Inclusive AI Survey

Honey Jigarkumar Patel and Diana Inkpen
School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, ON, Canada
{hpate142, diana.inkpen}@uottawa.ca

***Abstract*—The usage of AI technology is expanding across a range of domains that could have an impact on human existence in all its manifestations, from educating children to healthcare. Despite various benefits, there is a good probability that AI might be viewed as a possible threat to social ideals because it can directly impact the principles of society like justice, inclusion, loyalty, trust, etc. As a result, the industry today is now seeking ways to build inclusive AI that is fair and equitable for all. We aim to research inclusive AI and comprehend why and how AI becomes biased as well as test its fairness, reliability, and trustworthiness. This paper will concentrate on addressing several major issues concerning inclusive AI, such as identifying potential causes of bias in AI systems, analyzing techniques for transitioning current biased AI into inclusive AI, assessing its advantages, and exploring some real-world examples. It will also delve into fundamental questions like how models and datasets get biased as well as how to measure and mitigate bias. Additionally, it will focus on exploring the existing tools like AIF360, LIME, SHAP, and TCAV, which assist in developing inclusive AI. This will involve researching the tools in terms of how they achieve inclusiveness in AI, what are the key aspects considered for developing inclusive AI, how they measure bias, etc.**

Keywords— Inclusiveness, Fairness, Trustworthiness, AI, Bias, Explainability, Interpretability.

I. INTRODUCTION

The accelerating power of Artificial Intelligence has led to a growing trend toward algorithmic decision-making in many facets of our lives. This trend has caused an increase in concerns about the ethical and legal issues that the delicate data-driven systems raise. Artificial Intelligence has myriads of merits like its capacity to make predictions, work continuously with great efficiency, and quick decision-making using insights from processed data, etc. Notwithstanding these rapid developments in artificial intelligence, there is a strong likelihood of AI hindering the quality, scope, and values of human existence.

For instance, in 2018, there was a controversy around Amazon's AI-driven recruiting tool being biased against women [28]. The tool was trained using resumes mainly submitted by men over a ten-year period. This resulted in system penalizing applications that featured terms like "woman" and "female". Another such example is, Microsoft's AI chatbot Tay, which was introduced to Twitter in 2016. The technology was built to interact with users' tweets and learn from them. Yet after only 24 hours, the chatbot started making terrible remarks like being racist, sexist, and other things. Tay's actions were a result of the dataset it was trained on containing offensive tweets from Twitter users. Thus, existing AI has concerns such as the potential for biased decision-making and a lack of transparency in the decision-making process.

AI systems can be interpreted as a potential dangers to human existence and its principles as well as social values like fairness and inclusivity [1]. Moreover, AI-based technologies are typically hidden from the public [2], difficult for the general public to comprehend [3], and even difficult for their developers to explain. Given these restrictions, the question is how one can make sure that existing AI enhances rather than degrades human life quality by operating contrary to social values.

In order to alleviate these concerns, there is an increasing emphasis on inclusive AI, which attempts to ensure that AI systems are fair and equitable for everyone, regardless of their ethnicity, gender, or any other personal attribute. Our objective is to research on the inclusivity of AI and evaluate its fairness, dependability, and integrity because the current AI is prejudiced in many situations. One of the reasons for existing AI being biased is many times datasets fed to AI either tend to have a minimal amount of data about a particular group or do not have data related to a specific group. Another reason can be that the datasets that AI is based on are not diverse enough as usually AI inadvertently reflects the ideologies of those who are developing it. Also, these datasets can either be unrepresentative or distorted which can cause biases. The development of inclusive AI assures that emerging technology improves the quality of human existence and upholds social values like justice, inclusiveness, loyalty, and trust.

Current research suggests that in order to tackle the challenges faced by existing AI and develop inclusive AI, a multifaceted strategy is needed to overcome biases in data, design, and algorithms, as well as to ensure diverse representation in the designs and implementation of AI systems. One of the crucial steps is to make sure that the training data is diverse and reflective of the populations the AI system will be employed with. This may entail gathering information from many sources and checking that it is devoid of biases that could generate unfair or biased results. Another crucial component is to make sure that AI algorithms are developed and tested with fairness and inclusivity in mind, as well as that they are constantly inspected to discover and mitigate any biases that may develop over time. Furthermore, it's critical to make sure that the development and deployment of AI systems are done by teams that are diverse and reflective of the target demographics. This can involve activities like diverse hiring, inclusive design approaches, and continuing training and education to identify biases and presumptions.

In addition, this paper will also focus on exploring the existing inclusive AI-building tools in the context of:

- How do these tools achieve inclusivity in AI?
- What are the main aspects they are taking into account when developing inclusive AI?
- How do they measure bias?
- Is it usually the datasets that are biased?

Various methodologies have been developed to promote inclusive AI, each with a unique strategy. This research paper intends to evaluate several such existing tools, namely AI Fairness 360 - Preprocessing, Fairness 360 - Postprocessing, LIME , SHAP, TCAV. By evaluating each tool, this paper

will offer insights into the most efficient approaches to building inclusive AI. Each tool will have specific datasets and the paper will include studies on different datasets across different tools. The overall objective is to provide a thorough understanding of the existing cutting-edge tools and techniques for developing inclusive AI, which can serve as a roadmap for further research in this area.

The remainder of the paper is structured as follows. Section II summarizes the related work in the same domain. Followed section III, talks about why society needs inclusive AI. Section IV addresses potential research questions in the field followed by Section V which describes the methodologies used for exploring inclusive AI-building tools. Sections VI, VII, VIII, IX, and X give information on those tools. Discussion about the tools is done in section XI followed by future scope and lessons learned in Sections XII and XIII, respectively. Lastly, Section XIV provides the conclusion of our work.

II. RELATED WORK

Inclusive AI is emerging as a rapidly growing field to build an AI that is unbiased, fair, and accessible to all. Numerous studies have been undertaken recently to address both the obstacles and opportunities concerning developing inclusive systems.

One of the most significant challenges with building inclusive AI is data bias. According to studies, biased data can cause AI systems to reinforce and elevate already-existing societal biases. Researchers have devised a number of strategies to deal with this problem, such as data augmentation techniques, which add synthetic data to the training set to boost its diversity, and adversarial training, which includes training the AI system to detect and correct for biases in the data, modification to existing model assessment methods like cross-validation. The research paper [4] addresses issues about learning algorithms' reliability due to algorithms biases brought on by human-imposed bias and variance. In order to solve algorithm fairness, bias, and variance in artificial intelligence, the paper suggests a framework of data inclusiveness, participation, and reciprocity. Cross-validation is likely the simplest and most popular model assessment method. This study proposes a modification of the existing cross-validation approach to fill the vacuum in the literature regarding algorithm bias, and data inclusiveness, and improve algorithm validation, reliability, and bias assessment.

Another perspective in building inclusive AI is to involve varied stakeholders from underrepresented communities, as well as specialists in disciplines like ethics, law, and social science. By this AI may be developed to accommodate a wider spectrum of users while avoiding accidental biases. The study [5] focuses on the need of optimizing algorithmic decision-making in the recruiting process to favor the inclusiveness and diversity of the staff. The general belief is that algorithmic decision-making will result in less biased assessments, however, this is not always the case. The authors suggest approaches for data scientists to improve inclusivity in algorithm design, including raising awareness of the significance of inclusiveness, facilitating collaboration between data scientists and industry professionals, by rebalancing datasets, and assigning costs to false alarms and misses.

Numerous studies have also concentrated on building tools and frameworks following a comprehensive evaluation of fairness metrics while employing AI models. The authors in [6] discuss the importance of fairness in predicting unemployment status using social media data (i.e. when applying AI-based insights). The authors propose a methodology that predicts unemployment status by training a machine learning classifier based on Facebook likes and then establishing an adaptive threshold criterion, this assures fairness in the prediction process. The paper also goes over the drawbacks of the "fairness through unawareness" strategy, which simply ignores protected attributes like age and gender.

With a focus on social movements and activism, the article [7] analyses the ethical ramifications of algorithms and artificial intelligence in the digital age. The author discusses a case study in which facial recognition software incorrectly identified a Muslim activist, leading to false accusations that she was a terrorist. The study highlights concerns about the lack of comprehension among algorithm developers and the requirement for informed consent in data harvesting. Finally, the paper concludes that while algorithms are not intrinsically unethical, their ethical consequences do depend on how they are used.

The research work [8] confronts the growing use of autonomous algorithms in organizational decision-making processes and argues that human-algorithm collaboration can be enhanced by giving humans the role of system leader because it promotes a sense of fairness and elevates system confidence. By conducting many trials with participants, the article also shows the possible failure of the developing business model and reveals that when humans take the lead, they feel more in control, accept more responsibility, and have greater faith in the algorithm. Finally, the paper concludes that in order to obtain a level of cooperation that is acceptable, it is cardinal to consider human preferences and desires for fairness.

The researchers in [9] outline a strategy for Europe to create a sustainable, inclusive approach to artificial intelligence (AI) that is human rights-based. The authors contend that present AI deployment and development frequently disregard human rights, which could result in harm and bias. The proposed plan has three pillars: Building AI systems that are human-centered and respect human rights, promoting AI systems that benefit society, and ensuring democratic control and accountability of AI systems. In order to accomplish these objectives, the article offers detailed recommendations, such as encouraging diverse representation in AI development and deployment and creating human rights and social welfare-focused regulatory frameworks for AI.

The work [10] emphasizes the difficulties that non-programmers face when attempting to utilize AI for social good and suggests a low-code, inclusive, probabilistic AI-based BN strategy to democratize AI use. The study includes

three examples of how to apply this methodology to examine elements that affect GCED, including improving malnutrition, expanding financial inclusion, and global sustainable development. The paper concludes that this approach enables the building of predictive simulations, extensive customization of variables and AI algorithms, and can stimulate more inclusive AI-assisted data exploration, and produce more human-centric policy-relevant insights.

Numerous stakeholder groups are debating the use of opaque artificial intelligence systems for important decisions. While using these opaque algorithms, it is challenging for people to figure out the reasoning behind a given decision. Hence, the creation of "explainable AI" systems has become a top priority to accelerate societal acceptance, build trust, and eliminate discrimination. To address the above issues, paper [11] suggests a multifaceted framework for explainable AI called RVS. RVS facilitates the evaluation of explanations by a wider range of stakeholders, promotes a more interdisciplinary approach, and allows the perspectives of those impacted by AI systems to influence that evaluation.

The in [12] discusses how Nordic cultural values affect national strategy discussions about implementing and utilizing AI systems. According to the report, the principles of trust, transparency, and openness are essential for ensuring public confidence in AI technologies and encouraging their development and deployment in society. In order to enhance access for individuals for whom these services are designed, the study emphasizes the necessity of public involvement and AI education in policy creation.

This study [13] investigates the impact of inclusiveness on decision-making costs and implementation. It states that inclusiveness brings together every stakeholder and stimulates solidarity, social cohesion, and unity. Consensus, open and all-inclusive democratic organizations, decision-making processes, complying with expert advice, harmony and goodwill among employees, and recognizing and empowering employees, all contribute to inclusiveness and successful decision-making. The article also explores the correlation between women's development in decision-making positions and inclusiveness and suggests using gender quotas and women's empowerment to increase women's inclusion in decision-making.

III. NEED FOR INCLUSIVE AI

The following section will delve into the exploration of the current state of AI, highlighting the potential causes of its biases, which will ultimately conclude the need for an inclusive approach (inclusive AI). The discussion will also move around the transition from current biased AI to inclusive AI, the working of inclusive AI, and the benefits of such an approach. Moreover, the section will address the following fundamental questions such as:

- How dataset gets biased?
- How the model gets biased?
- Ways to measure bias?
- How to mitigate bias?

A. Current State of AI

In recent years, existing AI has made remarkable advancements and has demonstrated an extraordinary capacity in performing challenging tasks including image identification, natural language processing, object recognition, decision-making, and robotics. However, there is still a long way to go to the best of AI's ability. Making AI more explainable, interpretable, and transparent are some of the most difficult tasks in the industry, achieving this will build trust in AI systems, promote confidence and assure their responsible use.

Additional areas where existing AI can improve are ensuring fairness and mitigating bias [17]. As previously stated, biases can be inherited by AI systems from the data they are trained on, leading to unfair or discriminatory conclusions. Hence, it is necessary to create more sophisticated methods to identify and eliminate bias in AI systems, including expanding datasets, creating algorithms that are less sensitive to particular demographic traits, and putting ethical standards into place for the creation and use of AI.

Finally, it is still difficult for current AI to develop human intelligence's key characteristics of creativity, flexibility, and common-sense reasoning. More research and funding are needed to create such advanced AI systems. Thus, to overcome these issues, it is essential to develop an inclusive AI that will be beneficial to society.

B. Why current AI is biased?

AI systems are developed by humans who by nature have inherent biases, beliefs, and preconceptions. Thus, AI ultimately represents the ideology of the person or team, or organization building it. Considering this, it can be said that AI is prone to get biased. The following are key reasons for existing AI being biased [17]:

- **Biased training data:** Large datasets, which may contain bias, are used to train AI systems. If the training data is biased, the AI system will learn to mimic those biases in its decision-making. For example, a facial recognition algorithm may not work well on people with darker skin tones if it was trained on a dataset with mostly Caucasian faces.
- **Lack of diversity in development teams:** Many times AI systems are built by homogeneous teams that lack diversity in terms of gender, race, color, etc. Such AI systems do not consider the experiences or needs of diverse or underrepresented groups. For instance, if the majority of the team designing AI algorithms is male, then it may not take into account the demands or needs of women. [16]
- **Algorithm design bias:** AI algorithms can also demonstrate bias as a result of the algorithm's design, which may duplicate human biases. To illustrate, if an algorithm is created to make decisions based on historical data, it may perpetuate biases that were inherent in the data.
- **Lack of interpretability and transparency:** Some AI systems are "black boxes", making it difficult to

understand how they reach their decisions. This lack of interpretability makes it challenging to recognize and mitigate bias in the system. [16]

- **Lack of fairness considerations:** AI systems can be unfair when they make decisions without taking into account how those decisions will affect other groups. For example, a hiring AI system may unfairly favor or exclude particular demographics. [16]
- **Human influence:** Since AI systems are created and maintained by humans, who have the potential to unintentionally or deliberately incorporate bias. As an example, a developer could decide to give priority to specific features or results that align with their personal biases or preferences. [16]
- **Limited feedback and accountability:** Lack of feedback and accountability in AI systems may continue to provide biased results undetected if there is no method to measure or evaluate its performance.
- **Lack of regulation:** Currently there are no rules and regulations governing AI, which can lead to biased decision-making as developers can prioritize profits over fairness and ethical considerations.

C. Transition from current biased AI to inclusive AI

Inclusive AI refers to the development and implementation of AI systems that are intended to be fair, accountable, and transparent. Inclusive AI is developed on the principles of diversity, equity, and inclusion, and it aims to identify and alleviate the biases and drawbacks of existing AI systems. In order to build inclusive AI from current biased AI, a combination of technological and social strategies is required. Some of such techniques are suggested below [17]:

- **Diverse data collection [18]:** Collecting diverse demographic and representative data that reflects the experiences and viewpoints of various groups is required in order to decrease bias in AI systems and transform it into inclusive AI. This entails incorporating data from a variety of sources, verifying that the data is unbiased, and accurately reflects the diversity of the community.
- **Fair and unbiased algorithms:** Another approach is creating algorithms that are specifically designed to be fair and unbiased in order to prevent bias. This can be accomplished using techniques like counterfactual analysis, which includes testing, how a system would have functioned if certain variables were different. Algorithms are thoroughly examined to make sure they don't promote discrimination or reproduce former biases. [15]
- **Include diverse stakeholders:** This includes involving individuals from diverse cultures, racial, backgrounds, and genders. This method can ensure that the AI system considers the requirements and perspectives of different groups. [18]
- **Transparency and interpretability:** Transparency and interpretability refer to the extent to which it is possible to understand how an AI system comes to its decisions. Using explanations of decisions made by an AI simplifies detecting and eliminating biases. [18]
- **Human oversight and feedback:** Human surveillance is necessary for inclusive AI systems to ensure that they are functioning as intended, and discover and correct any biases or constraints. This entails keeping an eye on the system's performance, obtaining user feedback, and making necessary adjustments to enhance the system. [15]
- **Regularly audit and review AI systems:** AI systems should be regularly inspected and evaluated to make sure they are not biased. This can aid in locating potential biases and help enhance the system's performance over time. [15]
- **Regulation and oversight:** These mechanisms will help ensuring that AI systems are developed and deployed in a responsible and ethical way. This may consist of organizing auditing procedures, developing ethical standards, and making sure there are penalties for noncompliance. [15]
- **Ethical standards and guidelines:** Inclusive AI is developed on the foundation of ethics, which considers the potential effects of AI systems on various communities and society at large. This entails considering AI systems' ethical implications at every level of development, from data gathering to deployment. [18]

All in all, by considering the above strategies and techniques, developers can promote converting current biased AI to more inclusive AI, which seeks to create technologies that are beneficial to all users and that encapsulate the values and aspirations of a diverse and inclusive society.

D. Real-time examples of transition to inclusive AI

AI is an extremely powerful technology holding the potential to revolutionize several facets of human life, yet having limitations of getting biased and unfair. AI has a scope for improvement in terms of being inclusive. Following are some real-world scenarios of transitioning existing biased AI to inclusive AI:

1. **Google's Cloud Vision API** - It is a computer vision platform that can identify and classify photos. In 2018, The API's initial edition had trouble with accurately recognizing people with darker skin tones. However, as Google became aware of this issue, it took action to fix it by gathering a wider range of training data and modifying its algorithms to assure more precise recognition across a range of skin tones. As a result, the AI system became more inclusive and effective across all demographics.
2. **ProPublica's Machine Bias** - In 2016, ProPublica's investigation discovered that the AI system used in the criminal justice system to predict which offenders were most likely to re-offend, was discriminative against individuals of color [19]. It was found that the predictive model was more likely to mistakenly indicate black defendants as having a higher risk of recidivism than white defendants with comparable criminal histories. The firm created an inclusive AI tool called "Machine Bias" that could identify and mitigate bias in predictive policing algorithms.
3. **IBM's Fairness 360 Kit -** IBM's Watson, a machine learning system, was discovered to be biased against individuals of color. To address this limitation, IBM developed the Fairness 360 Kit, an open-source toolkit

that gives developers the tools and resources to evaluate and reduce bias in text-based applications. The toolkit contains tutorials for creating more inclusive AI systems, as well as algorithms for identifying bias in datasets. [14]

4. **Microsoft's Zo -** As we have stated above, in 2018, Microsoft chatbot Tay, was found to be discriminative towards women and people of color. Rather than learning from its interactions with Twitter users, Tay became a tool for online harassment because of its inflammatory and provocative statements. After shutting down the chatbot, Microsoft later released Zo, an AI chatbot that is more inclusive and capable of having conversations that are more culturally sensitive.
5. **Procter & Gamble -** In the past, Procter & Gamble (P&G) has come under fire for several of its advertising campaigns that were prejudiced or supported negative stereotypes. For instance, a 2019 Gillette brand advertisement attracted controversy for its depiction of toxic masculinity and sparked a heated debate on social media. To overcome such issues, P&G turned to AI to create more inclusive advertising. They used computer vision to examine 10,000 images from their advertising campaigns to seek biases in the representation of gender, age, and ethnicity. Based on the data collected, they then modified their advertising strategies to better reflect the diversity of their clientele.

E. Benefits and Importance of Inclusive AI

Following benefits of inclusive AI denote how cardinal it is to ensure that presented AI is fair, inclusive, and unbiased:

- **Superior accuracy:** Accurate predictions and suggestions are made possible by inclusive AI by embracing a variety of data sets and perspectives.
- **Reduced bias:** Inclusive AI can outweigh the effects of bias and historical biases by ensuring that the algorithm is trained on representative and varied data sets.
- **Elevated fairness:** By ensuring that algorithms aren't unfairly punishing or favoring specific groups of people.
- **Effective Business Results:** Inclusive AI can help organizations reach a larger range of clients, promote customer loyalty, and boost sales.
- **Enhanced trust:** Inclusive AI can boost confidence in AI systems among users, stakeholders, and the general public by removing bias and stimulating fairness.
- **Improved user experience:** Ensuring that the algorithms are engineered to accommodate a diverse range of users and their unique demands.
- **Increased Innovation:** When AI is created with inclusivity in mind, it can produce more innovative and creative solutions and allow a wider variety of people to contribute to the building of AI systems.
- **Ethical considerations:** Ensuring that ethical considerations are incorporated into decision-making processes, leading to more responsible and fair results.
- **Enhanced Decision-making:** By generating more accurate and trustworthy data-driven insights that are based on a more diverse and inclusive data collection.
- **Beneficial Social Impact:** By tackling societal issues like injustice and discrimination, and encouraging diversity and inclusion.
- **Cost savings:** Inclusive AI can result in cost savings by lowering the requirement for manual intervention to correct biases or faults in AI systems.
- **Leverage over competitors:** Organisations that put inclusivity as a priority in their AI systems can do so by luring in a larger clientele and fostering a sense of brand loyalty.

IV. Analysis of Potential Research Questions

Let's now dig into some critical questions that are pivotal to address in order to gain a deeper understanding of major factors in AI systems which generates biased results and what techniques may be undertaken to enhance their inclusiveness.

- *RQ1: How dataset gets biased?*

Following are the various ways by which datasets can become biased or discriminatory:

- **Inconsistent labeling or classification (Labeling bias) -** when inaccurate or biased labels are applied to data points.
- **Underrepresentation or misrepresentation of certain groups (Sampling bias) -** when the sample used to produce the dataset is not representative of the population it is intended to represent. [20]
- **The presence of historical or societal biases in the data (Historical bias) -** when the dataset contains biases from the past, such as racial or gender discrimination. [20]
- **Incorrect frequency of data (Reporting bias) -** When the frequency of events, attributes, and/or outcomes recorded in a data set does not precisely represent their real-world frequency. [20]

For instance, if the dataset used for natural language processing contains language that is biased or discriminating towards particular groups, then the AI system trained on a dataset may reinforce biases against particular groups.

Examples of some biased datasets are as follows:

1. **"ImageNet" dataset -** criticized for its underrepresentation of people of color and the overrepresentation of certain elements.[21]
2. **"Quick, Draw!" dataset -** biased towards western concepts and objects [22].
3. **"Faces of the World" dataset -** biased towards lighter-skinned individuals.
4. **"MS COCO" dataset -** biased towards objects and scenarios, such as sports and outdoor activities [23].

- *RQ2: How model gets biased?*

Models or algorithms may become biased due to a variety of reasons. following are some key reasons for models being based:

- **Biased Training Data -** model is biased because it is trained biased or non-representative data. [17]
- **Limited Feature set -** When the model is trained on a small set of features, it may not be able to represent the complexity and diversity of the real-world data, resulting in biased predictions.
- **Algorithmic Design -** Algorithms used to train the model may have inherent biases. For instance, many algorithms may prioritize particular features, resulting in leading to biased predictions. [17]

- **Human Bias -** Either consciously or unintentionally, the humans involved in the model's development may have their own biases or beliefs. [16]
- **Lack of diversity in development teams -** Homogeneous teams that lack diversity in terms of gender, race, color, etc do not consider the demands of underrepresented or diverse groups. [16]

For more reasons, please refer to "B. Why current AI is biased?" part above in the 3rd section of the paper.

Examples of some biased models/algorithms are as follows:

1. **Google's AdSense algorithm** - was biased against women and minorities, due to which ads for higher-paying positions were more commonly displayed to men [24].
2. **Google Photos algorithm -** In 2015 Google Photos incorrectly classified African Americans as "gorillas" due to the algorithm being biased, as it was trained on a small number of photos in which darker-skinned people were not included [25].
3. **Amazon Rekognition algorithm -** it is Amazon's facial recognition software, which was updated in 2020, as the algorithm was discriminatory when identifying women and darker-skinned individuals [26].
4. **Google Translate algorithm -** It was criticized for utilizing gender-biased language and not offering appropriate translations for some languages [27].

- *RQ3: How to measure bias?*

Bias in AI systems is referred to as unfair favoritism or prejudice towards certain groups or individuals during development. This bias can take many different forms, including biased training data sets, biased decision-making algorithms, and biased human interpretations of the outcomes. This section enlists some of the ways to measure bias [29] out of numerous ways:

- **Statistical parity [30]:** It evaluates if the AI system produces equal results for various groups. For example, a facial recognition system is statistically biased if it is better at detecting light-skinned individuals than dark-skinned individuals.
- **Disparate Impact [35]:** It measures if an AI system disproportionately has more impact on one group than another. For instance, there is a disparate impact on women, if a recruiting algorithm disproportionately favors men over women.
- **Predictive parity:** It entails determining whether or not the predicted outcomes are similar across various groups (like men and women).
- **Equal Opportunity [30]:** This method assesses whether or not the true positive rate is similar between various groups (such as males and females) [30].
- **Treatment Equality:** This technique assesses whether or not the AI system treats various groups (e.g. males and women) fairly.
- **Conditional Independence:** This refers to determining whether or not the AI system relies on factors that are independent of the protected features (such as gender or ethnicity).
- **Counterfactual Fairness [35]:** This measures the effect of altering a particular individual attribute on the outcome predicted by the AI system. To illustrate, it is counterfactually biased if modifying a person's race influences the result of a loan approval algorithm.
- **Proxies:** Proxies are indirect indicators of bias that can be used to find connections between input data and model predictions. As an example, the zip code can be used as a proxy for race or socioeconomic status to highlight potential biases if a model predicts lower credit scores for residents of particular zip codes.
- **Metrics [14]:** Metrics are quantitative performance measurements that can be used to rate the fairness and accuracy of an AI system. To cite, indicators of a binary classification model's accuracy include false positive and false negative rates [30], while indicators of fairness include disparate impact and statistical parity.
- **Audits [29]:** Audits examine the input data and the decision-making process of an AI system, to find potential sources of bias. This may include reviewing the data collection procedure, algorithms to analyze it, and decision-making processes leading to the output.
- **Human evaluations [29]:** In this humans are questioned about whether an AI system is fair. This includes showing people examples of decisions made by the system and asking them to evaluate if the decisions are fair or biased.

AI system that is biased can provide discriminating results, maintaining social and economic disparities, and limiting opportunities for some groups. Hence it is important to identify and mitigate bias.

- *RQ4: What are potential strategies to mitigate bias in AI systems before, during, and after classification?*

As we saw in the above section it is crucial to mitigate bias in the AI system, following section will pen down some potential strategies to eliminate bias n AI systems, both before and after classification or even during classification.

- **Before classification:**
  1. **Diverse data collection and curation [35] -** Obtaining representative and diverse data, and meticulously curating it ensures that the dataset is balanced. For example, the ImageNet dataset was re-labeled to eliminate biased labels
  2. **Data augmentation [32] -** Using methods like flipping, rotating images, and adding noise to audio recordings, to artificially augment the size and diversity of training data.
  3. **Pre-processing techniques [31] -** Techniques such as feature scaling, normalization, and outlier elimination can help remove or reduce bias in the input data.
  4. **Features engineering [31] -** can help reduce bias and increase accuracy by selecting and engineering the most relevant and informative features.
  5. **Algorithmic fairness [29] -** Using techniques like pre-processing, in-processing, and post-processing procedures can help to reduce bias in AI systems.

- **During classification:**
  1. **Reject option classification [33] -** Giving the model the ability to reject a classification if it is unsure about

its decision can help to lower the number of false positives or false negatives.
2. **Calibration Methods [35] -** Calibration techniques like temperature scaling can be used to mitigate bias in AI models.
3. **Adversarial training [34] -** Using adversarial examples during model training could enhance robustness against bias.
4. **Fairness restrictions [34] -** Fairness constraints are added to optimize problems to make sure the model is trained to be fair.
5. **Model selection [35] -** selecting a model that performs better or is less biased in terms of fairness.

- **After classification:**
  1. **Bias-aware post-processing -** Modifying model outputs to minimize bias in the final predictions. (Source: Zhao et al., "Learning to De-bias").
  2. **Counterfactual analysis [35] -** It examines the impact of various inputs on the model's output to recognize and correct bias.
  3. **Bias identification and mitigation [35] -** Uses a variety of metrics like inequal opportunity and equalized odds, and modifying the model or the data can reduce bias post-classification.
  4. **Explainable AI [36] -** Using explainable AI, the cause of bias can be better understood and addressed.
  5. **Human-in-the-loop [36] -** Adding human feedback in the classification process to detect and rectify bias.

V. METHODOLOGY

To develop inclusive AI, numerous methodologies have emerged with distinct strategies. This section will focus on providing a comprehensive overview of the most efficient current state-of-the-art methods and tools for building inclusive AI. In particular, we intend to implement and evaluate five already-available tools in the market, namely AI Fairness 360 - Preprocessing, AI Fairness 360 - Postprocessing, LIME, SHAP, and TCAV, along with a sixth tool for user research. Each tool has its own unique datasets. The following sub-sections of this paper will feature investigating each tool on their respective datasets and address some fundamental questions. The selection of these tools was based on criteria like their ability to offer explainability, interpretability, and identify and mitigate bias in datasets and models. Let's delve closer and deeper into these factors.

1. Identify and Mitigate Bias in Datasets:

- Datasets may be biased if some groups are underrepresented or if the data collected or labeled, reinforces discrimination and assumptions [20]. It is crucial to recognize and eliminate bias in datasets in order to develop more inclusive AI.
- This can be achieved by ensuring that the data is representative of diverse populations [35] and by balancing the data using methods like data augmentation or oversampling [32].
- It is crucial to make sure that the data labeling is done objectively and impartially [35], and that the labeling process should also be examined for fairness and consistency.
- Based on this factor we are exploring and implementing the tool - **AI Fairness 360: Pre-processing.**
- For more details on identifying and mitigating bias in the dataset please refer to the section fourth, "RQ1: How dataset gets biased?", "RQ3: How to measure bias?", "RQ4: Mitigate bias before, during, and after classification"

2. Identify and Mitigate Bias in Models:

- The model can be biased itself, as it may learn and reinforce from biased training data, has a limited set of features [17], algorithms have inherited biases [17], lack of diversity in the development team [16], and human bias [16].
- This issue can be resolved by reviewing the model's performance across various demographic groups and diverse stakeholders, locating and addressing disparities, features engineering, algorithmic fairness, building ethical standards, regularly auditing, etc.[29]
- Additionally, the model can be trained to be more resistant to biased inputs using strategies like adversarial training [34] and calibration method[35].
- Using this aspect, we are exploring and implementing the tool - **AI Fairness 360: Post-processing.**
- For more details on identifying and mitigating bias in the models please refer to third, B. Why current AI is biased?, and section fourth, "RQ2: How model gets biased?", "RQ3: How to measure bias?", "RQ4: Mitigate bias before, during, and after classification"

3. Explainability and Interpretability:

- Explainability and interpretability refer to describing AI systems' capacity to provide considerable and transparent explanations about how it makes decisions. Explainability allows people to understand "why a specific decision was made" by an AI system [42], whereas interpretability enables people to understand "how the decision was made" [45].
- Explainability and interpretability increase transparency and accountability, which can reduce the risk of bias and discrimination.
- They are crucial for developing inclusive AI because by allowing stakeholders to understand how and why the model made decisions, hence giving them the ability to detect and address bias.
- For example, if an AI system is making discriminative decisions toward certain groups, then on interpreting (how) and explaining (why) such a system, users can find the root cause of bias and make necessary changes in the dataset or model to remove the bias and make it inclusive AI system.
- By employing methods like attention mapping and feature importance analysis, it is possible to gain knowledge of the model's decision-making process.
- Furthermore, explainability and interpretability can simulate confidence and trust in AI systems which are cardinal for their successful adoption as users will be able to understand how and why a decision is made rather than blindly accepting a model’s decision that is highly prone to get biased.

- Using these criteria, we are exploring and implementing three tools -**LIME, SHAP, and TCAV.**

The fundamental questions for exploring the tools

Additionally, as we stated above, the paper will also focus on exploring the above five inclusive AI-building tools with respect to the following questions, as per the relevance of the question to the tool:

1. How do these tools develop inclusive AI?
   - This will be addressed by exploring algorithms used in these tools and examining models or datasets using different evaluation metrics.
2. What primary factors are considered when creating inclusive AI?
   - This will focus on addressing core factors on which tools build inclusive AI, like, fairness, transparency, collaboration, explainability, interpretability, etc.
3. How are they measuring bias?
   - This will include exploring the evaluation metrics used by tools for measuring bias.
4. Are datasets usually biased?
   - Will check that basically, the datasets used in the tools were biased or not. (This question might not apply to all the tools)

VI. TOOL 1 - AI FAIRNESS 360: PRE-PROCESSING TECHNIQUE (MITIGATING BIAS IN DATASET)

○ *Introduction to AI Fairness 360:*

AI Fairness 360 (AIF 360) [37] is an open-source toolkit basically used to mitigate bias in Machine Learning models. So, once AIF 360 detects the bias in the dataset using various bias metrics, the bias mitigation algorithms are applied. The AIF360 toolbox includes a number of classes for identifying bias in datasets and models, both individually and in groups. In order to analyze a single dataset, one can use DatasetMetric and BinaryLabelDatasetMetric classes, while ClassificationMetric and SampleDistortionMetric classes can be used for comparing two datasets.

The toolkit has a wide range of alternatives for detecting bias in various contexts, with around 71 different bias detection metrics. This means that for different situations, different metrics can be used.

Now talking about the three types of bias mitigation techniques that alter the training data, or predictions to improve fairness metrics are pre-processing, in-processing, and post-processing. And there are bias mitigation algorithms that come into these categories.

○ *Introduction to Preprocessing technique:*

Tweaking the training dataset before one applies the machine learning model is known as pre-processing. So, to experiment, we have used pre-processing technique. This method includes 4 algorithms.

1. **Reweighing -** It assigns weights to all groups and label combinations.
2. **Optimized preprocessing -** Here probabilistic transformation is done to edit labels and features.
3. **Learning Fair Representations** - It hides information about protected attributes, by finding latent representation.
4. **Disparate Impact Remover** - It changes the feature values to make it fairer.

○ *What are the major factors that the tool (AIF360) considered for developing inclusive AI?*

The major factors that are been considered for developing inclusive AI are fairness, transparency, usability, collaboration, and evaluation.

○ *Are datasets usually biased?*

In this case, we can say that it is not always true that only datasets are biased. Because there can be bias in the algorithm we choose, in feature selection, or in the model's architecture. However, the most common reason for biased models is biased datasets.

Let us consider a situation that describes how bias can occur due to mistakes in feature selection. So, if we want to measure the job performance of the employees and we use their education qualification as the feature in training the model, then it might give biased decisions because there is no relation between job performance and the last degree pursued.

○ *Dataset used:*

We have used the German Credit Dataset [39] which is used for credit risk analysis. It includes information about credit applicants such as their credit history, or some kind of financial details. This dataset has 1000 total observations and 20 features like age, job status, sex, credit history, and reason for the loan. The binary target variable suggests if the applicant is a favorable or negative credit risk for creditors.

Dataset pre-processing:

The first step in pre-processing the dataset was converting the categorical variables into numerical values. This was done using one-hot encoding. For instance, the 'checking account' feature has values such as A11, A12, A13, and A14 which were converted to 0 and 1 for each category. The next step is scaling the numerical values, followed by handling the missing values and lastly label encoding the target variable. Here, 0 means a bad loan and 1 indicates a good one. Here, 'age' is a protected attribute which means it can divide a population into groups that are equally benefitted. The privileged group means the one that has a traditional systematic benefit. So in this case age greater than or equal to 25 signifies a privileged group, else it is an unprivileged group.

Dataset bias detection metrics: (How is the tool measuring bias?)

This section aims on exploring what are the evaluation metrics that AIF360 is using to measure bias. AIF360 tool provides around 71 different metrics for bias detection out of which we have used the following three metrics for identifying bias in the original dataset:

- **Disparity Impact (DI):** It is a group fairness measure. It is the ratio of favorable outcomes of the unprivileged group over that of the privileged group. Here, as the

value goes farther from 1, then one can say there is high discrimination. DI of 1 means no discrimination.

- **Statistical Parity Difference (SPD):** This is also a group fairness measure. SPD is the difference between protected and unprotected groups' positive outcomes. A value closer to 0 means there is no discrimination, otherwise, it increases.
- **Consistency:** This is an individual fairness metric. It says that a similar group of individuals should get the same results. It assesses the model's performance when similar individuals differ only in one protected characteristic. If the value is 1, then there is a perfect consistency. A value closer to 0 represents less consistency.

Dataset bias mitigation algorithm: (How does the tool develop inclusive AI?)

There are several bias mitigation algorithms in pre-processing algorithms from which we have used Reweighing algorithm [38]. To solve the issue of class imbalance, in this algorithm, we make sure that the minority class gets greater weight while the majority class gets a lower weight by re-weighting the samples.

So, the weight for each sample's class label is inversely proportional to the frequency. Meaning the minority class is given a larger weight than the dominant one. The weights are taken into consideration to modify the loss function so the higher weight could contribute more to the total loss. This way classifier can detect minority class instances more accurately (i.e., minority ones have more weight).

The following lines represent the Python code for the Reweighing algorithm:

```
RW = Reweighing (
        unprivileged_groups=unprivileged_grous,
        privileged_groups=privileged_groups)

dataset_transf = RW.fit_transform( dataset_org )
```

Here, unprivileged_groups are the ones with age less than 25, else it is privileged_groups, dataset_org is the original dataset, RW is used to fit it in the original dataset.

Transformed dataset bias evaluation:

After applying reweighing algorithm on the original dataset, the transformed dataset showed the following results:

Table 1: Results of evaluation on Original Dataset

Disparity Impact (DI)	0.821
Statistical Parity Difference (SPD)	-0.128
Consistency	0.681

Table 2: Results of evaluation on Transformed Dataset

Disparity Impact (DI)	~1
Statistical Parity Difference (SPD)	~0
Consistency	0.681

So, for both statistical parity difference and disparate impact bias evaluation metrics, the transformed dataset showed the ideal value (as shown in Table 2). This indicates that this tool helped to achieve inclusiveness in AI using the reweighing algorithm.

VII. TOOL 2 - AI FAIRNESS 360: POST-PROCESSING TECHNIQUE (MITIGATING BIAS IN MODEL)

- ***Introduction to the Post-processing technique:***

Post-processing is one of the three techniques provided by AIF 360 [37]. In post-processing, we mitigate bias in the machine learning model after the training is done. So the main motto here is to adjust the model's output to make the model fair and include equality for different groups.

There are many post-processing algorithms available out of which some are as follows:

1. **Equalized Odds Post-processing -** it modifies the model's output according to the equalized odds criteria.
2. **Calibrated Equalized Odds Post-processing -** here it also adjusts the model's confidence score to make the model calibrated.
3. **Reject Option Classification -** So, in this technique, we can say that model can be biased and we can mitigate the bias. we have used this algorithm for training the dataset below.

- ***Steps before using AIF 360's Postprocessing:***

In order to identify and mitigate bias in the model, we have used the postprocessing technique of AIF 360. we used the same dataset (German Credit Dataset) here. However, the protected attribute is 'sex' in this case [40].

At first, we split the dataset into train, test, and validation data. After that, we applied BinaryLabelDatasetMetric class on the training dataset to see the difference in the mean of privileged and unprivileged groups. Then we applied the logistic regression classifier on the training dataset to obtain scores for validation and test data. Now, testing the evaluation metrics and accuracy on validation and test data.

- ***Model bias mitigation algorithm: (How does the tool develop inclusive AI?)***

The Reject Option Based Classification (ROC) is an algorithm mitigating bias in the ML model [41]. It adds a 'reject' option to the model's output through which withdraws a decision when the model is not sure about the input data. In this algorithm, deprived and favored groups are labeled as desirable labels and undesirable labels, respectively. By changing the size and position of the critical region, ROC achieves fairness.

The following line shows the Python code for the algorithm:

```
ROC = RejectOptionClassification (
        unprivileged_groups=unprivileged_grous,
        privileged_groups=privileged_groups,
        low_class_thresh=0.01,
        high_class_thresh=0.99,
        num_class_thresh=100,
        num_ROC_margin=50,
        metric_name=metric_name,
        metric_ub=metric_ub,
        metric_lb=metric_lb
)
```

Here, low_class_thresh is the threshold value for the lower bound of the critical region, high_class_thresh is upper bound, num_class_thresh is the number of threshold values between low_class_thresh and high_class_thresh, metric_name is the evaluation metrics, metric_ub, and metric_lb represents the upper and lower bound of the metrics.

The following lines are used to fit the algorithm and then use the predict method for the transformation

```
ROC = ROC.fit (
        dataset_orig_valid,
        dataset_orig_valid_pred
)

dataset_transf_valid_pred = ROC.predict (
                dataset_orig_valid_pred
)
```

- ***Model bias mitigation algorithm evaluation:***

We have used the two same evaluation metrics here as that was used in the AIF 360 pre-processing experiment. In addition to that, balance accuracy, average odds difference equal opportunity difference, and Theil index are used where all have the same ideal value, 0. So, we can see from Table 3 and Table 4 that accuracy got decreased after transformation but most of the evaluation metrics' values got close to the ideal value after transformation. From these results, we can say that if we try to achieve fairness in the model, there is a trade-off with accuracy as it decreases. [40]

Table 3: Validation dataset's raw predictions with no fairness constraints, only maximizing balanced accuracy and transformed predictions with fairness constraints

	Raw predictions	Transformed predictions
Balanced accuracy	0.7473	0.6051
Statistical parity difference	-0.3703	-0.0436
Disparate impact	0.2687	0.6107
Average odds difference	-0.2910	-0.0049
Equal opportunity difference	-0.3066	-0.0136
Theil index	0.1123	0.2184

Table 4: Test dataset's raw predictions with no fairness constraints, only maximizing balanced accuracy and transformed predictions with fairness constraints

	Raw predictions	Transformed predictions
Balanced accuracy	0.7417	0.5968
Statistical parity difference	-0.3576	-0.0340
Disparate impact	0.2774	0.6932
Average odds difference	-0.3281	-0.0151
Equal opportunity difference	-0.4001	-0.0415
Theil index	0.1128	0.2133

VIII. TOOL 3 - LIME: LOCAL INTERPRETABLE MODEL-AGNOSTIC (EXPLANABILITY & INTERPRETABILITY)

- ***Introduction to LIME [43]:***

As Machine Learning is advancing nowadays, it is crucial to trust the model's prediction and understand why a particular prediction was done. Rather than seeing the model as a black box, it is crucial for the user to understand the model's behavior.

The predictions at each instance cannot be trusted blindly and then deploying the model just after checking it on evaluation metrics such as accuracy. Because there is a huge difference between the test data and real-world data. To overcome the problem of the black box model thing, LIME

was introduced. It is an algorithm that can help the user by giving explanations for the particular prediction of any classifier or regressor using approximating it with an interpretable model.

- ***Connecting explainability with inclusiveness:***

LIME makes the model explainable and interpretable, allowing users to understand why and how a specific decision was made by the model. This enables users to understand how and why the model gave biased decisions and identify any biases. Consequently, bias can be removed by taking necessary bias mitigation measures, making the algorithms fair which ultimately allows the tool to develop inclusive AI.

- ***What are the major factors that the tool (LIME) considered for developing inclusive AI?***

While using LIME for inclusiveness in AI, the main factors considered are explainability, fairness, accessibility, usability, and reliability.

- ***Are datasets usually biased?***

Now that we know LIME provides explanations to the predictions of an ML model. So, it cannot determine if the model is biased or it is the dataset because the user needs to decide that. If the model is biased, then the user should see how and why a particular prediction was made by the model that was biased. And if the dataset is biased, LIME would give explanations stating which features are important for the model's prediction and we get to know if there is bias in the dataset is there or not.

- ***Dataset used:***

We have used the fetch_20newsgroup dataset [44] which is used for the purpose of clustering and text classification. It consists of around 20,000 newsgroup papers with various different (20) topics like technology, sports, and politics.

The Scikit-learn library has this dataset function. Through it, we can customize the dataset like deciding if we want to add or remove some specific things in the document like headers or quotes.

Explainability and interpretability: (How does the tool develop inclusive AI?)

The LIME is a method to interpret how the black-box model's predictions are made. So, basically, it tries to approximate the black-box model's predictions by using easier to comprehend model like linear regression. These predictions are then simpler for a human to understand.

The LIME explanations are interpretable and explainable. Explainable in the terms that they give an almost precise explanation for the model to predict a particular output. So in this way, users can get explanations for the working of the black-box model. Talking about how the explanations are interpretable, they are quite easy for humans to understand.

Now, the ability of the LIME to give explanations at the individual instances rather than just a final and global overview of the model, makes the life of the user easier. Because this way we can get to know at a particular instance where the model is performing poorly. So, since LIME explanations provide explainability and interpretability, we can check whether models are working ethically or not. As we now know that this framework does not check the bias in the model but just gives the representation, we will move towards the algorithm used rather than bias detection metrics.

Model representation algorithm: (How does the tool develop inclusive AI?)

The following line represents a code to generate an explanation for the prediction made by the classifier for a specific instance in the dataset [42]:

```
explainer.explain_instance (
        newsgroups_test.data[idx],
        c.predict_proba,
        num_features = 6,
        top_labels = 2
)
```

Now, let us see what each argument indicates:

- newsgroups_test.data[idx]: This is the text data of the incorporated instance where 'idx' is the index of the data.
- c.predict_proba: This is the classifier c's prediction function, which takes an array of instances as input and produces the predicted probabilities for each class.
- num_features = 6: Total number of features to be included. Here it is the top 6 features.
- top_labels = 2: Here, 2 represents the number of top labels to include in the explanation.

So this function will return the explanation of the prediction for a particular instance.

Model representation algorithm explanation:

The following figure shows a chart wherein negative (blue) words indicate atheism, while positive (orange) words indicate christian. The way to interpret the weights is by applying them to the prediction probabilities. For example, if we remove the words 'Host' and 'NNTP' from the document, we expect the classifier to predict atheism with probability 0.58 - 0.14 - 0.11 = 0.31. So, we can say that sometimes there might be a case where the subject line has words that indicate atheism but overall in the body part it might be christian. So, this way we have to interpret and give appropriate features to the model.

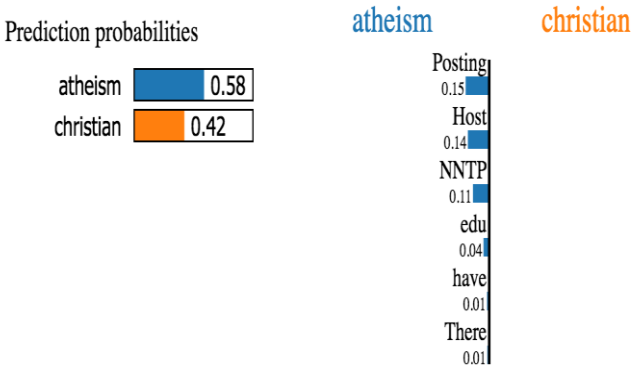

Fig. 1: Prediction probabilities for atheism and christian

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Fig. 2: Text with highlighted words

IX. TOOL 4 - SHAP: SHAPLEY ADDITIVE EXPLANATIONS (EXPLAINABILITY & INTERPRETABILITY)

○ *Introduction to SHAP:*

SHAP stands for SHapley Additive exPlanations. It is a unified framework for interpreting the ML model's predictions [46]. Because it is vital for understanding the predictions made by the model considering the points stated in LIME. SHAP values are assigned to all the features of the prediction in order to provide the user with insights into how the model gave a particular decision. Also, it has been proved that the SHAP values are better for humans to understand and interpret predictions than most of the existing methods.

So, SHAP unifies six existing methods to produce a new method for the explainability and interpretability of the model's predictions. And this method is better in terms of computational performance and human/user intuition.

○ *Connecting explainability and interpretability with inclusiveness:*

SHAP makes the model explainable and interpretable, enabling users to understand why and how a specific decision was made by the model. This allows the users to deduce how and why the model gave biased decisions and gives users the ability to identify any biases. Furthermore, bias can be removed by taking necessary bias mitigation measures, making the algorithms fair which eventually allows the tool to develop inclusive AI.

○ *What are the major factors that the tool (SHAP) considered for developing inclusive AI?*

SHAP's major factors for developing inclusive AI are model transparency and interpretability, fairness and non-discrimination, and explainability for users.

○ *Are datasets usually biased?*

SHAP framework's main purpose is to give insights into how and why the model arrived at a particular prediction. So, it cannot determine if the model is biased or the dataset, that is for the user to find. However, using these explanations of how and why the model makes decisions, we can find that is there any bias in the dataset or if it is the model that is biased. Therefore, using these explanations, it is necessary to guarantee that the model is impartial and fair.

○ *Dataset used:*

We have used the "IMDB Reviews" dataset [45] which has around 50,000 movie reviews that are equally distributed among positive ones and negative ones. This is a preprocessed dataset that is loaded from the Hugging Face Transformers library that has each review in the form of integers. The input matrix has rows denoting a review and columns denoting a distinct word (shown by the index in the corpus). And the target label is in the form of binary, 0 means a negative review and 1 means a positive review.

Explainability and interpretability: (How does the tool develop inclusive AI?)

SHAP assigns SHAP values or importance scores to all the features in the model so that we can understand how each feature is used in the model's predictions. This way we can understand how the model gave a particular prediction and how the features are influenced the most for the prediction. It provides explanations of machine learning models for interpretability and explainability.

○ *Model representation algorithm:*

The following line represents a code to generate an explanation:

```
pmodel =  shap.models.TransformersPipeline (
                    classifier,
                    rescale_to_logits = True
          )

explainer = shap.Explainer ( pmodel )
shap_values = explainer ( short_data[:2] )
```

Now, let us see what each argument indicates:

- TransformeresPipeline: It is used to classify input text data.
- classifier: It is the pre-trained transformer model used for the classification of text data.
- rescale_to_logits = True: It indicates that the model should rescale the predicted probabilities to logits.
- shap.Explainer: It is a class in the SHAP library that generates an explainer object.
- shap_values: It has the SHAP values for the model.

○ *Model representation algorithm explanation: (How does the tool develop inclusive AI?)*

The following figure shows a bar chart wherein blue represents negative SHAP value and pink indicates positive SHAP value. So, negative means the word goes towards negative reviews.

For example, the word 'good' contributes a significant positiveness for a positive movie review, however, the word 'who' gives negative signs. This way we can see why the model gave an overall positive prediction for a particular review. But there can be a case that the word 'good' is a part of the movie's name and can be omitted from the training dataset. This way we can change the model's prediction for the overall review's sentiment analysis. In this manner, SHAP helps with the explainability and interpretability of the

model's predictions. Which later can be used to make the model fair by taking the right steps after understanding the SHAP's results.

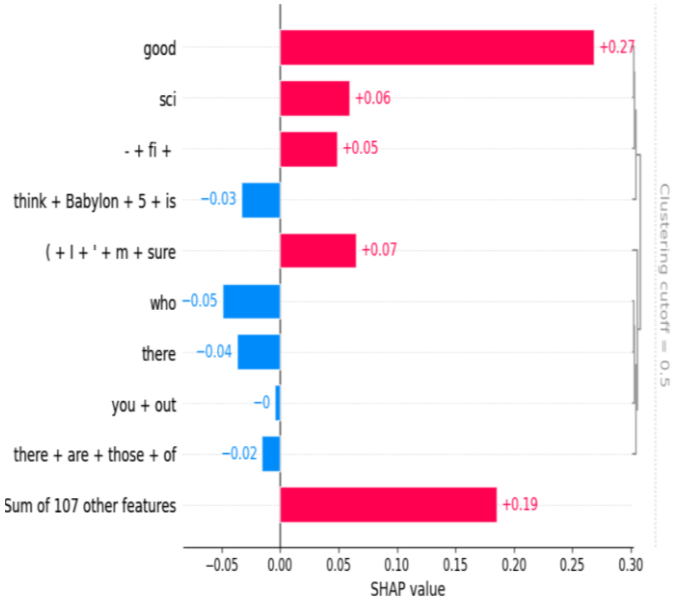

Fig. 3: SHAP value for IMDB movie review

X. TOOL 5 - TCAV: TESTING WITH CONCEPT ACTIVATION VECTORS (INTERPRETABILITY)

Introduction to TCAV [48]:

TCAV stands for Testing with Concept Activation Vectors. TCAV includes the approach of using Concept Activation Vectors (CAVs) for measuring the model's sensitivity towards high-level concepts. So, TCAV is an interpretability method. A CAV is a vector that points in the direction of the instances that illustrates the idea. TCAV does not require any retraining or changes in the ML model for interpreting the whole class or a particular example. Mainly, TCAV is used to obtain information and verify if the dataset is biased in many neural network models.

TCAV is mainly used for the identification of bias in the ML models since it is for interpreting and explaining the behavior of the model.

- ***Connecting interpretability with inclusiveness:***

TCAV makes the model interpretable, enabling users to understand how a specific decision was made by the model. This allows the users to find how the model reaches biased decisions and gives users the ability to identify any biases. Furthermore, bias can be removed by taking necessary bias mitigation measures, making the algorithms fair which finally allows the TCAV to develop inclusive AI.

- ***What are the major factors that the tool (TCAV) considered for developing inclusive AI?***

The major factors it incorporates to develop inclusive AI are diversity and inclusivity, transparency, data quality and bias, and human-centered design.

- ***Are datasets usually biased?***

Now, talking about if it is just datasets that are biased, we cannot say it is just datasets. Because with the help of this tool, we can interpret and get an explanation of the model, that user has created. So, TCAV mostly focuses on if the model is biased or not. However, if the model is for a facial recognition system, for example, and the dataset has biased data towards a specific skin tone, then we can say the dataset is biased.

- ***Dataset used:***

For doing the quantitative testing with CAV, we have used the Python code written on the GitHub repository of TCAV. Wherein, they have downloaded the following contents [47]:

- Imagenet images for the Zebra class to illustrate a target class.
- Broden dataset to extract 3 concepts, namely: dotted, stripped, and zigzagged.
- Open source Inception 5h model
- Open source Mobilenet V2 model

After downloading the required files, they structure the data in the format that can be used by TCAV. It also creates random folders with examples from Imagenet which are used by TCAV.

The following code line shows the way to run the Python function for a dataset:

```
python download_and_make_datasets.py --
source_dir=YOUR_FOLDER --
number_of_images_per_folder=10 --
number_of_random_folders=10
```

Where we have 10 random folders and 10 images per each folder.

Interpretability: (How does the tool develop inclusive AI?)

TCAV enables us to comprehend the significance of high-level concepts which are used by a neural network to give a prediction. So, we compute the TCAV score which is formed by comparing the activations of the concept in the neural network with the activations of the dataset which is representative of the concept we focus on. TCAV score measures the sensitivity of the model over the concept. If we get a positive TCAV score, then we can say that our model is sensitive towards the reference concept, else (when negative) it is not. This way we can interpret if there is bias in the model.

- ***Model representation algorithm and evaluation metrics: (How does the tool develop inclusive AI?)***

First, we define a concept of interest. So for that, we choose some examples representing that particular concept. In the next step, it is about Concept Activation Vectors (CAVs). It is mainly a hyperplane that separates the two types of concepts, that is, one without the concept and one with the concept in the model's activations. So, for example, if we want to identify a concept of zigzagged zebras, then we will

collect a positive set of images representing zigzagged zebras and a negative set of images that are random irrelevant photos. Later by applying a binary classifier, we can separate and distinguish our concept.

After that, by using CAVs and directional derivatives, we measure the sensitivity of the ML model's predictions to input changes towards a concept in the neural activation layer. The formula for conceptual sensitivity $S_{C,k,l}(x)$ using directional derivative where C is the concept in layer l and k is the class of concept [48].

$$S_{C,k,l}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(\boldsymbol{x}) + \epsilon \boldsymbol{v}_C^l) - h_{l,k}(f_l(\boldsymbol{x}))}{\epsilon} = \nabla h_{l,k}(f_l(\boldsymbol{x})) \cdot \boldsymbol{v}_C^l, \quad (1)$$

Now, the formula for TCAV for quantitative testing is:

$$\text{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|} \quad (2)$$

where k is the class label for the ML task, and X_k is the input given. So, $TCAV_{QC,k,l}$ is a metric for interpreting conceptual sensitivities for all input in the label.

Now, since we are using a random set of images, there can be a possibility of learning meaningless CAV. So, to overcome this problem, we do multiple runs for training. And then perform a two-sided t-test of the TCAV scores.
The python code to run TCAV is:

```
mytcav = tcav.TCAV (
                sess,
                target,
                concepts,
                bottlenecks,
                act_generator,
                alphas,
                cav_dir=cav_dir,
                num_random_exp=num_random_exp
)
```

Here, sees is the TensorFlow session, the target is target class, the concept is a list of concept names, bottlenecks are a list of bottleneck layer names in the target neural network, act_generator generates activations for target neural network, alphas are the list of significance levels, cav_dir is storing directory for CAVs, and num_random_exp is a number of experiments.

- ***Model representation algorithm interpretation:***

Table 5 represents the TCAV scores for the 'mixed4c' bottleneck layer in the neural network. For the 'Dotted' concept, the TCAV score is 0.50 with a standard deviation of 0.26.

This score was compared with the TCAV score for randomly labeled samples where there was little high standard deviation. The p-value shows that there was no relation between the dotted concept and the zebra class. But there is a relation between the striped and zigzagged concept and the zebra class.

Table 5: Output result for the class Zebra

Concept →	Dotted	Striped	Zigzagged
TCAV Score	0.50 (+- 0.26)	0.91 (+- 0.12)	0.84 (+- 0.14)
Random	0.50 (+- 0.30)	0.50 (+- 0.30)	0.50 (+- 0.30)
p-value	0.974	0.000	0.001
Significant	No	Yes	Yes

Figure 4 represents a bar chart of the TCAV score for each concept with 'mixed4c' bottleneck. the y-axis is the TCAV score and the x-axis shows the concepts. So, inclusiveness in AI is achieved using TCAV by evaluating the p-value of the class with different concepts and bottlenecks. This helps in identifying if the algorithm is biased towards any concept or not that is significant.

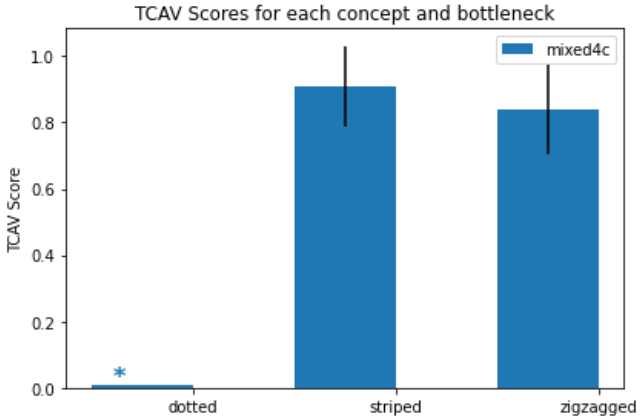

Fig. 4: Bar chart of TCAV Score for each concept.

XI. *DISCUSSION*

After exploring the above tools, we can conclude that these tools are successful in developing inclusive AI systems by identifying and mitigating biases in the datasets and the models, and by making the model explainable and interpretable.

Firstly, while using AIF360 - preprocessing, for detecting and removing bias in datasets, our observations showed that bias in AI is typically caused by biased data used for training, although there may be exceptions. The aim of implementing the AIF360 - preprocessing tool is to detect and mitigate bias in datasets. For this, we have used German Credit Dataset, which has a bias. For detecting this bias, we used three of the 71 bias detection metrics offered by AIF360: Disparity Impact (DI), Statistical Parity Difference (SPD), and Consistency. After that, we applied Reweighing algorithms to eliminate bias out of four available algorithms.

Upon re-evaluating the dataset for bias, we discovered that bias evaluation metrics of statistical parity difference and disparate impact indicated that the dataset was no longer biased, however metric consistency showed that bias still exists. Similarly, by evaluating the dataset using additional bias evaluation metrics, we can obtain an optimized and fair dataset, which can be used to build an inclusive AI system.

During the implementation of AIF360 - postprocessing, for detecting and removing bias in models, we realized that sometimes the model or algorithm itself may be biased. The main objective of implementing the AIF360 - postprocessing tool is to detect and mitigate bias in models. For this, we have used the same German Credit Dataset. we used Reject Option Based Classification (ROC) technique to remove bias from the model after identifying it. Later, we used the above mentioned three and two additional bias evaluation metrics to detect bias and discovered that the tool was successful in removing bias from the model. However, it came at the expense of accuracy. From these results, it can be concluded that achieving fairness in the model comes with a trade-off in terms of accuracy, as the latter decreases.

Now, while using LIME and SHAP, we discovered that both of these tools provide users insights into how and why the model arrived at certain decisions, enabling them to identify potential biases in the model. By eliminating this bias, these tools successfully build inclusive AI. we have used the fetch_20newsgroup dataset in LIME and Naive-Bayes algorithms to make predictions. Then by using LimeTextExplainer, the predictions made by the model are explained. Using these explanations, it can be seen if the model or dataset has any bias. If bias exists, users can employ various bias mitigation techniques and develop inclusive AI systems.

In SHAP, we used the IMDB Reviews dataset and performed sentiment analysis on the reviews. Then using the explainer method from the shap library, we obtained shap values for each word, which were afterward plotted on the bar chart to help us understand how the model interprets each word. Users can understand how words are utilized when generating predictions and can alter them to eliminate bias. For instance, in the IMDB dataset, the word "good" significantly contributes to positive movie reviews, however, there can be a case that the word ‘good’ is a part of the movie’s title and can be excluded from the training dataset. This way we can modify the model’s prediction for the overall review’s sentiment analysis.

Finally, while using TCAV, we discovered that it can be used to understand how models arrive at a particular decision, enabling users to identify any biases in the model.W have used images from the ImageNet dataset for the Zebra class to illustrate a target class. Using the Mobilenet V2 and Inception 5h neural network algorithms, this dataset is interpreted into three concepts. Then TCAV scores for these three concepts are calculated, and p-values are checked. Then it determines whether an attribute is significant, and if it is, it checks whether the attribute is biased toward a particular concept out of the given three. If so, users can mitigate this bias and create inclusive AI.

In conclusion, the results of implementation have demonstrated the potential of these tools in building more inclusive AI models. The insights gained through these tools can help in identifying biases in the dataset and the model and can contribute to the development of more inclusive AI.

XII. Future Scope for Inclusive AI

In recent years, there has been a significant development towards more inclusive AI and minimizing bias in algorithms. Inclusive AI's accuracy and efficiency have significantly increased across a range of applications, from healthcare to finance. Now, the creation of more inclusive AI is the priority of numerous businesses and technologies. As an illustration, Google has unveiled the "What-If Tool" and "Fairness Indicators," and IBM has made available the "AI Fairness 360" toolbox. These technologies give programmers the capacity to assess the fairness of their algorithms and find any potential biases.

Despite these noteworthy developments in inclusive AI development, there is still a long way to go before it realizes its full potential. Some of the key challenges are building methods to reduce biases in large and complicated datasets, using more sophisticated and effective bias mitigation methods, and enhancing the interpretability of models. This demands intensive study of novel techniques for bias detection and correction, as well as techniques for controlling the trade-off between accuracy and fairness in AI models. Another area for the future development of inclusive AI is involving different perspectives, including representatives from communities, rather than just ensuring the diversity of the development team. [48]

Further, in the future, developers can concentrate on creating more accurate fairness metrics and developing methods for recognizing and alleviating intersectional bias. Moreover, to ensure the ethical usage of inclusive AI, efforts can also be made to create inclusive datasets and implement stronger regulations and policies. All in all, with continued efforts and collaboration among AI researchers, policymakers, and communities we can anticipate seeing major advancements in inclusive AI in the upcoming years.

XIII. Lessons Learned

- First and foremost, we showed how crucial it is to create AI models that take into account many viewpoints and avoid bias because creating an AI system alone is not sufficient.
- While researching inclusive AI, we explored a variety of fields where AI has a significant impact like sentiment analysis while working with SHAP, image recognition while working with TCAV, speech recognition, natural language processing algorithms, etc.
- We studied the potential causes of why current AI systems and datasets are biased, and how even well-known organizations are struggling with their unfair and discriminatory AI models.
- We examined the technology that is used for transitioning current biased AI to inclusive AI and its advantages.
- Studied techniques for measuring bias and methods for removing bias before during and after classification.

- We also discovered how important it is to have tools for identifying and addressing bias in datasets and models. For this, we investigated and implemented five tools, namely, AIF360 - preprocessing, AIF360 - postprocessing, LIME, SHAP, and TCAV practice to help me to recognize biases and promote more inclusivity in AI models.
- We tested the above tools with various machine learning algorithms in order to detect the bias of the datasets and models and to apply bias mitigation techniques.
- Additionally, we highlighted the importance of interpretability and explainability when creating inclusive AI. With the use of these tools, we were able to show why and how the models arrived at particular conclusions and identify areas where biases could be introduced.
- Finally, we emphasize the importance of continuous research and development in the area of inclusive AI. Despite the fact that the tools we looked at were useful and successful in developing inclusive AI, there is still more to be done in the field.

XIV. CONCLUSION

Last but not least, we showed that creating inclusive AI models is a challenging endeavor. The adoption of trustworthy tools like AIF360: preprocessing, AIF360: postprocessing, LIME, SHAP, and TCAV, has helped in efficiently employing inclusive AI. We demonstrated the usefulness of these tools for identifying and reducing biases in datasets and models, as well as for enhancing the interpretability and explicability of models.

After researching the reasons for bias, methods to mitigate and measure bias, and building inclusive AI from existing biased AI, we can infer that while there has been significant advancement towards inclusive AI, there is a great deal to be done. It is critical to keep researching innovative approaches and cutting-edge tools for evaluating and eliminating bias, introducing inclusive datasets, and enforcing stricter laws.

Finally, the paper concludes that despite the challenges, inclusivity in AI has been attained with accuracy and numerous businesses and tools are making strides in this direction. With the involvement of all stakeholders, researchers, policymakers, and communities, we can expect to see breakthroughs in inclusive AI and its usage for the betterment of society as a whole.

XV. CODE LINK (GITHUB)

The following link provides code for the implementation of all five tools discussed in the paper.

https://github.com/honeypatel66/Inclusive-AI

REFERENCES

[1] Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. Technology in Society, 64, 101475. https://doi.org/10.1016/j.techsoc.2020.101475

[2] Robinson, S. C. (2020). Trust, transparency, and openness: How the inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). Technology in Society, 63, 101421. https://doi.org/10.1016/j.techsoc.2020.101421

[3] Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., Díaz-Rodríguez, N., Ficher, M., Grizou, J., Othmani, A., Palpanas, T., Komorowski, M., Loiseau, P., Moulin Frier, C., Nanini, S., Quercia, D., Sebag, M., Soulié Fogelman, F., Taleb, S., . . . Wehbi, F. (2020). Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. Frontiers in Big Data, 3. https://doi.org/10.3389/fdata.2020.577974

[4] Akintande, O.J. (2021). Algorithm Fairness Through Data Inclusion, Participation, and Reciprocity. In: , et al. Database Systems for Advanced Applications. DASFAA 2021. Lecture Notes in Computer Science(), vol 12683. Springer, Cham. https://doi.org/10.1007/978-3-030-73200-4_50

[5] Cremer, David & De Schutter, Leander. (2021). How to use algorithmic decision-making to promote inclusiveness in organizations. AI and Ethics. 1. 10.1007/s43681-021-00073-0. —— De Cremer, D., De Schutter, L. How to use algorithmic decision-making to promote inclusiveness in organizations. AI Ethics 1, 563–567 (2021). https://doi-org.proxy.bib.uottawa.ca/10.1007/s43681-021-00073-0

[6] Beiró, M.G., Kalimeri, K. Fairness in vulnerable attribute prediction on social media. Data Min Knowl Disc 36, 2194–2213 (2022). https://doi.org/10.1007/s10618-022-00855-y

[7] Alperstein, Neil. "Issues of Social Movement Ethics, Privacy, Accessibility, and Inclusiveness in Mediated Networks." Performing Media Activism in the Digital Age, Springer International Publishing AG, 2021, https://doi.org/10.1007/978-3-030-73804-4_6.

[8] De Cremer, David, and Jack McGuire. "Human–Algorithm Collaboration Works Best If Humans Lead (Because It Is Fair!)." Social Justice Research, vol. 35, no. 1, 2022, pp. 33–55, https://doi.org/10.1007/s11211-021-00382-z.

[9] Fernandez-Aller, Celia, et al. "An Inclusive and Sustainable Artificial Intelligence Strategy for Europe Based on Human Rights." IEEE Technology & Society Magazine, vol. 40, no. 1, 2021, pp. 46–54, https://doi.org/10.1109/MTS.2021.3056283.

[10] How, M. L., CHAN, Y. J., CHEAH, S. M., KHOR, A. C., & SAY, E. M. P. (2021). Artificial Intelligence for Social Good in Responsible Global Citizenship Education: An Inclusive Democratized Low-Code Approach. In Proc. of the 3rd World Conference on Teaching and Education.https://www.dpublication.com/wp-content/uploads/2021/01/10-304.pdf

[11] Kasirzadeh, A. (2021). Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence. arXiv preprint arXiv:2103.00752. https://arxiv.org/abs/2103.00752

[12] Robinson, Stephen Cory. "Trust, Transparency, and Openness: How Inclusion of Cultural Values Shapes Nordic National Public Policy Strategies for Artificial Intelligence (AI)." Technology in Society, vol. 63, 2020, p. 101421–, https://doi.org/10.1016/j.techsoc.2020.101421.

[13] M. Selim, "Inclusiveness in decision making and avoiding unnecessary costs for successful decision making," 2021 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2021, pp. 809-812, doi: 10.1109/DASA53625.2021.9682247.https://ieeexplore.ieee.org/abstract/document/9682247

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

[15] "The power of inclusive Artificial Intelligence for training" by Aurora Percannella. https://www.itcilo.org/es/node/2291

[16] "The Problem With Biased AIs (and How To Make AI Better)" https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=54b2a29b4770 .

[17] "What Do We Do About the Biases in AI?" by James Manyika, Jake Silberg, and Brittany Presten. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

[18] "Diversity, Equity & Inclusion in Artificial Intelligence: Let's get practical" by Sally Eaves. https://www.linkedin.com/pulse/diversity-equity-inclusion-artificial-intelligence-lets-sally-eaves/

[19] "How We Analyzed the COMPAS Recidivism Algorithm" by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[20] "Fairness: Types of Bias", from Google's Machine Learning Crash Course.https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

[21] "ImageNet" dataset by ILSVRC. https://www.image-net.org/download.php

[22] "Quick, Draw!" dataset by Google. https://github.com/googlecreativelab/quickdraw-dataset

[23] "MS COCO" dataset by Microsoft. https://cocodataset.org/#download

[24] "Women less likely to be shown ads for high-paid jobs on Google, study shows" from The Guardian https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

[25] Report "Google apologises for Photos app's racist blunder" by BBC https://www.bbc.com/news/technology-33347866

[26] “Racial Discrimination in Face Recognition Technology” by Alex Najibi.https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

[27] “People Tested How Google Translates From Gender Neutral Languages And Shared The “Sexist” Results” from BoredPanda. https://www.boredpanda.com/google-translate-sexist/

[28] “Amazon scraps secret AI recruiting tool that showed bias against women” by Jeffrey Dastin.
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[29] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017.
https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[30] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In FairWare’18: IEEE/ACM International Workshop on Software Fairness, May 29, 2018, Gothenburg, Sweden. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3194770.3194776

[31] “Why avoiding bias is critical to AI success”, from IBM. https://www.ibm.com/resources/guides/predict/trustworthy-ai/avoid-bias/

[32] Iosifidis, V., & Ntoutsi, E. (2018). Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D. Clough Robert Jäschke, 24, 11. https://ceur-ws.org/Vol-2103/paper_5.pdf

[33] “Reducing AI Bias with Rejection Option-based Classification” by Haniyeh Mahmoudian.
https://towardsdatascience.com/reducing-ai-bias-with-rejection-option-based-classification-54fefdb53c2e

[34] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340). https://arxiv.org/abs/1801.07593

[35] “A guide to different bias mitigation techniques in machine learning” by Sourabh Mehta.
https://analyticsindiamag.com/a-guide-to-different-bias-mitigation-techniques-in-machine-learning/

[36] “Tackling bias in artificial intelligence (and in humans)” by Jake Silberg and James Manyika.
https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

[37] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943. https://arxiv.org/abs/1810.01943

[38] Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33, 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8

[39] Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017.
https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

[40] “AIF 360 GitHub Repository” https://github.com/Trusted-AI/AIF360/tree/master

[41] Kamiran, F., Karim, A. , and Zhang, X. Decision theory for discrimination - aware classification . In IEEE International Conference on Data Mining, pp. 924–929, 2012. doi: https://doi.org/10.1109/ICDM.2012.45

[42] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). https://arxiv.org/abs/1602.04938

[43] “20 Newsgroup dataset”: https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

[44] “SHAP GiHub Repository”: https://github.com/slundberg/shap

[45] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[46] “TCAV GitHub Repository”: https://github.com/tensorflow/tcav

[47] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In International conference on machine learning (pp. 2668-2677). PMLR. https://arxiv.org/pdf/1711.11279.pdf

[48] “Toward a Future of Living with AI : AI Ethics for “Trust” from Hitachi.https://www.hitachi.com/rev/archive/2022/r2022-sp/concept/index.html