HIGH SCHOOL EDITION
Journal of Student Research

The Causes and Effects of Algorithmic Decision Making

Simran Saluja[1], Rajagopal Appavu[#], Jothsna Kethar[#] and Mariana Ahorta[#]

[1] Rocklin High School, Rocklin, CA, USA
[#] Advisor

ABSTRACT

The recommendations list that appears after watching a show on Netflix, the ads that come up on Instagram similar to other liked posts, these phenomenons are due to algorithmic decision making. Algorithms are able to help people do simple pattern-based tasks in day-to-day life but they also have implications on our society. Algorithms often make mistakes due to the human error in data they analyze. There are many different sources of algorithmic error in decision making, but the most significant cause by far is error in the data which algorithms are trained from. These errors affect advertisements, jobs, and also technological products. It is hard to get rid of these biases as many times it leads to underrepresentation, another cause of algorithmic error. The way an algorithm is trained has a large impact on the future decisions it makes.

Introduction

Algorithms have the ability to perform large computational decisions and can process more data than a human can grasp at once. Compared to humans, they can take more considerations into account when making decisions, and this is the reason why our society has grown to rely on them in many aspects of our day to day lives. Whether that be in banking and finance, product advertisement, healthcare, or research, they have become a part of us almost everywhere, somehow. Although, with these advantages, there are also many risks that we as a society have to take into consideration. Algorithms that help humans do pattern-oriented tasks that require analysis are becoming much more common today than ever before. Computer scientists that work in machine learning create and research algorithms. Social Scientists, on the other hand, are interested in understanding the social implications of the widespread use of algorithms and modern devices. The implications of algorithmic decisions make a mark on every aspect of our lives, just as humans leave a trace or footprint wherever they go. We just have to learn how we can integrate this new technology into our society in order to improve people's overall wellbeing. In the future, researching how algorithms can be improved, and new technologies that are coming to a market in this field of study, would be good to explore. Seeing where our new innovations are leading to and analyzing their possible implications is especially important, given that many of the technologies can make an impact in our lives on a worldwide scale. This paper has put together knowledge about algorithmic decision making to help readers understand how algorithms may affect their lives. It will cover an overview of algorithmic fairness and decision-making, algorithmic bias, its effects on sales, products, and jobs, and how we can improve these algorithms. What are the implications of algorithmic decision making on our society?

Literature Review

An Overview of Algorithmic Fairness and Decision-Making

People from this century's technological era believe algorithms can perform a multitude of decisions better than themselves because algorithms can integrate much more data than a human can grasp and take many more considerations into account (Pessach, D., & Shmueli, E., 2020). In addition, algorithms can perform complex computations much faster than human beings. Human decisions are subjective, and we often include biases in the decisions we make. A common belief in society is that if we use algorithms to make decisions, they will be more objective. According to Pew Research Center, “Roughly six-in-ten Americans (58%) feel that computer programs will always reflect the biases of the people who designed them, while 40% feel it is possible for computer programs to make decisions that are free from human bias. Notably, younger Americans are more supportive of the notion that computer programs can be developed that are free from bias” (Smith, A., 2018). This is where the study of algorithmic fairness comes into play: they evaluate whether algorithms and AI are really as “fair” as we believe them to be (Pessach, D., & Shmueli, E., 2020).

Many AI possess biases and inadvertent discriminatory “beliefs,” which could be used by companies during their implementation, making a much larger impact than we humans can on our own (Pessach, D., & Shmueli, E., 2020). According to Algorithmic Fairness by Pessach: “Machine learning and Artificial Intelligence systems can either introduce or amplify discriminatory effects that they learned through their training or experiences over time.” The common saying “unfair data leads to unfair models” explains this. If a system learns from unfair data, unfair people, or biased content, it will learn, integrate, explore, and expand that unfairness to whomever it’s the system. For this reason, machine learning and AI systems can be much more harmful than humans when it comes to making biased decisions, not because of their effect when a singular decision is made, but because AI systems have the computational capacity to make 100, 1000, or 10000 times the decisions and interactions that a single, biased human can in one day (Ashurst, C., Carey, R., Chiappa, S., Et al., 2022).

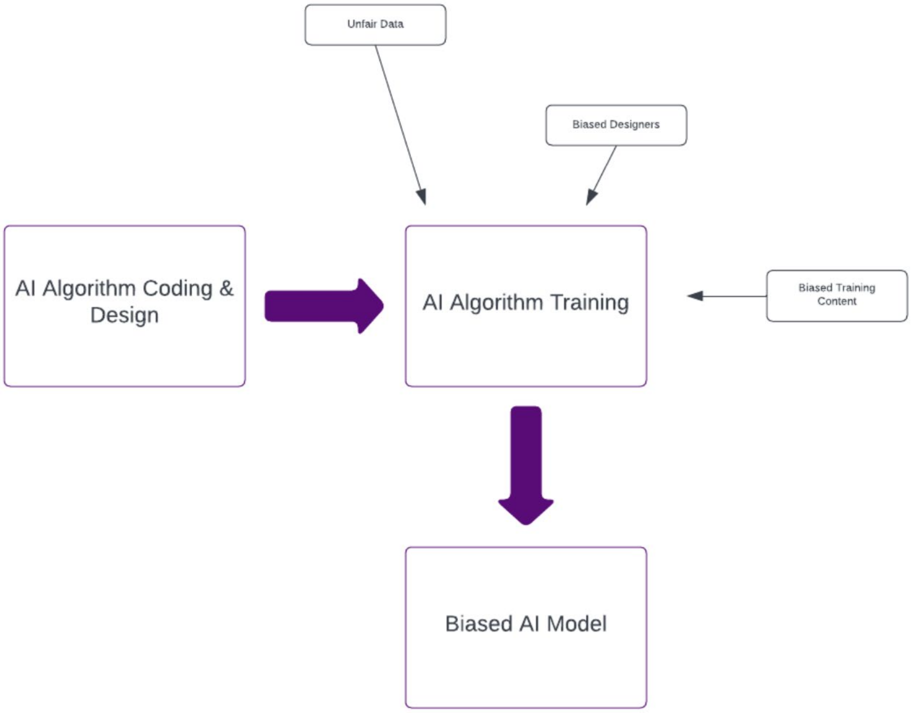

Figure 1. Unfair data leads to unfair models.

One example of this is natural language processing AI. According to Algorithmic Fairness by Kleinberg, language processing AI associated the word “nurse” with the pronouns “she” more than “he.” This is an example of how artificial intelligence and algorithmic systems can adopt inadvertent discrimination from the data they analyze. This stereotype that nurses are more likely to be female traces back to traditional and outdated beliefs not just from America, but from around the world, that women could only assist male doctors and couldn't become doctors themselves. Since the data used to train the algorithms and AI we use daily is ringed with certain stereotypes, biases, and discrimination, there are worries that our faults will start occurring in these devices, as presented in this example (Kleinberg, J., Ludwig, J., Mullainathan, S., Et al., 2018).

This same negative of algorithmic decision making appears in the criminal justice system, as well, where "recent news revealed that an algorithm used by the United States criminal justice system had falsely predicted future criminality among African-Americans' ' (Pessach, D., & Shmueli, E., 2020). One predicted reason for this to have occurred is because the algorithm was trained based on data sets that didn't have enough information about African Americans, and more data on other races, leading to bias. Not all African Americans may have had the same experiences shown to that specific algorithm (Pessach, D., & Shmueli, E., 2020)

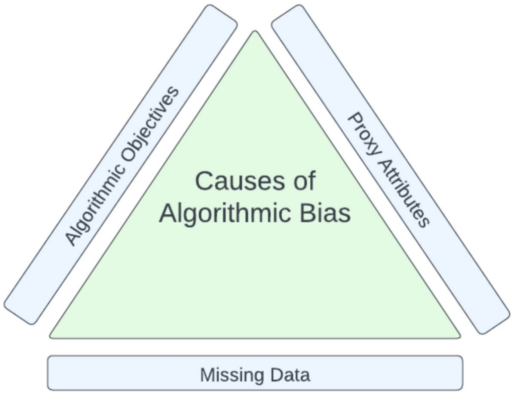

Figure 2. Causes of Algorithmic Bias.

As there are many cases of algorithms acting up in this way, "algorithmic fairness" by Pessach and Shemuli covers a few reasons why that may be. Biases are already included in the datasets used for learning, which are based on historically biased human decisions, and Machine learning algorithms are essentially designed to replicate these biases (Hanna, A., Denton, E., Smart, A., Et al., 2020). First is biases caused by missing data. These result in datasets that are not representative of the target population, which leads to situations where an algorithm is misrepresenting who would like a certain content, who would be more likely to commit certain crimes, etc. The second is biases that were formed from algorithmic objectives. These algorithms tend to focus on minimizing aggregated prediction errors and look at multitudes of past data. This leads the algorithm to tend to benefit majority groups of the area it is looking into rather than minority groups. Third, biases may be caused by proxy attributes instead of sensitive attributes. Sensitive attributes differentiate those who are privileged and not, such as race, gender, age, and more. Proxy attributes are non-sensitive attributes from data that, once derived, can be seen as sensitive. If a dataset that an algorithm is training from contains proxy attributes, the algorithm can make a decision based on that "sensitive" attribute but have it labeled as "legitimate," when in reality it is discriminatory towards a certain group (Pesach).

Types of Algorithmic Bias

Many of the issues stemming from algorithmic fairness and unfairness initially start from bias that is ingrained into the algorithm. These biases come in different shapes and forms, for example, underrepresented training samples, mislabelling outcomes, feedback loops, and biased programmers (Cowgill, B., & Tucker, C. E., 2019).

Under-represented training samples is a type of bias that is exactly what it sounds like: data sets that are used to train algorithms that do not represent all sides of an issue or topic. They may not have representation of all types of people or all of the options that could be involved in making algorithmic decisions later on. This results in the algorithm making flawed or incorrect decisions when it is used after training because, in reality, that algorithm was never trained to its fullest. Second, mislabelling outcomes: when an algorithm analyzes and labels a situation, it uses the data available to it to make an educated assumption and come up with a label for this new scenario. In some cases, though, the past data the algorithm is given during training is not enough, and situations can be mislabelled. For example, if an employee faces discrimination from a supervisor and quits, and an algorithm mislabels the employee as low-performing, the algorithm could be thinking that the employee just never put in the effort and it leads to a talk with their supervisor. Without training that gives an insight into almost every possible option, occurrences like this are bound to occur. Third, feedback loops: when an algorithm uses the predictions it makes based on other data and analyzes that prediction along with more data to conclude. This is biased since the algorithm is using a prediction to form another prediction, making the chance of both being correct decrease. Lastly, biased programmers: they are especially influential on the algorithmic bias present because they can make an algorithm based on a substantial level. According to Economics, fairness, and algorithmic bias by Cowgill, "Software engineers may unconsciously (or overtly) exhibit bias during development. According to the Bureau of Labor Statistics in 2017, software engineers are more often white, male, well-educated, and better-paid than America as a whole. These engineers may not be consciously biased, but their life experiences may influence their approach to developing an algorithm" (Cowgill, B., & Tucker, C. E., 2019).

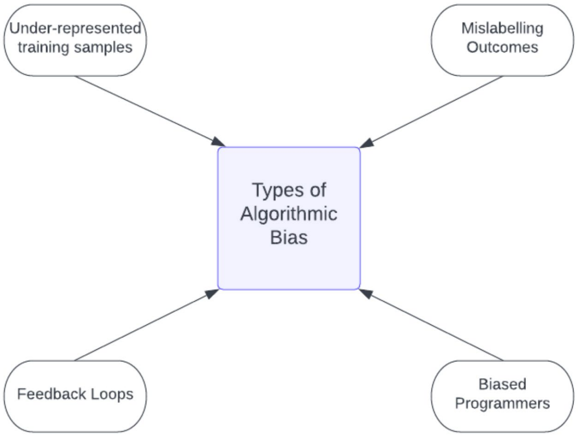

Figure 3. Types of Algorithmic Bias.

Overall, although we may not consciously insert bias into algorithms as trainers, programmers, or regular people, machine learning programs are trained to look for certain patterns in the data they are given and often, human bias can be mistaken as a pattern. This cannot only influence the future of artificial intelligence and algorithmic development, but the lives of those whom the algorithms make choices for (Cowgill, B., & Tucker, C. E., 2019).

How Does Algorithmic Fairness Affect Sales?

One way that algorithms biases can be integrated into the aspects of our lives we have let them take over is advertising, mortgage, and loans. Systematically coding algorithms to analyze a person's search history and give them similar promotional content can result in targeted advertisement. Many people believe that targeting certain advertisements to people based on what they do on the internet or in real life is a form of descrimination. For example, depending on a person's search history, an individual can receive ads targeted to a certain income group or set of people. This is also known as algorithmic discrimination.

This can be seen in algorithm integration in applications such as Google or other websites where, depending on what you've searched for or looked at previously, you will get similar ads as you scroll in the future. According to A qualitative exploration of perceptions of algorithmic fairness by Woodruff, many people complain about receiving only ads relating to "low income" or "cheap alternative" options after something similar had appeared in their search history. Ultimately, algorithms and artificial intelligence are coded to appeal to their audience based on past encounters, which can also lead to unfairness towards certain minority groups. This is because algorithms do not understand that humans are diverse, and cannot be understood solely from a single search (Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J., 2018).

Another effect of targeted advertisement through algorithms is that these algorithms can predict who the customer is. They can determine that "this ad thinks that you're male, actively consolidating your debt and are a high spender at luxury department stores. . . . This ad thinks you're female, a registered Democrat and are likely to vote for the sitting president." One prime example of this is Target. A father of a high school girl had made a complaint to the target HR stating that he was uncomfortable with his daughter receiving an ad about baby coupons and cribs, stating that they were encouraging her to get pregnant. The ads had been decided after an artificial intelligence algorithm analyzed the girl's history through Target's website, where she was already searching through baby items. It turned out that she was already pregnant (Kim, T. W., & Routledge, B. R., 2022).

Companies have started to share limited information about why you get the ads you do, and where they are giving your personal information, which only deepens the root of the issue many customers are facing due to artificial intelligence algorithms today. Knowing exactly who your customer is, is great for a company's sales. On the other hand, not knowing what you are categorized as and why can form a sense of panic in many customers if you make use of their favorite company's online options. This stems from a feeling of knowing whether those companies send their data to other entities with which they aren't comfortable and if they can even trust companies with their data online at all (Kim, T. W., & Routledge, B. R., 2022).

Another important part of algorithmic fairness in finances is mortgage lending and loans. Discrimination in finance can be divided into two types. Taste-based discrimination, which is when someone has a predetermined bias against a certain group/sect of people. Statistical discrimination, which is when people use facts like race, gender, or age, to determine things like creditworthiness or make decisions for loans, renters, buyers, etc. Many human-made decisions have biases and inaccuracies which can change the decision made from person to person, whereas an algorithm that makes a similar decision is capable of discriminating on a larger scale if there is a bias within its capabilities. Algorithmic decisions are made using a rules-based process, if x then y, whereas human decisions, and humans themselves, are complex. This is why algorithmic discrimination in finances can have a much larger impact than a human's discrimination within the financial world, in some instances. A primary example where this is shown is from a study in Chicago which found that black applicants on average were quoted lower loan amounts and received less information and help when applying for a mortgage (Lee, M. S. A., & Floridi, L., 2021).

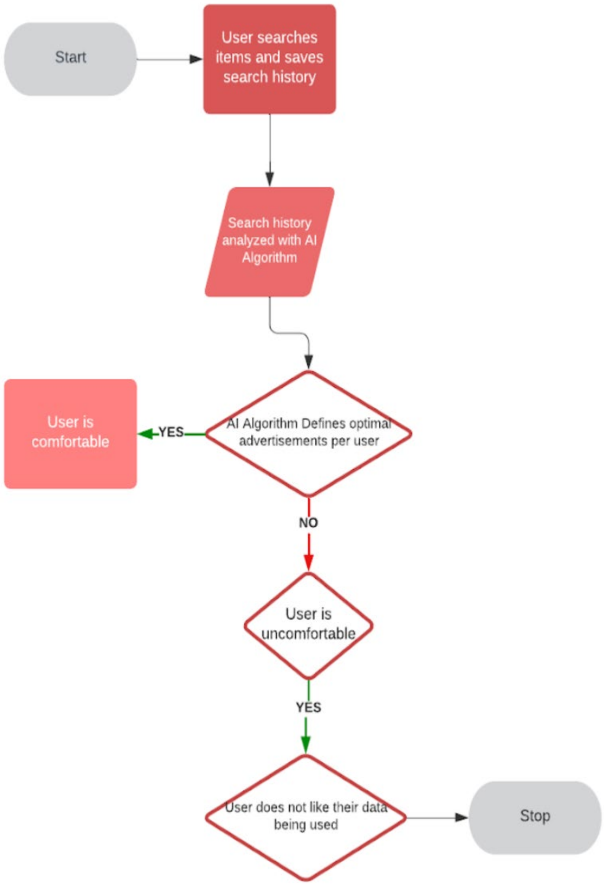
Start
User searches items and saves search history
Search history analyzed with AI Algorithm
AI Algorithm Defines optimal advertisements per user
YES
User is comfortable
NO
User is uncomfortable
YES
User does not like their data being used
Stop

Figure 4. Search history based advertising process.

Although algorithms can intend to follow a scorecard method (if income > X and if loan amount < Y, then approve), they are made to find patterns within the data they receive. The general worry among professionals is that information in the data can support inequalities and biases that are tended to be seen among humans. The data that algorithms analyze to achieve the cognitive capabilities which they are able is derived from humans themselves, and in us, humans come to both direct and indirect bias and discrimination. Algorithms simply analyze this data and make decisions similar to what they see, but if the data they see is biased in and of itself, it can result in the biased decisions made by that algorithm on a much, much larger scale (Lee, M. S. A., & Floridi, L., 2021).

How Does Algorithmic Fairness Affect Jobs Available to People?

Recently, Amazon discovered that its AI hiring system was discriminating against female candidates, particularly for technical job openings. One suspected reason for this is that most recorded historical data used during the training of the algorithm was for male software developers (Pessach, 2020). Biases already included in the datasets used for learning are based on historically biased human decisions and Machine learning algorithms are essentially designed to replicate these biases.

These biases can be caused by missing data, resulting in the underrepresentation of certain populations (Pessach, 2020). But how is this underrepresentation in data occurring? According to the Bureau of Labor Statistics in 2018, software engineers are more white, male, well-educated, and better-paid than those with other jobs in the United States. They may not be intentionally biased in their work, but their life experiences may influence how they develop algorithms, leading to unintentional bias which would be implemented on a much larger scale. Biased programmers can create under-represented training data sets for algorithms, furthering bias and misrepresentation of certain issues, thoughts, or communities (Cowgill, B., & Tucker, C. E., 2020).

These biases and issues can also stem from algorithmic objectives, aiming at minimizing prediction errors and therefore benefiting majority groups over minorities. Some algorithms have certain pre-coded methods making them believe that in some situations, the majority is preferred over the minority. If these algorithms aren't properly able to determine when they should use certain objectives, bias can present itself and lead to misfortune when making decisions (Pessach, 2020).

Biases are also caused by "proxy" attributes for sensitive attributes. Sensitive attributes differentiate privileged and unprivileged groups, such as race, gender, and age, and are typically not legitimate for use in decision-making. Proxy attributes are non-sensitive attributes that can be exploited to derive sensitive attributes. In the case that the dataset contains proxy attributes, the machine learning algorithm can make decisions based on the sensitive attributes under the cover of using presumably legitimate attributes. This can lead to cases where the algorithm is using an attribute originally seen as worthy, but in the context of use would be considered sensitive, resulting in a biased decision made (Pessach, 2020).

In another scenario in advertising, Google's ad-targeting algorithm had proposed higher-paying executive jobs more for men than for women. This goes to show that algorithms don't only make decisions affecting the direct hiring process, but can also make decisions influencing what jobs people end up applying to (Pesach 2020).

Overall, algorithmic fairness is an issue created by the presence of bias in an algorithm, underdeveloped methods, and underdeveloped objectives.

How Has Algorithmic Fairness Become an Issue in Products?

There are many examples where Algorithmic Fairness is an issue in certain products to be used for public or private use, specifically those involved in training an algorithm with a biased or incomplete dataset. For example, IBM's facial gender classification system worked poorly on people of dark skin, specifically women. This issue arose due to insufficient system training, as the algorithm did not have enough previous data on dark-skinned women to make future decisions based on. IBM resolved this issue by training the algorithm with more dark-skinned women (Bakalar, C., Barreto, R., Bergman, S., Et al.., 2021).

Another example is from researchers at Jigsaw. Many noticed their toxicity classifier labeled comments containing identity terms including "gay" or "Muslim" as toxic. This is likely either because of insufficient data training or poor data collection choice. The solution was to pick training data containing more neutral phrases, to show examples of what verbiage can be considered acceptable and not toxic. In the end, this improved their system overall (Bakalar, C., Barreto, R., Bergman, S., Et al.., 2021).

In healthcare, on the other hand, researchers found that when an algorithm was trained to find the cost of an outpatient service, African American patients incurred a lower cost than their white counterparts after analyzing need-based proxies among the data provided. This situation had no other solution other than having a better machine-learning algorithm, as it analyzed the data and came up with a cost that may not always be true for every scenario. Someone's race has nothing to do with whether someone is need-based or not, but the algorithm found a pattern and used it (Bakalar, C., Barreto, R., Bergman, S., Et al.., 2021).

What Methods Can Scientists Use to Create Fairer Algorithms and Machine Learning Models?

Humans have bias that affects them in every part of their lives. Humans have worldviews, and opinions, and use their judgment to make decisions that affect others. Even within all the data available, it is impossible to use data that is 100% unbiased, and that contains information on every option, every community, and every person that will be affected by that algorithm's future decisions. Realistically speaking, it is impossible to completely eradicate bias from algorithmic training. Although, there are a few ways we can try to significantly decrease the amount of bias an algorithm incurs Veale, M., & Binns, R., 2017).

The first is to design the algorithm to consider the pros and cons of all options before making a decision. Oftentimes, studying the pros and cons of a situation gives a deeper insight for the algorithm to make a decision based on rather than solely looking at a model or study provided to it. Models contain works that support a certain argument, sure, but they do not contain how the item the author is arguing for affects others Veale, M., & Binns, R., 2017).

The second way is to use discrimination-aware data mining (DADM) in machine learning. Discrimination aware data mining involves using data science processes to correct certain forms of bias within a model. They can operate at several stages, including pre-processing, in-processing, and postprocessing. This method aims to create patterns that do not lead to discriminatory decisions from the algorithm even though bias may be present in its training data Veale, M., & Binns, R., 2017).

Another option is to remove data that is biased or acknowledge which parts of the data set are biased. For example, in paper The long road to fairer algorithms, “For instance, several previous offenses can bear the stamp of historical racial bias in policing, as can the use of plea bargaining (pleading guilty being more likely to reduce a sentence than arguing innocence). This can leave researchers with a hard choice: either remove all data or keep biased data” (Kusner).

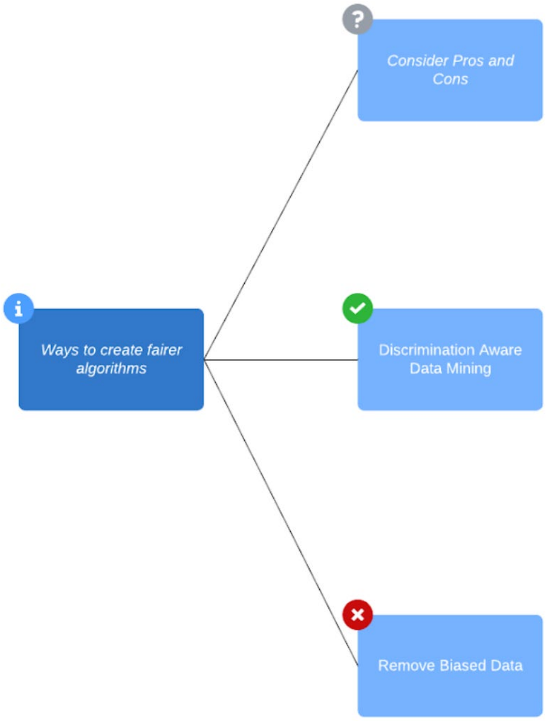
Consider Pros and Cons
Ways to create fairer algorithms
Discrimination Aware Data Mining
Remove Biased Data

Figure 5. Ways to create fairer Algorithms.

Casual models also solve issues. For example, if a person wants a loan and cannot receive one without paying back previous loans. Unfortunately, if discriminatory employers are likely to hire a person with a disability, it can make it harder for that person to pay back a loan. This bias in determining whether one can receive a loan or not can be solved using a casual model to quantify bias and decide what would be a more unbiased version of the conclusion. A causal model, “represents how data are generated, and how variables might change in response to interventions. This can be shown as a graph in which each variable is a node and arrows represent the causal connections between them. Take, for example, a data set about who gets a visa to work in a country. There is information about the country each person comes from, the work they do, their religion, and whether or not they obtained a visa" (Kusner, M. J., & Loftus, J. R, 2020).

Conclusion

The main focus of this paper was the effects and significance of algorithmic decision making. One important point that affects the way algorithms make decisions is the way they are trained. When an artificial intelligence algorithm is developed, it is often trained using large sums of data. Some developers and machine learning users say that since algorithms don't have biases, and make decisions based on data, it is impossible for them to make mistakes. In reality in many cases the data they are trained with is often biased. Humans developed the data, and put it into a written form, having both intentional and unintentional bias, which the algorithm can find as a pattern and base future decisions off of. Also, in some cases the programmers creating the algorithm itself are often biased, and it can affect the output. Misinterpretations and under-represented groups also affect algorithmic training. But why is training so important? Algorithmic training forms the basis of how algorithms make actual decisions in the future. Algorithms will have inevitable biases, just like humans. Algorithms are trained, they learn, based on data that we have discovered or put together in one place. They learn from analyzing our society, and our behaviors. Unless they can be trained using a method that eliminates all biased data, which may result in under-representation (another type of bias), it is very hard. There are methods to reduce bias, or try to keep it out of the algorithms training altogether, and over time more efficient methods will be developed, but we should make sure to know where the fine line between human and machine should be drawn. At what point is it too much?

Acknowledgments

First, I would like to thank Ms. Mariana Horta for guiding me through how to write a research paper. She helped me very much during the editing process and I am very thankful. I have learned a lot about the research and writing process under her guidance and I hope to continue learning from her. Secondly, I would like to thank Coach Jo and Mr. Rajgopal for guiding me thorough the publishing process.. Lastly, I would like to thank my parents, Rajinder Kaur and Gurmukh Singh, for supporting me all the way and providing me with beneficial resources.

References

Kleinberg, J., Ludwig, J., Mullainathan, S., Et al. (2018, May). Algorithmic fairness. In Aea papers and proceedings (Vol. 108, pp. 22-27). https://doi.org/10.1257/pandp.20181018.

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness. arXiv preprint arXiv:2001.09784. https://doi.org/10.48550/arXiv.2001.09784.

Ashurst, C., Carey, R., Chiappa, S., Et al. (2022). Why Fair Labels Can Yield Unfair Predictions: Graphical Conditions for Introduced Unfairness. arXiv preprint arXiv:2202.10816. https://doi.org/10.48550/arXiv.2202.10816.

Smith, A. (2018). Attitudes toward algorithmic decision-making. Pew Research Center. https://www.pewresearch.org/internet/2018/11/16/attitudes-toward-algorithmic-decision-making.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018, April). A qualitative exploration of perceptions of algorithmic fairness. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-14). https://doi.org/10.1145/3173574.3174230.

Lee, M. S. A., & Floridi, L. (2021). Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Minds and Machines, 31(1), 165-191. https://doi.org/10.1007/s11023-020-09529-4.

Kim, T. W., & Routledge, B. R. (2022). Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. Business Ethics Quarterly, 32(1), 75-102. https://doi.org/10.1017/beq.2021.3.

Hanna, A., Denton, E., Smart, A., Et al. (2020, January). Towards a critical race methodology in algorithmic fairness. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 501-512). https://doi.org/10.1145/3351095.3372826.

Cowgill, B., & Tucker, C. E. (2020). Algorithmic fairness and economics. Columbia Business School Research Paper. http://dx.doi.org/10.2139/ssrn.3361280.

Bakalar, C., Barreto, R., Bergman, S., Et al. (2021). Fairness on the ground: Applying algorithmic fairness approaches to production systems. arXiv preprint arXiv:2103.06172. https://doi.org/10.48550/arXiv.2103.06172.

Cowgill, B., & Tucker, C. E. (2019). Economics, fairness and algorithmic bias. preparation for: Journal of Economic Perspectives.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 2053951717743530. https://doi.org/10.1177/2053951717743530.

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. https://doi.org/10.1038/d41586-020-00274-3.