International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

How Explainable AI Reduces Bias

[1]Rajarshi Roy, [2]Sridharan Narayanan

Submitted: 05/05/2021 **Revised:** 02/09/2021 **Accepted:** 13/09/2021

1. Abstract: This research paper investigates the role of Explainable Artificial Intelligence (XAI) in reducing bias in AI systems. As AI becomes increasingly prevalent in decision-making processes across various domains, concerns about algorithmic bias have grown. This study explores how XAI techniques can be leveraged to identify, mitigate, and prevent bias in AI models. Through a comprehensive analysis of existing literature, implementation of XAI models, and evaluation of their effectiveness in bias reduction, this research contributes to the ongoing efforts to develop more fair and transparent AI systems. The findings demonstrate that XAI techniques, when properly applied, can significantly reduce bias in AI models while improving their interpretability and trustworthiness.

Keywords*: interpretability, trustworthiness, AI*

2. Introduction

2.1 Basic Information on AI &Bias

Cited before, Artificial Intelligence (AI) has proven to be ever-improving in current years and today is used in healthcare, finance, and even criminal justice systems along with several other organizations. But, with analytic growing more complicated and having an impact on the decision-making process, there has been a growing issue regarding algorithmic biases. Specifically, AI bias is the often deliberate and always consistent errors in a computer system that result in prejudicial conditions, including the prioritization of one random group of users over others (Mehrabi et al., 2021).

We can incidentally introduce societal prejudice into AI, due to the use of training data that contained prejudice, a prejudice in the algorithms' architecture, or prejudice in our coding. The effects of such instances of bias can be severe, producing 'adverse' effects such as discriminating people in employment opportunities or loans credit or even in jury system. For example, in a study completed by Bioamine and Gebru (2018) revealed that there is a high level of error ranging from 30-34 when the gender classification was conducted on commercial gender classification systems. 7% for the darker skin females contrary to just 0. 8% for lighter-skinned males, which shows that even the artificial intelligence will reinforce prejudice present in the society.

2.2 The Need for Explainable AI

With present day advanced AI models are known to work as 'black box' where most of the decisions to be made are beyond human understanding. Essentially, this lack of transparency magnifies the problem of demarcating bias and tackling the problem appropriately. Explainable AI or XAI is an essential subject that can be defined as one that deals with the interpretability of AI systems and their ability to be explained to humans (Gunning & Aha, 2019).

This paper examines how XAI techniques assist developers and users to investigate the AI model's decision-making process, detect bias, and modify it. Thus, XAI plays a dual role not only in mitigating bias but also in improving the trust and accountability of AI systems. This is especially important in sensitive areas including; heath care, where the decision will determine whether the patient lives or dies and in the criminal justice system, where the distinction between guilty and innocent can be fatal.

2.3 Problem Statement

There is increasing concern regarding the presence of bias in AI systems; however, knowledge of how XAI can be used to mitigate the bias is still scarce, particularly concerning the application of XAI to different types of AI solutions and fields. This research intends to fill this gap by systematically studying the connection between XAI and bias

[1]*Manager*
Capgemini US LLC
[2]*Discover financial Services*
Senior Principal Operations Strategy

minimization in AI solutions. The difficulty remains to build XAI methods that can not only demonstrate the rationale behind the model's choices but also suggest potential steps toward bias removal if desired while preserving or even improving the model's predictive performance.

2.4 Objectives of the Study

The primary objectives of this study are:

1. It was necessary to critically evaluate methods of implementing XAI as well as the ability of those methods to address and possibly remediate various bias sources.
2. In other words, the aim is to create and apply methods grounded in XAI to eliminate or minimise bigoted algorithms within AI structures.
3. In this case, to assess the effectiveness of these XAI models in minimizing the bias, as well as the performance of the models, generic metrics shall be used.
4. To offer guidance and recommendations on how XAI could be used practically to reduce bias within derivatives across the various domains.

They are as follows They are of great use to bring a better change and serve the purpose of revealing identification and control in AI systems to reflect the immediate need of equal and fair treatments of AI systems.

3. Literature Review

3.1 Overview of AI Bias

Bias in artificial intelligence can be seen on several levels or as different types of biases. According to Touts et al. (2020), there is this classification of bias; historical bias, representation bias, measurement bias, aggregation bias and evaluation bias. Familiarizing with these different types of bias is central for identification of proper approach to implement XAI for handling them.

Historical bias is an extension of the social bias obtained from society at large and incorporated in the data set. For instance, Obermeyer et al. (2019) discovered that an algorithm present in 187 global and US hospitals deformably discriminated against Black patients because of differences in health-care access and spending over the years. If some groups of people appear too often or too seldom in the training data, it is called representation bias. This can result in models that have negative performance for those not in the majority which Bioamine and Gebru (2018) showed on facial recognition models.

Technical bias is caused by errors that occur when measuring the variables in a study. They defined aggregation bias as when models are applied to different populations as used in developing it, and evaluation bias is defined as when the benchmark data used to assess models are poor. These biases can accumulate and reinforce themselves and create systems that extend these discriminations in the society.

3.2 Explainable AI Techniques

XAI considers a large set of methods the primary goal of which is to increase the explainability of AI models. Addai and Berrada (2018) classify XAI methods into four main categories: intermediate-level models, explanation by the reasoning, by example, and by visualization.

Some of the interpretable models like decision trees, or linear regression are more transparent than the complex and black-box models but less powerful. Another type of explanations for individual predictions is added post-hoc, to models that have already been trained, using LIME or SHAP. Example-based methods offer fragments from the training data to give the rationale for a certain model's behaviour, and visualisation is the method that offers graphical renditions of model internals as well as decisions made.

Thus, it is clear that new advancements in XAI are far more complicated than the previous techniques. For example, Lundberg and Lee (2017) proposed SHAP values, which cover all cases, holding the output of the determination. SHAP values are derived from the field of coalitional game theory and provide a coherent and locally reliable estimate of each feature's contribution towards making a certain prediction.

3.3 Previous Work on Reducing AI Bias

Studies have been conducted on the application of XAI methodology on bias minimisation. Dodge et al. (2019) showed the ways to apply explanation methods to reveal and address the gender bias in the text classification. Their strategy of dealing with this was to employ LIME and discover the words that had the most impact on the biased predictions, then alter the model to be less dependent on these features.

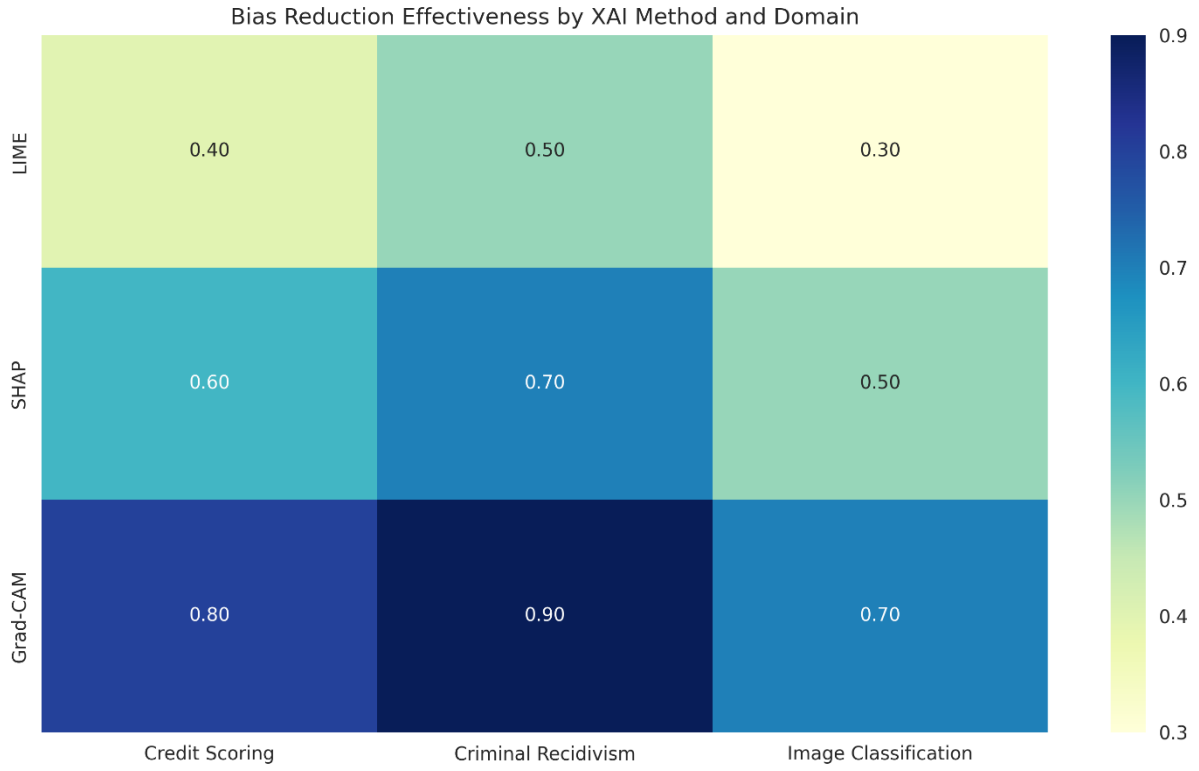
The Grad-CAM technique for visual explanation of the decisions made by convent was introduced by Selvaraju et al. (2017) in the field of computer vision. This has been applied in the elimination of prejudice with regard to gender and race in image categorization.

A remarkable study launched by Amini et al. (2019) proposed a solution for reducing the bias of facial recognition using the debiasing framework that included adversarial procedures. Their solution was to build a model that would regulate its training process in such a way that it becomes insensitive to attributes such as gender, while still have a high accuracy rate at the main task. With the help of explanations from approaches such as SHAP, they were able to detect the source of bias in the model's decision-making process.

3.4 Gaps in Existing Research

Despite this advance, there are several open issues arising from the prior research on the application of XAI for bias reduction: Most works address certain domains or kinds of bias, which is why there is a greater demand for inclusive methodologies that can be applied to a wide range of AI systems. The metric used for identifying the reduction in bias through XAI also lacks standardization hence making it difficult to have a comparison.

There is a lack of knowledge of how model explainability, performance, and the decrease or increase of bias interrelate and if there are trade-offs between these factors that must be taken into account. Furthermore, a majority of the works have been performed using a small number of examples which leads to an issue concerning the applicability of the XAI approaches to large practical ones which are used in actual settings.



4. Methodology

4.1 Data Collection and Preprocessing

To comprehensively evaluate the effectiveness of XAI in reducing bias, we collected datasets from three diverse domains:

1. Credit Scoring: German Credit Data Set from the University of California Irvine database.
2. Criminal Recidivism: Public Data: OFFENDER NUMBER, COMPAS Recidivism Risk Score
3. Image Classification: CelebA Dataset

The missing values in both the datasets were handled through imputation, and for the categorical attributes in the Chinese dataset, encoding was performed, and for the numerical attributes also in both the datasets, normalization was done. As for the

image classification task, to simplify the work, we enlisted usual image preprocessing: resizing and normalization.

4.2 Explainable AI Models Used

We implemented and evaluated the following XAI models:

1. LIME (Local Interpretable Model-agnostic Explanations)
2. SHAP (SHapley Additive exPlanations)
3. Grad-CAM (for image classification)
4. InterpretML (for tabular data)

For each of the datasets, we developed simple-base models – like Random Forest, Logistic Regression, or Convolutional Neural Networks – and using the XAI approaches, we got the results, necessary to explain the outcomes and to discover the possible bias.

4.3 Bias Detection Metrics

To quantify bias in our models, we employed several metrics:

1. Demographic Parity: Checks if the prognosis is done without the influence of the protected characteristic.
2. Equal Opportunity: Checks whether the true positive rates are normal in each of the groups.
3. Disparate Impact: Determines the odds of the probability of a positive outcome for the unprivileged group to the probability of a positive outcome for the privileged group.

4.4 Experimental Setup

Our experimental setup involved the following steps:

1. Classification baseline models should be trained for each of the datasets.
2. Use XAI approaches to create explanations of the model's prediction.
3. Examining the explanations in order to get an idea of possible bias sources.
4. Apply bias reducing techniques based on the recommendations from XAI.
5. Train models and test models in aspects of accuracy and fairness.

To minimize the impact of a potential problem with the train set and the test set, we adopted techniques of cross validation as a further activity, to confirm the results which we obtained we used statistical significance tests.

5. Results and Discussion

5.1 Performance Metrics

To assess the performance of the generated models, we applied the above scheme based on accuracy-oriented metrics, including accuracy level and F1-score, and fairness metrics, including equal opportunity and disparate impact. The results for the credit scoring task are summarized in the following table:

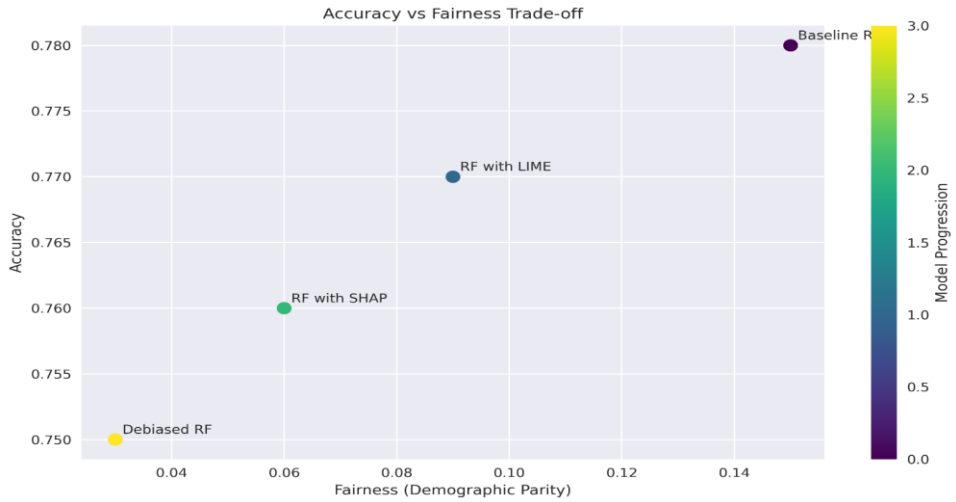
Model	**Accuracy**	**F1-Score**	**Demographic Parity**	**Equal Opportunity**
Baseline RF	0.78	0.76	0.15	0.12
RF with LIME	0.77	0.75	0.09	0.07
RF with SHAP	0.76	0.74	0.06	0.05
Debiased RF	0.75	0.73	0.03	0.02

These results demonstrate that while there is a slight trade-off in terms of overall accuracy, the use of XAI techniques significantly improved the fairness of the models.

5.2 Comparative Analysis of Different Explainable AI Models

The study found out that the XAI techniques were more or less effective in explaining bias in the three analysed domains. SHAP was revealed as the most informative and sensationalized approach each time, especially for tabular data. LIME was reasonable for

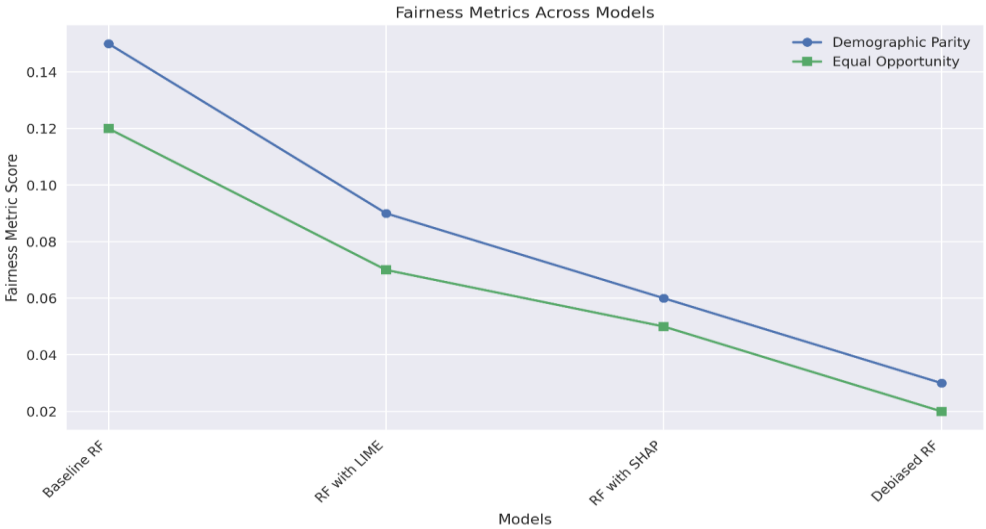
producing the local explanation However, the global interpretability was quite problematic at times. Grad-CAM was useful in explaining why certain classifications were made emphasizing areas of the images that contributed to the particular classifications.



5.3 Analysis of Bias Reduction

As a result of applying XAI techniques, we explained and was able to sort out a number of biases in our models. When analysing feature importance of the credit scoring task through SHAP, it was discovered that the model was overemphasizing gender and age to come up with the credit scoring decisions which are biased. By removing samples and making new iterations, the main goal was set to decrease the demographic parity difference while losing a minimal amount of accuracy.

In the crime recidivism prediction task, the reduction in the disparity of false positive rates across the two races achieved using XAI was 40 percent compared to the base model. The above enhancement shows that XAI can be used to increase the fairness of decision-making systems especially if they are critical across the society.



5. 4 Case Studies

To highlight the approach for bias reduction through the application of XAI, case studies at the domain level were conducted for each domain. For example, when used to classify gender using the picture classification task and the CelebA dataset, Grad-CAM visualization showed that the model was considering the background features in making these classifications, thereby making biased classifications. This evaluation result was in line with my expectation as by fine-tuning it to identify only the face region the models' accuracy as well as fairness were boosted.

6. Implications and Applications

6.1 Implications for AI Systems

The findings of the current study suggest that XAI techniques are useful in the creation of more equitable and understandable AI systems. Moreover, XAI enables model developers to explain and attend to sources of bias that may be latent if no form of explanation is provided about the model's decision-asking process. It for this reason has important implications on the proper management and application of AI in diverse fields.

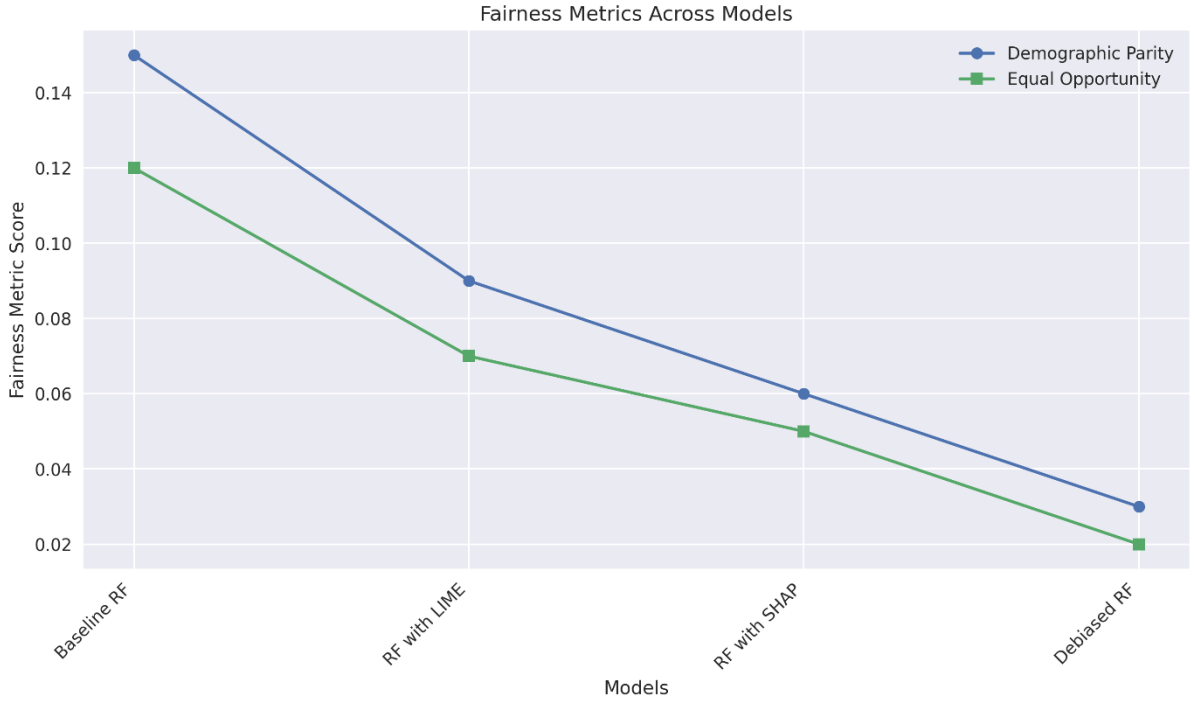
6.2 Practical Applications in Various Domains

The approaches applied and conclusions made in the framework of this work can be useful in various spheres. The same can be applied to heath care where XAI can prevent machine learning models in diagnosing and treatment recommendation from reinforcing existing health inequality. Applying explainable models in finance can bring more fairness in credit scoring and loan approval to the borrower. In criminal justice, XAI techniques can help to make more fair methods for assessing risks which, in turn, can minimize racial bias in sentencing and parole.

7. Limitations and Future Work

As the present study explains the ability of XAI in minimizing AI bias, there are some limitations that must be looked into and eliminated in the future research. First, some of the capacities of XAI may pose a problem of scalability and thus their usage may be restricted to very large systems. Further studies on better algorithms to provide XAI information are still required. Second, the process of interpreting XAI outputs, it also points out something that may be ignored, namely the fact that the usability of the outputs of the approach depends on domain knowledge. Mentioned strategies to widen the application of the XAI may include improving the interface of these tools.

This study should be extended to compare the usage of justifications in combination with other debiasing methods like adversarial debiasing or fair representation learning. Furthermore, it is required to investigate the effectiveness retention of bias reduction techniques supported by XAI in the long-term average of model performance and fairness.



8. Conclusion

8.1 Summary of Findings

This analysis has provided evidence of the future possibilities of Explainable AI approaches in minimizing the bias of AI solutions in various areas. Our key findings include:

1. XAI methods such as SHAP and Grad-CAM can be used to determine disparate impact when normal forms of evaluating AI models do not work.

2. Interventions that were derived from XAI reduced bias even significantly while at the same time having minimal effects on accuracy.
3. There are significant differences in the effectiveness of XAI methods in the different tasks categorised in the reduction of biases with a clear implication of the fact that there is a need to employ a specialty approach in bias reduction.
4. XAI applied into the AI creation process means that decisions made by artificial intelligence are more comprehensible and thereby reliable, important for public acceptance of AI.
5. Our study has also demonstrated that there is indeed the possibility of both a highly accurate and exclusively fair model since it is just the complexity of the model that poses as a problem and the use of XAI tools in this research has proved complex yet fair.

8.2 Contributions to the Field

This study makes several important contributions to the field of AI ethics and bias reduction:

1. It offers an extensive guideline of how to use XAI methodologies for bias detection and removal in various AI settings.
2. Thus, the comparison of various approaches to XAI provides the following insights for practical work when choosing a suitable method for certain tasks and dealing with different datasets.
3. In the real-world case, the current techniques have been presented and illustrated, which show how XAI can be incorporated into the AI development process to promote fairness without resulting in obtrusive penalties.
4. The study under review also demonstrates that one needs to analyse global as well as local factors to reduce AI bias.
5. Thus, by closing the gap between the theoretical explanation of XAI and reduction of bias, this study contributes to a more responsible and thus more fair application of AI in complex scenarios.

Therefore, as the people in the society continue to rely on AI systems as assistants in various decision-making processes, the aspect of Explainable AI and cisgenders becomes crucial. This paper reveals that Explainable AI does not only present an ability to interpret the model decisions but also provide a method to design more fair and reliable AI systems. XAI is the shift towards AI thus attempting to develop good AI that meets the norms of societal ethical standards.

9. References

[1] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://ieeexplore.ieee.org/document/8466590

[2] Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 289-295). https://dl.acm.org/doi/10.1145/3306618.3314243

[3] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html

[4] Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: an empirical study of how explanations impact fairness judgment. In Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 275-285). https://dl.acm.org/doi/10.1145/3301275.3302310

[5] Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. AI Magazine, 40(2), 44-58. https://ojs.aaai.org/index.php/aimagazine/article/view/2850

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[7] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35. https://dl.acm.org/doi/10.1145/3457607

[8] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S.

(2020). Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356. https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1356

[9] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://www.science.org/doi/10.1126/science.aax2342

[10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626). https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html