Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Algorithmic Bias and Fairness in Machine Learning: Two Sides of the Same Coin?

Merugu Bhuvana Naga Priya[1], Godavarthi Srujana[2], Angara Navya Sri Alekhya[3], Yamuna Mundru[4], Manas Kumar Yogi[5]

[1-4]CSE (AI&ML) Department, Pragati Engineering College (Autonomous), Surampalem, A.P., India
[5]CSE Department, Pragati Engineering College (Autonomous), Surampalem, A.P., India

E-mail: [1]bhuvana.merugu03@gmail.com, [2]srujanachoudhary3787@gmail.com, [3]navyaangara@gmail.com, [4]yamuna.lakkamsani@gmail.com, [5]manas.yogi@gmail.com

Abstract

The importance of counting for fairness has increased significantly in the design and engineering of those systems because of the rapid rise and widespread use of Artificial Intelligence (AI) systems and its applications in our daily lives. It is crucial to guarantee that the opinions formed by AI systems do not represent discrimination against particular groups or populations because these systems have the potential to be employed in a variety of sensitive contexts to form significant and life-changing judgments. Recent advances in traditional machine learning and deep learning have addressed these issues in a variety of subfields. Scientists are striving to overcome the biases that these programs may possess because of the industrialization of these systems and are getting familiar with them. This study looks into several practical systems that had exhibited biases in a wide variety of ways, and compiles a list of various biases' possible sources. Then, in order to eliminate the bias previously existing in AI technologies, a hierarchy for fairness characteristics has been created. Additionally, numerous AI fields and sub domains are studied to highlight what academics have noticed regarding improper conclusions in the most cutting-edge techniques and ways they have attempted to remedy them. To lessen the issue of bias in AI systems, multiple potential future avenues and results are currently present. By examining the current research in their respective domains, it is hoped that this survey may inspire scholars to amend these problems promptly.

Keywords: Artificial intelligence, bias, fairness, machine learning, models

1. Introduction

Machine Learning (ML) is the subset of Artificial Intelligence (AI) which is used to train and build models. ML algorithms work properly on small or intermediate level datasets.

DOI: https://doi.org/10.36548/jscp.2022.4.002
Received: 11.09.2022, received in revised form: 07.10.2022, accepted: 22.10.2022, published: 05.11.2022

ML provides training to machines based on following procedures: Supervised learning, Unsupervised learning, and Reinforcement learning. As machine learning models become inculcated within decision-making processes for a range of organizations, the content of bias in machine learning is the major consideration. The objective for any organization that deploys machine learning models should be to insure decisions made by algorithms are clear and independent of bias.

This study explores the content of machine learning bias, along with what it is, how it is detected, and threats of bias in machine learning. Recognising and resolving machine learning bias is vital so that the model outputs can be trusted and noticed as fair. It is associated with deliberations around model explainability i.e., the procedure of human understanding how machine learning model made its decision. Machine learning models study from the data itself, so it maps and learns the trend and pattern and are not elaborated directly by a human [1]. However, bias in machine learning can appear for a range of distinct reasons, if left unmonitored and unbounded. A familiar reason is that the sample of training data does not essentially define real world conditions faced by the model once deployed. If the training data is of high quality, it may contain historic bias from major societal influence which can affect the model. Once deployed, a biased model may favour groups or become less accurate with peculiar data subsets. This may lead to decisions which unfairly penalize a particular batch of people, which can have severe ramifications in a real-world setting.

1.1 Bias

A concept referred to as learning bias, also described as algorithm bias or AI bias, is when an automated system generates outcomes that are persistently skewed as a result of flawed assumptions generated during the training process. Machine learning bias typically results from issues that are brought about by the people who create and/or develop the ML models. These people may design algorithms that reflect unintentional cognitive biases or actual prejudices. Or the people could employ biased, inaccurate, or incomplete datasets to develop and/or validate the ML techniques, which would create biases. For example, data collected and used in the medical fields is frequently biased towards specific groups, which can have negative consequences for underprivileged groups [17]. They demonstrated how excluding African Americans caused them to be incorrectly labelled in medical trials, leading them to be champions for analyzing the genes from various groups are included in the records to protect underprivileged communities from harm. Paper [15], based on analysis of the 23 and Me genotyping dataset, discovered that out of 2,399 people, 2,098 (87%) of those who

freely disclosed their genotypes in public databases are European, whereas only 58 are Americans, 50 (2%) are African, and 2% are Asian. According to a similar research [16] undertaken elsewhere, the huge and well-known genetic database UK Biobank might not accurately depict the population under study. The studies revealed the proof of a selection bias of "healthy volunteers". Paper [18] contains additional instances of research on the biases currently present in data utilized in the medical field. Article [19] examined ML techniques and data used in the medical industry and discussed how not all sick people have benefited equally from AI in medical services.

1.1.1 Characteristics of a high bias model

- Failure to get correct data trends,
- Possible towards under fitting,
- Much generalized and excessively simplified,
- High error rate.

To avoid such situations, associations should audit the data which is being used to train machine learning systems for the need of comprehensiveness and cognitive bias. One of the biggest things of bias is prejudgment, which can affect in discriminatory trials. As some biases can be useful and are utilized in heuristic decision - making, it is essential to determine the balance between useful biases and negative prejudicial biases [2]. It can be almost insolvable to be fully unbiased. With the help of labelled data collected from the environment to develop models, ML enables us to anticipate and differentiate between processes. In previous years, an outburst of interest in algorithmic fairness from academics and the general public have been noticed. Although this interest and the recent increase in both the quantity and the speed of labour, the fundamental understanding of fairness in ML is emerging.

1.2 Fairness

As ML algorithms enhance outcomes in elevated situations like housing loans, hiring, jail punishments, etc., fairness is a crucial concern. Fairness is a complicated and multi-faceted theory that depends on context and culture. In simulation, a series of bias and fairness metrics characterize unique patterns where a machine might execute diversely in different batches of the given input. When such batches refer to groups of individuals, those individuals may be identified by securing or critical traits like religion, aging, and professional standing.

1.3 Algorithmic Fairness

This is the study of investigation which analyses and debugs prejudices [3]. This will be at the intersection of machine learning and ethics. Particularly, the study includes, researching the reasons for bias in models and algorithms, and declaring and applying measures of fairness. It is also essential to know that approaches to fairness aren't all quantitative. This is because the causes of unfairness go beyond data and algorithms. The exploration will also have an understanding and address the main reason for the unfairness. Although these biases are frequently not planned, the conclusions of their existence in machine learning models can be significant. Based on how the machine learning systems are utilized, such biases could affect in lower client service experiences, reduced deals and a profit, illegal or conceivably illegal conduct, and potentially dangerous conditions. By omitting human bias and concentrating exclusively on elements that can actually forecast a borrower's repayment capacity of a loan using historical data, automated mortgage choices have the capacity to become much more objective. However, it creates a lot of issues around what constitutes fairness and if a model built using skewed historical data can be fair.

1.4 Literature Study

Reference Number	Flaws due to bias
20	Limitations of Implicit Bias in Matrix Sensing: Initialization Rank Matters
21	The Dangers of Human-Like Bias in Machine-Learning Algorithms
22	Dissecting Racial Bias in an Algorithm used to Manage the Health of Populations
23	Ensuring Fairness in Machine Learning to Advance Health Equity

1.5 Defining and Adopting Measures of Fairness

Fairness is socially defined study where algorithmic bias is mathematically defined. They are employed in crucial situations like finance and employment. Algorithmic decision-making has pellucid advantages; unlike humans, machines don't get fatigued or exhausted and are able to consider orders of magnitude additional factors. Algorithms are nevertheless sensitive to biases that make the decisions they make "unfair," just like people are. Fairness in the context of decision-making is the absence of bias or preference toward an individual or a group based on their innate or learned attributes [4]. As a result, skewed algorithmic decisions are made in favour of an odd group of people.

2. Study of Bias in ML

2.1 Causes of Machine Bias

Algorithms create situations where justice is more difficult to achieve on a moral, lawful, mental, and social level due to the growth of AI. Individuals need to start taking those challenging scenarios in managing AI products and AI solutions carefully.

2.2 Different forms of biases

Can these societal biases be prevented from influencing machine learning models? Recognizing the processes that lead to these biases is the first step. Using skewed data to feed a biased model is one of the approaches. Here, various approaches that would possibly be abandoned with this skewed data are discussed.

2.2.1 Algorithmic Bias

Algorithmic bias is the term used to characterise consistent and recurring mistakes made by system programmes that lead to "unfairness", like favouring a class on any other method that goes against the algorithm's original purpose. Scholars are worried about the effects that unexpected output and data modification can have on the physical world as algorithms increase their capacity to govern communities, economies, organisations, and culture. Due to the part in psychological process of bias, algorithms that are frequently believed to be impartial can mistakenly be projected as having more authority than professional knowledge. In few instances, relying on techniques can also replace people obligation to their results. Bias can input into algorithmic systems due to pre-present communities, economies, or organizational norms, due to the boundaries imposed during structure, via means of being utilized in unanticipated contexts or via means of audiences who aren't taken into consideration within model’s preliminary architecture [5]. It is mentioned that instances starting to unfold the outline of offensive talk. It is additionally raised in judicial system, medical field, and selection procedure, in addition to the present ethnic, economic, and religious biases. Multiple erroneous arrests of black men have been connected to facial recognition technology's relative failure to recognise darker-skinned individuals; this problem results from imbalanced datasets. The unique form is that, it is normally dealt with confidentiality, which makes it difficult to comprehend, investigate, and identify this. Also, while complete transparency is offered, some algorithms' complexity makes it difficult to comprehend how they work. Programs might also additionally alter and react for intake and outtake for methods that are unpredictable and difficult to replicate for

study. There is neither any singular program for investigation, often inside one internet site, but rather a network of numerous interconnected programmes and data inputs among customers.

2.2.2 Sampling Bias / Selection Bias

It occurs if collecting accurately among subgroups is stopped. Considering the situation where there are more men's credentials than women's, women submissions are no longer accepted, and possibly given up on getting know women aspirants in favour of rejecting them. Some credentials are important for "data programs". A recent applicant with a focus on data systems might abandon detecting skewed selection among programmes because of the absence of illustration of this topic. Individual choice is an essential component to consider when speaking of decision biases. When considering conducting an internet study on PC utilization, trying to gather any examples from those who don't even use computers anymore is restrained.

2.2.3 Measurement Bias

This bias occurs as a result of erroneous entries entering the dataset. These are a few cases of size bias: defective tools, determining unsuitable values, and inaccuracy in recording data collected. A situation where the applicants are asked to complete a structure for the coronavirus indicators group, and the possibilities to treat a symptom like fever, is considered. If the patients doesn't know how high their temperature was, negative records for more temperature or less temperature can be entered! Actually, what you want is a "never clear" choice.

2.2.4 Inherited Bias

An ML algorithm should be fed with data from every other biased ML model's output, in an inherited bias. As a result, there will be skewed supplies and eventually skewed results. For example, it has been proven that the famous word embeddings include bias that results in comparisons just like "male: a web developer; lady: a housewife".

2.2.5 Prejudice Bias

The information employed for training the system in this instance represents active assumptions, judgements, defective sociological hypotheticals, and introducing identical actual bias from the ML process. In the distributed system, for illustration, employing

information on therapists that is solely comprised of primary care doctors and female nurses would perpetuate a real-world gender conception about medical professionals [6].

2.2.6 Exclusion Bias

It occurs whenever a major data value is omitted from records that are used. This occurs when evaluating a model. It consists of evaluation of programs using unrelated and unequal standards. These criteria which are employed for the assessment of identity verification structures, have been skewed in the direction of pores and melanin and the identification of gender, so that they could be used as instances of this type of prejudice [7].

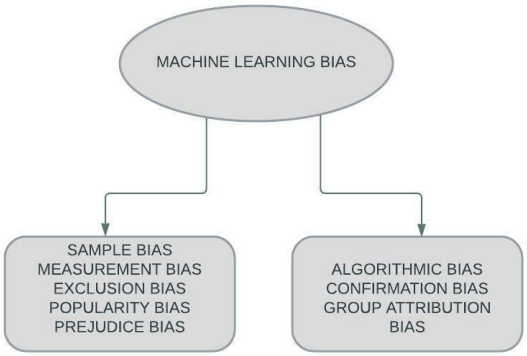

Figure 1. Types of Machine Learning Bias [7]

2.2.7 Popularity Bias

Items which are famous have a tendency of not to cover high. Furthermore, these measurements are vulnerable to alteration, such as faking evaluations or employing online networking robots. For example, this form of bias may be visible in Google or recommendation systems in which famous gadgets might be offered greater to citizens. These presentations, however, will never have an conclusion of high resolution; rather, they might be the product of any other biased variables.

2.2.8 Group Attribution Bias

This model effects from where a system is trained with data that contain an asymmetric outlook of a particular group. For illustration, in a particular sample dataset, if the maximum of a particular gender would be more successful than the other, or if the maximum of a certain race makes more than the another, the model will be inclined to accept these falsehoods. There is label bias in these cases. In reality, these kinds of labels should not be framed into a model in the first place [8]. The sample used to understand and consider the present scenario cannot just be used as training data without the applicable pre-processing to

regard for any implicit unjust bias. Machine learning systems are getting more essential in society without the common person indeed understanding, which makes group attribution bias exactly as likely to punish a person unjustly, because the compulsory path was not taken into account for the bias in the training data.

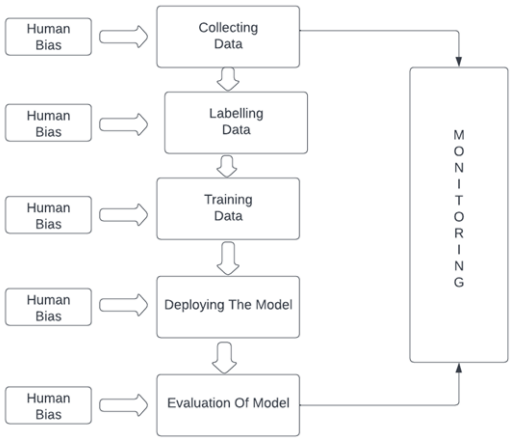

Figure 2. ML Bias Flowchart [9]

2.2.9 Confirmation Bias

This is a famous bias that has been studied within the subject of psychology and immediately applicable to how it could have an effect on a machine learning technique. If the humans of a particular use have a pre-current hypothesis that they would really like to verify with machine learning (there are possibly easy methods to do it, relying on the context), the humans worried within the modelling process is probably willing to deliberately manage the process closer to locating that solution. It could be far common than supposed, heuristically simply due to the fact that, people in an industry are probably compelled to get a positive solution earlier than even beginning the procedure than simply trying to see what the data is simply saying.

2.3 How prejudice in AI mirrors biases in society

Consequently, artificial intelligence isn't always secure for people's biased characteristics. If the effort to make AI fair is put forth, it can help individuals take decisions that are more objective. The source of artificial intelligence bias is frequently the fundamental information instead of the technique. In light of this, the following are some intriguing results from a report on overcoming artificial intelligence bias:

Models can be taught using information about people's picks or about societal or ancient inequalities. Word features, a group of NLP approaches taught from newspaper

stories, might also additionally replicate social and gender biases. Data can be biased via means of the manner they're collected or selected. It is considered that in judicial system artificial intelligence algorithms, unsampled specific regions might also additionally bring about extra information for crime in that area, which can cause greater enforcement. User-generated information might also result in a biased conclusion. This was observed while searching for names that identified African Americans. The term "arrest" came up more frequently than when searching for names that identified white people. The algorithm may have displayed this result extra regularly whether or not the word "arrest" was entered, depending on how frequently people may have clicked on different variants for other searches.

An ML device can identify correlation coefficients that are deemed to be immoral or illegal. For example, a housing finance version may decide that the aged humans have lower reliability and a higher likelihood. Illegal ageism might be dealt with, if the model only uses age to draw this result. It isn't tough that when biased algorithms are set up to resolve real-world issues, it is able to have unintentional consequences. For example, a facial recognition device can begin to be racially discriminatory, or a credit software assessment device can become gender-biased. There may be intense implications for those biased applications. A bias also can render software useless if utilized in a distinct context. For instance, a voice assistant is developed, but only taught with the voices of humans from a specific place, it can't be assumed that it would work perfectly if used in a distinct place, due to changes in voice tone, dialect, culture, etc. [9]

3. Study of Fairness in ML

3.1 Machine Learning Fairness

It is a department of AI that is related to the concept of computer systems studying information gathered to become aware of algorithms and giving conclusions that are as illustrated by humans. The process of addressing and eliminating algorithmic bias from ML systems is known as ML fairness.

3.2 What is needed to be known?

What to understand regarding this and implementing principles into a society that is becoming more and more computerized is that, our daily lives are being impacted by ML, a technique for building statistical models that get better (and learn) through the use of data and

skills. Examples include filtering resumes and college seats. Making sure that this data science is moral and honest, including the equipment and systems employed, is becoming more and more important. Whenever ML is unfair, clients and society may suffer as a result. For instance, popular algorithms that were designed to give individualized suggestions to users may have exacerbated social conflicts because of biased or compartmentalized news feeds (including fake news).

3.3 Why is it crucial to address fairness and ethics in machine learning?

One of the reasons it's important to address fairness and ethics is that ML algorithms can discriminate against the intended. Models and programs are utilized to help us choose furniture, find employment, hire new employees, apply to colleges, listen to music, acquire loans, get information, search on Google, focus on ads, and do a lot of other things. It has streamlined information and enhanced humans' ability to communicate with one another. But if the models don't support fair and honest behaviour, it may have catastrophic consequences.

Data scientists and ML experts must keep an eye out and address these potential biases in algorithmic models if they are to be avoided. However, machine learning tackles this issue by allowing a computer system to learn by doing rather than having given detailed instructions.

3.4 How to make machine learning fairer and much ethical?

There are some ways to ensure that ML is honest and ethical for those working in DS and AI with algorithms. One can:

- Analyze the algorithms' influence on individuals’ behavior to see if they are biased, and then develop algorithm strategies that avoid predicting future bias.
- Find any flaws or contradictions in open datasets and decide whether there has been a privacy infringement or not.
- Make use of tools that could help eliminate or reduce bias in ML.

3.5 Algorithmic Fairness

ML techniques like algorithm fairness aim to lessen the impact of prejudices in the data. Nevertheless, despite the wide variety of uses, only a small number of papers take the multi-class categorization setup into account from a fairness standpoint. Here, the notions of real and approximate fairness are expanded in the context of multi-class categorization for

ethnic equality. The best fair classifiers' related terms are listed. This demonstrates a data-driven plug-in process for which the security provisions are established. It has been demonstrated that the enhanced estimator mimics the behavior of the optimal rule in terms of uncertainty and equity. Fairness makes sure that there is no dispersion [10]. The strategy appears to be quite efficient in making decisions with a predetermined phase of injustice and has been examined on both synthesized and actual datasets. Additionally, this method competes favorably with cutting-edge fair learning processing in the particular binary classification situation.

3.6 Analysing and Measuring Unfairness

Most of algorithm fairness studies' objectives are at developing techniques to examine and measure unfairness. This can contain analysing information for the potential reasons for unfairness cited above. It additionally includes measuring unfairness in system predictions. For data analysts, scientists, and designers to explain their designs and comprehend the importance and correctness of their conclusions, ML system fairness and understandability are essential. In order to evaluate ML systems and make wise judgments about how to improve them, interpretability is also crucial.

3.7 Counterfactual Fairness

The fairness measurement determines whether or not the classifier produces the same output for a man or a woman as it does for every other man or woman who is similar to the first except in 1 or more attribute values. One method for exposing bias sources in a model's capability is to evaluate the classifiers for counterfactual fairness. Algorithm bias correction and elimination is the process of ML fairness [11].

3.8 Why is Fairness Essential in Machine Learning?

Fairness and understandability in ML programs are crucial for data scientists, analysts, and designers to explain their systems and comprehend the significance and correctness of their outcomes. Performance of the model is essential for debugging ML systems and for making informed judgments about how to improve them. The distribution of an algorithm's fairness risk is irregular. Another aspect of algorithm fairness that compares strategies based on knowledge to others is considered. A more thorough or consistent method is often assumed by lists, automatic testing, and process-based methods, in which you have a list of factors you are looking for and you will look for all of these matters in all products—

for instance, you will have to take note of boot camp variety or design parameters in all products. The issue with this is the fact that a threat is not distributed equally among specific problems inside a given product and it is not distributed equally across all product types. Instead, a small number of items bear most of the danger, and within each of those products, a handful of uncommon issues are likely to do so. Therefore, if you assume that there's an expense to researching problems (which is generally true at the moment), the goal should be to identify and test the issues that are most likely to result in harm rather than testing everything in a large combinatorial space. A vital competency that experts bring to the table is the ability to recognize such volatile issues, which they may eventually standardize.

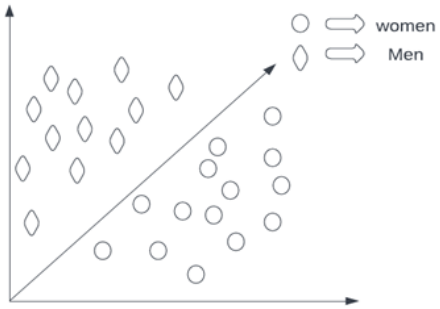

Figure 3. Is ML model fair? [11]

The issue of algorithm fairness is heinous. According to many factors, algorithm fairness is a serious problem, and it deserves to be treated with regard to its complexity. Different problems are complex and difficult systemic issues that frequently and uncleanly prevent a point, where attempts to address each aspect of the problem may also reveal or develop so many different issues, and decision makers have such divergent viewpoints that they cannot agree on what the question is, let alone what the solution might look like or who is responsible for the answer. Algorithm fairness is such a problem, and approaching it in that way, better equips us to handle it [12].

3.9 Different kinds of Fairness

Arguments about fairness are interminable in element due to the fact there are 3 specific kinds, making it easy for left and right to speak beyond each other. First, procedural fairness and distributive fairness ought to be distinguished. When making judgments that have an impact on other people's wellbeing, procedural fairness refers to whether or not objective and transparent methods are applied. Is the decision-maker objective? Is the game fixed? The health of a democracy depends on procedural fairness because when people have faith in the system, they are much more likely to accept outcomes that are unfavourable to

them. Additionally, they are much more likely to participate in populist uprisings once they suspect the system is corrupt.

Contrarily, distributive fairness refers to how things are divided, including benefits and costs. Is everyone receiving and acting in accordance with their truthful proportions? However, there are variations in distributive justice, including equality (everyone receives the same amount) and proportionality (all acquire rewards in proportion to their inputs; that are sometimes known as equity). This simple distinction can help us understand many of today's most difficult debates. Everyone supports proportionality, but the left also supports equality, even if it is at odds with proportionality. For its own purposes, the right has really no concern with equality. Conservatives like proportionality results in significant outcome disparities.

Based on the information provided, an example which contrasts equality with proportionality is considered: "Regardless of effectiveness, every professional in a given class should get the same pay". Thirty percent of respondents who identified as "extremely liberal" agreed. However, only 3% of "extremely conservative" issues did. Conservatives don't need to think about it; liberals do need to. A violation of fairness as proportionality is imposing equal effects in the absence of equal inputs. Conflicts between the left and the right show this difference on a worldwide level. For instance, a country's president might want to outlaw homework. His issue is that students from single-parent homes are significantly less likely to receive homework assistance from their parents than children from nuclear families, who are likely to be much better off financially. He has a tendency to slow down the education of a few children in order to reduce the inequality of outcomes brought about by the meritocratic institutions in their nation.

4. Performance Effects due to Bias and Fairness

When it involves overall performance reviews, biases have a massive effect. Particularly in overall performance management, biases can result in inconsistencies in aim, difficulty and consideration, training and remarks, improvement opportunities, and rewards.

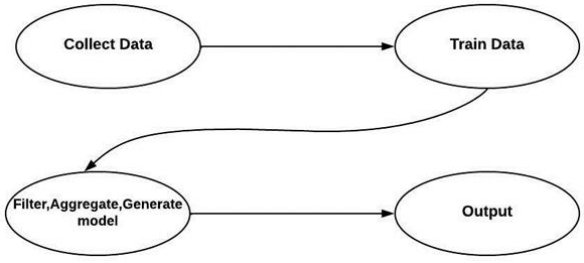

Figure 4. Evaluation Process of Fairness [14]

Given the capacity of bad effect of biases on workers, corporations can't simply accept that bias is only human and natural. In elevated circumstances where evaluations are involved, such as promotional offers, payment, recruitment, or even termination, biases can result in the increase or decrease of worker performance, which can have severe consequences.

4.1 Biases Affecting Overall Performance Evaluations

Few of the biases affecting the overall performance evaluations are:

a) Primacy Bias: In overall performance reviews, managers frequently fall for primacy bias once they permit a primary impact that have an effect on their overall evaluation of that mentee. Placing collectively a file of overall evaluation that consists of remarks from factors is primary bias.

b) Recency Bias: It refers to the ability to concentrate on the latest period of time rather than the whole term. To restrict the effect of this in evaluation facts, a routine of gathering workers' remarks at distinct points in time throughout the year is expanded. Did a person just complete a 3-month project? Great, give their friends a request for remarks so that you can get little information on how nicely they did. Did a person just complete inner training? Awesome, request remarks from the trainer about their participation. This manner, you have more common facts from the whole term on the end of the year.

c) Centrality Bias: It is the propensity to place the maximum objects at the center of the score range. While most situations benefit from development, taking a stand is frequently necessary during high-pressure conditions such as evaluations [13]. It might be challenging to separate low-performing employees from great performers if everyone is given the same rating. It is the best manner is to get rid of impartial choice. Examiners are forced to pick a side in this situation.

d) Confirmation Bias: The propensity to look for or interpret new data in a manner that supports an individual's established views, is confirmation bias. It is quite much like primacy bias however can generally tend to go a lot deeper. To reduce this, assume as a researcher. If scientists raise doubts that are trying to shape their conclusions they are trying towards finding out disconfirming other than verifying preliminary principles. Whenever there is influence on a particular person, leave and try to find proof that they may be the opposite or completely unique from what you suspect. When gathering remarks from others, pay careful interest to the remarks that goes against your beliefs.

e) Gender bias: When giving remarks, people generally tend to focus more on the character and attitudes of ladies and feminine-representing individuals. Contrarily, they focus more on the personality and achievements of men and masculine-representing people. Sometimes, unstructured remarks permit bias to creep in. Without some set standards, humans will possibly reshape the standards for achievement in their own image. Gender biases could have a massive effect on the studies and tests of non-binary and/or transgender folks. Although those biases may also happen in a slightly unique manner, it's crucial to stay alert and keep your eyes open.

4.2 Fairness Effects

Fairness in machine learning refers to the numerous tries at correcting algorithmic bias in automatic decision approaches primarily based totally on machine learning models. This brief explores "fairness" broadly, and then dives into the default fairness method in ML and related challenges. It ends with tools and concerns for the ones developing, dealing with and the usage of ML systems. It must be able to detect instances of unfairness in a particular dataset given a description of what constitutes fairness. In general, fairness and bias are taken into relevant consideration while the decision process affects people's lives. The concept of algorithmic bias in machine learning is well-known. Outcomes can be skewed through a variety of things and for that reason it is probably considered unfair with admiration to certain groups or individuals [13]. An increasing concern is that, this data science, along with the tools and processes employed, is to be ethical and just. The ultimate outcomes may be harmful to users and the community when machine learning is not always fair. Fairness and interpretability in machine learning models are crucial for data scientists, researchers, and developers to explain their models and comprehend the significance and correctness of their findings. In order to debug machine learning systems and make informed judgments about how to improve them, interpretability is also crucial.

- Participants will study how to use free and open-source fairness and interpretability packages.
- Generates attribute significant values and/or relevant data points for the existing model to better clarify model prediction.
- Attain model interpretability on a great scale over training and inference on different datasets.

- At the training stage, use an intuitive graphical dashboard to predict possible trends and causes.
- Utilize additional graphs and charts to analyze which user groups could be significantly affected by a model's implementation and compare other models for fairness and effectiveness.

4.3 Recent Approaches on Avoiding Bias

Machine learning bias is a major issue. Most of the time, individuals and processes are to blame when models don't work as intended. However, it is possible to use a "fairness by design" approach to machine learning that takes into account a few important aspects. Companies can do the following to achieve this: pair social scientists and data scientists, label with attention, combine fairness measures and typical machine learning metrics, combine representativeness and critical mass limitations when sampling, and maintain de-biasing in minds while developing models.

- Determine possible biases in your sources: Examining the data to understand how the various types of bias could affect the data being used to train the machine learning model is one technique to handle and reduce bias, using the aforementioned causes of bias as a guide.
- Establish guidelines and regulations to get rid of prejudices and practices: Organizations can take the appropriate steps to address issues with machine learning model bias by adopting these principles and sharing them in an open, transparent manner.
- Find reliable, representative data: Organizations should take steps to understand what a representative data collection should resemble before gathering and combining data for machine learning model training.
- How data is chosen and cleaned up should be documented and shared: Organizations should document their procedures for data selection and cleansing in order to ensure that minimal bias-inducing errors are made.
- In addition to performance, evaluate the model's performance and choose the one with the least bias: Before being implemented, machine learning models are frequently tested.
- Observe and evaluate the currently running models: Organizations should offer tools for tracking and reviewing models on an ongoing basis as they operate.

5. Conclusion

This survey has analysed and reviewed biases and fairness in Machine Learning (ML). After that, the topics of bias and fairness as laid out by experts are studied. The impact of bias on our society has also been explored. Further, algorithmic biases, and the different types of biases in machine learning such as, Prejudice bias, Historical Bias, Popularity Bias and so on are discussed. The intention is to enhance the consumer's perspectives so that, they may think deeply while utilising a product or a method to make sure that there is little likelihood of bias or possible harm being done to a particular community. Modern approaches to fairness machine learning typically focus on post-processing, model learning, or statistics-based interventions. This kind, explains the duties effectively of statistics specialists who are intended to carry out those procedures. There is a risk that this will result in a procedure that is focused on a limited, static set of covered instructions derived from regulation and is deficient in other areas, without considering the reasons behind the inclusion of those subjects and how they connect to the specific equity components of the utility under consideration. Given the increasing prevalence of ML in our society, it is vital that researchers approach this issue seriously and deepen their knowledge of the subject. Ideological analyses of fairness and bias encourage additional crucial issues on extra essential questions, and recommend avenues for further attention of what is probably applicable and why. This increases a sequence of sensible and demanding situations, which can restrict how powerful and standardized fair ML techniques may be in use. In this survey, how Biases and Fairness have an effect on the overall performance assessment and why fairness is crucial in machine learning have been categorized.

References

[1] Barocas Solon , Hardt Moritz, Narayanan Arvind, Fairness and Machine Learning https://fairmlbook.org/ , 2017.

[2] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in Machine learning. Commun, ACM, 63(5):82–89, April 2020.

[3] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi.”On the applicability of machine learning fairness notions”. SIGKDD Explorations,23(1), 2021.

[4] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. URL http://arxiv.org/abs/ 1908.09635. CoRR, abs/1908.09635, 2019.

[5] Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." ACM Computing Surveys (CSUR) 54.6 (2021): 1-35.

[6] Holstein, Kenneth, et al. "Improving fairness in machine learning systems: What do industry practitioners need?." Proceedings of the 2019 CHI conference on human factors in computing systems. 2019.

[7] Allison Woodruff. "10 things you should know about algorithmic fairness", Interactions, 2019.

[8] Verma, S., and Rubin, J. (2018). "Fairness definitions explained," in IEEE/ACM International Workshop on Software Fairness (fairware), Gothenburg, Sweden, May 29-29, 2018 (Newyork, NY: IEEE), 1–7.

[9] Ninareh Mehrabi, Fred Morstatter , Kristina Lerman ,Aram Galstyan, Nripsuta Saxena https://www.researchgate.net/publication/335420210_A_Survey_on_Bias_and_Fairness_in_Machine_Learning, 2019.

[10] Hellström, Thomas, Virginia Dignum, and Suna Bensch. "Bias in Machine Learning--What is it Good for?." arXiv preprint arXiv:2004.00686 (2020).

[11] R. K. E. Bellamy et al., "AI Fairness 360: an extensible toolkit for detecting understanding and mitigating unwanted algorithmic bias", arXiv e-prints arXiv:1810.01943, pp. 20, 2018.

[12] Yaniv Zohar - https://www.aporia.com/blog/machine-learning-bias-and-fairness/

[13] GoogleDevelopers-https://developers.google.com/machine-learning/crash-course/fairness.

[14] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).

[15] Joy Buolamwini and Timnit Gebru , 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[16] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. 2017. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. American Journal of Epidemiology 186, 9 (06 2017), 1026–1034. https: //doi.org/10.1093/aje/kwx246 arXiv:http://oup.prod.sis.lan/aje/article-pdf/186/9/1026/24330720/kwx246.pdf.

[17] Arjun K. Manrai, Birgit H. Funke, Heidi L. Rehm, Morten S. Olesen, Bradley A. Maron, Peter Szolovits, David M.Margulies, Joseph Loscalzo, and Isaac S. Kohane. 2016. Genetic Misdiagnoses and the Potential for Health Disparities. New England Journal of Medicine 375, 7 (2016), 655–665. https://doi.org/10.1056/NEJMsa1507092 arXiv:https://doi.org/10.1056/NEJMsa1507092 PMID: 27532831.

[18] Selwyn Vickers, Mona Fouad, and Moon S Chen Jr. 2014. Enhancing Minority Participation in Clinical Trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual. Cancer 120 (2014), vi–vii.

[19] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. 2019. Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA journal of ethics 21 (02 2019), E167–179. https://doi.org/10.1001/amajethics.2019.167.

[20] Eftekhari, Armin, and Konstantinos Zygalakis. "Limitations of Implicit Bias in Matrix Sensing: Initialization Rank Matters." arXiv preprint arXiv:2008.12091 (2020).

[21] Fuchs, Daniel J. "The Dangers of Human-Like Bias in Machine-Learning Algorithms." Missouri S&T's Peer to Peer 2, (1). https://scholarsmine.mst.edu/peer2peer/vol2/iss1/1, 2018.

[22] Obermeyer, Ziad et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366 (2019): 447 - 453.

[23] Rajkomar, Alvin et al. "Ensuring Fairness in Machine Learning to Advance Health Equity." Annals of Internal Medicine 169 (2018): 866-872.