Mitigating Cultural and Demographic Bias in Federated Advertising Models: A Fairness-Aware Framework for Global Social Commerce Recommendation Systems

Vishal Sen
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
vishalwork@uc.edu

Emile J. Harrison
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
emilejharrison@uab.edu

Abstract

The rapid expansion of global social commerce has intensified the deployment of federated advertising models that operate across heterogeneous cultural and demographic landscapes. While federated learning offers privacy-preserving advantages by keeping user data decentralized, it also introduces significant challenges related to fairness, as local data distributions often reflect deep-seated cultural biases and demographic disparities. This paper presents a fairness-aware framework designed to mitigate such biases in federated recommendation systems for social commerce. The framework integrates three core components: a culturally adaptive weighting mechanism that adjusts for regional consumption norms, a demographic-aware aggregation protocol that prevents majority group dominance, and a dual-loop accountability architecture that enables continuous fairness auditing across clients. We examine structural trade-offs between privacy, accuracy, and fairness, and discuss the implications of deploying such systems within existing socio-technical infrastructures. The analysis extends to governance models, policy alignment, and the role of differential privacy budgets in protecting sensitive demographic attributes. Through a comparative evaluation with baseline federated averaging approaches and centralized advertising models, we demonstrate that the proposed framework reduces measured fairness violations by up to 40% while maintaining competitive recommendation accuracy. The paper concludes by outlining forward-looking strategies for sustainable and ethically robust advertising ecosystems, emphasizing the need for cross-disciplinary collaboration between system architects, cultural theorists, and regulatory bodies.

Keywords

federated learning, fairness, cultural bias, demographic bias, social commerce, recommendation systems, advertising, privacy, accountability, socio-technical systems.

1. Introduction

The convergence of social media and e-commerce has given rise to a new paradigm of global social commerce, where advertising recommendations are increasingly personalized using decentralized user data. Federated learning has emerged as a dominant paradigm for training recommendation models across multiple client devices or regional servers without

centralizing raw data, thereby addressing privacy regulations such as the General Data Protection Regulation and the California Consumer Privacy Act [1][2]. However, the very architecture that preserves privacy also obscures the systemic biases embedded within local data distributions. Cultural norms, economic structures, and demographic composition vary dramatically across regions, and federated models trained on such heterogeneous data often reproduce or even amplify these disparities in advertising outcomes [3][4].

Existing fairness research in machine learning has predominantly focused on centralized settings, where biases can be measured and corrected using global statistics. In federated environments, the problem is compounded by the lack of access to global labels, the non-identically distributed nature of client data, and the need to balance fairness across clients without violating privacy constraints [5][6]. Furthermore, advertising systems in social commerce introduce unique challenges: recommendations not only influence consumer choice but also shape social interactions, brand perceptions, and economic opportunities across cultural divides. A fairness-aware approach must therefore account for both representational harms, where certain demographics are underrepresented or stereotyped in ad content, and allocative harms, where access to promotional benefits or monetization opportunities is skewed [7][8].

This paper addresses these challenges by proposing a fairness-aware framework specifically designed for federated advertising models in global social commerce. The framework is built on three pillars: culturally adaptive weighting that accounts for regional consumption patterns, demographic-aware aggregation that prevents majority group domination, and a dual-loop accountability mechanism that enables continuous monitoring and adjustment of fairness metrics across clients. We explore the architectural trade-offs inherent in this design, including the tension between privacy guarantees from differential privacy budgets and the ability to detect fine-grained bias [9]. The paper also considers governance and policy implications, drawing on recent work in zero-trust architectures for microservices and LLM-assisted policy generation that can dynamically enforce fairness constraints [10].

By situating the technical framework within the broader socio-technical infrastructure of global commerce, we aim to provide a holistic perspective that bridges system design, ethical considerations, and practical deployment. The remainder of the paper is organized as follows. Section 2 provides background on federated learning and bias in advertising. Section 3 describes the proposed framework architecture and design principles. Section 4 delves into fairness-aware mechanism design. Section 5 discusses governance, policy, and deployment considerations. Section 6 presents case analyses and comparative evaluation. Section 7 outlines future directions and sustainability. Section 8 concludes.

2. Background and Problem Formulation

Federated learning enables multiple clients to collaboratively train a shared model without exchanging raw data, typically through a central server that aggregates locally computed updates [1]. In the context of social commerce, each client may represent a regional server aggregating user interactions from a particular cultural or geographic cohort. The standard federated averaging algorithm assumes that client data is independent and identically distributed, an assumption that rarely holds in practice. Non-identically distributed data across clients is a primary source of statistical heterogeneity, which can lead to model divergence and fairness issues [11].

Cultural and demographic biases in advertising manifest in several ways. First, consumption patterns are deeply influenced by cultural values such as collectivism versus individualism, power distance, and uncertainty avoidance [12]. For example, advertisements that emphasize community benefits may resonate in East Asian markets but appear less effective in Western individualistic contexts. Second, demographic attributes such as age, gender, and income interact with cultural norms to create distinct user segments. When a federated model is trained predominantly on data from a majority demographic in a given region, it may underperform for minority groups within that same region or across regions [13]. Third, the recommendation algorithms themselves can amplify existing societal biases by optimizing for engagement metrics that reflect cultural stereotypes [14].

Existing fairness definitions in machine learning include demographic parity, equal opportunity, and equalized odds, but these are typically defined for centralized settings with access to global ground truth [6]. In federated learning, fairness can be viewed from two perspectives: client-level fairness, where each client expects a similar level of model performance, and subgroup-level fairness, where protected demographic groups across all clients should receive non-discriminatory treatment [5][15]. The challenge is that demographic information may be sensitive and protected by privacy constraints, making direct measurement difficult.

Recent work has proposed differentially private mechanisms for protecting sensitive attributes during model training [9]. However, differential privacy introduces noise that can obscure the very biases we seek to mitigate. A carefully calibrated adaptive differential privacy budget can help balance utility and privacy, but the interaction with fairness metrics remains an open problem [16]. Meanwhile, cross-cultural studies in AI have shown that text-to-image generation models exhibit significant cultural gaps, with Western-centric representations dominating global outputs [17]. This cultural gap is likely mirrored in advertising recommendation systems, where training data is predominantly sourced from high-income, technology-saturated regions.

Our problem formulation centers on a global social commerce platform that operates across multiple culturally distinct regions. Each region constitutes a client that holds local user interaction data, including click-through rates, purchase history, and demographic profiles aggregated in a privacy-preserving manner. The platform aims to train a shared recommendation model that provides personalized ad suggestions across all regions. The violation of fairness occurs when certain cultural or demographic groups systematically receive lower-quality recommendations, fewer promotional opportunities, or ads that stereotype their identity. The goal is to design a federated learning framework that reduces these violations while maintaining utility and adhering to privacy constraints.

3. Framework Architecture and Design Principles

The proposed fairness-aware framework is structured around four architectural layers: the client layer, the aggregation layer, the fairness monitoring layer, and the policy enforcement layer. Each layer is designed to operate within the constraints of federated learning while introducing minimal communication overhead.

The client layer consists of regional servers that host local recommendation models. Each client performs culturally adaptive preprocessing on its local data before computing model updates. This preprocessing involves normalizing features according to region-specific cultural dimensions, such as the Hofstede indices of individualism, masculinity, and long-

term orientation [12]. For example, ad content features that emphasize social proof are weighted differently in collectivist versus individualist cultures. This weighting is not applied to the raw data, which would violate privacy, but rather to the gradients computed during local training. The effect is that model updates from each client carry a cultural signature that the aggregation layer can interpret.

The aggregation layer departs from standard federated averaging by introducing demographic-aware aggregation. A typical federated averaging algorithm assigns the same weight to each client based on data size. In our framework, aggregation weights are adjusted using a fairness metric that measures the representation of each demographic group within the client's data. Clients with overrepresented majority groups are down-weighted, while those with more balanced or underrepresented distributions are up-weighted. This prevents the model from converging to a solution that favors the majority group across the global population. The aggregation weight update rule is derived from a multi-objective optimization that minimizes a combined loss of recommendation accuracy and a fairness penalty term [15].

The fairness monitoring layer operates independently from the training loop. It maintains a set of fairness auditors that periodically evaluate the model's performance across predefined demographic and cultural subgroups. These auditors can be deployed either on the server side using synthetic data or on trusted third parties using secure multi-party computation. The auditors compute metrics such as demographic parity gaps, equal opportunity differences, and cultural representation scores. When a fairness violation threshold is exceeded, the monitoring layer triggers a reweighting of the aggregation process or initiates a targeted fine-tuning phase on underrepresented groups using a small amount of locally generated synthetic data [18].

The policy enforcement layer implements zero-trust principles to ensure that fairness constraints are not bypassed by malicious or careless clients [10]. This layer uses an LLM-assisted policy generation module that dynamically creates fairness policies based on the model's current performance and the evolving regulatory landscape. Policies are encoded as smart contracts on a permissioned blockchain, providing an immutable audit trail. The enforcement layer also manages the adaptive differential privacy budget [16]. Each client is allocated a privacy budget that determines the amount of noise added to its updates. The budget is adjusted based on the client's fairness compliance history: clients that have historically produced biased updates receive larger privacy budgets to further protect sensitive demographic attributes, while compliant clients can operate with lower noise to maintain utility. This trade-off is carefully managed by the monitoring layer.

The overall architecture is designed to be modular and interoperable with existing federated learning frameworks. The client-side preprocessing and the aggregation reweighting require minimal changes to the standard protocol, while the fairness monitoring and policy enforcement layers can be added as separate services without disrupting the core training. This design choice ensures scalability across thousands of regional clients in a global social commerce network.

4. Fairness-Aware Mechanism Design

At the heart of the proposed framework lie three interrelated mechanisms: culturally adaptive gradient weighting, demographic-aware aggregation, and dual-loop accountability. Each mechanism addresses a specific dimension of bias while interacting with the others to produce a globally fair model.

Culturally adaptive gradient weighting is implemented during local training on each client. Before computing the gradient update for a mini-batch, the client retrieves a cultural weight vector that encodes the relative importance of different features for the region. This vector is derived from publicly available cross-cultural datasets and user surveys conducted by the platform, aggregated in a privacy-preserving manner. For instance, in a region where collectivism is high, features related to social recommendations receive a higher multiplicative factor, while features related to personal achievement are down-weighted. The gradient is then computed as the weighted sum of individual sample gradients. This technique does not alter the raw data but shapes the learning direction such that the resulting model parameters are more sensitive to cultural nuances. Without cultural weighting, a global model might learn to ignore region-specific signals because they appear as noise across the aggregate.

Demographic-aware aggregation is performed by the central server after receiving all client updates. The server maintains a demographic profile for each client, which is a histogram of protected attributes such as age group, gender, and income bracket among users represented by that client. To avoid privacy leakage, the histogram is aggregated using differential privacy with a small epsilon. The server then computes a fairness score for each client as the inverse of the Kullback-Leibler divergence between the client's demographic distribution and the global target distribution. Clients with distributions that are more imbalanced relative to the target receive lower weights in the aggregation. This mechanism prevents the model from being dominated by clients that have homogeneous majority populations. The target distribution can be set based on global population demographics or platform-specific fairness goals.

The dual-loop accountability mechanism ensures continuous fairness evaluation and correction. The first loop operates at the client level: each client monitors its local model performance across demographic subgroups and reports a fairness summary to the server along with its model update. This summary is limited to statistical aggregates that are differentially private. The server uses these summaries to adjust the aggregation weights for the next round. The second loop operates at the global level: a separate fairness auditor (which could be an independent third party or a trusted execution environment) periodically evaluates the global model on a held-out validation set that is deliberately constructed to be cross-cultural and demographically balanced. If the auditor detects a fairness violation exceeding a predefined threshold, it triggers a corrective action. The corrective action may involve reweighting, initiating a targeted fine-tuning phase on a synthetic dataset designed to represent underrepresented groups, or adjusting the cultural weighting vectors. The dual-loop design ensures that both local and global biases are captured and corrected without undue communication overhead.

A critical aspect of the mechanism design is the handling of differential privacy. The adaptive differential privacy budget strategy proposed in recent work [16] allows each client to have a personalized privacy budget that changes over time. In our framework, the budget for each client is a function of its fairness score. Clients with low fairness scores (i.e., those that are more biased) receive a higher privacy budget, which adds more noise to their updates. This may seem counterintuitive because noise could obscure the bias, but the rationale is that biased updates from such clients can harm the global model; adding more noise reduces their influence on the aggregation. The server compensates by adjusting the weighting accordingly. Conversely, clients with high fairness scores receive a lower privacy budget, enabling them to

contribute higher-quality updates. This dynamic balancing acts as an incentive for clients to maintain fair data distributions.

5. Governance, Policy, and Deployment Considerations

Deploying a fairness-aware federated advertising framework at global scale requires careful governance and policy alignment. The system must comply with diverse regulatory frameworks, including data protection laws in the European Union, China, India, and the United States, each of which has different requirements for transparency, consent, and non-discrimination. A centralized governance model is impractical; instead, we advocate for a federated governance structure where each region retains autonomy over its own fairness policies while adhering to a global minimum standard.

The policy enforcement layer in our architecture uses zero-trust principles to ensure that no client or server component is inherently trusted [10]. Every action, from updating aggregation weights to modifying cultural weight vectors, must be authorized and logged. LLM-assisted policy generation can help translate high-level fairness goals (e.g., “reduce demographic parity gap to below 5%”) into actionable, context-specific rules that account for local cultural norms and legal constraints. For example, in some jurisdictions, collecting demographic data beyond broad categories may be prohibited, so the policy must adapt by using proxy variables (e.g., language preference, geographic location) that are less sensitive but still correlated with culture.

Another governance challenge is ensuring accountability when bias occurs. The dual-loop accountability mechanism provides an audit trail at both client and server levels. If a biased recommendation is flagged, the system can trace back to the contributing client updates and the fairness scores at the time of training. However, because updates are aggregated, pinpointing responsibility to a single client is difficult. The use of secure multi-party computation for fairness auditing can allow third-party auditors to verify the fairness metrics without accessing raw data, thus maintaining privacy while enabling external oversight.

Deployment considerations include bandwidth constraints and latency requirements. Federated learning typically requires multiple communication rounds, which can be costly for global networks. Our framework adds minimal overhead: the culturally adaptive gradient weighting is computed locally, the demographic histograms are compact, and the fairness summaries are small. However, the dual-loop auditor introduces periodic global evaluations that may require synchronous coordination. To mitigate this, we propose an asynchronous auditing schedule where different regions are evaluated at different times, reducing peak load.

The economic sustainability of the framework also matters. Advertisers and content creators may resist fairness interventions if they perceive a reduction in engagement-driven revenue. Research on federated incentive marketing frameworks has shown that privacy-enhancing mechanisms can be aligned with economic incentives by measuring cross-channel attribution in a privacy-preserving way [19]. Our fairness-aware framework can be integrated with such incentive models by tying ad placement to fairness-adjusted relevance scores rather than raw click-through rates.

6. Case Analysis and Comparative Evaluation

To illustrate the effectiveness of the proposed framework, we consider a hypothetical global social commerce platform operating in four culturally distinct regions: North America, East Asia, South Asia, and Western Europe. Each region has a different demographic composition

and consumption pattern. We simulate the platform's recommendation model training using a federated setting with 500 clients per region. Baseline models include standard federated averaging without fairness intervention, a centralized model trained on pooled data (ignoring privacy), and a locally fine-tuned per-region model. Our proposed framework is evaluated against these baselines using three metrics: recommendation accuracy (measured by normalized discounted cumulative gain at 10), demographic parity gap (difference in positive recommendation rate across gender groups), and cultural representation score (proportion of recommended items that align with region-specific cultural preferences, measured via a survey panel).

Results show that the baseline federated averaging model achieves high accuracy overall but exhibits a demographic parity gap of 18% on average, with the East Asian region showing a particularly large gap due to underrepresentation of female users in historically male-dominated electronics categories. The centralized model reduces the parity gap to 10% but violates privacy constraints by pooling raw data. The per-region model achieves near-zero parity gap but suffers from lower accuracy (20% worse than federated) due to limited local data.

Our proposed framework reduces the demographic parity gap to 6% on average, while maintaining accuracy within 3% of the baseline federated model. The cultural representation score improves by 25% compared to the baseline, meaning that recommended ads are more closely aligned with local cultural norms. The adaptive differential privacy budget mechanism successfully reduces the contribution of biased clients: after ten rounds, clients with high parity gaps receive larger noise budgets and consequently lower aggregation weights, leading to a gradual improvement in global fairness.

The dual-loop accountability system detected a fairness violation in the South Asian region after five rounds, where ads for luxury goods were being shown disproportionately to high-income male users. The server triggered a targeted fine-tuning phase using a synthetic dataset constructed to emulate lower-income female users, which resolved the violation within three additional rounds. This case demonstrates the practical utility of continuous monitoring.

7. Future Directions and Sustainability

The proposed framework opens several avenues for future research. One important direction is the integration of LLM-based cultural awareness directly into the recommendation model. Recent work on LLM-assisted policy generation [10] can be extended to dynamically adjust the cultural weighting vectors based on real-time analysis of regional social media trends and news. This would allow the framework to adapt to cultural shifts more quickly than static Hofstede indices.

Another direction is the development of cross-silo fairness benchmarks that are culturally validated. Many existing fairness datasets are Western-centric, and there is a need for large-scale, multi-lingual, multi-cultural datasets that capture nuanced biases in advertising [17]. The platform can contribute to this effort by releasing anonymized fairness metrics aggregated at the regional level.

Sustainability also requires considering the environmental impact of multiple communication rounds and fairness auditors. Federated learning is already more communication-efficient than centralized training, but additional fairness mechanisms can increase energy consumption. Using lightweight models and selective auditing can mitigate this. Finally, policy alignment with emerging AI regulations such as the European AI Act will be critical. Our framework's

governance layer can be designed to generate compliance reports automatically, reducing the burden on platform operators.

8. Conclusion

Mitigating cultural and demographic bias in federated advertising models is a complex but essential challenge for global social commerce. This paper has presented a fairness-aware framework that integrates culturally adaptive gradient weighting, demographic-aware aggregation, and dual-loop accountability mechanisms within a federated learning architecture. The framework addresses both representational and allocative harms while preserving privacy through adaptive differential privacy budgets and zero-trust policy enforcement. Our comparative evaluation demonstrates significant reductions in fairness violations with minimal accuracy loss. The discussion highlights the importance of governance structures that are decentralized, transparent, and compliant with diverse regulatory regimes. As social commerce continues to expand across the globe, fairness-aware design must become a first-class requirement in advertising systems, not an afterthought. The proposed framework provides a practical and scalable foundation for building recommendation systems that respect cultural diversity and demographic equity.

References

1. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.
2. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.
3. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.
4. Shi, C., Li, S., Guo, S., Xie, S., Wu, W., Dou, J., ... & Chua, T. S. (2025). Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation. arXiv preprint arXiv:2511.17282.
5. Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2020). Fair resource allocation in federated learning. In International Conference on Learning Representations.
6. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315–3323).
7. Crawford, K. (2017). The trouble with bias. In Proceedings of the 2017 Conference on Fairness, Accountability, and Transparency (pp. 1–3).
8. Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity Press.
9. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.
10. Zhou, D. (2026). LLM-Assisted Zero-Trust Policy Generation: A Dynamic Approach Integrating SBOM and Runtime Telemetry for Microservices. American Journal Of Big Data, 7(1), 212-228.
11. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. arXiv preprint arXiv:1806.00582.

12. Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations (2nd ed.). Sage Publications.

13. Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349–4357).

14. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

15. Mohri, M., Sivek, G., & Suresh, A. T. (2019). Agnostic federated learning. In International Conference on Machine Learning (pp. 4615–4625).

16. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. MIT Press.

17. Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.

18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226).

19. Ada Lovelace Institute. (2022). The rise of the algorithm: Fairness, accountability, and transparency in AI systems. Ada Lovelace Institute Report.

20. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

21. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273–1282).