Building Preference Aware Trustworthy AI Through Fairness and Interpretability

Jingyu Hu
University of Bristol, BS8 1QU, United Kingdom
ym21669@bristol.ac.uk

Abstract

While AI has brought transformative changes to society, the opacity of current state-of-the-art models still raises ethical concerns in high-risk domains. It is necessary to improve the trustworthiness of these black-box models to make them more responsible. This paper presents five completed works on enhancing the trustworthiness of both traditional ML models and large language models (LLMs) from the perspectives of fairness and interpretability. Future work will extend the discussion of trustworthiness to multi-modal scenarios.

Introduction

AI and ML systems are pervasive. With the widespread adoption of these systems, governments, large corporations, industry, AI researchers and practitioners have begun to acknowledge their limitations in regard to ethics, transparency, privacy, and trust. In this context, my research focuses on improving AI trustworthiness by enhancing fairness and interpretability. Specifically, I propose to (1) explore bias mitigation techniques to improve model fairness so they can align with human values, and (2) explore interpretability techniques to make model decision-making processes more transparent, and provide explanations for users with different preferences.

Fairness can be discussed at individual and group levels. Individual fairness measures bias by checking whether similar predictions can be made for similar individuals. Group fairness compares the treatment of majority and under-represented groups, with fairness achieved when treatment is equal between groups. This paper introduces three completed work on group fairness enhancement, including improving fairness in both traditional ML models and Large Language Models (LLMs) through designed data augmentation and strategic data selection, as well as designing a hardness-aware loss function to address degree bias in graph contrastive learning.

Interpretability discussions can be divided into two categories based on user preferences: (1) Interpretation for AI experts to understand models' internal mechanisms, and better optimize models in a responsible direction, and (2) explanations for non-experts to increase public trust in AI. For the first category, our work uses representation learning for political opinions interpretation: learn, detect and intervene in the internal states of LLMs across each layer. For the second category, our work develops a healthcare interactive interface to visualize decision paths and generate counterfactual explanations to enhance non-experts' confidence in predictive models.

The following section introduces these work details on fairness and interpretability respectively, then drives our future research to further advance the field.

Completed Work on Model Fairness

Data is collected from the real world, and models can reflect the features they learn from data. Therefore, model bias mirrors the real-world bias to some extent. We propose three bias-mitigation methods with their core idea to improve the fairness of data usage and selection.

Pre-processing Bias Mitigation As ML models can deepen bias in biased datasets, an intuitive bias mitigation way is to augment the dataset itself to improve fairness. Our work proposes a pre-processing method called ProxiMix (Hu et al. 2024) to generate more synthetic samples for under-represented groups. ProxiMix is a variation of Mixup. Mixup is a data augmentation technique that linearly interpolates two samples to create synthesized data. However, we found its limitation is that if the original labels in the dataset are biased, this bias can be retained in the mixed samples, potentially introducing extra bias to the models. ProxiMix defines proximity samples to re-audit mixed labels, aiming to mitigate this bias in Mixup. The experiments discuss the trade-offs between ProxiMix and traditional Mixup, and our ProxiMix enables the model to achieve fairer performance across different ML models (decision tree, logistics regression, MLP)

In-processing Bias Mitigation SHARP (Hu et al. 2025a) is our proposed in-processing method to specifically address the degree of bias in graph data. The degree bias in the node classification task refers to models that tend to perform better in the high-degree nodes group than in low-degree nodes group. We assume it is due to that low-degree nodes are more likely to receive insufficient and noisy information from their neighbouring nodes. To enhance available information to low-degree nodes, we introduce a Hardness Adaptive Reweighted (HAR) graph contrastive loss. It as-

signs more nodes to low-degree based on label information and assigns adaptive weights to these nodes based on their learning hardness. This allows low-degree nodes learning from more informative nodes. To validate the effectiveness of HAR on broader scenarios, we extended HAR to semi-supervised scenarios to form SHARP framework. SHARP is compared with three baselines, and results show that our approach improves both the global prediction performance and subgroup-level performance of the low-degree nodes.

Training-free Bias Mitigation Both of the above methods require retraining the model. Our third work (Hu, Liu, and Du 2024) explores a training-free approach to mitigate bias in large language models (LLMs). LLMs are pre-trained on massive corpora and can exhibit gender bias in their responses. However, fine-tuning LLMs is computationally expensive. Therefore, we propose demonstration selection strategies to enhance LLM fairness in in-context learning. Overall, we find that increasing the number of demonstrations from minority groups can improve fairness. Building on this, we propose a Fairness via the Clustering-Genetic algorithm. It uses clustering and evolutionary ideas to select diverse and representative demonstrations. The experiments demonstrate that deliberately selected demonstrations can enhance both LLMs prediction and fairness performance.

Completed Work on Model Interpretability

The interpretability work mainly focuses on providing explanations for both AI experts and non-experts. AI experts refer to people with AI-related backgrounds (e.g., computer science, machine learning). Non-experts are those from non-AI backgrounds (e.g., philosophy, sociology, and cognitive sciences). People with different backgrounds can have opposite preferences for explanations. For example, experts prefer to understand the internal structure of models and decision boundaries to think about further model optimization. This usually includes extensive case enumeration and the generation of numerous charts. Non-experts prefer to present the explanation in an intuitive way and engage in communication to lead their own understanding of models. Our work conducted explanation research based on their respective preferences.

Model Interpretability for Non-experts. Our work (Hu et al. 2023) presents an interactive and dynamic explainable AI (XAI) interface tailored for non-expert users. It presents a case study of predicting the 10-year risk of coronary heart disease using the decision tree classifier. Instead of overwhelming users with opaque model outputs, the proposed interface integrates global explanations, local explanations, and counterfactual explanations, offering a comprehensive view of the model's decision-making process. We evaluated the interface with 200 participants and the questionnaire confirmed the overall good experience. One interesting insight is that non-AI experts reported higher satisfaction than those already familiar with AI. It shows our designs considered the preferences of the target user group.

Model Interpretability for Experts. Many current ethical issue discussions of LLMs are based on their explicit responses (Bender et al. 2021). However, recent work has shown the LLMs' responses and their internal states can be misaligned. To explore LLMs' internal states, our work (Hu et al. 2025b) investigates political opinion learning, detection, and intervention within the internals of LLMs. To avoid political concept confounds like 'left is right', we define a fine-grained interpretability framework for the experiments. Our results show that different political leanings show different feature distributions within LLMs' internal representations. By intervening in these internal distributions, LLMs can be guided to generate statements aligned with different political stances. We also discover that political opinions embedded in internal layers can be inconsistent with the model's final layer outputs. These findings provide valuable insights for further researchers aiming to improve the trustworthiness of LLMs.

Future Work

Current completed work has discussed some approaches to improve model fairness and interpretability. These studies provide valuable insights for enhancing trustworthiness in both traditional machine learning models and large models.

Nevertheless, one limitation is that these methods are mostly based on single modalities, while data is often multimodal in reality. To broaden the scope of trustworthy research, my next step is to explore fairness and interpretability issues in multimodal models. One challenge in multimodal trustworthiness is handling information conflicts across modalities. For instance, when predicting a person's risk of heart disease, there may be X-ray images, tabular data, heartbeat sounds, and other types of patient records. Some data modalities may indicate the person is ill, while others suggest they are healthy. Determining the relative contribution of each modality is essential for trustworthy decision-making. Therefore, we will consider methods to align different modalities and develop a unified framework to assess the trustworthiness of multimodal models in the future work.

References

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21*, 610–623.

Hu, J.; Bo, H.; Hong, J.; Liu, X.; and Liu, W. 2025a. Mitigating Degree Bias Adaptively with Hard-to-Learn Nodes in Graph Contrastive Learning. *arXiv preprint arXiv:2506.05214*.

Hu, J.; Hong, J.; Du, M.; and Liu, W. 2024. ProxiMix: Enhancing Fairness with Proximity Samples in Subgroups. In *AEQUITAS@ECAI 2024*.

Hu, J.; Liang, Y.; Zhao, W.; McAreavey, K.; and Liu, W. 2023. An Interactive XAI Interface with Application in Healthcare for Non-experts. In *World Conference on Explainable Artificial Intelligence*, 649–670. Springer.

Hu, J.; Liu, W.; and Du, M. 2024. Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning. In *EMNLP 2024*, 7460–7475.

Hu, J.; Yang, M.; Du, M.; and Liu, W. 2025b. Fine-Grained Interpretation of Political Opinions in Large Language Models. *arXiv preprint arXiv:2506.04774*.