HAL
open science

A Path Towards Fair and Equitable AI Advancing Bias Mitigation in Federated Learning

Ferraguig Lynda, A Benoit, Faiza Loukil, Mickaël Bettinelli, Christophe Lin-Kwong-Chon

To cite this version:

Ferraguig Lynda, A Benoit, Faiza Loukil, Mickaël Bettinelli, Christophe Lin-Kwong-Chon. A Path Towards Fair and Equitable AI Advancing Bias Mitigation in Federated Learning. The IEEE International Symposium on Women in Services Computing (WISC 2023), Jul 2023, Chicago, United States. , 10.13140/RG.2.2.23701.29921 . hal-04158784

HAL Id: hal-04158784

https://hal.science/hal-04158784

Submitted on 11 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Path Towards Fair and Equitable AI
Advancing Bias Mitigation in Federated Learning

Ferraguig Lynda, Benoit Alexandre, Loukil Faiza, Lin-Kwong-Chon Christophe, Bettinelli Mickaël

LISTIC Laboratory, Université Savoie Mont Blanc

Introduction

- **Machine learning** is nowadays used in practically every domain to analyze data and guide **the decision-making process**.
- With the advent of **big data**, machine learning has evolved towards **decentralized solutions** to be more efficient.
- As technology advances, several regulations have been introduced with **respect to data privacy**, such as the **GDPR**.
- To address the security and privacy issues surrounding data, Google introduced Federated Learning (FL) in 2016 [McMahan et al., 2016].
- FL is a **promising** approach for **privacy preserving ML** but also brings various challenges, notably bias and fairness in AI models.

Federated Learning An Overview

Federated Learning Definition. (FL) is a machine-learning setting where multiple entities (clients) collaborate in solving a machine-learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective. [Kairouz et al., 2021]

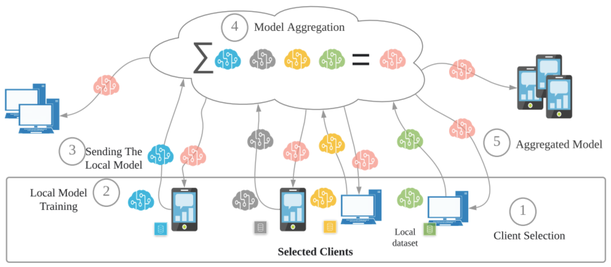

Figure 1. Federated Learning Steps [Ferraguig et al., 2021]

FL Workflow:

- (1) Client Selection by the server
- (2) Local Training at each Client's level
- (3) Sending Focused Updates by each Client
- (4) Model Aggregation by the server
- (5) Clients receiving an aggregated global model

Fairness And Bias in Machine Learning

- **Fairness, Sensitive Attributes.** The technical definitions of fairness in machine learning are based on the concept of **sensitive attribute**, also known as a **protected attribute**, which are characteristics that divide the population into groups based on certain common traits, such as **gender, religion**, etc. A machine learning model is considered unfair if it systematically exhibits biases that result in unjust treatment or disparate impact towards certain individuals or groups [Mehrabi et al., 2021].
- **Bias.** is the inclination of a model to favor or discriminate against an individual or group in a manner considered unfair [Mehrabi et al., 2021].

Sources And Impact Of Bias

Bias in machine learning can stem from various **sources** [Lambrecht, 2018] such as:

- **Historical data** reflecting past biases.
- **Unintentional or intentional human biases** in data collection and annotation
- **Measurement biases** during data collection.
- **Representation bias** caused by inadequate or underrepresented training data.

Negative Impact Of Bias Machine learning may learn how to be racist, sexist, and discriminatory [Mehrabi et al., 2021].

Figure 2. Impact of Bias In ML

Bias Mitigation Techniques in Classical ML

1. **Pre-processing.** involve modifying the training data by applying techniques, such as data augmentation, data balancing, etc., to address bias before training the model. [Calders et al., 2009]
2. **In-processing.** focus on adjusting the learning algorithm to reduce bias during the training process [Kamishima et al., 2012].
3. **Post-processing.** involve applying fairness-aware algorithms to modify the predictions generated by the trained model to ensure fairness [Louppe et al., 2017].

Fairness And Bias in Federated Learning

Why Is It Harder To Alleviate Bias In FL? *[Chang and Shokri, 2023]*

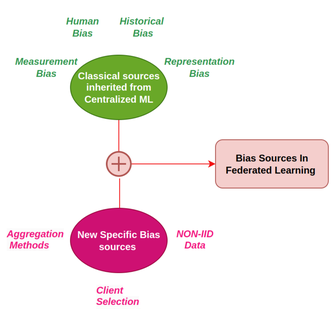

Figure 3. Source of Bias In FL

Bias Mitigation In Federated Learning

1. Existing Works

Paper	Main Findings
Mitigating Bias in Federated Learning (Abay et al., 2020)	-Explores causes of bias in Federated Learning (FL). -Test classical bias mitigation techniques in FL. -Proposes bias mitigation methods, and demonstrates their effectiveness in addressing data heterogeneity and ensuring fairness without compromising data privacy.
Towards Bias Mitigation in Federated Learning (Djebrouni et al., 2022)	-Characterize the impact of Federated Learning (FL) settings on bias. -Propose novel FL selection and aggregation methods for bias mitigation, without compromising privacy.
Bias Propagation in Federated Learning (Hongyan Chang et al.,2023)	-Bias in Federated Learning exceeds that of centralized training on the combined dataset. -Biased parties inadvertently encode their bias in a small number of model parameters, gradually increasing the model's reliance on sensitive attributes . -Auditing group fairness in Federated Learning is crucial, and robust learning algorithms are needed to mitigate bias propagation.
Mitigating Group Bias in Federated Learning (Ganghua Wang et al.,2023)	-Investigates the relationship between local and global model fairness in Federated Learning. -It recognizes the issue of group fairness and the limitations of centralized learning methods -The study proposes locally fair training to mitigate bias at the client level. - The researchers establish that global model fairness can be obtained using summary statistics from local clients.

2. Promising Approaches

Clustered Federated Learning (CFL). addresses suboptimal results in FL due to diverging data distributions by grouping clients based on similar data and leveraging geometric properties for joint training. [Sattler et al., 2020, Duan et al., 2021]

Personalized Federated Learning (PFL). the training process is customized for each client based on their preferences. This helps overcome the problem of using a single model that may not work well with different types of data. As a result, performance is improved, and the model is better aligned with each client's specific goals. [Deng et al., 2020]

Ongoing Work

1. Propose a **novel clustering method (CFL)** to **identify under and over-represented demographic** populations.
2. Propose innovative bias mitigation techniques that **leverage** Clustered Federated Learning **(CFL)** and Personalized Federated Learning **(PFL)** while considering FL privacy constraints.
3. Evaluate the proposed approaches using real-world datasets to assess their effectiveness.

References

[Calders et al., 2009] Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 Ninth IEEE International Conference on Data Mining Workshops*, pages 13–18.

[Chang and Shokri, 2023] Chang, H. and Shokri, R. (2023). Bias propagation in federated learning. In *The Eleventh International Conference on Learning Representations*.

[Deng et al., 2020] Deng, Y., Kamani, M. M., and Mahdavi, M. (2020). Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*.

[Duan et al., 2021] Duan, M., Liu, D., Ji, X., Wu, Y., Liang, L., Chen, X., and Tan, Y. (2021). Flexible clustered federated learning for client-level data distribution shift.

[Ferraguig et al., 2021] Ferraguig, L., Djebrouni, Y., Bouchenak, S., and Marangozova, V. (2021). Survey of bias mitigation in federated learning. In *Conférence francophone d'informatique en Parallélisme, Architecture et Système*.

[Kairouz et al., 2021] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*.

[Kamishima et al., 2012] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer.

[Lambrecht, 2018] Lambrecht, A. C. E. T. (2018). Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *same*.

[Louppe et al., 2017] Louppe, G., Kagan, M., and Cranmer, K. (2017). Learning to pivot with adversarial networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 982–991.

[McMahan et al., 2016] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2016). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Mehrabi et al., 2021] Mehrabi, N., Morstatter, et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

[Sattler et al., 2020] Sattler, F., Müller, K.-R., and Samek, W. (2020). Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3412.