Fairness and Inclusion in AI

An Interactive Qualifying Project Report:

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

Submitted by:

Sizhe Li

Project Advisors:

Professor Therese Mary Smith

Submitted: 10 April 2023

Abstract

The advent of artificial intelligence has ushered in a fourth revolution in human cognitive abilities. However, despite the enormous potential, AI systems can exhibit algorithmic biases that can lead to unfair competition, algorithmic discrimination, and abuse. These biases come from different sources, including subjective factors and cognitive biases of algorithm designers, data bias, and the opacity of the algorithm. In addition, AI algorithms are not value-neutral; they reflect the values of their creators.

To address algorithmic bias, while going through the technical aspects of optimizing algorithms, we must also hold AI algorithm designers and users accountable for their actions, ensure the accuracy of data mining, and improve the transparency and interpretability of algorithms.

Acknowledgments

Thank you to my advisor Professor Therese Mary Smith for her support and guidance through this project. I would like to thank the Computer Science departments, as well as WPI for their role in supporting this project. I would also like to thank Professor Shue for reviewing this report.

Contents

1.1 What is AI?

The core of artificial intelligence is big data and algorithms. Artificial intelligence is based on algorithm-based big data analysis to discover the structure or patterns hidden behind the data, thus realizing data-driven artificial intelligence decisions. The development of artificial intelligence such as face recognition, language recognition, machine learning, machine translation, deep learning, and driverless cars all rely on algorithms. Artificial intelligence technology affects both the mental and external worlds of people and algorithm-dependent AI plays a revolutionary role in self-awareness and social progress.

Machine learning and deep learning are the most popular branches of AI, and deep learning algorithms depend on data, of which there are many. Add images to millions of tagged cat pictures, for example, using an image classification algorithm, which can tell you if a previously unseen photo contains a cat. Provide a speech recognition algorithm with millions of speech samples and their corresponding written words, and it will be able to transcribe spoken words faster than most people. The more labeled data the algorithm sees, the better it performs the task. The trade-off with this approach, however, is that deep learning algorithms will develop blind spots based on missing or overly rich data in the training data. Finding and responding to biases becomes more difficult in other situations where AI is processing vast amounts of available data in an endless sea of online information.

1.2 Analysis of the performance of artificial intelligence algorithm bias

According to artificial intelligence algorithmist Thomas H. Cormen and others, "an algorithm takes a value or a set of values as input and generates a value or a set of values as output. Thus, an algorithm is a series of computational steps that convert an input into an output" [1]. Algorithms as computer programs are driven by pure mathematical logic, and they are impartial and objective, but there are algorithmic biases in the development of artificial intelligence. There is no uniform definition of algorithmic bias in the academic community, but algorithmic bias generally refers to the unfairness in the output of a computer program or artificial intelligence system due to the inclusion of implicit human values in the process of data collection, data selection, and use. The algorithmic bias is implicit in the algorithm, and we can grasp the algorithmic bias through the external manifestation of the algorithmic bias, which is mainly manifested in the following aspects:

1.2.1 Algorithm bias manifests itself as algorithmic discrimination

Discrimination refers to the unfair treatment caused by the bias or prejudice of the cognitive subject, for example, different race, gender, religion, and residence may lead to racial discrimination, gender discrimination, religious discrimination, and geographical discrimination. The phenomenon of algorithmic discrimination caused by algorithmic bias is widespread, and the artificial intelligence algorithm may also imply racial discrimination, class discrimination, and gender discrimination.

For example, in 2016, Tay, a chatbot developed by Microsoft that mimics human conversations by capturing data from user interactions, spread racist messages; in 2012, a Harvard University study published in the Journal of Social Issues showed that searches for traditional African American names were more likely to show arrest records; and in 2016, a study published in the Journal of Social Issues showed that searches for traditional African American names were more likely to show arrest records. In 2016, a ProPublica survey found that artificial intelligence tools used by law enforcement showed that blacks were more likely to commit crimes than whites. All of these artificial intelligence algorithms contain some sort of racial discrimination.

If you're a dark-skinned person, you may be more vulnerable to self-driving cars than your white friends, according to a new study from the Georgia Institute of Technology. That's because autonomous vehicles can better detect pedestrians with lighter skin tones. The authors of the study began with a simple question: How accurate are state-of-the-art object detection models, such as those used by self-driving cars, at detecting people from different demographic groups? To find the answer, they looked at large datasets of images containing pedestrians. They segmented people using the Fitzpatrick scale, a system that classifies human skin color from light to dark. The researchers then analyzed how often these models correctly detected the presence of people in the light-skinned group, and how correctly they related to people in the dark-skinned group. The results? For the dark-skinned group, the detection was on average 5 percentage points less accurate. This difference persisted even when the researchers controlled for variables such as time in the image or occasionally obstructed pedestrian sight lines. [2]

One of the most well-known effects is still on facial recognition algorithms: a facial recognition algorithm may be trained to recognize whites more easily than blacks because this type of data is used more frequently in training. This can have a negative impact on people from minority groups, as discrimination prevents equality of opportunity and perpetuates oppression. The problem is that these biases are not intentional and it is difficult to know them until they are programmed into the software.

Another example is the introduction of UBI (Usage Based Insurance) insurance by European insurance companies, which is suspected of class discrimination [3]: UBI car insurance is an algorithmic analysis based on the actual driving time, location, mileage, and specific driving behavior of the driver, and introduces a differentiated insurance plan.

Algorithm bias can also lead to gender and even age discrimination: A 2015 study by Carnegie Mellon University's International Computer Science Institute (ICSI) showed that gender discrimination lurks in the ad targeting algorithm created by Google. Men are more likely to receive higher-paying ads, while women are far less likely to receive recommendations for higher-paying careers than men. Researchers at Microsoft and Boston University also found sexism in some artificial intelligence algorithms, with words such as "programming" and "engineering" often associated with men, and "housewife" and "domestic worker" often associated with men. Words such as "programming" and "engineering" are often associated with men, and words such as "housewife" and "housework" are often associated with women. Machine learning systems may potentially detect statistical associations that are considered socially inappropriate or illegitimate. For example, a mortgage model may determine that older people have a greater probability of default, reducing their creditworthiness. If the model draws this conclusion based on age alone, then we may see illegitimate age discrimination. [4]

Algorithmic discrimination is an outward manifestation of algorithmic bias, and while algorithmic discrimination is many times not intentional for the designer, it is undeniable that algorithmic discrimination is primarily triggered by the subjective cognitive biases of the algorithm designer.

1.2.2 Algorithm Bias Leads to Unfair Competition

Artificial intelligence algorithms rely on big data, and data resources have become capital with strong economic value, which has become a key factor for the profitability of data companies, and data owners can generate more economic benefits than those who do not own data, leading to unfair competition. Data is generated by users, but data is often not possessed and used by data users. The algorithm bias based on unfair ownership and unfair use of data is manifested as unfair data possession and unfair data mining. For example, many Internet vendors now have big data-enabled price discrimination against existing customers: old customers see the same goods or services at a much more expensive price than new customers. For example, in the process of booking a room or ticket, users with Apple phones pay more than those with Android phones, older users pay more than new users, and VIP users pay more than regular users. Different operators have different levels of data mastery, data analysis capabilities and forecasting levels, resulting in unfair competition among companies. Computers can analyze users' consumption behavior, habits, and ability through algorithms, calculate their price sensitivity and dependence, and predict the price ceiling of consumers, thus raising prices for old customers and pricing differently for different customers.

1.2.3 Algorithm bias manifests itself as algorithm abuse

Algorithm abuse occurs when a person or organization uses an algorithm that harms others, violates ethical principles, or goes against the intended purpose of the algorithm. Here are some examples of algorithm abuse in various areas of life: Algorithm abuse occurs when a person or organization uses an algorithm that harms others, violates ethical principles, or goes against the intended purpose of the algorithm. Here are some examples of algorithm abuse in various areas of life:

Algorithmic trading has been blamed for exacerbating market volatility and contributing to financial crises. In 2010, high-frequency trading algorithms contributed to the "flash crash" that saw the Dow Jones Industrial Average plunge nearly 1,000 points in a matter of minutes.[5] Algorithmic risk assessments used in criminal justice have been criticized for perpetuating racial bias and leading to discriminatory outcomes. In 2016, ProPublica found that a widely used algorithm for predicting future criminal behavior was more likely to flag Black defendants as high risk than white defendants.[6] In 2020, an algorithm used in healthcare in the UK to predict which patients were at risk of developing sepsis was flawed, leading to potentially fatal consequences.[7]

The algorithm is the basis of AI technology development, but the bias of the algorithm is not objectively neutral, it contains the subjective bias and objective bias of the algorithm designer, and the algorithm should not be applied as an objective criterion, and the application of AI technology should not be expanded blindly due to superstition. The misuse of algorithms can neither guarantee the objectivity of algorithms nor the accuracy of their results.

2. Traceability of causes of artificial intelligence algorithm bias

The causes of AI algorithm bias include subjective reasons of algorithm designers, data-dependent reasons, and objective reasons such as the opaqueness of algorithms. In the following, we analyze the causes of algorithm bias from three aspects.

2.1 Algorithm bias due to subjective reasons

Bias in artificial intelligence algorithms is caused by subjective bias, cognitive bias, or cognitive limitations of the designer, which are subjective factors that cause algorithm bias. In the editorial "More accountability for big-data algorithms" in Nature, it is suggested that "bias in, bias out" [8], biased subjects or data lead to biased algorithms. The bias caused by subjective factors is manifested in the following ways:

The first is the bias caused by the subjective bias of the algorithm designer: the subjective factor of algorithm bias is mainly caused by the algorithm designer, who artificially designs biased and discriminatory algorithms in order to achieve certain purposes and obtain certain benefits, thus leading to algorithm bias. Any algorithm is designed by the designer to achieve a certain purpose, and it reflects the designer's intention to avoid the algorithmic bias of artificial intelligence. For example, in 2016, Cambridge Analytica obtained Facebook data and used "precision marketing" algorithms to influence voters' political choices and help the Trump team win the U.S. presidential election. Technical bias in AI algorithms is easy to avoid, but it is difficult to eradicate the influence of the algorithm designer's own bias or social bias, which makes it difficult to exclude bias in AI algorithm design.

Second, the algorithm bias is caused by the cognitive bias of the algorithm designer. A cognitive bias refers to the distortion of a subject's perception of himself/herself, other people, or the external environment due to self or contextual reasons. For example, Survivorship Bias is a common logical fallacy in which the algorithm designer only looks at the statistical results of the data and ignores the data mining process, leaving out valid or critical data. This also occurs when the algorithm designer has a biased perception of things, or the algorithm operates with biased data, resulting in deviations from objective facts.

Third, the subjective cognitive limitations of the algorithm designer lead to bias. The algorithm designer also has knowledge blindness and cognitive limitations, and such cognitive limitations will accompany the whole process of algorithm design and inevitably lead to algorithm bias. The design purpose, data use, and result representation of AI algorithms reflect the value choices of the algorithm designers, who may unconsciously embed their own biases into the AI algorithms. The algorithms are not value-neutral, because different designers have different upbringings and educational backgrounds, and have formed their own relatively independent values, which determines that different algorithm designers have different interests and value judgments, and form deep-rooted value biases and habitual views, which lead to the deviation of different algorithms designed. The algorithm is influenced by the designer's subjective factors, the algorithm design is biased, and the algorithm is loaded with the designer's values and biases.

2.2 Algorithm bias caused by bias on data.

Artificial intelligence algorithms are data-driven, and accurate and comprehensive data can guarantee the authenticity and accuracy of AI algorithm results; conversely, false data, missing data, and data pollution can lead to data bias. The existence of discrimination is partly due to the bias of decision makers, but it should not be overlooked that much of the inequality in modern society should be attributed to "data discrimination", which can easily lead to algorithmic bias if data is contaminated or incomplete. If the data-driven data sources are contaminated, inaccurate, and biased, then the algorithms will also present biased results.

As mentioned above, "bias in, bias out", if the data itself is biased, the algorithm will only produce biased results, no matter how objective and fair the designer is. Artificial intelligence systems learn from historical data, and artificial intelligence algorithms may encode historical bias. When the data being mined is itself the result of intentional discrimination in the past, there is usually no obvious way to adjust the historical data to remove this contamination, and the data is mined to inherit the biases of previous decision makers or to reflect the biases prevalent in society. When data is incomplete or inaccurate, it may reflect societal biases, and the algorithm may similarly perpetuate the biases of previous decision makers. For example, if an employer has consistently declined to hire female or black applicants in previous hiring decisions, the computer may algorithmically discriminate against some female or black applicants; or if the data show that long commutes are associated with high turnover, the hiring algorithm may negatively rate people who commute longer.

The above data discrimination is almost an unintentional feature of algorithm use rather than a conscious choice by the algorithm programmer, but it also results in discrimination. Data mining algorithms eliminate human bias but introduce new algorithmic biases. In a data-driven world, algorithms are a source of both strengths and weaknesses. Artificial intelligence algorithms rely on data, and the quality of the data extracted from a large collection of data greatly affects the accuracy of the algorithm.

2.3 Algorithmic bias due to opacity of algorithms

The opacity of an algorithm means that the designer converts input into output using a process that is not transparent to anyone but the developer. Users have no way of knowing how their data is being processed, and they can only unilaterally trust that the developer has designed their program in a fair and equitable manner. This is known as the Algorithmic Transparency dilemma in artificial intelligence technology.

A variety of algorithms are used in artificial intelligence, such as pattern recognition algorithms, natural language processing algorithms, data mining algorithms, and so on, which follow a set of instructions or learn to accomplish a goal. In theory, AI algorithms can help humans make unbiased analyses and decisions by reducing human short-sightedness and bias, so they have been considered impartial and objective, but the opacity of algorithms makes it difficult to guarantee objectivity. Transparency as a private company or state secret; opacity as a non-technical person; and opacity due to the nature of the machine learning algorithm and the scale required. [9] Public information often allows unscrupulous individuals to have a very bad impact on a company or its business or even reputation, so they are unlikely to choose to keep their systems transparent, especially since many of these systems touch on corporate privacy or even trade secrets: many search engines, such as Google, can effectively provide personalized human recommendations to people when they look up an entry, and most companies prefer to keep the Most companies tend to hide the inner workings of their personalized push algorithms as trade secrets, resulting in websites with algorithmic biases in their human push messages. The complexity of artificial intelligence algorithms and the bias with which they are applied make them difficult to interpret.

In AI systems, human subjective biases and cognitive limitations, inaccurate and incomplete data sources, and opaque algorithms can lead to algorithmic biases. Therefore, algorithmic biases need to be eliminated and avoided from both subjective and objective dimensions.

3. Avoidance measures of artificial intelligence algorithm bias

To avoid algorithmic bias, first, it is important to acknowledge the limitations of AI, which is not omnipotent and not absolutely neutral. Second, AI algorithm bias is not inherent, but AI algorithms can intentionally or unintentionally embed human bias in the process of data mining, acquisition, and application. If we want to avoid AI algorithm bias, we should emphasize the responsibility of AI algorithm subjects for both subjective and objective reasons AI algorithm bias, ensure the accuracy and completeness of AI algorithm mining data, as well as improve the transparency and interpretability of AI algorithm.

3.1 Emphasis on the responsibility of the AI algorithm subject

Technology neutrals argue that algorithms are value-free entities, that algorithms as technologies are merely tools for human transformation and do not carry human values, that they are neutral, and that the fact that an algorithm or system produces a value judgment as its output does not prove that the algorithm used to produce the value judgment is inherently valuable.

But an emerging school of thought offers a relatively pointed view: they argue that algorithms are loaded with human value and that both the designers of algorithms, data owners, and users of data and algorithms are driven by value claims and interests, which leads to the idea that if algorithms themselves are valuable, then those who design them should have a corresponding moral responsibility for the algorithms they design, even if this means that those who design them should have a minimum level of judgment. that those who design algorithms should have the most basic ethical standards for judging right and wrong. [10] This school of thought is known as algorithmic ethics. This school of thought also proposes that avoiding algorithmic bias requires the promotion of effective Algorithmic accountability.

Algorithmic accountability refers to the idea that those who develop and deploy algorithms are responsible for ensuring that those algorithms are transparent, explainable, and accountable. This means that those who create and use algorithms must be able to explain how they work, how they make decisions, and how they affect individuals and society as a whole.

Algorithmic accountability is particularly important when algorithms are used to make decisions that have a significant impact on individuals or groups. For example, algorithms used in hiring, credit scoring, or criminal justice may have a significant impact on people's lives and therefore require a high degree of accountability.

To ensure algorithmic accountability, there are several principles that should be followed. First, algorithms should be transparent, meaning that their inputs, outputs, and decision-making processes should be clear and understandable. Second, algorithms should be explainable, meaning that their decision-making processes should be able to be explained in a way that is understandable to non-experts. Third, algorithms should be auditable, meaning that they should be subject to testing, evaluation, and verification to ensure that they are functioning as intended.

3.2 Guarantee the accuracy of the acquired data

Even if we take all the human subjectivity out of the equation, the algorithm itself processes the data - then the ultimate effectiveness of the algorithm depends heavily on the data it processes. Data is critical in today's digital age. It underpins artificial intelligence (AI), machine learning, and other technologies that have transformed the way we live, work and communicate with each other.

Data provides the information that is needed to make informed decisions. Whether it's a business decision, a medical diagnosis, or a policy decision, data is essential for making decisions that are accurate, effective, and reliable. Data fuels innovation by providing the raw material that is needed to develop new products and services. Companies can use data to identify trends, patterns, and insights that can be used to create new products and services that meet the needs of their customers. Data can help to improve efficiency by providing insights into how processes can be optimized and streamlined. This can lead to cost savings, improved productivity, and better customer experiences. Data can be used to personalize products and services to the specific needs and preferences of individual users. This can lead to better customer experiences, increased engagement, and greater loyalty.

Given the importance of data, the accuracy and comprehensiveness of the data used to train AI can also effectively circumvent algorithmic errors and thus ensure algorithmic objectivity: AI algorithms often rely on large amounts of data for training, and the data collection is a conscious choice by the algorithm designer. Algorithm bias is inevitable when the data set reflects the implicit value of the humans involved in data collection, selection, or use. Algorithms are data-based and rely on human-selected data; therefore, they are formative rather than descriptive, and we can avoid algorithmic bias by ensuring the accuracy of the data.

3.3 Improving the transparency and interpretability of algorithms

In many cases, algorithms operate as "black boxes," meaning that it is difficult or impossible to understand how they arrived at a particular decision or recommendation. In many cases, algorithms operate as "black boxes," meaning that it is difficult or impossible to understand how they arrived at a particular decision or recommendation. For AI technologies to gain the trust of industry and government, they need to be more transparent and explain their decision-making processes to users. For algorithmic applications to be secure, it is also necessary to provide accurate explanations of algorithms, to give clear reasons for algorithmic decisions, to provide explanations of algorithmic processes and specific decisions by the systems and institutions that use them, and to break down technical black boxes through technical means to ensure the interpretability and security of algorithms.

For algorithmic transparency, as Diakopoulos Nicholas proposed in "Algorithmic Transparency in New Media", "we define algorithmic information disclosure as allowing interested parties to monitor, inspect, criticize, or intervene" [11], the improvement of algorithmic transparency as a way to achieve The ethical assumptions in algorithms should be at least transparent and easily recognized by users, eliminating unfair, discriminatory or biased effects of algorithmic applications through human intervention and control of algorithms, reducing information asymmetry, accurately describing the decision-making process and learning mechanism of algorithms, and ensuring the objectivity of algorithms.

Explainability of an algorithm is "the explanation of why certain characteristics of an AI algorithm's inputs cause a particular output result". Researchers and practitioners of explainable AI are encouraged to collaborate with researchers and practitioners in the social and behavioral sciences to inform model design and human behavior experiments. The Defense Advanced Research Program Agency (DARPA), part of the U.S. Department of Defense, has been working on Explainable Artificial Intelligence (XAI), and this research is seeking to develop new machine learning systems that will be able to explain their rationale. Importantly, they will also translate models into understandable explanations that are useful to end users. [12] Such research emphasizes collaboration between researchers and practitioners in artificial intelligence and those in the social and behavioral sciences to inform model design and human behavior experiments.

"Human in the loop" is also a solution that allows regulators to intervene in the algorithm: refers to a process or system in which a human operator or decision maker is involved to provide oversight, guidance, or intervention at some point during the process. It is often used in the context of artificial intelligence (AI) and machine learning (ML), where human expertise and oversight are required to ensure the accuracy and fairness of automated decision-making processes.

In a "human in the loop" system, an algorithm or machine learning model is used to analyze and process data, but a human is also involved to review the results and provide feedback or make decisions based on the output. For example, in a customer service chatbot, the chatbot may be able to handle simple queries, but a human operator may be needed to handle more complex or sensitive issues that require empathy and nuanced understanding.

The "human in the loop" concept is often used in contexts where automation is not sufficient or where the consequences of incorrect or biased decisions could be severe. It allows for the benefits of automation to be combined with human judgment, expertise, and empathy.

As Arne Wolfewicz mentions in "Human-in-the-Loop in Machine Learning: What is it and How Does it Work?": the purpose of humans in the loop is to achieve goals that neither humans nor machines can achieve on their own. When a machine is unable to solve a problem, humans need to step in and intervene. The result of this process is the creation of a continuous feedback loop. With constant feedback, the algorithm learns and produces better results each time. [13] Typically, there are two machine learning algorithms that can integrate HITL methods. These include supervised learning and unsupervised learning. In supervised learning, experts use labeled datasets to train the algorithm to produce appropriate features. These can then help map new examples. Doing so will allow the algorithm to correctly determine the function of the unlabeled data. In unsupervised learning, the unlabeled datasets are fed into the algorithm.

Therefore, they need to learn on their own to find a structure in the unlabeled data and remember it accordingly. This belongs to the "human in the loop" deep learning approach.

4. Conclusion

The widespread use of artificial intelligence (AI) has given rise to concerns about its potential to perpetuate and even amplify human biases. AI algorithms are designed to make decisions based on patterns and trends in data, and their accuracy is crucial in ensuring the fairness and objectivity of AI output. However, the very data that these algorithms rely on is often biased, reflecting the prejudices and assumptions of those who collect and interpret it. Moreover, the complexity of AI algorithms and the "black box" nature of their decision-making processes make it difficult to identify and correct for bias.

One of the main challenges in developing ethical and unbiased AI is addressing the fundamental assumption that algorithms are objective and free of bias. Many people trust the judgments of AI algorithms, assuming that they are purely rational and unbiased. However, as with any technology, the accuracy of AI algorithms depends on the quality of the data they are trained on and the biases of the humans who design and implement them. This means that even the most sophisticated AI algorithms can be vulnerable to the same biases and prejudices that affect human decision-making.

To address these challenges, there is a need for greater transparency and accountability in AI development. This includes ensuring that the data used to train AI algorithms is diverse and representative and that the algorithms themselves are designed with ethical and moral principles in mind. It also means making AI algorithms more interpretable and understandable, so that their decision-making processes can be scrutinized and audited for potential bias.

Another important step in addressing algorithmic bias is to shift the focus from individual algorithms to the broader social and ethical contexts in which they operate. This means acknowledging that AI is not a neutral technology, but one that is shaped by social, cultural, and economic factors. It also means recognizing that AI has the potential to perpetuate and exacerbate existing inequalities and injustices, particularly in areas such as criminal justice, employment, and healthcare.

Ultimately, addressing algorithmic bias requires a multi-disciplinary approach that draws on the expertise of computer scientists, ethicists, social scientists, and other stakeholders. It also requires a commitment to ongoing monitoring and evaluation of AI systems to ensure that they are operating in ways that are ethical, transparent, and aligned with human values. By taking these steps, we can create a future in which AI is used to promote human well-being and social justice, rather than perpetuating and amplifying existing biases and inequalities.

References

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Introduction to Algorithms," 3rd ed. Cambridge, MA: MIT Press, 2009, pp. 5-6.

[2] S. Samuel, “Study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians,” *Vox,* Mar. 05, 2019. https://www.vox.com/future-perfect/2019/3/5/18251924/self-driving-car-racial-bias-study-autonomous-vehicle-dark-skin

[3] M. Nørskov, J. Seibt, and O. S. Quick, *Culturally sustainable social robotics: Proceedings of Robophilosophy 2020, August 18-21, 2020, Aarhus University and online*. Amsterdam: IOS Press, 2020, pp. 395-406.

[4] Z. Larkin, “AI bias - what is it and how to avoid it?,” *levity.ai,* Nov. 16, 2022. https://levity.ai/blog/ai-bias-how-to-avoid

[5] Leinweber, D. J. (2011). The flash crash: The impact of high frequency trading on an electronic market. Journal of Portfolio Management, 37(1), 19-24.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine Bias,” *ProPublica,* May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[7] B. Clover, D. West, and J. Illman, "Oxygen supply problems 'the new PPE', warn hospital bosses," Health Service Journal, Apr. 6, 2020. [Online]. Available: https://www.hsj.co.uk/quality-and-performance/flawed-algorithm-led-to-deaths-in-sepsis-detection-tool-trial/7027333.article. [Accessed: Apr. 13, 2023].

[8] “More accountability for big-data algorithms,” *Nature*, vol. 537, no. 7621, pp. 449–449, Sep. 2016, doi: https://doi.org/10.1038/537449a.

[9] Burrell J. , How the Machine"Thinks" : Understanding Opacity in Machine Learning Alooritms, Social Science Electronic Publishing ,3 (1) ,2015 ,pp. 1 - 12.

[10] F. Kraemer, K. van Overveld, and M. Peterson, “Is there an ethics of algorithms?” *Ethics and Information Technology*, vol. 13, no. 3, pp. 251–260, Jul. 2010, doi: https://doi.org/10.1007/s10676-010-9233-7.

[11] Diakopoulos N. , Koliska M. , “Algorithmic Transparency in the News Media”, 2016, https://doi.org/10.1080/21670811.2016.1208053

[12] M. Turek, “Explainable Artificial Intelligence,” *Darpa.mil*, 2018. https://www.darpa.mil/program/explainable-artificial-intelligence

[13] A. Wolfewicz“Human-in-the-loop in machine learning: What is it and how does it work?,” *levity.ai*. https://levity.ai/blog/human-in-the-loop