# Investigation and Mitigation of Bias in Explainable AI

Muhammad Suffian[1,*], Alessandro Bogliolo[1]

[1]Department of Pure and Applied Sciences, University of Urbino Carlo Bo, Urbino, Italy

Abstract

Fairness in artificial intelligence (AI) has recently gained more attention since decisions made by AI applications could unfavourably affect individual groups and have ethical or legal consequences. Making sure that AI-based systems don't show bias toward particular individuals or groups is crucial. In the field of explainable artificial intelligence (XAI), counterfactual explanations are considered user-friendly and provide counterfactual data points to the user to achieve the desired outcomes. The implementation of suggestions by the user could even lead to unfavourable results (unfortunate explanations) in many cases, and this cyclic process (suggestion implementation) could cause frustration for the end user. Meanwhile, assigning a negative label to each implemented data point could imbalance the repository of all implemented data points. The learning from new data streams undergoes the problems of data skew such as *concept-drift* and *data-drift*. Most existing approaches retrain the models offline by including the new data in old data without considering the data skew and class imbalance. In this position paper, we propose designing a fairness-aware mechanism for user-friendly explanatory systems (counterfactual explanation-based systems). This mechanism encompasses an incremental learning approach for the underlying machine learning (ML) model to retrain it on the implemented data points suggested by the counterfactual explanatory system during online interaction. We propose an experiment to investigate the bias (fairness issues) and bias mitigation with our strategy in interactive explanatory systems. The contribution of this work is two-fold. First, we process the new data to evaluate and compare it with old data for class imbalance and data drift. Second, we introduce an incremental learning-based ensemble of ML models to improve the performance and use them for class-label prediction of the new data points.

**Keywords**

Bias, Transparency, Fairness, Counterfactual Explanations, Explainable AI

## 1. Introduction

Artificial Intelligence (AI) plays a significant role in governing and shaping our lives, from informing decisions regarding offenders in criminal justice systems, approvals in lending applications to recommending which books to read, movies to watch, and websites to browse [1, 2]. The notion of bias and biased decisions are well-known in the AI community, while their competing definitions are still inconsistent [3]. Nevertheless, there is a substantial discussion about the central aspect of AI algorithms and whether they are fair and transparent in their predictions [4]. The ability to explain and justify their decisions is critical for a fair and transparent system. There are three unanimously agreed types of bias in the literature: data

*1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA 2022, University of Udine, Udine, Italy, 2022*

*Corresponding author.

m.suffian@campus.uniurb.it (M. Suffian); alessandro.bogliolo@uniurb.it (A. Bogliolo)

0000-0002-1946-285X (M. Suffian); 0000-0001-6666-3315 (A. Bogliolo)

bias, algorithmic bias, and human bias [5]. We argue that data bias leads to algorithmic bias and could be mitigated by ensuring data collection and processing techniques. At the same time, the problem of human bias needs special attention from the developers and designers of the algorithm. Apart from bias mitigation processes, biased decision-making persists in live[1] predictive (decision-making) platforms where decision-making is performed, such as social networks. Similarly, automated data-driven systems do discrimination against a specific community or individuals sharing the same attributes. For example, bias in face and voice recognition systems [6] and Microsoft's AI chatbot (Twitter taught chatbot had become racist in less than 24 hours after its launch) [7].

Explainable Artificial Intelligence (XAI) has emerged to provide more transparent and fair solutions [8] by explaining how an intelligent system viz-a-viz ML model came to a particular decision [9]. There has been a vast amount of research on XAI; the literature on XAI uncovers its potential to explain how algorithms make decisions and predict how they will act in the future [10]. Empirical research is being done on human-in-the-loop for XAI-based systems [1, 11, 12]. Wachter et al. [10] argue that in an automated interactive explanation system, an explanation should achieve one of the multiple goals, which is to know what would need to change to obtain a desired outcome in the future based on the current decision-making model. Also, she claimed that counterfactual explanations (CE) could produce explanations by adhering to this explanation goal [10]. The ML model retrained on the biased and imbalanced data that could mislead the CE system to provide biased explanations for a specific individual or community. In post-hoc XAI techniques, users are offered explanations based on outcome to make changes in the input, although these changes do not guarantee that explanations are in line with them [13].

## 1.1. Motivation

Counterfactual explanations can be misguided by the underlying ML model; nevertheless, these can be utilized in investigating the bias and fairness of the same model as an interactive channel. Counterfactual explanations offer a human-interpretable explanation of a machine learning outcome and provide a "suggestion" to reach a different, perhaps more advantageous result by offering a "what-if" scenario. This feature sets CE apart from other explanation techniques like Local Interpretable Model-agnostic Explanations (LIME) [14] and Shapley Additive Explanations (SHAP) [15]. A CE refers to the counterfactual scenario as a "hypothetical point" that is classified differently from the actual point. We also term it as new data point[2]. The new data points emerge while the user implements multiple suggestions to get desired results. The various attempts of implementations lead to the creation of new data that are assigned class labels (in case of classification) from the same ML model and stored in the production data. After a time 't', the new data is included in the old data to retrain the models (to update the models) without considering any fairness-aware strategy [16]. Thus, a model providing negative outcomes keeps

[1] we refer to the term 'live' as an interchangeable term with 'online' and 'real-time', in this study, we are considering the already deployed systems which provide ML predictions and explain their outcomes simultaneously, which are responsive to multiple attempts of user based on the provided suggestions. Also, which are retrained on updated data after specific time intervals.

[2] In the rest of the article, we use 'new data point', 'hypothetical point', and 'suggestion' as interchangeable terms.

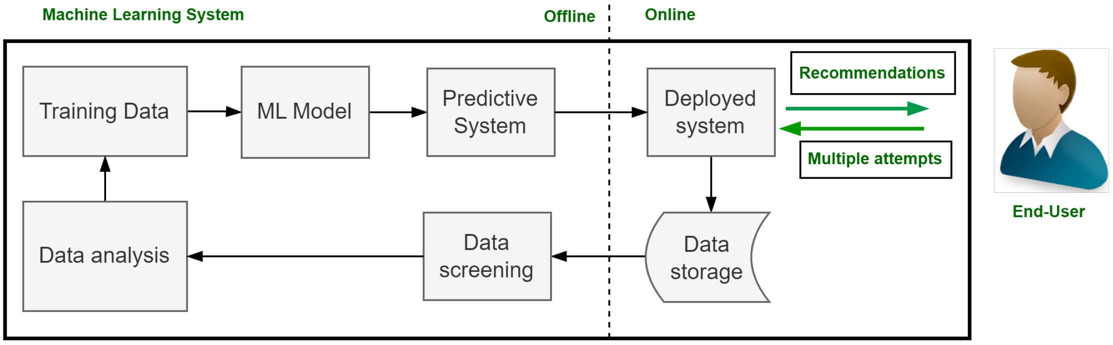

**Figure 1:** A high-level overview of the functional schema of fairness-aware machine learning-based XAI system.

learning the same concept from the new data turning biased towards a specific group. The motivation to use CE as an explanatory system is its explanation format that can easily be interpreted whether there is a bias. The objective is to analyze and compare the explanations generated for the retrained ML model with and without any fairness-aware strategy. This comparison will help maintain the counterfactual situation if the sensitive information in the input is changed, which will evaluate the proposed strategy's performance (See section 2).

## 1.2. Research Problem

The data streams evolve, and some data characteristics might lead to changes in the distribution, termed data drift [16]. Another factor of evolving data streams is class imbalance. In the case of classification, the positive class is considered a minority and is overlooked in the retraining of the models. To deal with bias problems in the data and algorithms, the researchers have advocated for increasing algorithms' fairness by introducing techniques that focus on sensitive attributes (protected) in the data [17]. A fairness-aware adversarial perturbation method (FAAP) [18] perturbs the input data related to fairness-related features to blind the model without changing the model structure. Another work presents an approach for fairness-aware machine learning to mitigate the algorithmic bias in law enforcement technology while relying on biased data to recognize violent activities [19]. Generally, most ML systems do not unfold their fairness mechanism. Every record in real-time predictive systems is essential. It helps to improve the underlying predictive systems with growing data [20].

Figure 1 presents a high-level functional schema of ML systems (online and offline). We can observe that if the user implements suggestions (new data points), in that case, there is no mechanism to scrutinize the inclusion of data points in the old data. A common perception is that the data is screened and analyzed to make it compatible with the old data to include in training data. We draw two assumptions about these systems: the new data points are included in the old data without ensuring the data imbalance problem, and the new data is not analyzed to confirm data drift. We investigate these cases with our proposed solution. The counterfactual explanation system produces suggestions, which by definition, may correspond to fictional people (such as clients, patients, and recidivists). However, we indicate that even though these

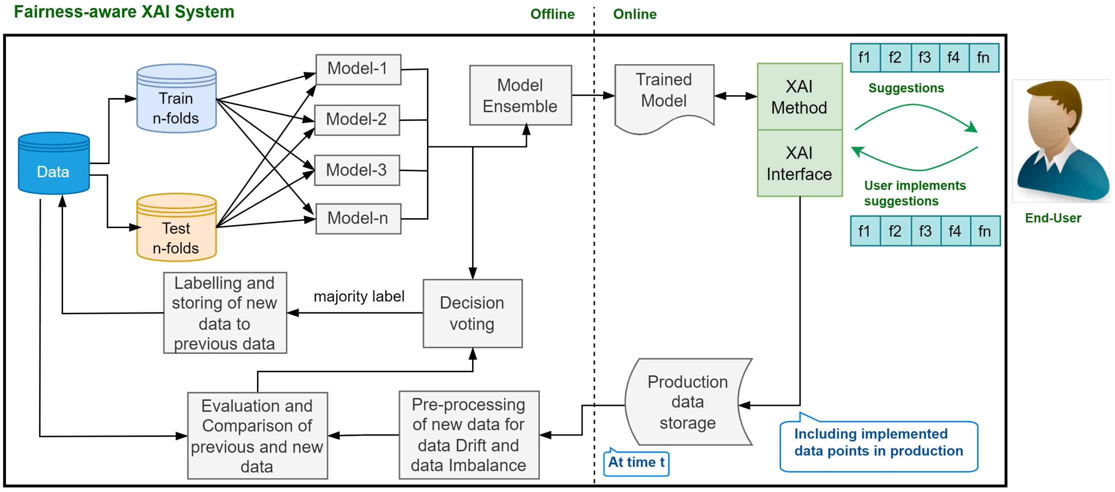
Fairness-aware XAI System
Offline
Online
Data
Train n-folds
Test n-folds
Model-1
Model-2
Model-3
Model-n
Model Ensemble
Trained Model
XAI Method
XAI Interface
f1 f2 f3 f4 fn
Suggestions
User implements suggestions
f1 f2 f3 f4 fn
End-User
Labelling and storing of new data to previous data
majority label
Decision voting
Evaluation and Comparison of previous and new data
Pre-processing of new data for data Drift and data Imbalance
Production data storage
At time t
Including implemented data points in production

**Figure 2:** A high-level overview of the functional schema of fairness-aware machine learning-based XAI system.

are hypothetical situations, the data points used to encode the counterfactual situations could represent real people. The data points could have biased outcomes for individual groups, which are generated during multiple implementations of the suggestions. The cycle of implementations and generation of new data points may impact the distribution and composition of data, i.e., the cohorts of patients, customers, or criminals used to retrain the model. If such data points are included in the data storage, these will imbalance the whole data and cause biased learning for the ML model. Accordingly, we assume that most explanation systems providing post-hoc explanations are not flexible to the formal processes of data balancing, data drift and bias removal. In such systems, the underlying ML models are retrained with spurious data. Such retrained models ultimately misguide the explanatory systems to produce unethical and biased explanations.

## 2. Proposed Approach

Our proposed fairness-aware approach ensures the new data's imbalance and drifts before including it in retraining data, which is presented in Fig. 2.

We enhance the general functional schema of ML systems by introducing a mechanism that analyzes data drift and imbalance. We include an ensemble of ML models trained and tested with n-folds. The best model from the ensemble models is used as a trained model in the XAI system, which explains its decisions to the end user in counterfactual explanations (suggestions). An explanation interface helps to present counterfactual suggestions to the end user. Multiple hypothetical situations could be suggested to the end user. In response, if the end-user acts to implement those suggestions, then such implemented data is stored in the production data. At time 't', the production data could be subjected to analysis with a fairness-aware mechanism. The new data is analyzed, compared and evaluated with the old data to identify the imbalance and drift properties of the data. We use drift handling techniques if the data is imbalanced and owning drift. Several data drift techniques are available concerning specific problems of statistical properties of data [16]. We use an adaptive windowing algorithm (ADWIN) [21] to detect data drift. ADWIN uses data streams to account for the different statistical properties between the old and new data, and detects drifted data points accordingly. For the case of imbalanced data, we use ensembles of ML models with balanced class n-folds retraining.

After analysis and evaluation, the new data is predicted from the ensemble models, and class labels are assigned with the majority votes for a specific label. Thus, newly processed and labelled data is included in the old data. As part of our fairness-aware pipeline, we periodically call the ADWIN algorithm and other analyses to retrain the models after time ‘t’. For example, suppose an implemented data point is situated near the decision boundary; then, the voting mechanism helps to assign the label. In other cases, if the new data show different statistical properties regarding distribution, then ADWIN can identify such data points. In the case of data streams (reflecting data drift), the underlying distribution of data can be adjusted and adapted accordingly.

We propose an experiment to perform an investigation relating to bias mitigation. Our approach provides a fairness-aware mechanism for the drifted and imbalanced data points in the retraining process. We will focus on two types of user studies: recidivism racial bias (compass[3] data) and credit lending (german[4] credit data). We will analyze the results in terms of presented suggestions (counterfactual cases) using benchmark counterfactual explanation techniques. We will highlight the improvements produced by the proposed approach in response to ethical events. The differences in the explanations with and without a fairness-aware strategy could help evaluate the proposed mechanism.

## 3. Discussion

Our investigation and mitigation strategy leaves several key issues unanswered. What are the high-stakes applications that require online retraining in XAI systems, and how can our approach be scaled? How frequently should the synchronization of data and retraining happen during explanations? At what level human-bias could be involved in such systems? What level of algorithmic rationale ought to be visible in the design? We hope to explore and discuss these questions during the workshop.

## References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[2] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106.

[3] O. B. Deho, C. Zhan, J. Li, J. Liu, L. Liu, T. Duy Le, How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics?, British Journal of Educational Technology (2022).

[3] https://www.kaggle.com/datasets/danofer/compass

[4] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[4] D. Varona, J. L. Suárez, Discrimination, bias, fairness, and trustworthy ai, Applied Sciences 12 (2022) 5826.

[5] J. E. Fountain, The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms, Government Information Quarterly 39 (2022) 101645.

[6] A. Howard, J. Borenstein, The ugly truth about ourselves and our robot creations: the problem of bias and social inequity, Science and engineering ethics 24 (2018) 1521–1536.

[7] M. J. Wolf, K. W. Miller, F. S. Grodzinsky, Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications, The ORBIT Journal 1 (2017) 1–12.

[8] T. Speith, A review of taxonomies of explainable artificial intelligence (xai) methods, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2239–2250.

[9] I. Ahmed, G. Jeon, F. Piccialli, From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where, IEEE Transactions on Industrial Informatics 18 (2022) 5031–5042.

[10] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[11] M. Suffian, P. Graziani, J. M. Alonso, A. Bogliolo, Fce: Feedback based counterfactual explanations for explainable AI, IEEE Access 10 (2022) 72363–72372.

[12] J. Zhou, A. H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (2021) 593.

[13] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, K. D. Mandl, Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency, NPJ digital medicine 3 (2020) 1–5.

[14] M. Tulio Ribeiro, S. Singh, C. Guestrin, " why should i trust you?": Explaining the predictions of any classifier, ArXiv e-prints (2016) arXiv–1602.

[15] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[16] I. Khamassi, M. Sayed-Mouchaweh, M. Hammami, K. Ghédira, Discussion and review on evolving data streams and concept drift adapting, Evolving systems 9 (2018) 1–23.

[17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[18] Z. Wang, X. Dong, H. Xue, Z. Zhang, W. Chiu, T. Wei, K. Ren, Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10379–10388.

[19] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez, S. Aidinlis, T. Marquenie, A. Gurzawska, D. Tzovaras, Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2302–2314.

[20] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and information systems 33 (2012) 1–33.

[21] A. Bifet, R. Gavalda, Learning from time-changing data with adaptive windowing, in: Proceedings of the 2007 SIAM international conference on data mining, SIAM, 2007, pp. 443–448.