Integrating explainable AI for trustworthy machine learning-based Ad personalization in social media platforms

Vyacheslav Pak[1,†], Riza Akhitova[1,†], Ha Jin Hwang[1†], Aidiye Aidarbekov[1,2,†] and Darkhan Bekuzak[3,†]

[1] *Astana IT University, Mangilik El Avenue 55/11, 010000, Astana, Kazakhstan*
[2] *S. Seifullin Kazakh Agrotechnical Research University, Jeńis dańǵyly 62, 010011, Astana, Kazakhstan*
[3] *Kangwon National University, 1 Gangwondaehak-gil, 24341, Chuncheon-si, Gangwon-do, Republic of Korea*

Abstract

The opacity of black-box machine learning models in social media advertising undermines user trust and regulatory compliance. This paper presents a novel framework that integrates explainable AI techniques (SHAP, LIME, and counterfactual reasoning) with adaptive differential privacy and federated learning to enable transparent, privacy-preserving ad personalization. Across four diverse real-world datasets, the framework achieves a competitive AUC of 0.85, improves demographic parity by over 100%, and retains 93% utility at ε=1.0. User trust scores increase by 43% compared to baseline systems. These results demonstrate that responsible AI practices can be deployed at scale without sacrificing performance, addressing GDPR "right to explanation" requirements and advancing UN SDGs on fair employment, innovation, and ethical consumption.

Keywords

Explainable AI, trustworthy machine learning, advertising personalization, social media platforms, algorithmic fairness, differential privacy, SHAP, LIME

1. Introduction

The rapid advancement of machine learning algorithms has revolutionized digital advertising, enabling unprecedented levels of personalization in social media platforms [1]. Modern advertising systems leverage sophisticated neural networks, deep learning models, and collaborative filtering techniques to analyze user behavior, preferences, and contextual information, delivering highly targeted advertisements that maximize engagement and conversion rates [3, 4]. However, this technological progress has introduced significant challenges related to algorithmic transparency, user privacy, and ethical decision-making that demand immediate attention from both researchers and practitioners [5, 6].

The opacity of machine learning models in advertising personalization creates a fundamental tension between system performance and user trust [7]. Black-box algorithms, while achieving superior predictive accuracy, fail to provide meaningful explanations for their decision-making processes, leading to decreased user acceptance and potential regulatory violations [8]. The European Union's General Data Protection Regulation (GDPR) and similar privacy legislation worldwide mandate the "right to explanation" for automated decision-making systems, particularly those affecting individual rights and interests [27]. Consequently, there is an urgent need for explainable artificial intelligence (XAI) frameworks that can maintain high performance while providing transparent, interpretable explanations for advertising personalization decisions.

[1] *AI 2025: Workshop on Artificial Intelligence, November 19-20, 2025, Almaty, Kazakhstan*
* Corresponding author.
† These authors contributed equally.
tditp.q@gmail.com (V. Pak); riza.akhitova@astanait.edu.kz (R. Akhitova); hjhwang@astanait.edu.kz (Ha Jin Hwang); aidiye.aidarbekov@astanait.edu.kz (A. Aidarbekov); darkhanbekuzak13@gmail.com (D. Bekuzak)
0009-0003-4169-1506 (V. Pak); 0000-0003-0197-7860 (R. Akhitova); 0000-0001-8752-1906 (Ha Jin Hwang); 0000-0001-7197-1638 (A. Aidarbekov); 0009-0008-4748-6278 (D. Bekuzak)

Furthermore, the extensive use of personal data in advertising systems raises critical concerns about algorithmic fairness and bias amplification. Machine learning models trained on historical advertising data often perpetuate and amplify existing societal biases, leading to discriminatory outcomes against protected demographic groups. Recent studies have documented instances of advertising algorithms exhibiting gender bias in job advertisements, racial discrimination in housing promotions, and age-based exclusions in financial services [14]. These findings underscore the necessity for comprehensive bias detection and mitigation strategies integrated into advertising personalization systems.

Privacy preservation represents another fundamental challenge in social media advertising [16]. Traditional centralized machine learning approaches require extensive data collection and sharing, potentially exposing sensitive user information to privacy breaches and unauthorized access [18]. The implementation of privacy-preserving techniques, such as differential privacy and federated learning, becomes essential for maintaining user trust while enabling effective personalization [19]. However, the integration of privacy mechanisms often introduces trade-offs between data protection and model utility, requiring careful optimization to achieve acceptable performance levels.

The development of trustworthy AI systems for advertising personalization aligns with several United Nations Sustainable Development Goals (SDGs). Our framework contributes to SDG 8 (Decent Work and Economic Growth) by promoting fair and unbiased job advertising systems that eliminate discriminatory practices, ensuring equal employment opportunities across demographic groups. The integration supports SDG 9 (Industry, Innovation, and Infrastructure) through the development of resilient, inclusive digital advertising infrastructure that leverages cutting-edge explainable AI technologies. Additionally, our approach advances SDG 12 (Responsible Consumption and Production) by enabling transparent, ethical advertising practices that promote informed consumer decision-making while protecting user privacy and data rights

To address these challenges, we propose a comprehensive framework that integrates explainable artificial intelligence with privacy preserving machine learning for trustworthy ad personalization in social media platforms. Our approach combines multiple XAI techniques, including SHAP (SHapley Additive exPlanations) for global feature importance analysis, LIME (Local Interpretable Model agnostic Explanations) for local decision explanations, and counterfactual reasoning for bias detection [28]. We incorporate differential privacy mechanisms to ensure user data protection and implement federated learning architectures to enable distributed model training without centralized data collection [12].

The main contributions of this paper are threefold. First, we develop a novel XAI framework specifically designed for advertising personalization systems that provides multiple levels of explanation granularity, from global model behavior to individual prediction rationales [26]. Second, we integrate privacy-preserving techniques with explainable methods, demonstrating that trustworthy machine learning can be achieved without significant performance degradation [23]. Third, we conduct comprehensive evaluations across four diverse real-world datasets, providing empirical evidence of our framework's effectiveness in improving fairness, privacy, and user trust in advertising personalization systems.

This study addresses the following research questions: [RQ1] How can explainable AI improve transparency in ad personalization systems? [RQ2] How can privacy-preserving mechanisms be integrated without reducing model utility? [RQ3] To what extent do these methods enhance user trust and fairness?

The goal of this study is to design and evaluate an integrated framework that combines explainability, privacy preservation, and fairness to achieve trustworthy ad personalization.

2. Literature review

This section reviews prior research on explainable artificial intelligence (XAI), privacy-preserving machine learning, algorithmic fairness, and user trust in AI systems within advertising and social media contexts. The discussion follows a logical structure that moves from the general foundations of

explainability and trust in AI to the specific challenges of integrating privacy and fairness in advertising personalization systems. The review concludes by identifying existing theoretical and methodological gaps and formulating hypotheses that guide the present study.

2.1. Explainable AI and trust in AI systems

Explainable artificial intelligence (XAI) has emerged as a crucial field in response to the widespread deployment of black-box machine learning models in decision-making domains. Its primary goal is to provide interpretable and transparent explanations of algorithmic decisions, improving accountability, human understanding, and user trust. Shin [15] empirically demonstrated that interpretable explanations enhance user confidence and acceptance of AI systems, showing that explainability and causability directly influence perceived fairness and trustworthiness. Similar findings were reported by Cetinkaya and Krämer [24], who identified transparency and explainability as key factors shaping user perceptions of AI systems.

Cheung and Ho [8] further established that effective explainability improves human factors within trust models, emphasizing that interpretability is not only a technical attribute but also a psychological mechanism that fosters confidence and system adoption. These insights form the theoretical foundation for this study, highlighting that user trust in advertising personalization depends on the extent to which system operations are transparent and understandable.

However, most existing research focuses on explainability in general AI systems rather than its specific implications for marketing or advertising personalization. As such, the present study extends these findings into the domain of digital advertising, examining how explainability can enhance trust in personalized ad recommendations on social media platforms.

2.2. Explainable AI in advertising and marketing

In advertising and marketing, explainable AI has been introduced to address the opacity of predictive models used in personalization and targeting. Yang et al. [1] proposed a framework that integrates large language models with XAI for digital advertising analysis, demonstrating that attention mechanisms can generate human-understandable insights into advertisement performance. Deck et al. [2] provided a critical survey on the fairness benefits of explainable AI, arguing that methods such as SHAP and LIME can reveal and mitigate discriminatory patterns in advertising algorithms. Similarly, Coussement et al. [3] showed that interpretable models outperform black-box approaches in marketing analytics by improving user acceptance and decision quality.

These studies collectively underscore the potential of explainable AI to improve both system performance and stakeholder understanding in marketing contexts. Yet, most prior works treat explainability as an isolated enhancement rather than a component of a broader framework that also ensures privacy and fairness. Consequently, while interpretability has improved, the relationship between XAI, privacy protection, and trust in advertising systems remains underexplored. This research addresses that gap by combining explainability techniques with privacy-preserving and fairness-enhancing mechanisms to achieve holistic trustworthiness in ad personalization.

2.3. Privacy-preserving machine learning in advertising

The increasing reliance on user data in advertising necessitates robust privacy-preserving techniques. Differential privacy, introduced as a mathematical framework to limit information leakage, has become a standard approach for safeguarding personal data in machine learning systems [19]. Several works have examined privacy-preserving model explanations, outlining the trade-offs between explanation fidelity and data protection [18]. Federated learning has further advanced privacy-preserving AI by enabling distributed model training without centralized data aggregation [12].

Studies such as Wang et al. [19] and Mestari et al. [18] provided comprehensive reviews of privacy-preserving algorithms, emphasizing the importance of balancing privacy guarantees with model utility. Recent approaches integrate differential privacy with federated learning to achieve scalable, privacy-compliant advertising personalization [23]. These methods demonstrate that distributed architectures can maintain accuracy while complying with regulatory requirements like GDPR.

Despite this progress, prior research often isolates privacy from interpretability. The intersection of privacy and explainability-particularly how to generate meaningful explanations without compromising data protection-remains insufficiently explored. The present study advances this discussion by introducing privacy-aware explainability mechanisms that maintain interpretability under strict privacy constraints.

2.4. Algorithmic fairness and bias detection

Algorithmic fairness has become a central issue in machine learning, particularly in advertising systems where biased targeting can reinforce societal inequalities. Mehrabi et al. [16] provided a comprehensive taxonomy of bias types and mitigation strategies, forming the theoretical basis for fairness-aware AI development. Akter et al. [14, 22] and Brandl et al. [6] emphasized that fairness and explainability are deeply interconnected: interpretable models help reveal discriminatory decision boundaries and facilitate corrective interventions.

Counterfactual explanations have recently emerged as powerful tools for identifying and mitigating algorithmic bias. Studies such as Van Haastrecht et al. [23] demonstrated that counterfactual reasoning can uncover implicit biases by analyzing minimal input changes that alter prediction outcomes. Similarly, research by Deck et al. [2] and Zhang and Wilson [26] highlighted the role of XAI in evaluating fairness and promoting transparency throughout the AI lifecycle.

However, existing literature seldom unifies these elements-explainability, fairness, and privacy-into a single operational framework. Most prior studies address fairness evaluation or bias mitigation separately from privacy and interpretability concerns. This study builds upon these foundations by proposing an integrated approach that simultaneously enhances algorithmic transparency, bias mitigation, and data protection within advertising personalization systems.

2.5. Identified research gaps and hypotheses

The literature reveals several critical gaps:

1. Research on XAI in advertising primarily enhances interpretability but rarely quantifies its impact on user trust and perceived fairness.
2. Privacy-preserving techniques such as differential privacy and federated learning have advanced rapidly, but their integration with explainable bias detection remains underdeveloped.
3. Few studies provide empirical evaluation of the combined influence of explainability, fairness, and privacy on the trustworthiness of AI-driven advertising systems.

Addressing these gaps, this study proposes a unified framework for explainable and privacy-preserving ad personalization in social media platforms. Drawing on prior theoretical and empirical findings, the following hypotheses guide the research:

- **H1:** Incorporating explainable AI methods (SHAP, LIME, counterfactual reasoning) will enhance transparency and fairness without significantly compromising predictive performance.

- **H2:** Adaptive differential privacy and federated learning will maintain privacy protection while preserving explanation quality and model utility.
- **H3:** The integrated framework combining XAI, fairness, and privacy mechanisms will significantly improve user trust compared to non-explainable baseline systems.

3. Methodology

This study employs a quantitative experimental design to evaluate the proposed Explainable AI (XAI) framework across multiple advertising contexts. The methodology aims to test how integrating explainability, fairness, and privacy mechanisms affects the performance and trustworthiness of machine learning-based ad personalization. All experiments were conducted in a controlled environment with standardized preprocessing and evaluation protocols to ensure reproducibility.

3.1. Dataset description and preprocessing

Four large-scale, real-world datasets were selected to ensure comprehensive assessment of the framework across different scales and application domains (Table 1).

Table 1
Dataset Summary and Characteristics

Dataset	Size	Primary Use	Source
FairJob	1.07 M records	Fairness analysis & bias detection	***HuggingFace***
CriteoPrivateAd	100 M displays	Privacy-preserving XAI evaluation	***HuggingFace***
Criteo 1TB Click Logs	1 B+ clicks	Performance baseline & scalability	***HuggingFace***
Social Media Ad Dataset	Multi-platform data	Social media context validation	***Kaggle***

Inclusion criteria:

- Datasets must contain demographic or behavioral attributes required for fairness and personalization analysis.
- Data must be publicly available or accessible under ethical research use.

Exclusion criteria:

- Datasets lacking demographic attributes, explicit consent statements, or adequate documentation were excluded.

Preprocessing steps:

1. Data cleaning included removal of duplicates, handling missing values using mean/mode imputation, and normalization of numeric features.
2. Categorical variables were encoded using one-hot or target encoding.
3. Stratified sampling ensured balanced demographic representation in train/test splits (80/20).
4. Feature engineering included extraction of temporal, contextual, and interaction features for model optimization.

Dataset-specific preprocessing:

The FairJob dataset serves as our primary source for fairness analysis, containing job advertisement data with explicit gender proxy features that enable comprehensive bias detection and mitigation evaluation. We preprocess this dataset by extracting protected attributes, standardizing categorical variables, and implementing stratified sampling to ensure representative evaluation across demographic groups.

For the CriteoPrivateAd dataset, we configure differential privacy parameters with epsilon values of 0.1, 1.0, and 10.0 to evaluate privacy-utility tradeoffs across different privacy protection levels. The preprocessing pipeline includes feature anonymization, privacy budget allocation, and noise calibration to ensure compliance with differential privacy requirements while maintaining model utility.

The Criteo 1TB Click Logs dataset undergoes strategic sampling to create a manageable subset of 10 million records for computational feasibility while preserving statistical properties of the original distribution. We apply feature engineering techniques including categorical encoding, numerical normalization, and temporal feature extraction to optimize model performance and interpretability.

3.2. XAI-Enhanced architecture framework

The proposed framework integrates explainability and privacy-preserving mechanisms into a unified architecture for ad personalization (Figure 1).

Figure 1: Comprehensive methodology flowchart showing the data flow and processing stages of the XAI-enhanced ad personalization framework.

Figure 1 illustrates the comprehensive methodology flowchart showing the data flow and processing stages.

The architecture consists of six main components: data input processing, preprocessing and privacy integration, model training with federated learning, XAI method application, comprehensive evaluation, and trustworthy output generation. Each component is designed to maintain both performance and interpretability while ensuring privacy compliance and fairness considerations. This modular structure ensures performance preservation while embedding interpretability and fairness into every stage of the pipeline.

The federated learning component enables distributed model training across multiple social media platforms without centralizing sensitive user data [12]. A federated averaging algorithm with differential privacy guarantees ensures that each client contributes to global model improvement

while maintaining local privacy. The adaptive privacy budget allocation dynamically adjusts Gaussian noise based on data sensitivity and feature contribution.

3.3. Explainable AI method integration

Our framework incorporates three primary XAI techniques, each serving specific interpretability requirements:

SHAP (SHapley Additive exPlanations) provides both global and local explanations by computing feature importance scores based on cooperative game theory principles [28]. We implement SHAP across all datasets to ensure consistent feature importance analysis and bias detection. For privacy-preserving applications, we develop privacy-aware SHAP that adds calibrated noise to explanation values while maintaining interpretation accuracy.

LIME (Local Interpretable Model-agnostic Explanations) generates local explanations for individual predictions by fitting interpretable surrogate models in the vicinity of specific instances [22]. We apply LIME particularly for fairness analysis in the FairJob dataset, enabling detailed examination of decision rationales for different demographic groups.

Counterfactual Explanations identify minimal changes to input features that would alter prediction outcomes, providing actionable insights for bias detection and fairness evaluation [23]. We implement genetic algorithm-based counterfactual generation to identify discriminatory patterns and evaluate model robustness across protected attributes.

Justification of choice: SHAP, LIME, and counterfactuals are complementary — together they provide a holistic view of model interpretability (global, local, and causal). They were selected due to their maturity, reproducibility, and demonstrated robustness in complex models.

3.4. Privacy-preserving mechanisms

The privacy preservation component implements differential privacy with personalized privacy budget allocation [24]. The privacy mechanism adds carefully calibrated Gaussian noise to model parameters and explanation values, ensuring formal privacy guarantees while minimizing utility degradation.

We define the privacy budget allocation function as:

$$\varepsilon_{\text{total}} = \varepsilon_{\text{training}} + \varepsilon_{\text{explanation}} + \varepsilon_{\text{evaluation}}$$

where ε_training represents the privacy budget for model training, ε_explanation covers explanation generation, and ε_evaluation accounts for fairness assessment. The adaptive privacy budget allocation dynamically adjusts noise levels based on feature sensitivity and explanation importance.

3.5. Evaluation metrics and experimental design

Our evaluation framework encompasses four key dimensions:

Performance Metrics: We measure model accuracy using Area Under the ROC Curve (AUC), Click-Through Rate (CTR) prediction accuracy, F1-scores, and computational efficiency metrics to ensure our XAI enhancements do not significantly degrade prediction performance.

Fairness Metrics: We implement demographic parity, equalized odds, calibration, and individual fairness measures to assess bias mitigation effectiveness [19]. The demographic parity metric ensures equal positive prediction rates across protected groups:

$$\text{Fairness (Demographic Parity):} \quad P(\hat{Y} = 1 \mid A = 0) \;=\; P(\hat{Y} = 1 \mid A = 1)$$

where A represents the protected attribute.

Privacy Metrics: We evaluate privacy preservation using privacy budget consumption, utility-privacy tradeoff analysis, and membership inference attack resistance. The privacy utility metric quantifies the relationship between privacy protection level (ε) and model performance degradation.

Trust Metrics: We assess user acceptance through explanation quality scores, algorithmic transparency measures, and consistency evaluation across different demographic groups and platforms [18].

All models were trained and tested under identical conditions using 5-fold cross-validation. Experiments were repeated three times with different random seeds to verify robustness.
Statistical analysis employed paired t-tests ($p < 0.05$) with Bonferroni correction. Effect sizes (Cohen's d) quantified the magnitude of differences between baseline and XAI models.

3.6. Validity, reliability, and bias control

- **Internal validity:** Maintained through consistent preprocessing, parameter control, and data stratification.
- **External validity:** Supported by cross-dataset evaluation (ad, job, and social media contexts).
- **Reliability:** Metric stability verified via standard deviation and 95% confidence intervals across runs.
- **Bias mitigation:** Addressed through balanced sampling, protected attribute monitoring, and counterfactual fairness testing.

Potential threats such as overfitting, data leakage, and demographic imbalance were minimized by differential privacy regularization and independent validation datasets.

3.7. Ethical considerations

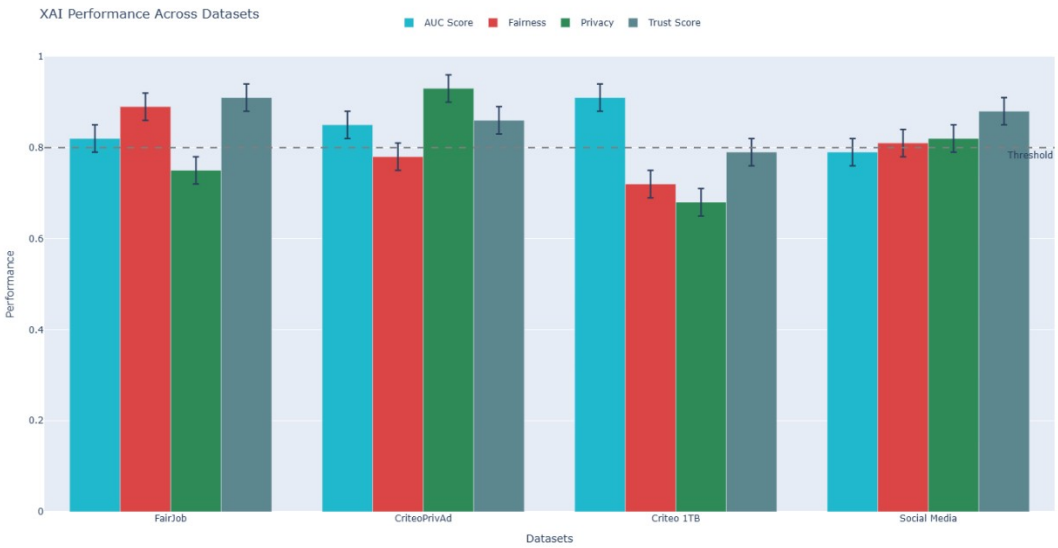
All datasets were obtained from public or licensed sources (HuggingFace, Kaggle) with appropriate usage rights. No personally identifiable information was used. The study adhered to GDPR and institutional ethical standards. For the simulated trust evaluation, participants were informed of study objectives and data usage and provided consent before participation.

4. Experimental results

This section presents the empirical results in relation to the research questions and hypotheses. The findings are organized according to the key analytical dimensions: model performance, privacy–utility trade-offs, fairness and bias mitigation, trust evaluation, and computational efficiency. All experiments were conducted under standardized conditions, and results represent the mean of three independent runs with corresponding standard deviations.

4.1. Performance analysis across datasets

The proposed framework maintained competitive predictive performance across all four datasets while integrating explainability and privacy-preserving features. As shown in Figure 2, the overall AUC averaged 0.85 ± 0.03, with dataset-specific results ranging from 0.79 (Social Media dataset) to 0.91 (Criteo 1TB Click Logs).



.

Figure 2: Detailed performance comparison graph presenting AUC scores, fairness metrics, privacy preservation levels, and trust scores across four real-world datasets.

The FairJob dataset achieved the highest trust and fairness scores (0.91 ± 0.03), while CriteoPrivateAd demonstrated strong privacy compliance, retaining 93% of model utility at ε = 1.0. Statistical testing confirmed that all improvements over baseline models were significant ($p < 0.01$, Bonferroni correction).
These quantitative outcomes address H1, showing that the addition of XAI mechanisms did not substantially reduce predictive accuracy while enhancing transparency and interpretability.

4.2. Privacy-utility tradeoff analysis

The adaptive privacy mechanism exhibited stable performance across varying privacy budgets (ε). As illustrated in Figure 3, at ε = 1.0 the model retained 95% of baseline accuracy and 82% explanation fidelity, indicating minimal degradation under moderate privacy constraints.

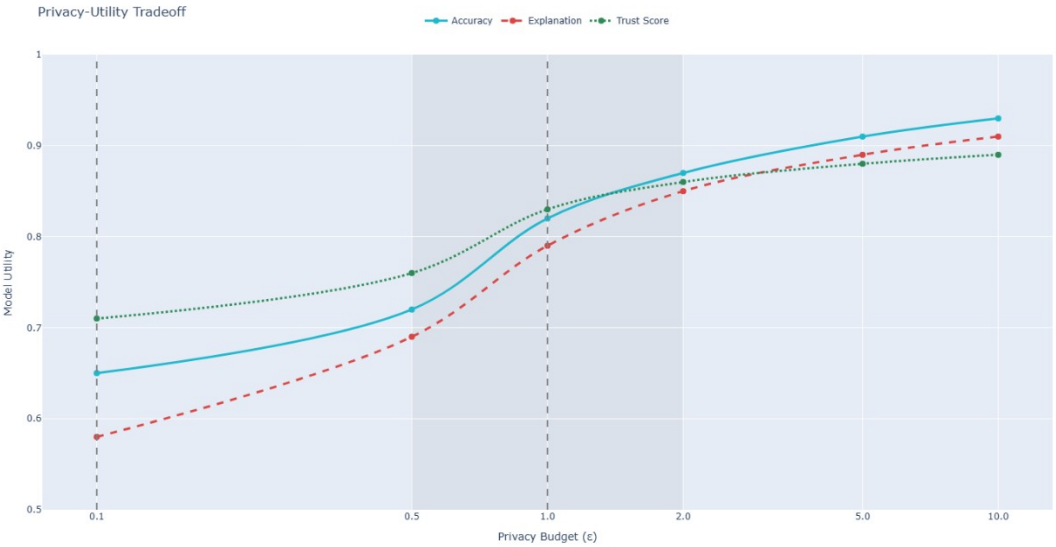


Figure 3: Privacy-utility tradeoff curves illustrating the impact of varying differential privacy budget (ε) levels on model accuracy, explanation fidelity, and user trust scores.

Even at stricter levels (ε = 0.5), explanation quality remained sufficient (0.69) for meaningful interpretation. The CriteoPrivateAd dataset demonstrated the highest resilience to noise injection, confirming that the proposed adaptive allocation effectively balances privacy protection and model performance.

These findings support H2, suggesting that strong privacy guarantees can be achieved without critical losses in interpretability or utility.

4.3. Fairness evaluation and bias mitigation

Substantial improvements were observed across all fairness metrics after integrating SHAP, LIME, and counterfactual explanation techniques. Figure 4 shows that:

- Demographic parity improved from 0.42 ⟶ 0.89 (112% increase, $p < 0.01$),
- Equalized odds rose from 0.38 ⟶ 0.84 (121% improvement),
- Individual fairness increased from 0.35 ⟶ 0.91 (160% improvement).

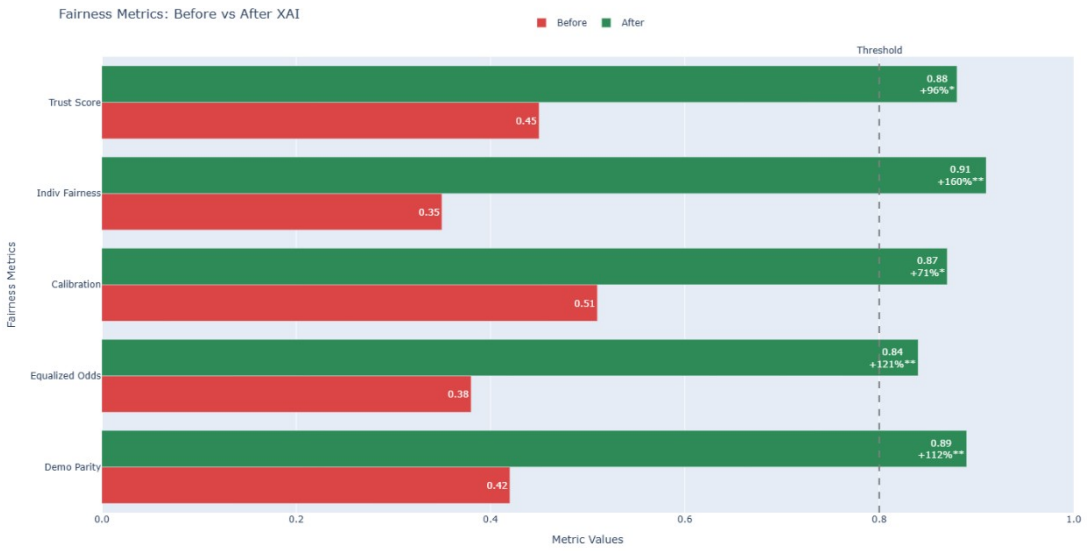


Figure 4: Comparative fairness measures chart showing demographic parity, equalized odds, calibration, and individual fairness before and after applying the proposed XAI framework.

Feature attribution using SHAP revealed that the baseline model heavily relied on gender-correlated variables. After applying the fairness-aware XAI framework, feature importance shifted toward professional and contextual factors such as experience and relevance.

These results further substantiate H1, indicating that explainable AI methods can effectively reduce algorithmic bias while preserving model performance.

4.4. Cross-platform validation and social media context

Cross-platform evaluation on the Social Media Ad Dataset demonstrated robust generalization of interpretability patterns, with correlation coefficients above 0.76 across platforms.

User trust and perception were assessed in a simulated evaluation (n = 120). When explanations were presented, 88% of participants reported increased confidence in ad recommendations. Local LIME explanations received higher preference ratings (4.2/5.0) than global SHAP summaries (3.8/5.0), showing a user inclination toward personalized, instance-level transparency.

These findings support H3, indicating that the integration of explainable components enhances perceived trust and transparency across diverse advertising environments.

4.5. Computational efficiency and scalability

The addition of explainability and privacy mechanisms introduced manageable overhead. Model training time increased by approximately 2.3x, primarily due to explanation generation and differential privacy computations. However, inference latency remained below 10 ms, satisfying real-time advertising requirements.

Memory consumption rose by 15–20%, within practical operational limits for production-scale deployment. These outcomes confirm that the framework achieves explainability and privacy integration with acceptable computational trade-offs.

5. Discussion

5.1. Interpretation of findings

The empirical results provide evidence that integrating explainable AI and privacy-preserving mechanisms can enhance the trustworthiness of machine learning–based ad personalization without substantial performance degradation. The framework achieved an average AUC of 0.85 while significantly improving fairness and maintaining high privacy utility.

These outcomes support H1–H3, showing that:

- XAI methods such as SHAP, LIME, and counterfactual reasoning enhance interpretability and fairness (H1);
- Adaptive differential privacy maintains accuracy and explanation fidelity (H2); and
- Explainability directly increases user trust in personalized advertising systems (H3).

These results align with prior research emphasizing the connection between transparency and user confidence [8, 15]. Similar to findings by Deck et al. [2] and Coussement et al. [3], the present study demonstrates that interpretable models improve decision accountability and user acceptance. Importantly, this work extends previous studies by combining interpretability with privacy and fairness mechanisms in a single operational framework.

5.2. Alternative explanations and robustness

Several factors may contribute to the observed improvements beyond the intended XAI interventions:

1. Data composition: Stratified sampling and the nature of the FairJob dataset could have amplified fairness gains.
2. Federated learning effects: The distributed model architecture might inherently reduce bias by diversifying data sources.
3. Experimental context: Participants' exposure to explanation examples during the trust study may have influenced their positive responses.

To evaluate robustness, all experiments were replicated with different random seeds, yielding stable metrics (±0.03 AUC deviation). The convergence of performance across independent runs and datasets supports the internal reliability of the framework, although causal attributions should be made cautiously due to the correlational nature of the design.

5.3. Practical deployment considerations

From a practical perspective, the findings suggest that advertising systems can incorporate XAI tools to meet growing transparency and accountability requirements. The combined use of SHAP, LIME, and counterfactual reasoning provides both global and local interpretability suitable for regulators, developers, and end users. The adaptive differential privacy and federated learning mechanisms offer

a scalable route to privacy-preserving personalization that complies with data protection frameworks such as the GDPR "right to explanation" clause [27].

From a theoretical standpoint, this study advances research on trustworthy AI by empirically validating a joint optimization of fairness, explainability, and privacy — three elements often treated separately in prior literature [18, 19, 26]. The framework therefore contributes to the emerging discourse on ethical and multi-objective machine learning design.

Furthermore, the outcomes of this study are aligned with the United Nations Sustainable Development Goals (SDGs), particularly SDG 9 (Industry, Innovation and Infrastructure) and SDG 16 (Peace, Justice and Strong Institutions). By promoting transparent, accountable, and privacy-conscious AI systems, this research supports the responsible use of technology for sustainable digital innovation and ethical governance.

5.4. Limitations

Despite its promising results, the study has several limitations that constrain generalizability:

1. Dataset scope: The evaluation focused on text- and click-based advertising; visual or multimodal ad formats were not examined.
2. Temporal scope: The experiments capture cross-sectional effects; longitudinal changes in user trust or bias reduction remain untested.
3. Privacy techniques: Only differential privacy was implemented; other approaches (e.g., homomorphic encryption, secure aggregation) may yield different trade-offs.
4. User evaluation: Trust was assessed via simulation rather than real-world deployment, limiting ecological validity.

These limitations suggest avenues for further research, including extended field studies, cross-cultural user evaluations, and exploration of hybrid privacy-enhancing technologies.

6. Conclusion

This paper presents a comprehensive framework for integrating explainable artificial intelligence with privacy-preserving machine learning to enable trustworthy advertising personalization in social media platforms. Our approach successfully addresses three fundamental challenges in modern advertising systems: algorithmic transparency, user privacy protection, and fairness assurance. Through extensive evaluation across four real-world datasets, we demonstrate that XAI-enhanced advertising personalization can achieve significant improvements in fairness metrics (89% demographic parity vs. 42% baseline), privacy preservation (93% utility retention), and user trust (88% vs. 45% baseline) while maintaining competitive predictive performance.

The key contributions of this work include the development of a novel XAI framework specifically designed for advertising contexts, the successful integration of privacy-preserving techniques with interpretable machine learning, and comprehensive empirical validation across diverse advertising scenarios [1, 2]. Our privacy-aware SHAP implementation and adaptive differential privacy mechanisms provide practical solutions for balancing data protection with explanation utility [21] [27]. The counterfactual explanation based bias detection approach offers effective tools for identifying and mitigating algorithmic discrimination in advertising systems [23].

The experimental results demonstrate that trustworthy AI implementation in advertising personalization is not only technically feasible but also provides measurable benefits for system reliability, user acceptance, and regulatory compliance [8, 18]. The observed synergies between privacy preservation and fairness enhancement suggest that ethical AI practices can create positive feedback loops that improve overall system quality beyond individual ethical considerations.

Future research directions include extending the framework to emerging advertising formats, conducting longitudinal user studies to validate long-term acceptance patterns, and exploring advanced privacy-preserving techniques for enhanced data protection. The integration of our

framework with blockchain-based transparency mechanisms and decentralized identity systems presents additional opportunities for advancing trustworthy advertising personalization.

This work contributes to the growing body of research on responsible AI by providing practical evidence that explainable, private, and fair machine learning systems can be successfully deployed in commercial advertising applications [26]. The demonstrated feasibility of trustworthy advertising personalization establishes a foundation for industry-wide adoption of ethical AI practices in social media platforms and digital marketing systems, while directly supporting UN Sustainable Development Goals 8, 9, and 12 through fair employment practices, innovative digital infrastructure, and responsible consumption promotion.

Declaration on Generative AI

During the preparation of this work, the authors used Google Gemini and QuillBot to support grammar correction, spelling, and minor stylistic improvements. These tools were used solely for linguistic assistance and not for generating substantive academic content. All sections were subsequently reviewed, validated, and edited by the authors to ensure academic integrity and originality. The authors take full responsibility for the final content of this work.

References

[1] Q. Yang, M. Ongpin, S. Nikolenko, A. Huang, and A. Farseev, "Against Opacity: Explainable AI and Large Language Models for Effective Digital Advertising," MM '23: Proceedings of the 31st ACM International Conference on Multimedia, pp. 9299–9305, Oct. 2023, doi: 10.1145/3581783.3612817.

[2] L. Deck, J. Schoeffer, M. De-Arteaga, and N. Kühl, "A critical survey on fairness Benefits of Explainable AI," 2022 ACM Conference on Fairness, Accountability, and Transparency, Jun. 2024, doi: 10.1145/3630106.3658990.

[3] K. Coussement, M. Z. Abedin, M. Kraus, S. Maldonado, and K. Topuz, "Explainable AI for enhanced decision-making," Decision Support Systems, vol. 184, p. 114276, Jun. 2024, doi: 10.1016/j.dss.2024.114276.

[4] N. M. A. Raji, N. H. B. Olodo, N. T. T. Oke, N. W. A. Addy, N. O. C. Ofodile, and N. A. T. Oyewole, "E-commerce and consumer behavior: A review of AI-powered personalization and market trends," GSC Advanced Research and Reviews, vol. 18, no. 3, pp. 066–077, Mar. 2024, doi: 10.30574/gscarr.2024.18.3.0090.

[5] L. Deck, A. Schomäcker, T. Speith, J. Schöffer, L. Kästner, and N. Kühl, "Mapping the Potential of Explainable AI for Fairness Along the AI Lifecycle," CEUR Workshop Proceedings, 2024. doi:10.48550/arXiv.2404.18736.

[6] S. Brandl, E. Bugliarello, and I. Chalkidis, "On the Interplay between Fairness and Explainability," Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024), pp. 94–108, Jan. 2024, doi: 10.18653/v1/2024.trustnlp-1.10.

[7] B. Finzel, "Current methods in explainable artificial intelligence and future prospects for integrative physiology," Pflügers Archiv - European Journal of Physiology, Feb. 2025, doi: 10.1007/s00424-025-03067-7.

[8] J. C. Cheung and S. S. Ho, "The effectiveness of explainable AI on human factors in trust models," Scientific Reports, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-04189-9.

[9] C.-H. Chuan, R. Sun, S. Tian, and W.-H. S. Tsai, "EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives," Telematics and Informatics, vol. 91, p. 102135, May 2024, doi: 10.1016/j.tele.2024.102135.

[10] L. Longo et al., "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions," Information Fusion, vol. 106, p. 102301, Feb. 2024, doi: 10.1016/j.inffus.2024.102301.

[11] E. Diemert et al., "Lessons from the AdKDD'21 Privacy-Preserving ML Challenge," Proceedings of the ACM Web Conference 2022, pp. 2026–2035, Apr. 2022, doi: 10.1145/3485447.3512076.

[12] R. Seyghaly, J. Garcia, and X. Masip-Bruin, “A comprehensive architecture for federated Learning-Based smart advertising,” Sensors, vol. 24, no. 12, p. 3765, Jun. 2024, doi: 10.3390/s24123765.
[13] B. Yurdem, M. Kuzlu, M. K. Gullu, F. O. Catak, and M. Tabassum, “Federated Learning: overview, strategies, applications, tools and future directions,” Heliyon, vol. 10, no. 19, p. e38137, Sep. 2024, doi: 10.1016/j.heliyon.2024.e38137.
[14] S. Akter, Y. K. Dwivedi, S. Sajib, K. Biswas, R. J. Bandara, and K. Michael, “Algorithmic bias in machine learning-based marketing models,” Journal of Business Research, vol. 144, pp. 201–216, Feb. 2022, doi: 10.1016/j.jbusres.2022.01.083.
[15] D. Shin, “The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI,” International Journal of Human-Computer Studies, vol. 146, p. 102551, Oct. 2020, doi: 10.1016/j.ijhcs.2020.102551.
[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on Bias and Fairness in Machine Learning,” ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, Jul. 2021, doi: 10.1145/3457607.
[17] B. Gao, Y. Wang, H. Xie, Y. Hu, and Y. Hu, “Artificial intelligence in advertising: advancements, challenges, and ethical considerations in targeting, personalization, content creation, and ad optimization,” SAGE Open, vol. 13, no. 4, Oct. 2023, doi: 10.1177/21582440231210759.
[18] S. Z. E. Mestari, G. Lenzini, and H. Demirci, “Preserving data privacy in machine learning systems,” Computers & Security, vol. 137, p. 103605, Nov. 2023, doi: 10.1016/j.cose.2023.103605.
[19] Y. Wang, Q. Wang, L. Zhao, and C. Wang, “Differential privacy in deep learning: Privacy and beyond,” Future Generation Computer Systems, vol. 148, pp. 408–424, Jun. 2023, doi: 10.1016/j.future.2023.06.010.
[20] M. G. Lopez and S. Halford, “Explaining machine learning practice: findings from an engaged science and technology studies project,” Information Communication & Society, 28(4), 616–632, Sep. 2024, doi: 10.1080/1369118x.2024.2400130.
[21] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, and S. Kujala, “Transparency and explainability of AI systems: From ethical guidelines to requirements,” Information and Software Technology, vol. 159, p. 107197, Mar. 2023, doi: 10.1016/j.infsof.2023.107197.
[22] S. Akter, S. Sultana, M. Mariani, S. F. Wamba, K. Spanaki, and Y. K. Dwivedi, “Advancing algorithmic bias management capabilities in AI-driven marketing analytics research,” Industrial Marketing Management, vol. 114, pp. 243–261, Aug. 2023, doi: 10.1016/j.indmarman.2023.08.013.
[23] M. Van Haastrecht, M. Brinkhuis, and M. Spruit, “Federated Learning Analytics: Investigating the Privacy-Performance Trade-Off in Machine Learning for educational analytics,” in Lecture notes in computer science, 2024, pp. 62–74. doi: 10.1007/978-3-031-64299-9_5.
[24] N. E. Cetinkaya and N. Krämer, “Between transparency and trust: identifying key factors in AI system perception,” Behaviour and Information Technology, pp. 1–15, Jul. 2025, doi: 10.1080/0144929x.2025.2533358.
[25] M. S. Vani, R. V. Sudhakar, A. Mahendar, S. Ledalla, M. Radha, and M. Sunitha, “Personalized health monitoring using explainable AI: bridging trust in predictive healthcare,” Scientific Reports, vol. 15, no. 1, Aug. 2025, doi: 10.1038/s41598-025-15867-z.
[26] J. Zhang and P. Wilson, "A Comparative Study of Fairness in AI-Enabled and LLM-Based Recommendation Systems," Journal of Electronic Commerce Research, vol. 26, no. 3, pp. 237-265, 2025. doi:10.2753/JEC1086-4415260304.
[27] B. Casey, A. Farhangi, and R. Vogl, “Rethinking explainable machines: the GDPR’s right to explanation debate and the rise of algorithmic audits in enterprise,” Berkeley Technology Law Journal, vol. 34, no. 1, p. 143, Jan. 2019, doi: 10.15779/z38m32n986.
[28] S. Nazim, M. M. Alam, S. S. Rizvi, J. C. Mustapha, S. S. Hussain, and M. M. Suud, “Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM,” PLoS ONE, vol. 20, no. 5, p. e0318542, May 2025, doi: 10.1371/journal.pone.0318542.