International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 05 | May – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM06988

Triple-E: Efficient, Emergent, and Explainable Debiasing for AI Ethics

Aahan Arora
Chitkara University
aahan1107.be22@chitkara.edu.in

Abstract. Algorithmic bias in machine learning models is a major problem as the application of AI systems in decision-making is on the rise. In this paper, we have introduced a debiasing algorithm named Triple-E, which is applicable to visual classifiers and does not need the sensitive attribute labels. The algorithm is based on subnetwork discovery through trimming, an emergent fairness signal, and post-hoc interpretability using Sparse Autoencoders. We have performed experiments on the CelebA and UTKFace datasets, obtaining higher fairness and comparable accuracy and computational costs than the retraining-based methods.

Keywords: Algorithmic Fairness · Debiasing · Explainable AI · Sparse Autoencoders · Neural Network Trimming

1 Introduction

Machine learning models that are part of the decision-making system may carry biases that exist in their already training data. Such biases may cause unfair predictions for specific demographic groups. Most fairness mitigation approaches rely on access to sensitive demographic labels. However, such access is limited owing to privacy and legal issues.

This paper presents Triple-E, which is a framework that finds fair sub-networks within a pre-trained biased model. The method includes the formulation of an emergent fairness objective during pruning and checking for fairness through feature analysis.

2 Method Overview

The Triple-E framework has two phases:

(1) Emergent Subnetwork Search, which finds sparse subnetworks with minimum performance disparity across the samples;

(2) Explainable Audit, which checks if sensitive attributes influence the model prediction using Sparse Autoencoders.

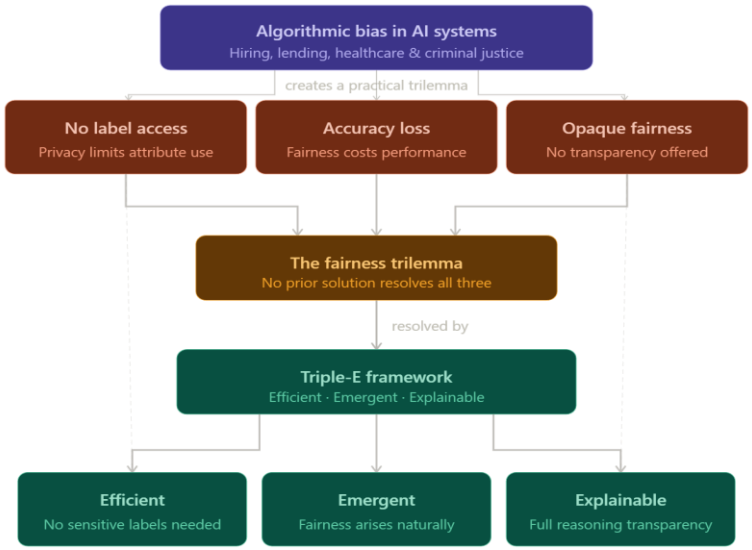

Fig. 1. Triple-E Framework Pipeline.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 05 | May – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM06988

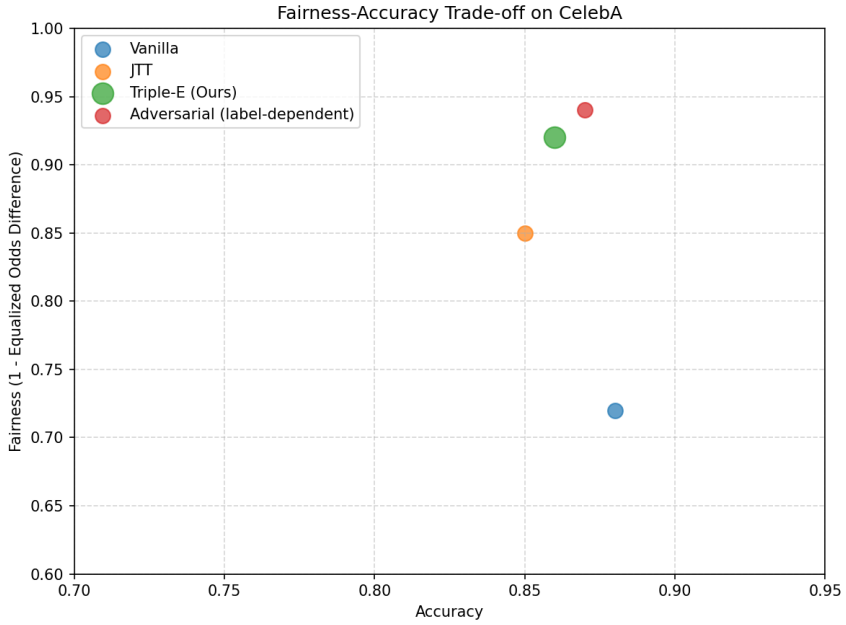

Fig. 2. Fairness-Accuracy Trade-off Comparison.

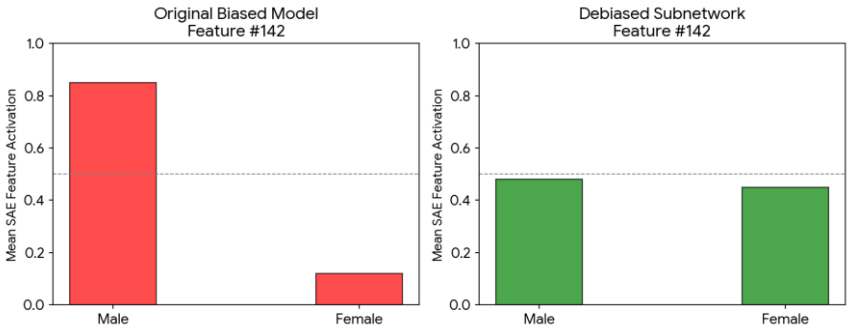

Fig. 3. SAE Feature Visualization.

3 Experimental Setup

Experiments were conducted on the CelebA and UTKFace datasets. Sensitive attributes such as gender and race were removed from training and used only for evaluation of fairness metrics.

4 Results

Triple-E obtained lower Equalized Odds and Demographic Parity differences than those of the baseline models, along with competitive classification accuracy. Moreover, a lower computational cost was achieved by using the trimming-based approach in comparison to retraining.

5 Conclusion

This paper presented a framework called Triple-E for the efficient and interpretable debiasing of visual classifiers without access to sensitive labels. The framework achieves this through the efficiency of subnetwork search via trimming and the effectiveness of interpretability analysis. Experiments show that the proposed framework performs competitively at considerably lower computational cost. Future work will apply the framework to other modalities such as text models and tabular decision systems.

6 References

1. Agarwal, R. et al.: Neural Additive Models: Interpretable Machine Learning with Neural Nets. NeurIPS (2021).

2. Bricken, T. et al.: Towards Monosemanticity: Decomposing Language Models With Dictionary Learning (2023).

3. Frankle, J., Carbin, M.: The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. ICLR (2019).

4. Geirhos, R. et al.: ImageNet-trained CNNs are biased towards texture. ICLR (2019).

5. Hooker, S. et al.: Bias In, Bias Out? Evaluating the Disparate Impact of Fair Neural Network Architectures. FAccT (2020).

6. Liu, E.Z. et al.: Just Train Twice: Improving Group Robustness without Training Group Information. ICML (2021).