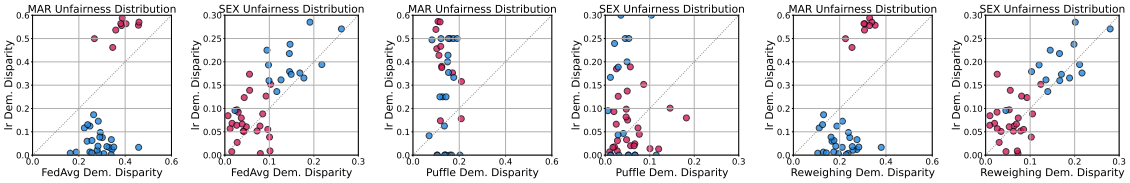


(a) **Attribute-silo:** XGB vs. FedAvg

(b) **Attribute-silo:** XGB vs. PUFFLE

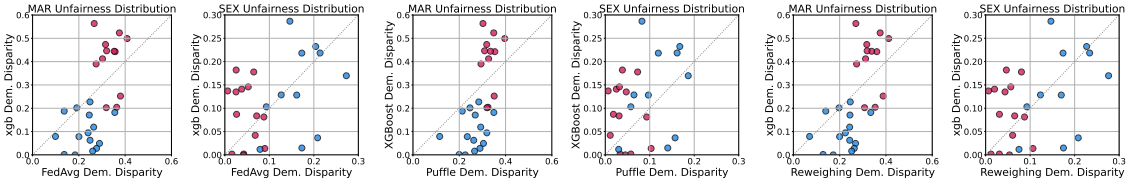
(c) **Attribute-silo:** XGB vs. Reweighing

Fig. 23. Attribute-Bias toward Mar and Sex measured with Demographic Disparity on the local XGBoost models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the attribute-silo Dutch Census dataset.



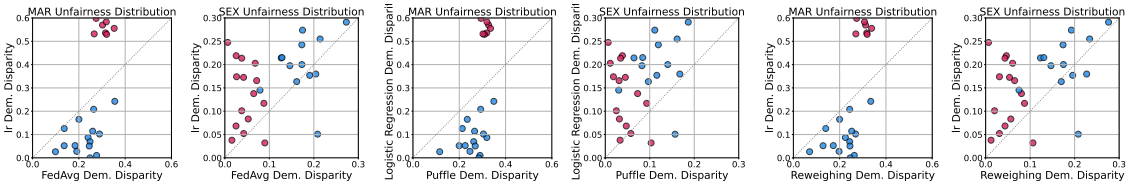
(a) **Attribute-silo:** LR vs. FedAvg (b) **Attribute-silo:** LR vs. PUFFLE (c) **Attribute-silo:** LR vs. Reweighing

Fig. 24. Attribute-Bias toward Mar and Sex measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the attribute-silo Dutch Census dataset.



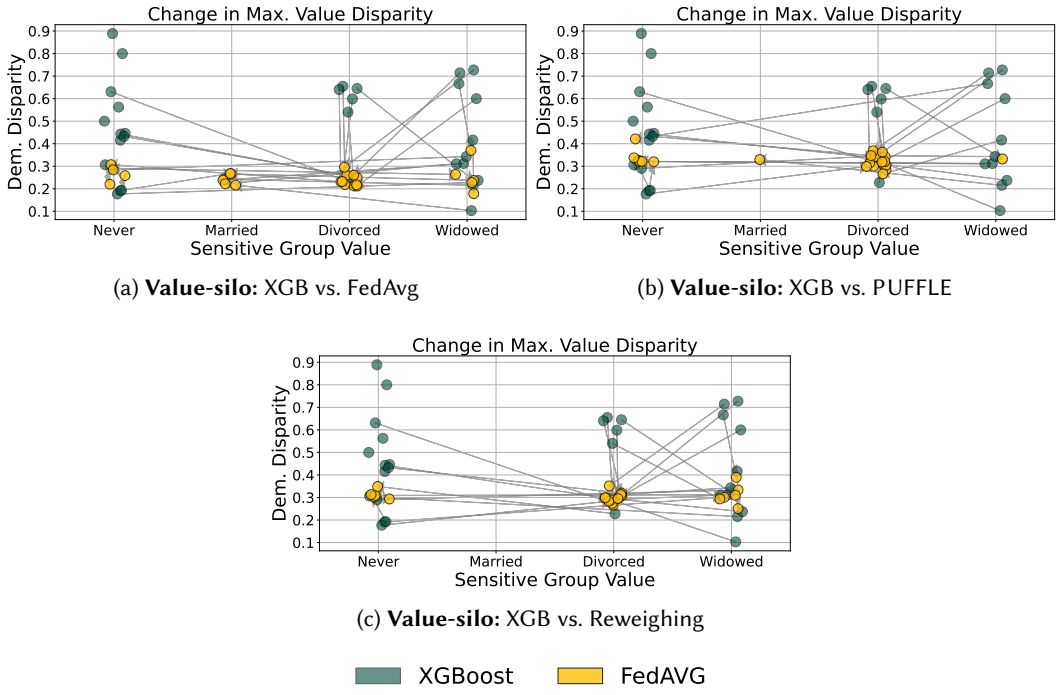
(a) **Attribute-device:** XGB vs. FedAvg (b) **Attribute-device:** XGB vs. PUFFLE (c) **Attribute-device:** XGB vs. Reweighing

Fig. 25. Attribute-Bias toward Mar and Sex measured with Demographic Disparity on the local XGBoost models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the attribute-device Dutch Census dataset.



(a) **Attribute-device:** LR vs. FedAvg (b) **Attribute-device:** LR vs. PUFFLE (c) **Attribute-device:** LR vs. Reweighing

Fig. 26. Attribute-Bias toward Mar and Sex measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the attribute-device Dutch Census dataset.

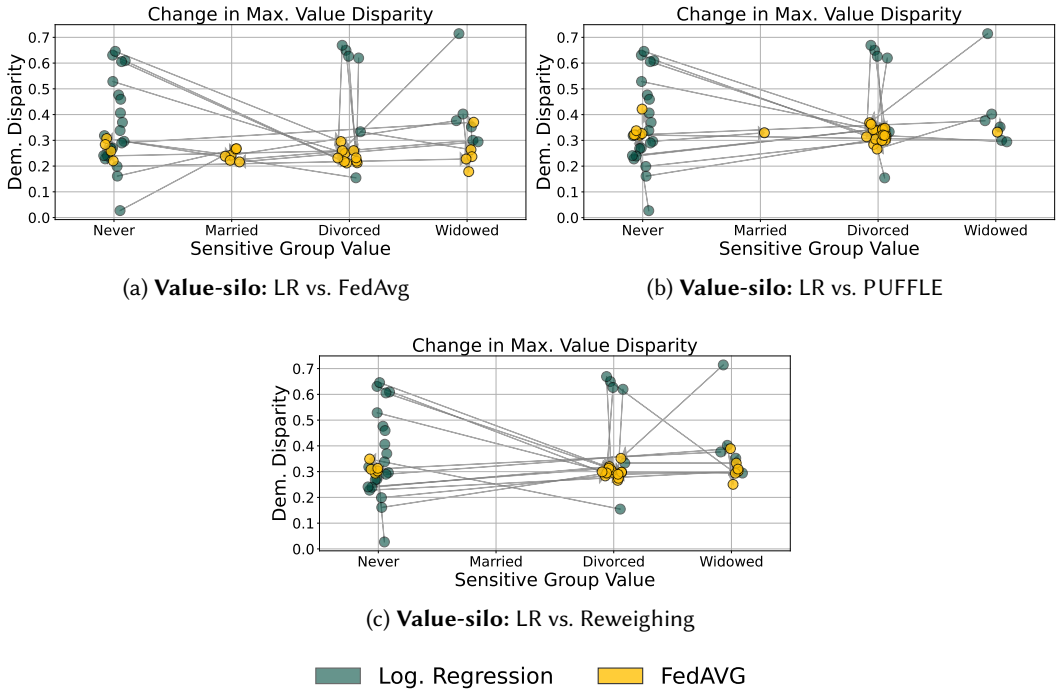


(a) **Value-silo:** XGB vs. FedAvg

(b) **Value-silo:** XGB vs. PUFFLE

(c) **Value-silo:** XGB vs. Reweighing

Fig. 27. Value-Bias toward `Mar` as well as value changes measured with Demographic Disparity on the local XGBoost models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-silo `Dutch Census` datasets.

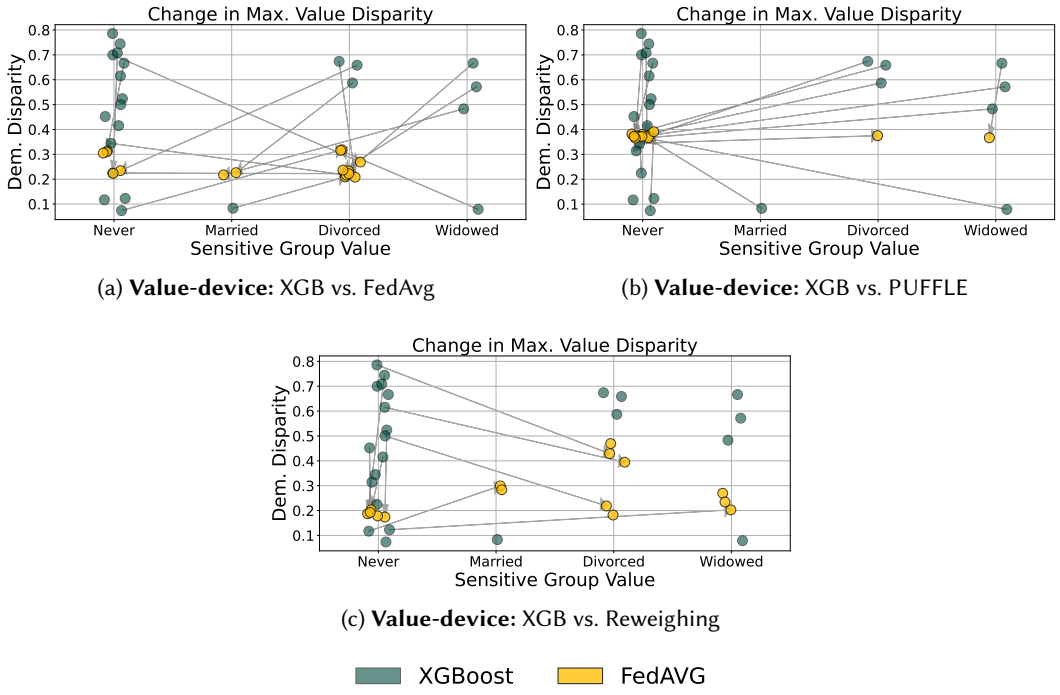


(a) **Value-silo:** LR vs. FedAvg

(b) **Value-silo:** LR vs. PUFFLE

(c) **Value-silo:** LR vs. Reweighing

Fig. 28. Value-Bias toward Mar as well as value changes measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-silo Dutch Census datasets.

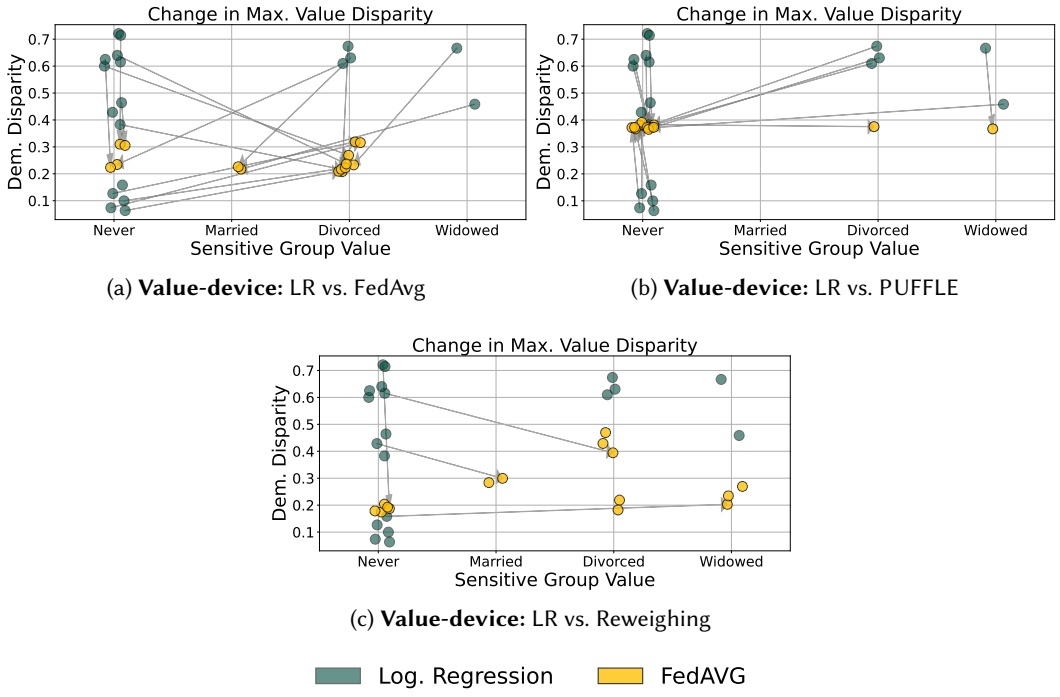


(a) **Value-device:** XGB vs. FedAvg

(b) **Value-device:** XGB vs. PUFFLE

(c) **Value-device:** XGB vs. Reweighing

Fig. 29. Value-Bias toward `Mar` as well as value changes measured with Demographic Disparity on the local XGBoost models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-device `Dutch Census` datasets.



(a) **Value-device:** LR vs. FedAvg

(b) **Value-device:** LR vs. PUFFLE

(c) **Value-device:** LR vs. Reweighing

Fig. 30. Value-Bias toward `Mar` as well as value changes measured with Demographic Disparity on the local Logistic Regression (LR) models versus the FedAvg model, the PUFFLE model, and the Reweighing model for the value-device `Dutch` Census datasets.

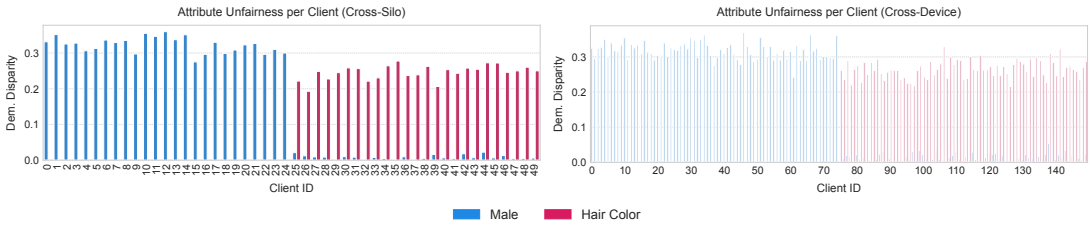


Fig. 31. Histogram showing the Attribute-Bias distribution, measured with Demographic Disparity on the `CelebA` dataset in the cross-silo and cross-device scenario. The first half of the clients are unfair toward `Sex` while the second half toward the attribute `Hair Color`.

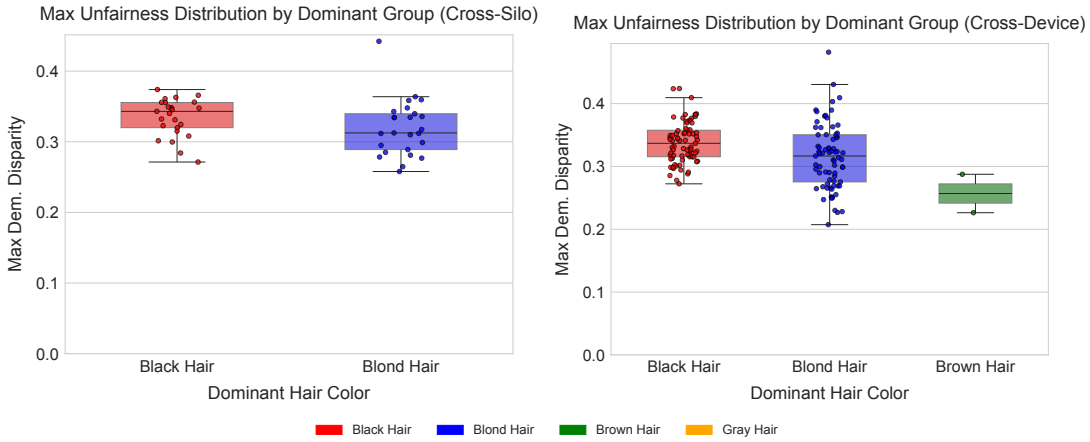


Fig. 32. Unfairness distribution of the Value-Bias experiment in both cross-silo and cross-device on the CelebA dataset.