The quiet revolution of artificial intelligence looks nothing like the way movies predicted; AI seeps into our lives not by overtaking our lives as sentient robots, but instead, steadily creeping into areas of decision-making that were previously exclusive to humans. Because it is so hard to spot, you might not have even noticed how much of your life is influenced by algorithms.

Picture this — this morning, you woke up, reached for your phone, and checked Facebook or Instagram, in which you consumed media from a content feed created by an algorithm. Then you checked your email; only the messages that matter, of course. Everything negligible was automatically dumped into your spam or promotions folder. You may have listened to a new playlist on Spotify that was suggested to you based on the music that you'd previously shown interest in. You then proceeded with your morning routine before getting in your car and using Google Maps to see how long your commute would take today.

In the span of half an hour, the content you consumed, the music you listened to, and your ride to work relied on brain power other than your own — it relied on predictive modelling from algorithms.

Machine learning is here. Artificial intelligence is here. We are right in the midst of the information revolution and while it's an incredible time and place to be in, one must be wary of the implications that come along with it. Having a machine tell you how long your commute will be, what music you should listen to, and what content you would likely engage with are all relatively harmless examples. But while you're scrolling through your Facebook newsfeed, an algorithm somewhere is determining someone's medical diagnoses, their parole eligibility, or their career prospects.

At face value, machine learning algorithms look like a promising solution for mitigating the wicked problem that is human bias, and all the ways it can negatively impact the lives of millions of people. The idea is that the algorithms in AI are capable of being more fair and

efficient than humans ever could be. Companies, governments, organizations, and individuals worldwide are handing off decision-making for many reasons — it's more reliable, it becomes easier, it is less costly, and it's time-efficient. However, there are still some concerns to be aware of.

Defining Bias in General

Getty Images

Bias can be defined broadly as a deviation from some rational decision or norm, and can be statistical, legal, moral, or functional. We see bias in our everyday lives as well as on a societal scale. Oftentimes, one perpetuates the other.

For example, on your way home, you may choose to take a route that is in a “safer” neighbourhood— what determines this? Maybe the area is home to those who are lower on the spectrum of socio-economic privilege. While it’s not necessarily the case that the less privileged are more likely to participate in criminal activities, your bias, whether explicit or implicit, urges you to take a different route. On a larger scale, these areas may be more heavily patrolled by police, which, in turn, could lead to a higher arrest rate than a more affluent neighbourhood, giving off the illusion of a higher crime rate, regardless of the actual amount of crime that goes on there. This vicious cycle only seems to reinforce our initial biases.

Algorithms and Machine Learning

Let’s first differentiate between classic algorithms and machine learning. Algorithms are often described as input-output machines. Traditional programming relies on functions that are rooted in logic — IF x, THEN y. Algorithms are rule based, explicit, and hard-wired. Machine learning is more complicated than that. Learning algorithms make their decisions not by a pre-programmed condition that their data must meet, but through the auditing and statistical analyses of hundreds or thousands of datasets in the realm that it makes the decision in.

For example, in a hiring learning algorithm seeking candidates that are most likely to succeed, the training dataset may be fed with data of 200 resumes from the top-performing candidates in the company. The algorithm then seeks out patterns and correlations, which contribute to their predictive power when analyzing the likelihood of success in a new candidate, based on their resume. Handing decision-making over to machine learning algorithms has many benefits for the humans in question, including saving time, money, and effort. However, when it

comes to the ethics and responsibility of the decision, the lines become blurred. Because we aren't able to understand exactly why a machine may have made the decision that it did, we aren't always able to detect and evade bias when it happens.

Bias in Machine Learning

retrieved from www.mathwashing.com

Mathwashing (Bias in Favour of Algorithms)

Mathwashing is a term coined to represent the societal obsession for math and algorithms, and the psychological tendency to believe the truth of something more easily if there is math or jargon associated with it — even if the values are arbitrary. Humans have a tendency to assume that the involvement of mathematics automatically renders something objective, since mathematical objects seem to be independent of human thought. Arguments against this is rooted in the very existence of mathematics, which was based on human thought. Math as a construct, along with its properties, exist as a product of human thought, which leaves it vulnerable to human subjectivity just the same as other measures.

Training Data 'Fairness in Classification'

We'll start with how algorithms are trained — machine learning algorithms are trained based on datasets that are chosen by the programmers. With this training data, they recognize and leverage patterns, associations, and correlations in the statistics.

For example, an algorithm can be trained to distinguish between a cat and a dog by being fed thousands of pictures of different cats and dogs. Classification is the easier of the tasks; applying an algorithm to a judgement call based on a *human* is much more multifaceted than that. For example, in the case of AI in the criminal justice system, specifically assisting judges in making a decision whether or not to grant parole to an offender — engineers can feed thousands of decisions and cases that were made by humans in the past, but all the AI can understand from that is the outcome of a decision. It still does not possess the sentience to understand that humans are influenced by so many variables, and rationality is not always the top tier of human decision-making. This is a problem coined by computer scientists called 'selective labelling.' Human biases are learned throughout many years of societal integration, cultural accumulation, media influences, and more. All of these learned biases seep into the algorithms that learn — just as humans, they don't start off biased. However, if given a flawed dataset, they might end up as such.

Societal Reflection

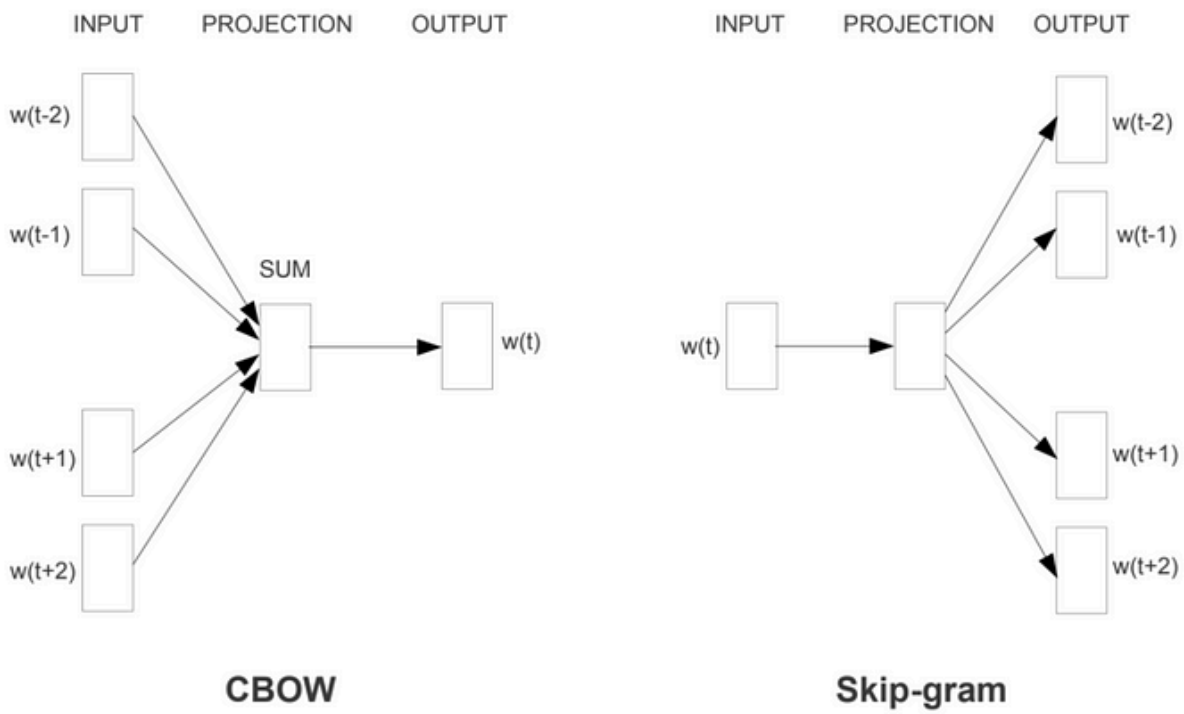
Algorithms are taught to make predictions based on information fed to it and the patterns it extracts from this information. Given that humans show all types of biases, a dataset representative of the environment can learn these biases as well. In this sense, algorithms are like mirrors — the patterns they detect reflect the biases that exist in our society, both explicit and implicit.

Tay, the Artificial Intelligence chatbot designed by Microsoft in 2016.

Take Tay, the original Microsoft chatbot, for example. Tay was designed to simulate the tweets of a teenage girl from interactions with Twitter users — however, in less than 24 hours, the internet saw Tay go from tweeting innocent things like “humans are super cool” to quite worrisome ones, such as “Hitler was right I hate the jews,” simply in virtue of the surrounding tweets on the internet. Microsoft removed the tweets, explaining that Tay had shown no issues in the initial testing phase, which had a training data set that featured filtered, non-offensive tweets. Clearly, filtering had gone out the window when Tay came online. This seems indicative of a possible method of bias

alleviation, which would be to monitor and filter incoming data as algorithms are put into use and engagement with the real world.

Word Embedding



Taken from “Efficient Estimation of Word Representations in Vector Space,” 2013

Word embedding is a technique used in machine learning in which words are translated to a vector — these vectors make up the dictionary of words for algorithms. Word embedding is widely used in many common applications, including translation services, search,

and text autocomplete suggestions. Depending on the angle of the vector, the machine would be able to understand the meaning of the word, in addition to commonly associated words and correlations. For example, the words king and queen were associated with prince and princess. The level of understanding of word embedding is capable of can be quite complex, making it a great tool to analyze things like SAT tests, job applications, cover letters, and so on.

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Taken from Bolukbasi et. al, 2016.

However, a problem with word embedding is that it has the potential to amplify existing gender associations. One study done by Bolukbasi

et. al at Boston University explored word embedding used in Google Translation services. The training period seldom involves many human engineers, and instead are trained based on libraries of natural language content such as news articles, press releases, books, etc. Bolukbasi investigated the relationship between Turkish to English translations, as Turkish phrases use gender neutral pronouns. In the translation, Google would be forced to choose a pronoun. The study found Google's sexism bias, as it translates "o bir doktor" to "he is a doctor," and "o bir hemsire" to "she is a nurse."

An 'aware' algorithm vs. an 'unaware' algorithm

At face value, the most simple approach to conquering the issue of fairness is to withhold the information that creates the bias in the first place; for example, in an algorithm that reviews resumes, eliminating the name and gender from the resume conceptually sounds like it could prevent gender bias from happening. If there is no information on what gender the person is, then the machine cannot possibly treat men and women differently, right?

It’s more complicated than that. What I just described above is called the ‘unaware’ approach to algorithm building. By removing this attribute, the premise is that gender will be a negligible factor when it comes to job competency. However, because algorithms are trained to identify patterns within the statistics, the existing correlations, stereotypes, and inequalities that are so embedded into society emerge wherever we go; they exist in reality, so they exist in the datasets that we train algorithms in too. Machine learning will be able to pick up on observable features associating gender that are not explicitly stated. For example, a hiring classifier may place weight on the length of ones military service and associate that with competency or loyalty, when in Israel, men typically serve 3 years, while women serve 2. Now you have an attribute that is closely correlated with gender, but having removed the essential information, you remove the context that is necessary to make an objective decision. For this very reason, an unaware algorithm can sometimes be more biased than its fully-informed counterpart.

On the other hand, the ‘aware’ approach does use gender information and takes into account the tendency for a shorter military term to be served by women. Mitigating these problems about accuracy and fairness often involve a trade-off — they cannot exist perfectly in the same realm. The unaware approach is a more fair process — it does not take into account sensitive attributes during its training phase. However, this can lead to a biased outcome. The aware approach uses a process that is more unfair — it takes into account sensitive classifications and informations, but can end up with a more objective outcome.

Feedback Loops/Self-Perpetuation

Furthermore, machine learning is prone to being stuck in feedback loops, which can end up perpetuating bias. For example, when machine-based prediction is used in criminal risk assessment, someone who is black is more likely to be rated as high-risk than someone who is white. This is simply due to the disparity in criminal records between black and white people, which unfortunately reflects

human bias in race. And because the machine has labelled yet another black person as high-risk, this new addition to the collection of data further tips the scale to be biased against black defendants. In this case, the system has not only reflected patterns learned from human bias but has also reinforced its own learning.

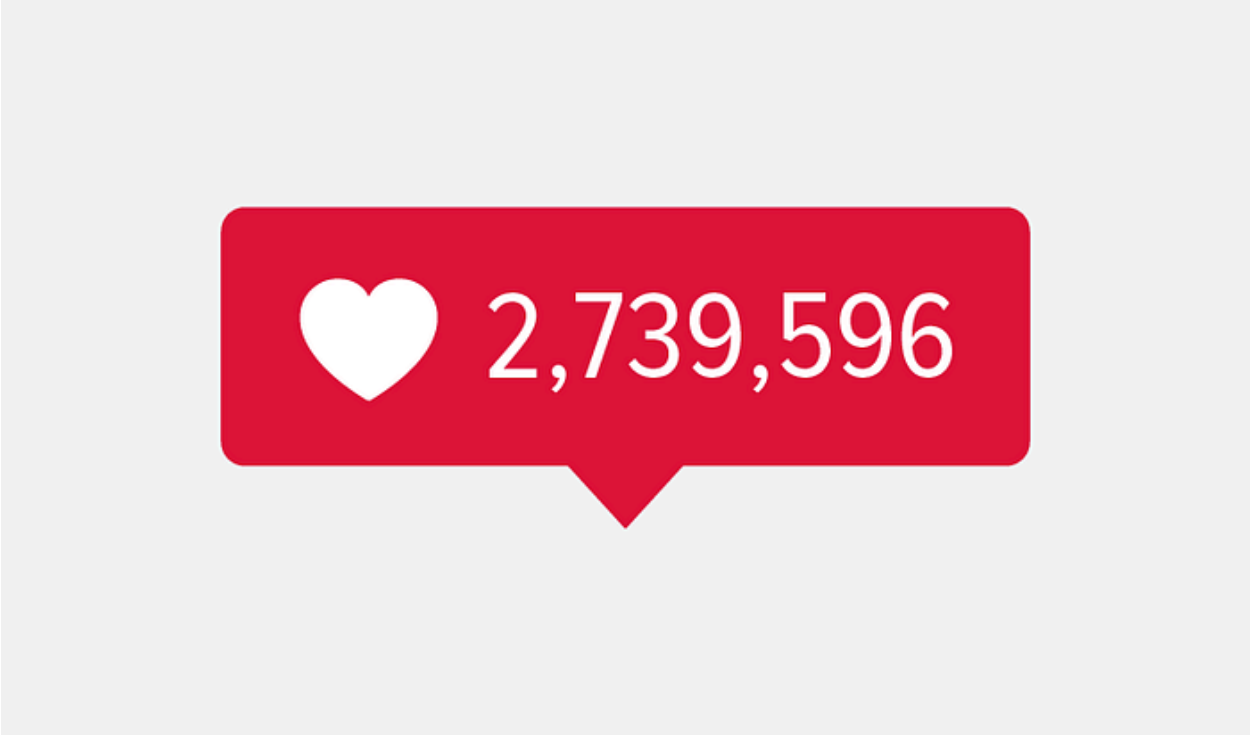
Surrogate Objectives

Besides problems within the training data, there are many ways in which bias can make its way into the process of an algorithm. Our next exploration is concerning the construct validity of the measures that propagate algorithms — is what you're trying to measure *actually* measuring what you need it to? And when it doesn't accurately measure, what are the consequences?

Social media algorithms no longer show posts based on chronological order, but rather, a machine learning algorithm filters through everything you have ever engaged with. The goal is to measure engagement — based on your previous interest, it will then show you

more content that it believes you would be likely to engage with. The higher the engagement rate on a piece of content, the more likely that algorithm is to take the piece of content and pop it on to others newsfeeds — in a perfect world, this makes sense. Posts that are popular should in theory be better content — otherwise, why would they perform so well?

Unfortunately, humans are not as smart as we need them to be in order for this algorithm to work the way it should. The content that performs the best consistently can be composed of fake news, celebrity gossip, political slander, and many other things that serve no purpose to the betterment of the world. But because these algorithms can't understand that, these echo chambers form, and it continues on.



Many of the decisions in the process of hiring practices are also being handed off to AI, in areas such as resume screening, job aptitude analyzing, and comparison. Job recruiting is an extremely timely process and has high costs for everyone involved — even higher if a mistake is made. The National Association of Colleges and Employers estimated the cost of hiring an employee to be around $7,600 at a medium sized company of 0–500. By letting an algorithm do the heavy lifting, a company can devote much of its resources and funds elsewhere, and hopefully end up with a successful choice.

However, surrogate objectives become a problem in this process, as many desirable job traits are very difficult to operationalize. Some of the industry buzzwords these days include ‘creativity,’ ‘communication,’ and ‘productivity,’ all of which are incredibly hard to measure. The most common test for measuring creativity is the alternative uses test, in which one comes up with unconventional uses for common items. Based on this measure, an employee may be assigned a ‘creativity aptitude’ score, which then is part of a training dataset that screens prospective employees for the same trait. The problems is that the alternative uses test only tests one aspect of creativity — divergent thinking. It neglects all other aspects of creativity, some which may be very valuable for company culture. You end up with a staff of creatives that are all creative in the same way — ironically boring.

As much as we romanticize the possibility of crediting machine learning algorithms for making important decisions, the truth is, they can’t understand objectiveness, truth, neutrality, or equality. All of

these traits are important considerations when human lives are at stake. Where do we go from here?

Conclusion

Although we've illuminated many of the problems that AI models can introduce, there are a multitude of reasons that companies may make the switch from a human-centered decision-making approach. As previously mentioned, despite all of its flaws, artificial intelligence is still more objective than humans. Because of this, we see a continued use of artificial intelligence in decision- and prediction-based tasks. But less biased is not equivalent to unbiased — what happens when an algorithm makes a biased decision? How do we decide who should take responsibility? It is not as if we can punish an algorithm for making a biased prediction (what would we do, erase it?).

Arguably, the best way to keep track of accountability is to keep accurate and detailed records of the processes of AI decision-making.

That is, the processes and data by which the decisions come to be made need to be transparent, so that if anything should go wrong, some third-party auditor is able to retrace the steps leading up to the outcome to locate the source of the problem. Bills and laws have been established to keep practices transparent for this purpose.

Of course, this method is not without problems of its own. Audit is not always feasible for artificial intelligence featuring big data, which are extremely large data sets, nor is it always applicable to systems engaged in deep learning, which feature large data sets as well as complex networks. Algorithmic autonomy and transparency seem have an inverse relationship — as these algorithms become increasingly better at 'learning' and adjusting, it becomes more difficult to understand where the biases occur. While auditing is effective for simpler models, we may need a different way to alleviate bias for complex algorithms.

Another way of mitigating bias is aimed at the the trainers and creators of the AI. By making them aware of their own prejudices, we have a better chance of keeping it out of the algorithms. It's important to note that human bias exists and is hard to mitigate due to it being an evolutionary trait, but we are becoming increasingly more aware of the biases that our own brains are susceptible to. To conclude, algorithms can be part of alleviating institutional bias — if we remain educated, aware, smart, and selective.

"The best thing to do is to keep trying to make culture better, and to keep updating AI to track culture while it improves." Joanna Bryson

References:

Abate, Tom., Krakovsky, Marina. "Which is more fair: a human or a machine?" Stanford Engineering, January 31, 2018.

Bornstein, Aaron M. "Are Algorithms Building an Infrastructure of Racism?" Nautilus, December 21, 2017.

Bright, Peter. "Microsoft Terminates Its Tay AI Chatbot After She Turns Into a Nazi." Ars Technica, March 24, 2016.

Courtland, Rachel. "Bias Detectives: the researchers striving to make algorithms fair." Springer Nature, Macmillan Publishers, June 21, 2018.

Miller, Alex P. "Want Less-Biased Decisions? Use Algorithms." Harvard Business Review, July 26, 2018.

Schep, Tijmen. "What is Mathwashing?" Mathwashing, 2018.

Shapiro, Stewart. "The Objectivity of Mathematics." *Synthese*, vol. 156, no. 2, 2007, pp. 337–381.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai, A. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." Microsoft Research New England, 2016.

Yona, Gal. "A Gentle Introduction to the Discussion on Algorithmic Fairness." Towards Data Science, Medium. October 5, 2017.