DOI: 10.5281/zenodo.13294739

Interrogating Algorithm Bias in Machine Learning and Moral Relativism with Kantian Principle of Universality

Saheed A. Agunbiade

ABSTRACT

One of the contemporary ethical issues in artificial intelligence has to do with the problem of algorithm bias. Algorithm bias involves favouritism or unfairness towards certain groups of people through computational means; it happens when the data used to train the algorithm is biased or when the algorithm itself has biases. When an algorithm is built on biased data and/or in biased environments, their outputs can put certain groups of individuals at a disadvantage. Because the machine learning process is long and complex, bias can creep into the system at every stage and can be difficult to identify even by artificial intelligence practitioners. These biases have been relatively described as that which only acknowledged the dignity of a group, not humanity. It is on the basis of this that the paper critically interrogates the ethical relativistic problems that are embedded in the algorithm bias. It uses Kant's moral principle of universality as a theoretical framework to address the problem. While following Kant's principle of universality, the paper holds that the establishment of a machine learning algorithm should be based on the dignity of the individuals, not on the basis of a group. It argues that algorithm biases can be mitigated in artificial intelligence if not abolished, when attention is being placed on human dignity, not on related interests of a group.

Keywords: *artificial intelligence, machine learning, algorithm, algorithm bias, Kant, moral relativism.*

INTRODUCTION

The growth of artificial intelligence and other technological entities has transformed the face of the global world both directly and indirectly, causing attention to shift away from solely focusing on the moral issues pertaining to humans. Artificial intelligence is being used to solve both human and business problems, and these have impacts not only on industries but on the sociocultural and moral lives of the people. Artificial intelligence presents chances that are both transformational and sustain-

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

able. Examples of these opportunities include personalizing social media advertisement content, automating administrative activities, and enhancing financial models (Hao, 2020). One of the subfields of artificial intelligence is machine learning, which involves analyzing vast amounts of data to identify patterns.

The system is then able to classify and predict data by recognizing these patterns and repeatedly "learning" from it. But, biased information based on sociocultural biases and ideas at a certain period permeates much of human history. Machine learning has a tendency to yield biased outcomes if historically biased data is employed in the system. Also, algorithms may become biased as a result of the individuals and places in which they are developed. "Garbage in, garbage out" is a well-known saying that statistics students have likely heard before. Systems built on machine learning can be found everywhere in daily life. Perhaps more importantly, systems built on machine learning can be found in high-stakes decisionmaking processes such as medical diagnoses, job candidate selection, and the criminal jus-tice system (Yapo, 2018: 5365).

However, the biases found in the machine learning algorithm have been attributed to the morally relativistic principles that are embedded in the sociocultural and geographical differences. These principles are basically reflected in the output of the machine learning algorithm. It is on the basis of these problems, that the paper at-tempts to examine the problems of moral relativism in the algorithm bias. It attempts to show that algorithms should be established on the dignity of the individuals, not on the sociocultural influences or contribution; the ontological nature of man should play the key role in the establishment of machine learning algorithms since all societies develop at their own pace. Nevertheless, this paper argues that the problem of algorithm biases can be mitigated if not abolished, when the individuals are treated as ends in themselves, not as mere means; the problem can be mitigated when the dignity of the individual is placed ahead of the socio-cultural differences which reflect in the establishment of machine learning algorithms.

RESEARCH METHODOLOGY

In this paper, two philosophical methods are employed, which are comparative and evaluative methods. The former is used to compare the historical development of algorithms in order to show how moral biases have been developed in machine learning algorithms. This is used to justify the need for the mitigation, if not abolished, of algorithm biases in machine learning. While the latter is used to evaluate the algorithm biases in order to find a lasting solution to the problem; the method is used to show that the sources of algorithm biases should be traced to the differences in the sociocultural values of the societies and the way this can be addressed is by treating the individuals as an end in themselves not as a mere means.

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

ALGORITHM AND MACHINE LEARNING

The word "algorithm" actually comes from the name of a Persian mathematician named Al-Khwarizmi. He lived around the 9th century and made significant contributions to mathematics and astronomy. The term *algorithm* is derived from his name, Al-Khwarizmi, and it refers to the systematic process of solving problems using a set of defined rules or steps (George et al., 2000: 114). So, we can say that algorithms have been around for a long time and continue to shape our modern world. Algorithms are like step-by-step instruc-tions that computers follow to solve problems. It helps us to automate tasks, analyze data, and make decisions faster. For example, algorithms are used in search engines, recommendation systems, and even social media feeds to show relevant content (George et al, 2000: 114). Hence, an algorithm is like a set of instruc-tions that tells a computer what to do. It is like a recipe, but for computers instead of cooking, algorithms are used in all sorts of things we do every day, like searching for information on the internet, recommending movies or music, or even driving directions. They help computers solve problems and make decisions quickly and ac-curately.

The concept of algorithms dates back thousands of years. Ancient civilizations, like the Egyptians and Babylonians, used algorithms to solve mathematical problems and perform calculations. However, it was in the 9th century that Persian mathematician Al-Khwarizmi introduced the term "algorithm" in his book, which was a translation of Greek and Indian mathematical ideas. Fast forward to the 20th century, the development of computers brought algorithms into the digital realm. In the 1930s, mathematician Alan Turing made significant contributions to the theory of computation and introduced the idea of a "universal machine" that could execute any algorithm. This laid the foundation for modern computing and algorithmic thinking. As computers advanced, so did algorithms (George et al., 2000: 115). In the 1950s and 1960s, pioneers like John McCarthy and Edsger Dijkstra developed algorithms for tasks like artificial intelligence and graph theory. The 1970s saw the rise of algorithms for sorting and searching data, such as quicksort and binary search efficiently. With the advent of the internet and the digital revolution, algorithms became even more prominent. Search engines like Google rely on complex algorithms to provide relevant results. Social media platforms use algorithms to curate personalized feeds. And machine learning algorithms have revolutionized fields like image recognition and natural language processing (Cormen, 2009). However, artificial intelligence refers to any smart technology that can simulate human cognitive processes, including learning, problem-solving, and planning. Artificial intelli-gence, which specializes in the process of absorbing vast amounts of data and using it to learn in ways that are beyond the capability of human beings, is called machine learning (Cormen, 2009).

Within the field of artificial intelligence, machine learning involves the use of so-

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

phisticated neural network algorithms to educate computers to think like humans while handling large amounts of data and doing com-putations quickly, accurately, and ostensibly without prejudice. It is on this basis that Bernard Marr (2021: Par 13) holds that:

> The development of neural networks has been the key to teaching computers to think and understand the world in the way we do while retaining the innate advantages they hold over us, such as speed, accuracy and lack of bias. A Neural Network is a computer system designed to work by classifying information in the same way a human brain does. It can be taught to recognize, for example, images, and classify them according to elements they contain. Essentially, it works on a system of probability—based on data fed to it, it is able to make statements, decisions or predictions with a degree of certainty. The addition of a feedback loop ena-bles "learning"—by sensing or being told whether its decisions are right or wrong, it modifies the approach it takes in the future.

A lot of areas of our everyday lives already involve machine learning algorithms, which have an invisible impact on the decisions that are made about and for us with little to no transparency. Machine learning, which is beyond human capacity, mimics human learning through data and algorithms, growing more accurate over time as it continues to learn. Classifications and predictions derived from machine learning provide insights that support business decision-making (Yapo, 2018: 5360). One crucial machine learning activity that requires pattern recognition is supervised classification. One way to do this would be to search for patterns in text, hu-man faces, or the DNA of diseases. The objective of supervised classification is to identify a mapping between the class label Y and the input data X. To correctly anticipate the label of data that has not yet been seen, the system learns from training data (Ratsch, 2004). With machine learning approaches, artificial intelligence (AI) systems can either support human decision-making or completely replace them. Three popular uses for artificial intelligence systems are cognitive process automation, which automates administrative duties; cognitive insights, which finds meaning in data patterns; and cognitive engagement, which engages staff members. machine learning is believed to improve analytical and decision-making skills in workers and assist them in making better decisions (Duan & Wu, 2019: 23-30).

The following steps are usually involved in the machine learning process: data preparation, collecting, evaluation, development, post-processing, and deployment of the models. The first step is to determine the target population, after which data must be either acquired or an existing dataset can be chosen for use. Prepro-cessing operations are carried out and the data is divided into training and testing data during the data preparation step. The training data is then used to build the model, and an optimization objective is used to select the final model. Following the model's selection, the testing data performance is assessed to determine how accurately the model represents the unknown

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

data. Any post-processing actions are completed after the model evaluation and before the model is deployed (Suresh & Guttag, 2021: 17).

Because machine learning analyses a larger volume of data than a human could, it enables organizations to react quickly to societal and market trends. Numerous facets of daily life now involve the application of ma-chine learning. For instance, chatbots employ machine learning to provide real-time answers to website users' inquiries. Machine learning is used by e-commerce sites to better anticipate the products that a consumer might be interested in by analyzing their unique buying preferences. Algorithms in email are used to identify spam and facilitate user filtering through emails. Virtual assistants can also learn user preferences and make predic-tions with the aid of machine learning (Suresh & Guttag, 2021: 20).

Machine learning algorithms are already interwoven into many aspects of our daily lives, influencing in unseen ways as decisions are made for us and about us with little to no transparency. Web search and recommendation machine learning algorithms drive relevant search results and product recommendations from the likes of Google, Netflix, and Amazon. Facebook's facial recognition uses a machine learning algorithm to auto-matically identify and tag friends when uploading a photo and the news feed algorithm prioritizes posts for what it thinks we would most like to see. Machine learning in medicine is providing important advances in health care and treatment decisions while AI in computational biology/drug discovery is increasing the pace of research. The Finance industry utilizes machine learning algorithms to uncover credit card fraud, make predic-tions about creditworthiness, and identify trends in the stock market (Yapo, 2018: 5366). However, the possible questions that can be raised from the above views are: what is the underlying moral principle that guides algo-rithms in machine learning? On what ground the moral judgement of algorithms is based since artificial intel-ligence both directly and indirectly affects human beings who are moral agents? Is there any connection be-tween morality and algorithms in machine learning? Before the questions are examined, an idea of moral rela-tivism will be discussed first in the next section.

AN IDEA OF MORAL RELATIVISM

Moral relativism is one of the ethical theories that attempts to address ethical issues in moral philosophy. Ethics is one of the core branches of philosophy that examines norms, principles, and moral issues that revolve around humans and non-humans. Ethics is not only interested in humans as moral beings but also concerns itself with the non-humans' actions that affect humans. It is on this basis that the idea of ethics is conceived. Ethical issues, therefore, are not restricted to the issue that affects the individuals alone but also look at the issues that both directly and indirectly affect the moral and identical lives of the people in the societies. One of the ethical theories that attempt

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

to resolve ethical issues is moral relativism. Moral relativism holds that morality is relative, not objective; it believes that what is right or wrong is de-termined by a society or individuals. It kicks against the establishment of universal moral principles and the use of this principle globally on the ground that there are there are different societies in the world with different cultures and experiences (Omoregbe, 2002: 66).

Also, moral relativism holds that different societies have different morality; morality should not be determined through a single approach since there are different cultures in the world, and these cultures serve as the basis of morality. Hence, society determines what is right for them and what is wrong for them; they are responsible for the establishment of their moral principles. It is on the basis of this that Protagoras argues that "Man is the measure of all things; of things that are, that they are, and of things that are not, that they are not" (Russell, 1945: 73-75). This assertion suggests that truth and morality are subjective and vary from person to person or culture to culture. According to Protagoras, what is true or right for one individual or society may not be true or right for another. This idea highlights the importance of understanding and respecting different cultural perspectives and values. In the same vein, some Anthropologists like Franz Boas, Ruth Benedict, and Clifford Geertz have contributed to the development of cultural relativism as a framework for understanding and studying different societies and cultures.

Frantz Boas, a prominent anthropologist, is known for his influential work in cultural relativism. Boas (1982) believed that each culture is unique and should be studied within its own historical, social, and envi-ronmental context. He emphasized the importance of understanding cultural practices and beliefs from an insider's perspective, rather than imposing external judgments or value systems. Boas argued against the idea of biological determinism and racial hierarchy, advocating for the recognition of cultural diversity and the equal worth of all cultures. His research and teachings have had a lasting impact on the field of anthropology and continue to shape our understanding of cultural relativism today. However, one of his well-known ideas is, "Civilization is not something absolute, but . . . is relative, and . . . our ideas and conceptions are true only so far as our civilization goes" (Boas, 1982). This quote reflects Boas' belief in the importance of cultural relativism and the understanding that different societies have their own unique ways of perceiving and interpreting the world.

Ruth Benedict, a prominent anthropologist, made significant contributions to the field of cultural relativism. Her work emphasized the importance of understanding and appreciating different cultures within their own unique contexts. Benedict believed that each culture has its own distinct patterns of behaviour, beliefs, and val-ues, which shape the lives of its members. She argued against the idea of imposing external judgments or values on other cultures, advocating instead for a more empathetic and open-minded approach to studying human diversity (Benedict, 1989: 45). One of Benedict's key ideas

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

was that cultural relativism allows us to gain a deeper understanding of the complexity and richness of human experiences. She believed that by studying and comparing different cultures, we can uncover the diverse ways in which people perceive and interpret the world around them. Benedict argued that cultural practices and beliefs should not be evaluated based on our own cultural standards, but rather on their own merits within their specific cultural contexts (Benedict, 1989: 45).

In her influential book, *Patterns of Culture* (1989), Benedict explored the concept of cultural relativism through the study of three different societies: the Zuni, the Dobu, and the Kwakiutl. By examining these cultures, she aimed to demonstrate that each society has its own unique system of values and beliefs, which guide their behaviour and shape their worldview. Benedict highlighted how these cultural patterns influence various aspects of life, such as social organization, family structures, and even personality development. Benedict's work challenged the prevailing notion of cultural superiority and ethnocentrism that was prevalent during her time. She believed that understanding and appreciating cultural differences is crucial for fostering tolerance, respect, and cooperation among different societies. By recognizing the value and validity of diverse cultural practices, Benedict argued that we can move away from judgment and prejudice, and instead embrace the richness of human cultural heritage. It is important to note that while Benedict emphasized cultural relativism, she did not advocate for moral relativism. She believed that certain universal ethical principles, such as the value of human life and the importance of basic human rights, should be upheld across cultures.

However, she argued that the interpretation and expression of these principles may vary across different societies, and should be understood within their specific cultural contexts. Benedict's ideas continue to shape the field of anthropology and have had a lasting impact on our understanding of cultural diversity. Her work has influenced subsequent generations of anthropologists, who have built upon her ideas and developed further theories and methodologies for studying cultures around the world. Benedict's commitment to cultural relativism has paved the way for a more inclusive and nuanced approach to understanding human societies (Bene-dict, 1989: 45-50). Nevertheless, the extreme aspect of moral relativism is individual relativism. It holds that morality is a product of individual affairs; what is right and wrong are determined by individuals, not society. Individuals determine the moral principles that should guide their lives. The advocate of this idea holds that individuals are the starting point of the establishment of moral values and principles in society. Some of the advocates of this ideology are Fredrick Nietzsche and Jean-Paul Sartre.

Nietzsche had complex views on moral relativism. He argued that morality should be seen as a product of human culture and history. Nietzsche believed that moral values are not universal truths but rather expressions of power and dominance. He emphasized

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

the importance of individual perspectives and the need for individu-als to create their own values. Nietzsche's ideas challenge the notion of objective moral standards and encour-age us to critically examine and question societal norms and values. His moral relativism is embedded in his assertion: "Absolute values devalue themselves" (Nietzsche, 1883). He is suggesting that the belief in absolute and fixed moral values undermines their own authority and significance. Nietzsche believed that traditional moral systems, which rely on the notion of absolute values, restrict individual freedom and hinder personal growth. According to Nietzsche, these absolute values, such as good and evil, are human constructs that have been imposed on society. However, he argues that by recognizing the subjective nature of these values and em-bracing the idea that they are created by individuals, we can reclaim our own agency and create our own values. In essence, Nietzsche is highlighting the potential for personal growth and self-empowerment when we question and challenge the notion of absolute values. Nietzsche offers a solution to the devaluation of absolute values by encouraging individuals to embrace their own will to power and create their own values. He believed that by recognizing the subjective nature of values and embracing our own desires and instincts, we can over-come the limitations imposed by traditional moral systems. Nietzsche's solution involves a reevaluation of val-ues and a rejection of absolute truths in favour of personal growth and self-empowerment. It's about finding meaning and value in our individual experiences and embracing the chaos and complexity of life (Leiter & Neil, 2007).

Also, Sartre, a prominent existentialist philosopher, explored the concept of individual relativism in his philosophical works. Sartre believed that each individual has the freedom to create their own meaning and values in life, and that there are no objective or universal standards to judge one's choices or actions. In this sense, Sartre's philosophy can be seen as embracing a form of individual relativism. According to Sartre, human existence precedes essence, meaning that we are born into the world without predetermined purposes or mean-ings. Instead, it is up to each individual to define their own existence and give meaning to their life through their choices and actions. Sartre (1984) famously stated, "Existence precedes essence," emphasizing the primacy of individual freedom and responsibility. Sartre's concept of individual relativism is closely tied to his idea of rad-ical freedom. He argued that individuals are completely free to choose their actions, and that this freedom comes with the burden of responsibility for the consequences of those choices. Sartre rejected the idea of predetermined human nature or fixed moral standards, asserting that individuals must take full responsibility for their own lives and the values they choose to uphold (Omoregbe, 2002).

In his influential work, *Being and Nothingness* (1984), Sartre explores the concept of bad faith, which he sees as a form of self-deception. Bad faith occurs when individuals deny their own freedom and responsibility by conforming to societal expectations or

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

adopting external values without critically examining them. Sartre believed that embracing individual relativism requires individuals to confront the anxiety and uncertainty that come with the realization of their own freedom. It is important to note that Sartre's philosophy of individual relativism does not imply a complete moral relativism or an absence of ethical considerations. While he rejected the idea of objective moral standards, Sartre argued that individuals should act in accordance with their own authentic values and take responsibility for the impact of their choices on others. He believed that individuals must engage in a continual process of self-reflection and self-examination to ensure that their actions align with their own sense of integrity and authenticity. Sartre's philosophy of individual relativism has had a significant impact on existentialist thought and has influenced various fields such as psychology, literature, and ethics. His emphasis on individual freedom and responsibility continues to resonate with those who seek to navigate the complexities of human existence and find meaning in a world devoid of pre-established values (Sartre, 1984).

It can be deduced, therefore, that moral relativism acknowledges the intersubjectivity of ideas and society should be treated as such. By virtue of this, when moral relativism serves as the basis of cultural, societal and individual differences, such perspectives tend to reflect in an idea or principle like algorithms in machine learning, introduced not only for the benefit of one society but also for others in the world. It is on the basis of this that one can classify moral relativism as a discriminatory theory, since it does not recognize universal principles as the basis of moral judgement. Nevertheless, the fundamental questions that can be raised, therefore, are: what then is the connection between moral relativism and algorithm bias? Can an algorithm be biased independently of external interferences? What roles does artificial intelligence play in the moral issue of algorithm bias? These questions will be answered in the next section.

ALGORITHM BIAS AND MORAL RELATIVISM

AI systems are often viewed as capable of making smarter and more objective decisions. However, it is im-portant to remember that these systems are built using data and parameters defined by humans. Unfortunately, historical data can contain biases, including discrimination against minority groups. When AI learns from biased data, it can produce biased results and even perpetuate prejudices from the past (Marr, 2021). AI tools can be biased not only technically but also based on the environments they are built in and the people who build them. Currently, many large AI systems are developed by a small group of elite universities that have a history of discrimination and exclusion. These places often lack diversity, are predominantly white, affluent, technically oriented, and male. This lack of diversity can contribute to biases in AI systems (Yapo, 2018: 5365).

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

Bias can indeed enter the realm of machine learning at various stages, both through the data itself and the individuals overseeing the machine learning process. When specialists develop a machine learning algorithm, they first need to define the objective. Unfortunately, if the company's focus is solely on profits, it can unintentionally lead to discriminatory decision-making through the algorithm. Bias can also emerge during the data collection process, where data that does not accurately reflect reality can result in prejudiced outcomes. Additionally, bias can be introduced during data preparation, as the attributes chosen to build the model can impact its potential bias (Hao, 2020). While there are many advantages that machine learning algorithms can offer to people, investors, companies, consumers, the government, and society at large, recent studies have found numerous instances of bias in machine learning algorithms that have unsettling implications and negative out-comes (Yapo, 2018: 5365).

For instance, there were reports of apparent gender bias in Google searches in 2015 from academic institu-tions and the media. The top images found while searching for an image of a "CEO" were all of white men. Soon after, research from Carnegie Mellon University showed that if Google believed a woman was searching for a job, it would show much fewer advertisements for high-paying executive positions. The researchers discovered that, whereas Google displays the advertising for high-paying executive jobs 1,852 times for men, it only dis-plays them 318 times for women. On this basis, Professor Annupam Datta believes human discoveries indicate that certain areas of the advertising system are starting to show signs of discrimination, and the lack of transparency is worrisome. This is a concern from a societal perspective. He further argues that these days, algo-rithms are making a lot of important decisions in society. But the thing is, we do not have access to the nitty-gritty details of how these algorithms work. The whole idea behind this project was to take a peek inside that box and see if there are any negative consequences happening as a result (Carpenter, 2015: 41). In a similar vein, it has recently been determined that digital photo technology contains racist algorithms. Flickr's image-recognition tool allegedly produced discriminatory results in May 2015, labelling black persons as "apes" or "animals." Webcam software from Hewlett-Packard had trouble distinguishing dark skin tones, and camera software from Nikon was mistaking Asian persons for blinking.

In "Artificial Intelligence's White Guy Problem," a 2016 *New York Times* story, Kate Crawford attributed the bias to incomplete data and a lack of inclusion. Engineers frequently select certain photographs to feed algorithms, and the system uses those images to create a representation of the outside world. A system will struggle to identify non-white faces if it is trained on images of a predominantly white population.

More recently, the *New York Times* revealed that the morning following the 2016 presidential election, Facebook executives thought internally about the role the firm had

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

played in the outcome of the vote. Public charges from various sources intensified that its news feed algorithm disseminated intentional disinformation and fake news articles that unfairly skewed voters against Hillary Clinton and helped Donald Trump win the election, all the while widening the gulf between Americans through "filter bubbles." While instances of gender and racial bias in digital photo technology and Google Search were more subtle, Facebook's algorithm regularly exposes billions of people to overtly racist, sexist, and alt-right content, normalizing it through its circulation on trustworthy and mainstream platforms (Yapo, 2018: 5366).

Nonetheless, it can be deduced from the foregoing that algorithm bias can be traced to the related decisions of certain set of people which affects large number of people in the world at large. It is on this basis that Kantian principle of universality will be considered as a remedy to the relativistic fallacy that is committed in the machine learning algorithms.

MACHINE LEARNING AND ALGORITHM BIAS: A KANTIAN RESPONSE TO THE PROBLEM OF MORAL RELATIVISM

One of the key figures of the Enlightenment was the German philosopher Immanuel Kant, who lived from 1724 to 1804. Kant, who was born in Königsberg, is regarded as the father of modern ethics, the father of modern aesthetics, and the father of modern philosophy because he combined rationalism and empiricism. His extensive and methodological works in epistemology, metaphysics, ethics, and aesthetics have made him one of the most significant and contentious figures in contemporary Western philosophy (Kant, 2024: 21). His main works on ethics are the *Groundwork of the Metaphysics of Morals* (1785), *Metaphysics of Morals* (1797) and *Critique of Practical Reason* (1781).

Kant talks about the need to have a universal moral principle. Kant argues that the only thing that is good without qualification is goodwill; everything else that we typically think of as good is not necessarily good; ra-ther, their goodness needs to be qualified because they can turn bad when misused. For instance, intelligence is good, but it can turn bad when misused (e.g., by being used to commit crimes); courage is also good, but it can turn bad when misused; wealth is good, but it can turn bad when misused, etc. According to Kant, a goodwill is the will that is inherently good; it is an intrinsic will that is good in itself. What then is good will? A goodwill, in Kant's view, is one that acts out of duty. There is a distinction made by Kant between "acting ac-cording to duty" and "acting for the sake of duty." Behaving out of reverence for the moral rule is the only reason to behave, rather than out of a desire to benefit from the activity, a sense of obligation, or a natural tendency. Put in another way, it refers to acting on moral principles alone, even when doing so could result in material loss. Nonetheless, acting in accordance with one's obligation entails acting with caution and in the best interests of oneself. Not be-

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

cause such acts are wrong; rather, Kant argues that they are morally worthless and unworthy of praise. Actions carried out in line with innate proclivities or impulses are equally applicable. While such deeds may be desirable, they are morally worthless, according to Kant. An action must only be taken strictly out of duty in order for it to have moral significance (Kant, 1978).

How can I determine if what I am about to do is morally right or wrong? What standard is used to differentiate between activities that are right and wrong? The universalization principle is the yardstick, in Kant's opinion. You should seek to universalize the action's maxim, or underlying principle, to determine if the action you plan to take is morally right or wrong. Would you want your action's maxim to become a rule for everyone to follow? Put otherwise, do you think it would be preferable if everyone in a circumstance comparable to yours carried out the same action that you now plan to carry out? Do you want this to become a national or interna-tional law? If your response is affirmative, it indicates that the behaviour in issue is morally correct; conversely, if it is negative, it indicates that the activity is morally wrong. For instance, Kant uses the example of a man who is having financial difficulties. He chooses to borrow money in order to get out of this situation, even though he is fully aware that he will not be able to repay it. According to Kant (1795), the maxim of his action is this: Whenever I believe myself short of money, I will borrow money and promise to pay it back, though I know that this will never be done.

It is now for him to decide if he wants this maxim to become a global law. What would happen if my maxim was made into a worldwide law? That this maxim cannot become a universal norm without contradicting itself would become evident right away because no one would take such a promise seriously or lend money to anybody on the basis of such a promise if it were to become a universal law that anybody in financial difficul-ties should borrow money with an untrue promise to pay it back even though the money will really not be returned. This would put an end to borrowing and lending.

Also, Kant talked about the difference between a principle and a maxim. According to Kant's ethics, a principle is an objective rule of morality that every man should follow, whereas a maxim is a subjective principle that a person is de facto acting on. Men would not have moral issues if they were purely rational, according to Kant, as everyone would act in accordance with the objective and universal rules of morality, which are founded on practical reason. However, mankind has sentiments, passions, emotions, feelings, inclinations, natural tendencies, and other things; they are not entirely rational. Because of this, morality is a constant source of struggle and the moral rule is not easily followed by humans. This would not be the case if human beings were only logical beings; in such a case, morality would not exist, and people would act in accordance with their normal moral standards (Kant, 1978). He, therefore, established an imperative upon which an

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

action should be based, identifying two types of imperatives: hypothetical and categorical imperatives. While the former is based on a type of conditional imperative that directs someone to take an action that serves as a means to an end, the latter is an imperative that is absolute. It does not give instruc-tions that are merely means to an end; rather, what it demands is desirable in and of itself. Acts are commanded not as means to an end but as virtues in and of themselves. It has no "if" clauses or conditions attached, hence, it accepts no exceptions. As a result, it requires all men, without exception (Kant, 1978).

According to Kant, morality is a self-imposed law and morality originates from man's ra-tional will. The moral code is imposed upon man by his rational will. The autonomy of the will is referred to by Kant as "the principle of the autonomy of the will." Because the will is autonomous, it obeys the moral law because it is its own law that it imposes on itself as a rule of practical reason. It is not driven or determined by anything outside of itself. Nevertheless, what then is the response of Kant to the relativistic issue embedded in algorithm bias? From the foregoing, it can be argued that Kant's theory attempts to address the humanistic issue of unfair treatment of certain groups. The Kantian principle of universality, also known as the categorical imperative, basically says that we should act according to moral principles that can be applied universally to all rational beings. When it comes to machine learning algorithms, this principle can guide us in tackling bias and ensuring fairness. Machine learning algorithms can be biased because they are trained on data that reflects the biases and prejudices present in society. This can lead to unfair outcomes, like discrimination against certain groups. To combat this, we can apply the Kantian principle of universality by striving to develop algorithms that treat everyone fairly and equally, regardless of their race, gender, or any other characteristic. One way to do this is by making sure that the data used to train the algorithms is diverse and representative of the population. By including a wide range of perspectives and experiences, we can reduce the risk of biased outcomes. It is also important to regularly check and evaluate the algorithms to identify and correct any biases that may arise. For example, let us say we have an algorithm that determines credit scores for loan applications. If the algorithm unfairly discriminates against certain demographics, it can perpetuate social inequalities. By incorporating ethical guidelines that prioritize fairness and equal treatment, we can mitigate the biases and ensure that the algorithm promotes equal opportunities for all applicants. Also, we can adhere to Kant's ethical principles and aim for fairness and justice in algorithmic outcomes by consistently checking and assessing algorithms for biases. This involves regularly assessing the outcomes and impact of algorithms, and making necessary adjustments to mitigate biases and ensure fairness. For instance, consider a social media platform that uses an algorithm to curate users' news feeds. If the algorithm starts prioritizing certain types of content and excluding others, it can create echo chambers and limit users' exposure to

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

diverse perspectives. By continuously monitoring the algorithm's performance and user feedback, the platform can identify and rectify any biases that may arise, promoting a more inclusive and diverse user experience.

Furthermore, in Kant's ethical theory, the idea of promoting collaboration and dialogue among different stakeholders like developers, users, and regulatory bodies aligns with the concept of universalizability and treating individuals as ends in themselves. By involving various perspectives and expertise, we can collectively work towards addressing algorithm biases and ensuring that the technology serves the best interests of society. For example, imagine a government implementing an algorithmic decision-making system for public services. By engaging with citizens, advocacy groups, and experts, the government can gather valuable insights and feedback to improve the system's fairness and effectiveness, while also ensuring that it aligns with ethical and legal standards. Another approach is to involve a diverse group of people in the design and development of these algorithms. By including individuals from different backgrounds, we can bring in various viewpoints and challenge any potential biases that might emerge. For instance, in the development of facial recognition technology, a company forms a diverse team consisting of individuals from different racial, ethnic, and cultural backgrounds. By including people with diverse lived experiences, the team can bring in various viewpoints and challenge any potential biases that might emerge in the algorithm. This helps ensure that the technology is fair and accurate for all individuals, regardless of their racial or ethnic background.

Also, a social media platform should want to improve its algorithm to avoid echo chambers and promote diverse perspectives. They can create a task force that includes individuals from different age groups, genders, socioeconomic backgrounds, and political beliefs. By involving this diverse group in the design and development process, the platform can gather insights and perspectives that challenge potential biases and help create a more inclusive and balanced algorithm. These examples demonstrate how involving a diverse group of people in algorithm design and development can help identify and address biases, leading to more equitable and inclusive outcomes. Transparency and accountability are also important in addressing algorithmic bias. By providing clear explanations of how the algorithms work and making the decision-making process transparent, we can identify and rectify any biases. It also allows people to understand the factors that influence the outcomes and hold the developers accountable for any unfairness; for instance, a ride-sharing platform that uses an algorithm to determine pricing for different routes. To address concerns of bias in pricing, the platform provides clear explanations of how the algorithm calculates fares, considering factors such as distance, demand, and time of day. By making this information transparent to both drivers and riders, the platform allows them to understand the pricing decisions and identify any po-

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

tential biases. This transparency fosters trust and enables users to hold the platform accountable for fair and equitable pricing.

More so, let us consider an online job recruitment platform that uses an algorithm to match candidates with job opportunities. To ensure fairness and mitigate bias, the platform discloses the key factors and criteria used by the algorithm to make job recommendations. This transparency allows job seekers to understand how their skills, experience, and qualifications are evaluated. It also enables them to identify any potential biases in the algorithm's decision-making process, such as gender or racial biases. By holding the platform accountable for fair and unbiased recommendations, users can advocate for equal opportunities in the job market. In both examples, transparency and accountability play a crucial role in addressing algorithmic bias. By providing clear explanations of how algorithms work, making the decision-making process transparent, and allowing people to understand the factors that influence outcomes, we can identify and rectify biases. Additionally, this transparency holds developers and platforms accountable for any unfairness, fostering trust and promoting fairness in algorithmic systems.

DISCUSSION

This paper discussed the problem of algorithm biases in machine learning. It believes that there are biases in algorithms that affect the moral lives of some individuals in the world. The basis of these problems is traced to the differences in the socio-cultural activities of the world, which serves as the basis of moral relativism. While showing the historical issues that justified the biases in machine learning algorithms, through evaluative and comparative approaches, the paper employed the Kantian principle of universality as a theoretical framework to address the problems. It argues that the issue of the biases in machine learning algorithms can be mitigated when attention is mainly directed to human individuality or dignity. It argues that when the individuals in the world are treated as ends in themselves not as mere means, it will reduce the idea of placing one society above others. The problem of algorithm bias will be reduced, if not abolished, when the programmers of algorithms treat everyone as an end themselves, not placing one society above others.

CONCLUSION

Kant's ethical theory can contribute to reducing algorithmic biases in the world by emphasizing the importance of treating individuals as ends in themselves, rather than mere means. By adopting this principle, developers can prioritize the ethical implications of their algorithms and ensure that they respect the dignity and autonomy of all users. Also, Kant's emphasis on universalizability can guide developers to create al-

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

gorithms that apply moral principles consistently and without bias. By considering the moral implications of their algorithms and striving for fairness and equality, developers can work towards reducing biases and promoting ethical decision-making in the design and implementation of algorithms. Kant's emphasis on the categorical imperative can guide developers to create algorithms that treat all individuals as equals, regardless of their personal characteristics. By applying this principle, developers can strive for fairness and impartiality in algorithmic decision-making. Kant also emphasizes the importance of treating human beings as ends in themselves which can encourage developers to design algorithms that allow users to have control over their data and decisions. By integrating these ethical principles into algorithm development, we can work towards minimizing biases and promoting a more equitable digital landscape.

However, the paper recommends that to address the problem of algorithm biases which have been affecting not just Africans but also other races in the world, the focus of the programmers of algorithms in machine learning should be directed to the protection of human dignity or the individualities of the individuals in order to reduce the role of sociocultural factors in the development machine learning algorithm.

REFERENCES

Benedict, R. (1989). *Patterns of Culture*. Preface by Margaret Mead; foreword by Mary Catherine Bateson. Houghton Mifflin. ISBN 978-0-395-50088-0.

Buroker, J. (2006).*Kant's Critique of Pure Reason: An Introduction*. Cambridge University Press

Carpenter, J. (2015). Google's algorithm shows prestigious job ads to men, but not to women. The Independent - News. http://www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-0372166.html

Cormen, H. (2009). *Introduction to Algorithms (3rd ed.).* Cambridge, Mass.: MIT Press.

Crawford, K. (2016). *Artificial intelligence's white Guy Problem*. Sunday Review. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

De Vos, T. (2016). Cool machine learning examples in real life. http://itenterprise.co.uk/cool-machine-learning-examples-real-life/

George Ofrah, David Bellos, E.F. Harding, Sophie Wood & Ian Monk (2000). *The Universal History of Numbers: From Prehistory to the Invention of the Computer.* Weley.

Kant, I. {(1785) 1996}. "Groundwork of the Metaphysics of Morals." In *Practical Philosophy*, 37-108.

Kant, I. {(1797) 1996}. "On a Supposed Right to Lie From Philanthropy." In *Practical Philosophy*, 605-615.

Kant, I. {(1978) 1996}. "The Metaphysics of Morals". In *Practical Philosophy*, 353-603.

Kant, I. (1999). *Critique of Pure Reason (The Cambridge Edition of the Works of Immanuel Kant)*. Translated and edited by Paul Guyer and Allen W. Wood. Cambridge University Press.

DOI: 10.5281/zenodo.13294739

DOI: 10.5281/zenodo.13294739

Kant, I., and Allen W. W. (1996). *Practical Philosophy*. Edited by Mary J. Gregor. The Cambridge Edition of the Works of Immanuel Kant. Cambridge University Press.

Leiter Brian and Neil Sinhababu (eds.), 2007. Nietzsche and Morality. Oxford University Press.

Marr, B. (2021). What is the difference between artificial intelligence and machine learning? Forbes.Retrieved from: http://www.forbes.com/sites/bernardmarr/2024/06/13/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#3a953a62687c

Omoregbe, J.(2001). *A Simplified History of Western Philosophy.* Joja Press.

Omoregbe, J.(2002). *Ethics: A Systematic and Historical Study.* Joja Press.

Ratsch, G. (2004). *A Brief Introduction into Machine Learning*.Tubingen:Friedrich Miescher Laboratory of the Max Planck Society.

Russell B. (1945). *The history of Western philosophy.* Simon and Shuster.

Sartre, J. P. (1984). *Being and Nothingness*. Trans. Hazel E. Barnes. Washington Square Press. pp. 93–94.

Suresh, H. and Guttag, J. (2021). "A Framework for Understanding Sourcesof Harm throughout the Machine Learning Life Cycle". *Equity and Access in Algorithms, Mechanisms, and Optimization*.10.1145/3465416.3483305

Wu, B. and Duan, T. (2019).*The Advantages of Blockchain Technology in Commercial Bank Operation and Management. Proceedings of the 2019 4th International Conference on Machine Learning Technologies*, Nanchang, 21-23 June 2019, 83-87.https://doi.org/10.1145/3340997.3341009

Yapo, Adrienne& Joseph Weiss. (2018). "Ethical Implications of Bias In Machine Learning"in*Proceedings of the 51st Hawaii International Conference on System Sciences*. URI: http://hdl.handle.net/10125/50557.

DOI: 10.5281/zenodo.13294739