**ABSTRACT**

*Algorithmic bias is a tendency to not merely neutrally transform or extract information from data but to operate on it in ways that deviate from a normative (moral, statistical, social, etc.) standard such that one kind of individual or group is unfairly privileged over another based-on aspects of their social identity. Countless instances of algorithmic bias against specific population groups result in incorrect outcomes. For example, Amazon scrapped an AI algorithm for job recruitment that systematically downgraded women’s CVs. Elsewhere, an algorithm (COMPAS - Correctional Offender Management Profiling for Alternative Sanctions) that predicted whether defendants would re-offend gave higher risk scores to African-Americans than Whites. However, both groups were equally likely to re-offend. Relatedly, some algorithms that powered facial recognition AI systematically misclassified people of colour or mislabelled Black men as ‘primates.’ The underlying issue is not the technology itself but the data using which the models are trained. Mislabelling of AI training data and unrepresentative sampling were identified as prime factors for bias. Algorithmic bias stems from existing human biases. Research suggests that algorithms mirror existing social inequalities and stereotypes, resulting in bias. This bias can have impacts ranging from inadvertent privacy violations to reinforcing social biases of race, gender, sexuality, and ethnicity. Hence, building ethics by design throughout the AI development lifecycle is paramount to mitigation. Several algorithmic techniques like adversarial de-biasing, using semi-supervised variations encoders, dynamic up a sampling of training data, and distributionally robust optimization can also mitigate algorithmic bias. This area of research is gaining tremendous attention as underrepresented communities worldwide seek adequate and accurate representation not just in technology but also in entertainment, politics, and more.*

1. **INTRODUCTION**

The trajectory of machine learning (ML) is on course to achieve a growth target of $20.83 billion in 2024 from $1.58 billion in 2017, with a compound annual growth rate of 44.06% [4]. Today, machine learning algorithms influence our lives in various ways. Computer algorithms that make decisions and predictions are often viewed as inherently fair and objective. But in recent years, a competing perspective has emerged -- the perspective that algorithms often encode the biases of their developers or the surrounding society, producing predictions or inferences that are discriminatory towards specific groups. This is also called algorithmic bias. Examples of algorithmic bias cross contexts, from criminal justice, medicine, computer vision, and hiring. These limitations appear -- and are particularly salient -- for high-stakes decisions such as predicting recidivism or administering anaesthesia [2]. Scholars increasingly caution that emerging research grapples with many algorithmic biases from training data, analytics models, and socio-cultural sources [3]. As such, biased algorithmic decision-making may result in unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization. Different types of biases exist in Machine Learning. They are: - [1]

Social Bias:

Available data reflects an existing bias in the relevant population before the creation of the ML model. It is synonymous with Historical Bias, Societal Bias, Individual Bias, and Pre-existing Bias.

Measurement Bias:

Chosen features and labels are imperfect proxies for the actual variables of interest. Synonyms include Linking Bias, Omitted Variable Bias.

Representation Bias:

The input data must represent the relevant population, leading to systematic errors in ML model predictions. Synonyms are Temporal Bias, Longitudinal Data Fallacy, Emergent Bias, Population Bias, Group Bias, Aggregation Bias, Behavioural Bias, Sampling Bias, Content Production Bias, (Self) Selection Bias, and Availability Bias.

Label Bias:

Labelled data systematically deviate from the underlying truth categories.

Algorithmic Bias:

Inappropriate technical considerations during modelling lead to systemic deviation of the outcome. Synonyms are Statistical Bias and Technical Bias.

Evaluation Bias:

A non-representative testing population or inappropriate performance metrics are used to evaluate the ML model. Synonyms include Observer Bias, Funding Bias

Deployment Bias:

The ML model is used and interpreted in a different context than it was built for. Synonymous to Cause-Effect Bias

Feedback Bias:

The outcome of the ML model influences the training data such that a feedback loop can reinforce a slight bias. Synonyms are Presentation Bias, User Interaction Bias, Popularity Bias, Ranking Bias, and Second Order Bias.



**Figure 1 Symbolic depiction of Algorithmic Bias**

1. **LITERATURE SURVEY**

This section of the report includes the existing research performed on Algorithmic Bias in Machine Learning.

**2.1 Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods.**

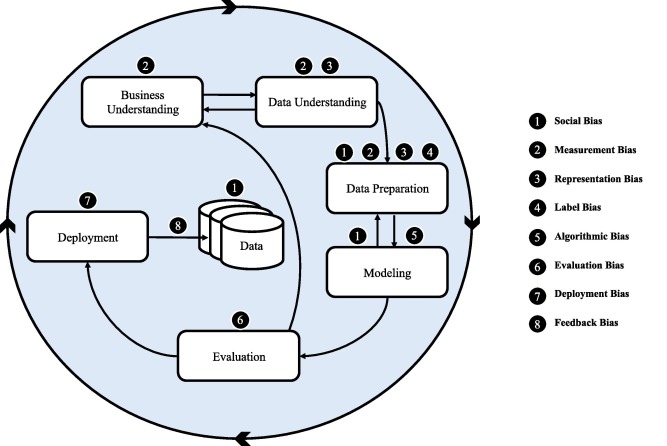
[1] In 2022, Benjamin van Giffen, Dennis Herhausen, Tobias Fahse, performed a systematic, interdisciplinary literature review of machine learning biases as well as methods to avoid and mitigate these biases. They identified eight distinct machine learning biases, summarized these biases in the cross-industry standard process for data mining to account for all phases of machine learning projects, and outlined twenty-four mitigation methods. They further contextualize these biases in a real-world case study and illustrate adequate mitigation strategies. These insights synthesize the literature on machine learning biases in a concise manner and point to the importance of human judgment for machine learning algorithms.

Like humans, ML algorithms are vulnerable to biases that make their predictions and decisions “unfair”. In the context of ML decision-making, fairness is the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics. Thus, a biased and unfair ML algorithm makes decisions that are skewed toward a particular group of people. Although ML algorithms operate in the digital domain, ML biases have many real-world consequences and may cause substantive harm to both consumers and companies. A famous example relates to the Apple credit card, launched in partnership by Apple and Goldman Sachs, which offered lower lines of credit to women than to men of equal or even lower financial standing.

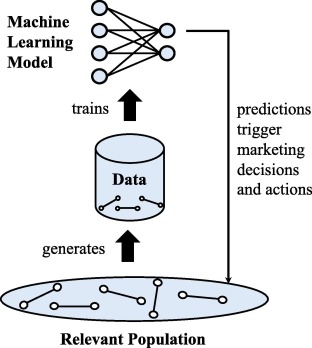
Specifically, the decision-making of algorithms is very different to human decision-making in two important ways: algorithms are extremely literal and they are black boxes. First, while humans understand soft goals and trade-offs, algorithms will pursue a specified objective single-mindedly. For example, Ukanwa and Rust show that for loan decisions, discriminatory results can occur even if there is no bigotry programmed into the algorithm because the algorithm only seeks to maximize profit. Second, algorithms are black boxes in the sense that they can often form predictions with great accuracy, but they do not provide causes or reasons for an event. Thus, algorithms often lack interpretability, in terms of having a transparent model structure and clear linkage between variables.

Moreover, ML algorithms differ substantially from deterministic, rule-based algorithms that have been used in the past for decision support in the organizational context. ML algorithms, such as neural networks, follow a probabilistic approach in which decisions are not made by following programmed rules but by learning patterns from historical data and applying these to new input data. The decision support from ML algorithms is provided in the form of probabilities, leading to different levels of uncertainty and therefore increased susceptibility to systematic biases. For instance, Lambrecht and Tucker have shown that gender bias can occur without any conscious (or unconscious) attempt to produce a biased outcome—using only an unbiased algorithm.

Finally, a series of subjective choices must be made in the process of any ML project, and all these choices may introduce biases and lead to unwanted outcomes. For example, not all relationships within the relevant population necessarily generate data, human coding might determine the data generation process, and the impact of the ML model may reinforce certain patterns in the data generation. As a result, the data does not represent the “whole” relevant population because not all observations and relevant variables are recorded. Even if the data is perfectly unbiased, the decision on how to build and train the model can introduce biases (e.g., selection of unsuitable variables and over- or underfitting during the model training). Even if one assumes the resulting ML application is free from bias introduced through data or design decisions, an inappropriate context of use may nevertheless lead to a bias.



**Figure 2 The different types of biases in different stages of an ML project**



**Figure 3 The origin of Algorithmic bias**

Fig. 2 displays the six phases of the CRISP-DM process model that can be used to plan, organize, and implement an ML project. Anticipating their findings, they also embed the eight ML biases from their review into this figure. Published in 1999 to standardize data mining processes across industries, the CRISP-DM has since become the most common process model for data mining, data analytics, and data science projects.

They conducted a systematic, problem-centred literature review to integrate existing knowledge about ML biases through the conceptual lens of the CRISP-DM model. First, different types of ML biases are identified and consolidated into distinct categories. Second, possible mitigation methods that address these biases are grouped. Third, both ML biases and mitigation methods are incorporated into the different phases of the CRISP-DM model.

 They used the empirical-to-conceptual approach meaning that they started with empirical data clusters and deductively conceptualized the nature of each cluster afterwards. The *meta*-characteristic used for distinguishing the identified biases is the origin (i.e., the occurrence in the CRISP-DM process model) and cause of the bias (i.e., in the data, through humans, in the algorithm, etc.). They stopped iterating when the objective (no bias was merged with a similar bias or split into multiple biases in the last iteration) ending conditions were met.

**2.2 Algorithmic Bias in Education**

[2] In 2022, Ryan S. Baker, Aaron Hawn reviewed algorithmic bias in education, discussed theoretical work on the root causes of algorithmic bias, and reviewed the existing empirical literature on the specific ways that algorithmic bias is known to have manifested in education. In doing so, they distinguish themselves from more algorithmically-focused reviews of algorithmic bias in education. There is a great deal of merit to reviewing mathematical definitions of fairness and algorithmic approaches to reducing bias. Their review focuses instead on understanding exactly who appears to be impacted, and the impact played by the context surrounding the algorithms themselves. They focus in particular on biases emerging from how variables are operationalized and which data sets are used. This review is also distinct from broader discussions of how artificially intelligent technologies can be biased, including in the processes of the design of these technologies, focusing on the narrower issue of bias in the algorithms used to assess and make decisions.

Across categories, the findings of these studies seem to suggest that models trained on one group of learners perform more poorly when applied to new groups of learners. This is not universally true -- for example, there have been conflicting results for urban/rural learners, and the studies conducted across several nationalities often find different nationalities being disadvantaged in different analyses. But in aggregate, the findings suggest that it is problematic to ignore group differences when applying models. The simple expedient of collecting a diverse sample, and training on all students, seems to provide benefits in some cases. It may be that emerging methods for fairness-aware machine learning will lead to considerable improvements, once a representative sample is collected. Even if these methods are highly successful, we will also need to figure out how many members of an underrepresented group are necessary for a combined model to be valid, which remains a challenge in machine learning. The trend in machine learning over the last few decades has largely been to consider ever-larger data sets rather than minimum data set sizes needed.

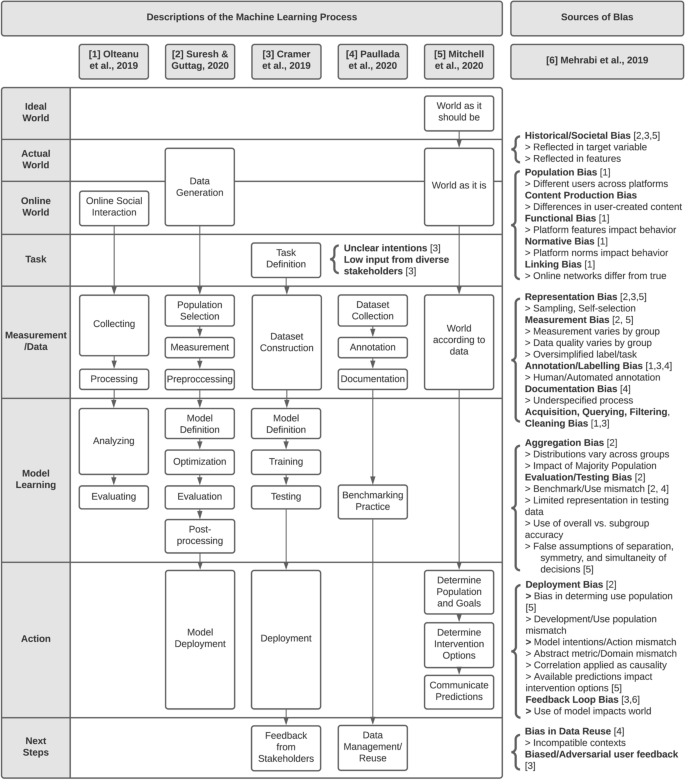
While not discounting the “unreasonable effectiveness of big data”, they note that it is still necessary to determine how many learners of a specific group need to be in a

training set before the model can generally be expected to be reliable for that group.

Another factor that is quickly apparent in looking across studies is the idiosyncrasy of the categories that have been studied. Three core categories have received most of the attention from researchers: race/ethnicity (but not indigenous learners), gender (but not non-binary or transgendered students), and nationality (for a small number of nationalities; in terms of learners’ current locations). A handful of other categories have been discussed in one or two papers. The list of categories that have been studied seems idiosyncratic. To some degree it is based on convenience -- U.S. census categories are relatively likely to be collected, and a learner’s current national location is likely to be known. To some degree it is based on the categorizations that are societally or politically important. To some degree it is based on the biases in what students even make it into the samples -- this may explain to a large degree why indigenous learners are omitted.

Even within the categories that have been relatively heavily studied, there is still considerable idiosyncrasy in the contexts where these categories are studied. For instance, despite the existence of large multi-national datasets involving MOOCs and considerable recent research using MOOCs as a context for conducting research on the differences between learners in different countries, MOOCs have not yet become a widely-used context for studying algorithmic biases involving national difference. Many of the findings discussed above were inconsistent across different studies. It is not yet clear whether this is simply due to noise and random factors, or whether some differences matter more in specific contexts than in other contexts. Fully understanding not only which categories matter, but what their characteristic manifestations are in different contexts, will need to wait until a much larger number of studies have occurred, conducted across a range of contexts.

It is not immediately obvious why some categories have been studied and other categories have not been studied. However, there seem to be effects showing up for a range of groups, suggesting that algorithmic bias likely impacts other groups as well. A broader range of groups need to be more explicitly studied. For instance, children of migrant workers experience many of the same challenges that military-connected students do, such as high personal mobility and concerns about the safety of family members abroad, but have not been studied. Religious minorities have not been studied. Age has not been studied as a factor in undergraduate courses, graduate courses, or professional learning.

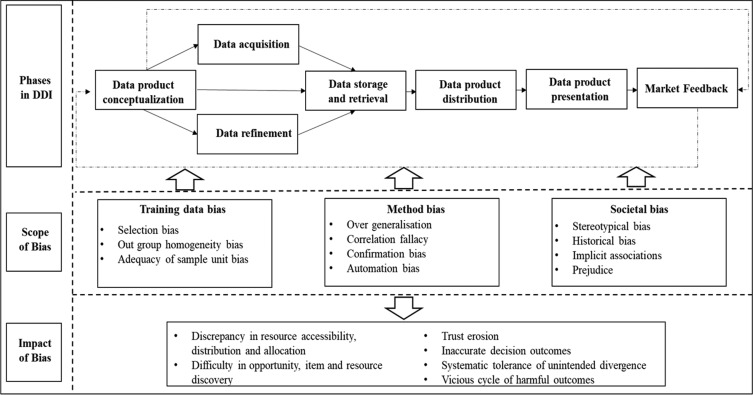


**Figure 4 Descriptions of the Machine Learning Process and Possible Sources of Bias**

**2.3 Algorithmic bias in data-driven innovation in the age of AI**

[3] In 2021, Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K. Dwivedi, John D’Ambra, K.N. Shen aimed to explore the sources of algorithmic biases across the DDI process using a systematic literature review, thematic analysis and a case study on the Robo-Debt scheme in Australia. Their findings show that there are three major sources of algorithmic bias: data bias, method bias and societal bias. Theoretically, the findings of their study illuminate the role of the dynamic managerial capability to address various biases. Practically, they provided guidelines on addressing algorithmic biases focusing on data, method and managerial capabilities.

Advances in algorithmic bias research offer avenues to unmask DDI black-box in the age of AI. The findings of their study show that biases may originate from various sources, specifically data, method and societal factors. Understanding the nature and type of these biases opens exciting research avenues for DDI scholars in developing transparent, explainable and auditable algorithms. Leveraging such algorithms across the development, deployment, and use of data products can help establish.

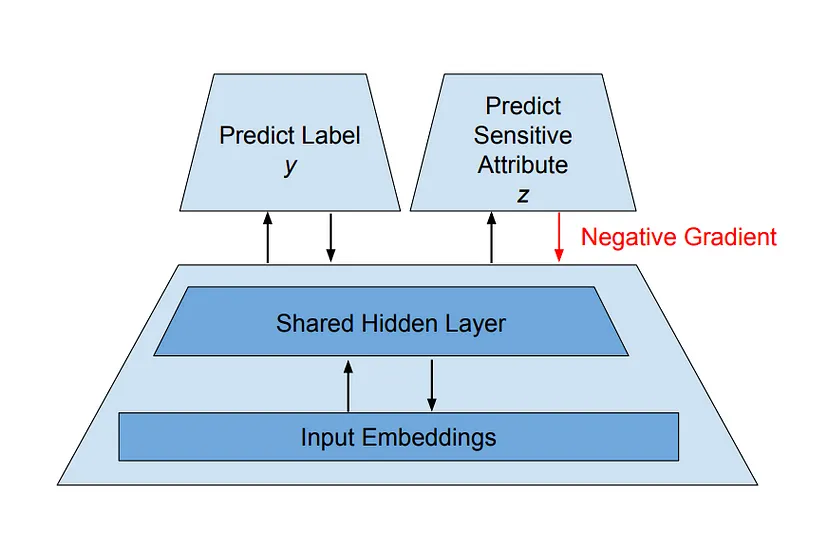


**Figure 5 Phases in Data Driven Innovation, Scope of Bias, Impact of Bias**

1. **METHODOLOGY**

There are four proposed solutions for overcoming algorithmic bias in Machine Learning.

* 1. **ADVERSARIAL DE-BIASING:**
* The technique of adversarial de-biasing is currently one of the most popular techniques to combat bias.
* It relies on adversarial training to remove bias from latent representations learned by the model.
* Let Z be some sensitive attribute that we want to prevent the algorithm from discriminating on, e.g., age or race. It is typically insufficient to simply remove Z from our training data, because it is often highly correlated with other features. What we really want is to prevent our model from learning a representation of the input that relies on Z in any substantial way. To this end, we train our model to simultaneously predict the label Y and prevent a jointly-trained adversary from predicting Z.



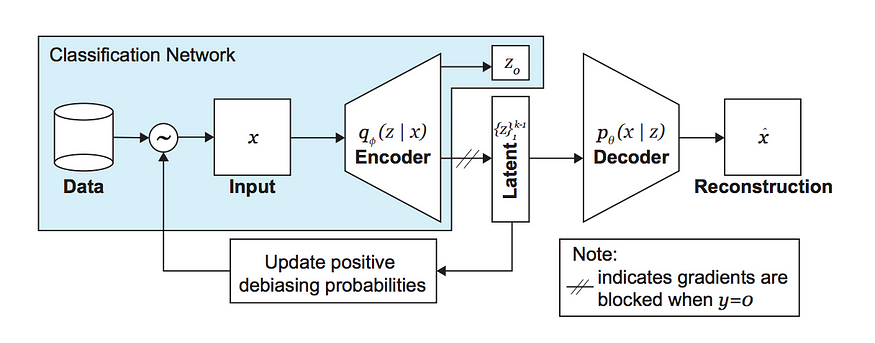
**Figure 6 Architecture used in Adversarial De-biasing**

**3.2 VARIATIONAL “FAIR” AUTOENCODERS**

* Another technique that learns a “fair” representation of data is the Variational Fair Autoencoder (VFAE). VFAE characterizes “fairness” as a representation that is explicitly invariant with respect to some known aspect of the dataset.
* This technique very neatly ties together ideas in ML bias and the broader field of unsupervised and semi-supervised representation learning. Practically, because it is a semi-supervised approach, it can be especially useful in taking advantage of unlabelled data.
* Intuitively, we can think of the problem of learning an “unbiased representation” as recovering the underlying, probabilistic sources of information behind this input, in a way that explicitly separates out the sensitive sources (s) from the invariant ones (z).

**3.3 DYNAMIC UPSAMPLING OF TRAINING DATA**

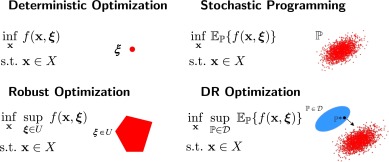
* The first two techniques both modify the learned representation of models. They use outcomes — the ability to either predict or correlate with sensitive attributes — to regularize the learned representation of the data.
* This approach, known as the Debiasing Variational Autoencoder, actually uses learned representations to rebalance the training data. Its premise is quite intuitive: since many modern ML systems fail on certain demographics due to a lack of appropriate representation in training data, let the model learn which inputs come from under-represented groups and sample those inputs more frequently during training.
* The greatest advantage of this approach is that unlike the first two, it does not require you to know or specify the sensitive attributes in your data; the model will learn them automatically as it trains. As a result, the model will also be free to learn more complex and nuanced sources of “under-representation” than a human annotator could easily specify.



**Figure 7 Architecture used in Dynamic Up-sampling of Training Data**

**3.4 DISTRIBUTIONALLY ROBUST OPTIMIZATION**

* This method proposes the use of distributionally robust optimization (DRO), which essentially minimizes the worst-case loss of each group in the dataset, in place of the status quo of empirical risk minimization (ERM), which minimizes average loss in the dataset.
* It is probably the most general of all approaches described: it is model-agnostic, and does not require us to know the identity of the protected group(s) nor their proportions in the dataset.



**Figure 8 Mathematical depiction of Distributionally Robust Optimization**

1. **APPLICATION & DRAWBACKS**

**4.1 Applications:**

Researching on algorithmic bias in machine learning has several important applications, including:

* Mitigating discrimination: One of the most important applications of researching on algorithmic bias in machine learning is to identify and mitigate discrimination in decision-making systems. By understanding the sources of bias in algorithms, researchers can develop strategies to reduce or eliminate this bias, thereby ensuring that these systems are fair and equitable for all users.
* Improving accuracy: Another key application of researching on algorithmic bias in machine learning is to improve the accuracy of predictive models. Bias in algorithms can lead to errors in predictions, which can have serious consequences in fields such as healthcare, finance, and criminal justice. By addressing bias in these models, researchers can improve their accuracy and reliability.
* Enhancing transparency and accountability: Researching on algorithmic bias in machine learning can also help enhance transparency and accountability in decision-making systems. By identifying the sources of bias in these systems, researchers can help make them more transparent and explainable, enabling users to better understand how decisions are made and holding those responsible for these decisions accountable.
* Ensuring ethical AI: Finally, researching on algorithmic bias in machine learning is crucial for ensuring the development of ethical AI. Bias in algorithms can have significant ethical implications, such as reinforcing societal inequalities, perpetuating stereotypes, and violating privacy rights. By addressing bias in AI systems, researchers can help ensure that they are developed and used in ways that are consistent with ethical principles and values.

**4.2 Drawbacks:**

Algorithmic bias in machine learning can have several drawbacks, including:

* Discrimination: One of the biggest drawbacks of algorithmic bias in machine learning is that it can lead to discrimination against certain groups of people. Biased algorithms can perpetuate stereotypes and reinforce societal inequalities, resulting in unfair treatment and opportunities for certain individuals or groups.
* Inaccurate Predictions: Bias in algorithms can also lead to inaccurate predictions, which can have serious consequences in fields such as healthcare, finance, and criminal justice. For example, if an algorithm is biased against a particular demographic, it may not accurately predict their likelihood of developing a disease, their creditworthiness, or their risk of reoffending.
* Lack of Transparency: Biased algorithms can also lack transparency, making it difficult for users to understand how decisions are being made. This lack of transparency can make it difficult to identify and address bias in algorithms, which can lead to further discrimination and inaccuracies.
* Limited Scope of Data: Algorithms that are biased may only be trained on a limited set of data, leading to incomplete or inaccurate models. This can result in the algorithm missing important features or factors that should be considered when making a decision, leading to suboptimal outcomes.
* Negative Impact on Society: Finally, algorithmic bias can have a negative impact on society as a whole, perpetuating and amplifying systemic inequalities and reinforcing discrimination. This can lead to a lack of trust in technology and can hinder progress towards creating a more equitable and just society.

Overall, it is important to address algorithmic bias in machine learning to ensure that these systems are fair, accurate, and equitable for all users.

1. **CONCLUSION**

Machine Learning sometimes incorporates inadequate characteristics that produce outcomes that are both technically and societally unsatisfactory. Aside from performance standards like accuracy, dependability, and efficiency, bias reduction ought to be a key component of any ML application. Emphasis is also given to the economic aspect of bias, which is different from the predominant discussion on bias, which focuses mostly on discrimination based on race, gender, religion, or belonging to a social minority. There is an urgent need to manage bias in ML projects. It involves more than just ensuring fairness; it also entails ensuring that ML in commercial settings generates sustained economic value.

**5.1 Future Scope:**

The future scope of research on algorithmic bias in machine learning is vast and expanding rapidly. Some potential areas of focus for future research include:

* Intersectional Bias: One area of future research could be to investigate the intersectional nature of bias in machine learning algorithms. Intersectionality refers to the multiple dimensions of identity that overlap and intersect, such as race, gender, and socioeconomic status. Researchers could explore how these multiple dimensions interact to create bias in algorithms and how to address them effectively.
* Fairness Metrics: Another potential area of research could be to develop better fairness metrics for evaluating algorithms. Currently, there is no universally agreed-upon definition of fairness in machine learning, and different fairness metrics may prioritize different aspects of fairness. Developing better metrics for evaluating fairness could help researchers and practitioners better understand and address bias in algorithms.
* Human-in-the-Loop Approaches: Human-in-the-loop approaches involve incorporating human feedback into machine learning algorithms to improve their accuracy and reduce bias. Future research could focus on developing more effective and efficient ways of incorporating human feedback into algorithms and evaluating the impact of these approaches on reducing bias.
* Explainability: Explainability refers to the ability to understand and interpret the decision-making process of machine learning algorithms. Future research could focus on developing more transparent and interpretable algorithms that can help users better understand and address bias in their systems.
* Ethical Considerations: Finally, future research on algorithmic bias in machine learning could focus on the ethical considerations of using these technologies. As AI becomes increasingly integrated into our lives, there will be a need to ensure that these technologies are developed and used in ways that are consistent with ethical principles and values. Research in this area could explore the ethical implications of bias in algorithms and how to develop more ethical AI systems.

1. **REFERENCES**

 [1] *Benjamin van Giffen, Dennis Herhausen, Tobias Fahse, “Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods”, Elsevier - Journal of Business Research Volume 144, May 2022 - https://doi.org/10.1016/j.jbusres.2022.01.076.*

[2] *Ryan S. Baker, Aaron Hawn, “Algorithmic Bias in Education”, International Journal of Artificial Intelligence in Education (2022) 32:1052–1092 - https://doi.org/10.1007/s40593-021-00285-9.*

[3] *Grace McCarthy, “Algorithmic bias in data-driven innovation in the age of AI”, Elsevier - International Journal of Information Management Volume 60, October 2021.*

[4] *Shahriar Akter, Yogesh K. Dwivedi, Shahriar Sajib, Kumar Biswas, Ruwan J. Bandara, Katina Michael, “Algorithmic bias in machine learning-based marketing models”, Elsevier - Journal of Business Research Volume 144, May 2022 - https://doi.org/10.1016/j.jbusres.2022.01.083*