Machine Learning And Discrimination

Isaac Speed

Truman State University

Introduction

The world has hit the new Industrial Revolution: big data. Machine learning has exploded, solving problems once thought to be impossible for computers, from beating the world's best Go players to diagnosing cancer better than doctors could. These algorithms can run through hundreds of scenarios in a second. Between this efficiency and all the data available today, companies are automating more processes than they could even five years ago. These changes will revolutionize our world as the field develops, but there is a darker side. There is little regulation on where these algorithms are used and what data can be learned from. Often, data scientists do not apply proper vigilance to their algorithms; they source data that contains discrimination which then becomes encoded in their algorithms. With how widespread these algorithms have become, it can greatly increase the inequality present in modern society. Machine learning can, has, and will affect systemic discrimination.

Understanding and solving this problem requires two different approaches: looking at these algorithms and data from the perspective of society, and looking at them from the perspective of computer science. In order to understand how data affects the algorithms and vice versa, we must consider them in the context of the society in which they are run. To more deeply understand how the algorithms learn to discriminate (and what we can do about it), we must look at machine learning from its native field, computer science. This paper examines both sides, starting with computer science.

Computer Science

In the age of the internet, algorithms govern our lives in hidden ways. From applying to a credit score to being granted parole from a judge, computers influence countless decisions. Machine learning algorithms have made processes exponentially more efficient, but they've obfuscated the details of those processes. When those algorithms learn from biased data and go onto make decisions for millions of people, however, it can cause problems. Most applications of machine learning have grown complex enough that it is difficult to understand why they make the decisions they do, making it near impossible to audit the innards of an algorithm for discrimination. When discrimination does slip through the cracks, as it often does, the algorithm creates more biased data, reinforcing the discrimination.

It is almost impossible to live in the modern Western world without algorithms affecting your daily life. Whenever you browse the internet, the ads you see are determined by algorithm. While this may seem innocuous, it can also further discrimination. In one study, it was discovered that males are nearly six times more likely to receive ads for career coaching for high-paying executive positions compared to women when controlling for all other aspects (Datta, Tschantz, & Datta, 2015). More concerningly, algorithms are being used to influence the decisions of judges. In the judicial system, companies like Northpointe have created software to produce a score that predicts a criminal's chance of recidivism (Liptak, 2017). This score is then given to a judge when deciding whether to offer the accused parole or early release from prison. These algorithms, despite their effect on people's lives, are black boxes to the public. They are the intellectual property of the companies, hidden for the most part from third-party vetting. We can only examine the results of the decisions, not what led to the decisions (Liptak).

One of the most inscrutable, yet also most popular, classes of algorithms is something new to mainstream use: machine learning. Rather than programming specific rules into a computer--in this case, do this; in that case, do that--an engineer can give the computer a goal, a supply of data, and let it learn from that data. It can complete tasks that are far too complex to program explicit rules for. This has led to incredible jumps in what computers can do: from painting pictures and creating music, to driving cars and detecting cancer, code is moving into fields that were once thought to be suitable solely for humans. The technology holds huge promise, not only for automating tasks and saving lives, but for reducing discrimination. They hold the promise of life-changing decisions being made by objective algorithms, rather than biased individuals.

This, however, is a false promise. While algorithms themselves are objective, the data they are given (and the rules they learn) are not. Machine learning algorithms are only as good as the data they are given. In a perfect world, they would learn from unbiased data that fairly represents all demographics, and those algorithms would make perfect extrapolations form that data. We do not live in that world. Data for many important tasks, like selecting who gets a good loan or which neighborhoods police should patrol, come from decisions that have previously been made by subjective, biased human beings. This historical data has been tainted by both implicit biases and outright discrimination by those in charge of those decisions. Take, for example, selecting which neighborhoods a police department should patrol (as done by software such as PredPol). You could collect data from police departments on which neighborhoods they feel are the most dangerous, and which they patrol most often. You could then give this data to a machine learning algorithm. It would learn from this data how to predict which neighborhoods

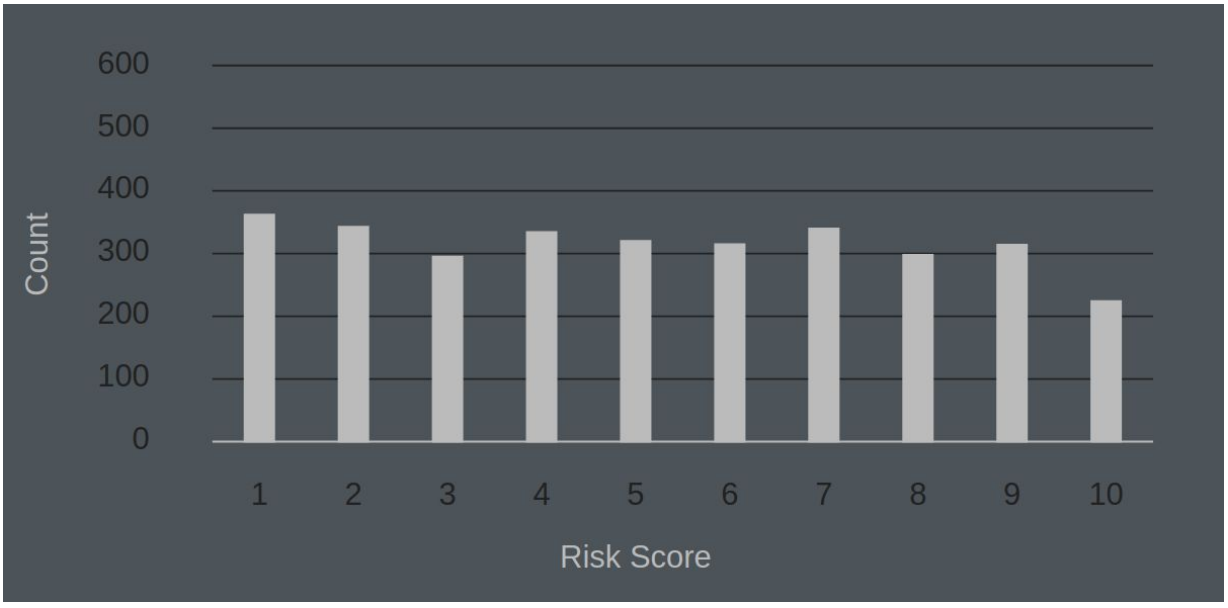
should be patrolled most--as far as the data tells it. This data, however, will be contaminated. It is quite possible that poorer neighborhoods with a large minority population will be selected more often than mainly white neighborhoods, even if they have the same crime rate. The algorithm will then mimic those decisions, sending the police to those neighborhoods more often. In turn, this will generate more data targeted against poorer locales, feeding the problem.

Despite how machine learning may seem magical, it is powered by mathematics and statistics. The process is an algorithm, albeit a very effective and widely applicable one. Numerical 'weights' are created randomly and associated with each feature of the data, such as one's income or a portion of an x-ray. These weights are then layered on more weights, allowing the algorithm to make more complicated solutions. Once all those parameters are created, the training begins. The engineer feeds data into the algorithm, which makes predictions and then compares it to the desired output. If the output is wrong, it tweaks the weights so that next time the result will be closer. This is repeated millions of times, if not more, until it can calculate the desired output for the training data and, hopefully, out in the field. By multiplying these weights together, we can do remarkable things like build self-driving cars.

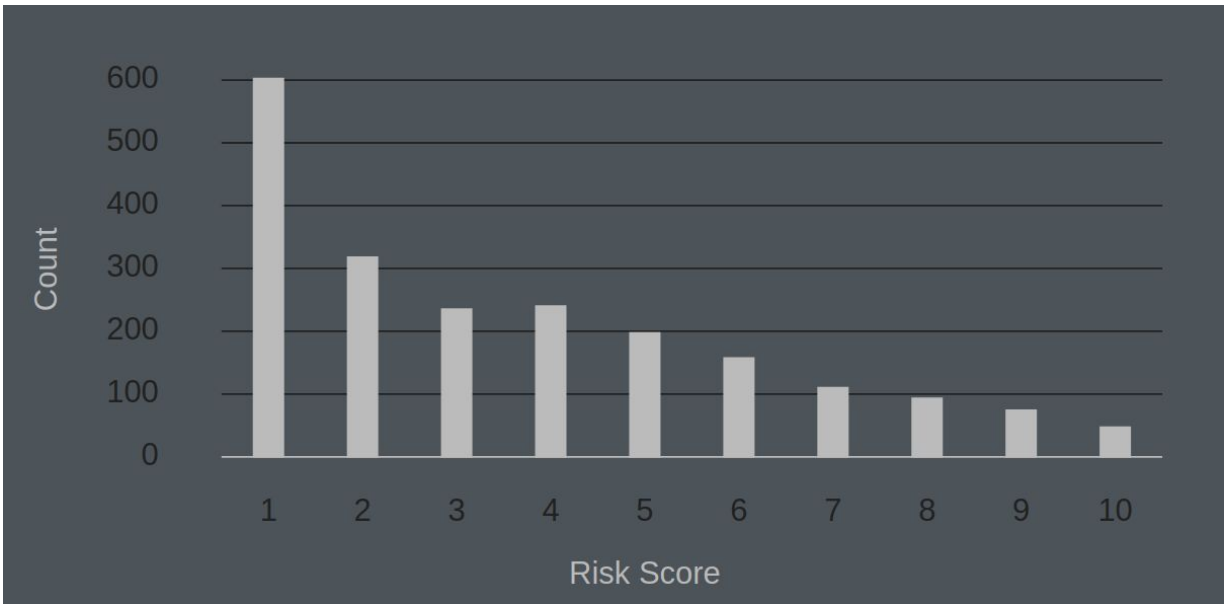
The rub, however, is the vastness of those numbers. AlexNet, a famous neural network created to recognize one thousand different objects in pictures, has sixty million parameters (Krizhevsky, Sutskever, & Hinton, 2017). When a neural network is trained for anything beyond the most trivial tasks, the number of weights grows far too large for a human to intuit how it makes the decisions it does. Because of this, an engineer cannot tell at a glance if a neural network is correct, let alone if it discriminates between demographics. It is not enough to simply

train an algorithm, glance at the accuracy, and decide it is finished. Yet all too often, that is what happens.

One of the most striking examples of these failures comes from Northpointe's COMPAS software for determining rates of recidivism. Using data pulled from criminal records and a questionnaire, it uses an algorithm to create a score suggesting how likely the person in question is to re-offend in the next two years. It boasts a 61% accuracy rate for predicting recidivism. This leaves the other 39% percent of defendants to be incorrectly classified. While rates of misclassification for both white and African-American defendants are similar, the way they are misclassified is different. ProPublica examined the results for more than 7,000 people in Florida, isolating race from other factors like criminal records, and found that African-American people are 77% more likely than white people to be flagged as high risk for violent crime. They are 45% more likely to be flagged as high risk for any kind of crime. White defendants were more likely to incorrectly be given a lower score (Mattu, 2016).



Scores for African-American defendants (Mattu, 2016)

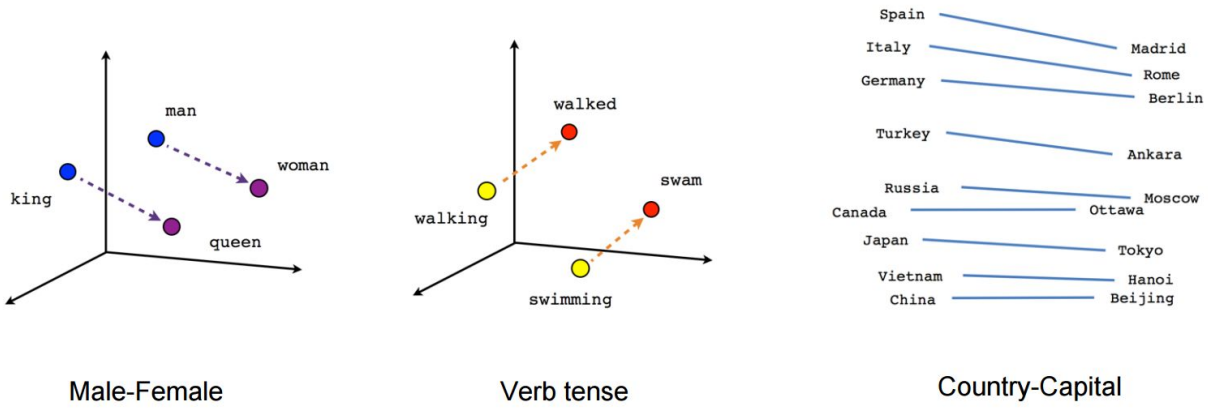


Scores for white defendants (Mattu, 2016)

Because the algorithms are proprietary, it is difficult for defendants to challenge their score. They cannot point out where in the algorithm discrimination may have occurred. One defendant, Eric Loomis, launched a lawsuit against the state of Wisconsin, claiming that the assessment was unconstitutional for that reason, as well as for considering men and women differently. While they did agreed that the process was entirely opaque, the courts rejected his lawsuit as he could not show the assessment was inaccurate (Liptak, 2017). This gave legal precedence to allowing black-box algorithms to make decisions unquestioned.

In a completely separate field, algorithms are reinforcing other social biases. One field that machine learning has had remarkable success with is natural language processing. Natural language processing uses computer programs to process human languages. One of the most

prominent examples of this is machine translation (e.g., Google Translate). Other uses of natural language processing include sentiment analysis (determining the emotion of a text) and using a program to generate text. A common technique for improving performance is generating word embeddings: a machine learning algorithm looks at a large collection of text and maps the meanings of words to a vector space (similar to an XY plane from algebra, but with many more dimensions). Each word can then be represented as a coordinate in that vector space, grouped with similar concepts. Because the meaning is encoded in the vector representation of the word, one can do a sort of arithmetic of analogies. In the leftmost graph below, for example, ‘king’ - ‘man’ + ‘woman’ = ‘queen’.



Graphical examples of word embeddings ("Vector Representations of Words").

This is an incredibly useful technique, driving innovations in natural language processing. However, it too learns from biased sources. The text it learns from was written by human beings, and thus reflects the biases of the society it was written in. “As machines are getting closer to acquiring human-like language abilities, they are also absorbing the deeply

ingrained biases concealed within the patterns of language use," writes the Guardian, examining recent research into how word embeddings can encode bias by Caliskan, Bryson, & Narayanan (Devlin, 2017). The research examined pre-trained word embeddings from GloVe, an algorithm for creating word embeddings developed by researchers at Stanford. The embeddings were created based on text gathered from the internet; the training set contained roughly 840 billion words. Through their research, they found that names associated with being European were significantly more associated with pleasant rather than unpleasant concepts, compared to African-American names. They also found that female words (such as 'she' and 'woman') were more strongly associated with family than career words, compared to male words. Similarly, they were more closely associated with the arts than science or mathematics (Caliskan, Bryson, & Narayanan, 2017). A previous study corroborated gender bias in word embeddings. They used word embeddings from word2vec, created by Google and trained on text from Google News. In this embedding, they found that 'man' - 'woman' is nearly the same as 'computer programmer' - 'homemaker' (that is, 'man' is to 'woman' as 'computer programmer' is to 'homemaker'). Later on in the paper, they propose a method to debias word embeddings by removing certain associations from words. They demonstrate this technique by removing gender biases, while still preserving appropriate gender analogies. Before debiasing, 19% percent of the analogies were found to be gender biased by ten humans evaluating them; after debiasing, only 6% were found to contain gender bias (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016).

Extreme she words	**Extreme he words**
Homemaker	Maestro

Nurse	Skipper
Receptionist	Protege
Librarian	Philosopher
Socialite	Captain
Hairdresser	Architect
Nanny	Financier
Bookkeeper	Warrior
Stylist	Broadcaster
Housekeeper	Magician

Occupations most strongly associated with femininity and masculinity. Words such as businesswoman, where gender is part of the definition, are excluded (Bolukbasi, et al., 2016)

A more striking example of how natural language processing can go wrong is Tay, a chatbot released onto Twitter by Microsoft in 2016. Tay was designed to learn from the users it interacted with and reply with seemingly intelligent messages. On the website created for her, Microsoft wrote, “The more you chat with Tay the smarter she gets, so the experience can be more personalized for you” ("Meet Tay", n.d.), Microsoft, however, did not anticipate what she would learn from the internet: within twenty-four hours, she was besieged by trolls sending racist and offensive messages. While at first her messages were wholesome, soon she began to tweet politically charged messages. While many of the worst were simply her parroting what users said

to her (using a “repeat after me” function), the tweets she generated were still appropriate (Price, 2016).

Tweets from Microsoft’s Tay; on the left, she claims to support genocide; on the right, she declares the Holocaust was made up (Price, 2016).

In addition to tweets supporting genocide and denying the Holocaust, she also voiced agreement with the Fourteen Words, a popular white supremacist slogan (Price, 2016). Sixteen hours after her release, the experiment was suspended by Microsoft. Many of the most offensive tweets were deleted, and Microsoft offered a public apology. However, it is important to note that Tay did not truly support genocide. She was a chatbot that generated replies based on previous conversations she had seen: she had no concept of what the Holocaust or genocide are. “When Tay started training on patterns that were input by trolls online, it started using those patterns. This is really no different than a parrot in a seedy bar picking up bad words and repeating them back without knowing what they really mean,” Louis Rosenberg , founder of Unanimous AI, said on the topic. (Reese, 2016) The most important lesson for Microsoft to learn

from the situation was not that artificial intelligence can recite racist propaganda, but that you must be careful of the environment your algorithms learn from. As Joanna Bryson said in response to the bias of word embeddings, “A lot of people are saying this is showing that AI is prejudiced. No. This is showing we’re prejudiced and that AI is learning it” (Devlin, 2017).

Not only can machine learning applications learn from bias in their environment, they can also perpetuate it. When an algorithm is trained on a biased data set, it learns to produce similar results. If that algorithm is then used in the real world, it will then generate more data based on its biased results. This creates a feedback loop, reinforcing biases. The White House gives an example in a report they released on big data and discrimination:

“Unintentional perpetuation and promotion of historical biases, where a feedback loop causes bias in inputs or results of the past to replicate itself in the outputs of an algorithmic system. For instance, when companies emphasize ‘hiring for culture fit’ in their employment practices, they may inadvertently perpetuate past hiring patterns if their current workplace culture is primarily based on a specific and narrow set of experiences. In a workplace populated primarily with young white men, for example, an algorithmic system designed primarily to hire for culture fit (without taking into account other hiring goals, such as diversity of experience and perspective) might disproportionally *[sic]* recommend hiring more white men because they score best on fitting in with the culture” (Muñoz, Smith, & Patil, 2016).

Algorithms can also perpetuate discrimination on a larger scale, with higher stakes, as in the case of Northpointe’s COMPAS. If engineers and data scientists are not careful when collecting their data and implementing their algorithms, they may end up doing just that. In many cases, protected attributes such as race have been removed from the data sets algorithms are

trained on, with the intention of preventing bias. However, this is not always enough. Often, those protected attributes can be redundantly encoded in the data; even if the attribute is not explicitly included, it can be inferred from other data (e.g., one's gender can often be determined from one's name) (Hardt, 2016). Companies are collecting massive amounts of data, redundantly encoding even more attributes.

In many applications, data scientists are not doing enough to make machine learning a force for equality, rather than against it. The issue is pressing enough that the White House released two reports on how big data can affect civil rights. In the first report, they provide some examples of discrimination, but affirm that machine learning can protect civil rights. They acknowledge "it is implausible for consumers to be presented with the full parameters of the data and algorithms," but defend a consumer's right to have some insight into the process when their personal information influences an algorithm. They conclude, "Big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace. Americans' relationship with data should expand, not diminish, their opportunities and potential" (Podesta, Pritzker, Moniz, Holdren, & Zients, 2014).

In the second report, they delve deeper into the details. They thoroughly dismantle the assumption that because big data relies on algorithms, it is objective. Having dismissed that misconception, they list some case studies of how big data can be used fairly and where it can go wrong. They examine how it could be used in law enforcement, employment, and access to higher education. Throughout, the writers stress how important it is to be cognizant of biases and to question the algorithms. It is not enough to just feed data into the program and hope for the

best. The data must be examined for biases and the developers must work to counteract them. In the design process, they must consider how groups could be marginalized and why. The report concludes with a list of recommendations, such as supporting transparency and continuing research into mitigating discrimination (Muñoz, et al., 2016).

However, it is not enough to simply make algorithms transparent. One of the difficulties with designing fair algorithms is defining what is fair. Describing what 'fair' is in plain English is hard enough; it is even more difficult to design an algorithm to decide. Simply removing the protected attributes is not always enough, due to redundant encoding. Another proposed solution is called *demographic parity*, which means there should be no correlation between the output of the algorithm and a protected attribute. This fails on three accounts: it does not ensure fairness and it can decrease accuracy. For one demographic, it could provide accurate, useful results, while for a minority it could guess at random. In some cases, there should be a correlation between protected attributes and the results. Biological females are more likely to buy feminine hygiene products than biological males, but demographic parity would not allow an algorithm to model that. Finally, it does not account for combinations of protected attributes (Hardt, 2016).

Another approach to deal with classification (such as whether someone will default on a loan or not, or whether they will commit another crime) that has been suggested is equalized odds. Similar to demographic parity, it pushes the algorithm to match probabilities of certain outcomes between demographics. However, rather than ensuring there is no correlation, it incentivizes the algorithm to have equal true positive rates and equal false positive rates across demographics. This allows the classification to depend on the protected attribute, but only if the true classification does. By focussing on the true positive and false positive rates, it enforces

equal accuracy between demographics. By applying this criterion, the algorithm will avoid classifying the majority well while neglecting the minority. Many algorithms are vulnerable to this, as there tends to be less data on minorities for algorithms to learn from. One failing with this approach, however, is that it cannot filter out bias from the data. It can compensate for one demographic having less data, but it cannot recognize discrimination against a specific demographic if it is part of the data it learns from. Because of this and other limitations, the researchers acknowledge that their criterion should not be considered a proof of fairness, but a measure of unfairness (Hardt, Price, & Srebro, 2016).

Despite the difficulty of defining fairness, progress is being made. Discrimination in machine learning is coming into the public eye, putting companies under greater scrutiny. Research continues on reducing discrimination in algorithms and using machine learning to detect discrimination. As the field continues to evolve and become more accessible, solutions will be proposed and new algorithms will be developed, slowly pushing back against the bias of historic data.

Machine learning is a young field, but it has already had a substantial impact on our society. Ensuring fairness in machine learning is an even younger field. Unfortunately, these algorithms have already begun to solidify discrimination in areas like the judicial system; however, they have the potential to increase opportunities for minorities, removing explicit human bias and making discrimination more quantifiable. There is much work to be done, but at least it is easier to permanently change an algorithm than to change a person. The White House assessed the situation well when they said, “Big data is here to stay; the question is how it will be used: to advance civil rights and opportunity, or to undermine them” (Smith, Patil, & Muñoz,

2016). Machine learning, too, is here to stay. We are only beginning to see the effects it will have on our world.

Social Science

The 21st century is quickly becoming the century of data. More and more of the world is being collected and organized into data sets, then used to fuel algorithms. These algorithms are then used to make decisions, ranging from recommending a movie on Netflix to recommending which prisoners receive parole (Winerip, M., Schwirtz, M., & Gebeloff, R. 2016). Many of these algorithms are black boxes: it can be nearly impossible to decipher how they decide what they do. This makes it difficult to question the validity of its conclusion. It is especially troublesome given that these algorithms learn from historical data and the biases embedded in it. Even if they are trained on perfect data, containing no biases, it is still possible it will put minorities at a disadvantage due to lack of data. While these algorithms can perpetuate discrimination if careful thought is not given to their design, they also have great potential to fight implicit bias and improve equality of opportunity for all demographics.

A common saying in learning algorithm circles is, "Your algorithm is only as good as your data." Many, however, do not consider the society they live in when collecting data. The United States, since its inception, has had a long history of racism, sexism, homophobia, and systematic discrimination. Nearly all data, from news articles to arrest records, has been contaminated by that bias. Those practitioners' algorithms then learn from that bias and reproduce it. Since the majority of workers in the technology industry are affluent white males, discrimination is often the last thing on their mind. This is dangerous when the algorithms they

create can change the course of thousands of lives. Cathy O'Neil (2016, September 12), author of *Weapons of Math Destruction*, described in an interview the kinds of algorithms that can be the most dangerous:

"They have three characteristics. The first is that they're high-impact, they affect a lot of people. It's widespread and it's an important decision that the scoring pertains to, so like a job or going to jail, something that's important to people. So it's high-impact. The second one is that the things that worry me the most are opaque. Either that means that the people who get the scores don't understand how they're computed or sometimes that means that they don't even know they're getting scored. Like if you're online, you don't even know you're scored but you are. And the third characteristic of things that I care about, which I call weapons of math destruction, the third characteristic is that they are actually destructive, that they actually can really screw up somebody's life. "

Despite how opaque the algorithms are and the lasting effects they can have, people do not question them enough, Cathy O'Neil (2016, October 13) says. "There are a lot of issues, but the most obvious one is the trust itself: that we don't push back on algorithmic decisioning, and it's in part because we trust mathematics and in part because we're afraid of mathematics as a public." Machine learning is sold to the public as something almost magical, a divine combination of linear algebra, calculus, and statistics. For that reason, questioning the process can seem beyond a lay person, especially when even experts cannot determine how an algorithm lands at a decision. When the public does not question the algorithms, it makes it all the harder to discover which algorithms treat demographics unfairly.

The majority of these algorithms learn from historic data, aiming to imitate the decisions made previously. In many important areas, such as employee and parole records, that data has been tainted by systemic biases. The algorithms then do exactly what they were instructed to and learn to imitate those biases. Cathy O'Neil (2016, September 12) in an interview explained a simple example of how it can occur:

"An engineering firm that decided to build a new hiring process for engineers and they say, OK, it's based on historical data that we have on what engineers we've hired in the past and how they've done and whether they've been successful, then you might imagine that the algorithm would exclude women, for example. And the algorithm might do the right thing by excluding women if it's only told just to do what we have done historically."

Similarly, an algorithm trained to decide who should receive parole may be biased against African-Americans. In an analysis by The New York Times of thousands parole decisions over the course of several years, they found that fewer than one in six Hispanic or black men were released at their first hearing, whereas one in four white men were released at their first hearing (Winerip, M., Schwirtz, M., & Gebeloff, R. 2016). Even if their race is stripped from the data, it can be inferred from other characteristics such as their name or zip code. The algorithm can then learn to grant parole as the historical data does: favoring white men, even with race removed from the data.

The potential for machine learning to perpetuate and even increase inequity exists beyond the criminal justice system. In the private sector, they affect how mentally ill people obtain jobs. Kronos, a company that creates management software, developed a questionnaire for job applicants. The potential employee's answers are fed through an algorithm that outputs an

assessment score. Companies can use this in conjunction with other methods, like interviews, to evaluate applicants, or to screen them out entirely. Some believe that the algorithm discriminates against those with mental illness. Cathy O'Neil (2016, October 06) elaborates, "So we have reason to believe that some of these algorithms discriminate based on mental health status. This is illegal under the Americans with Disability Act. You're not allowed to give somebody a health exam, including a mental health exam, as part of hiring. *[...]* And if some of those personality tests filter out people with mental health problems, then that's a real problem. So not only does it destroy their chances of getting a job, but it is exactly creating that feedback loop that the ADA is meant to stop, which is actually isolating that group of people from normal society."

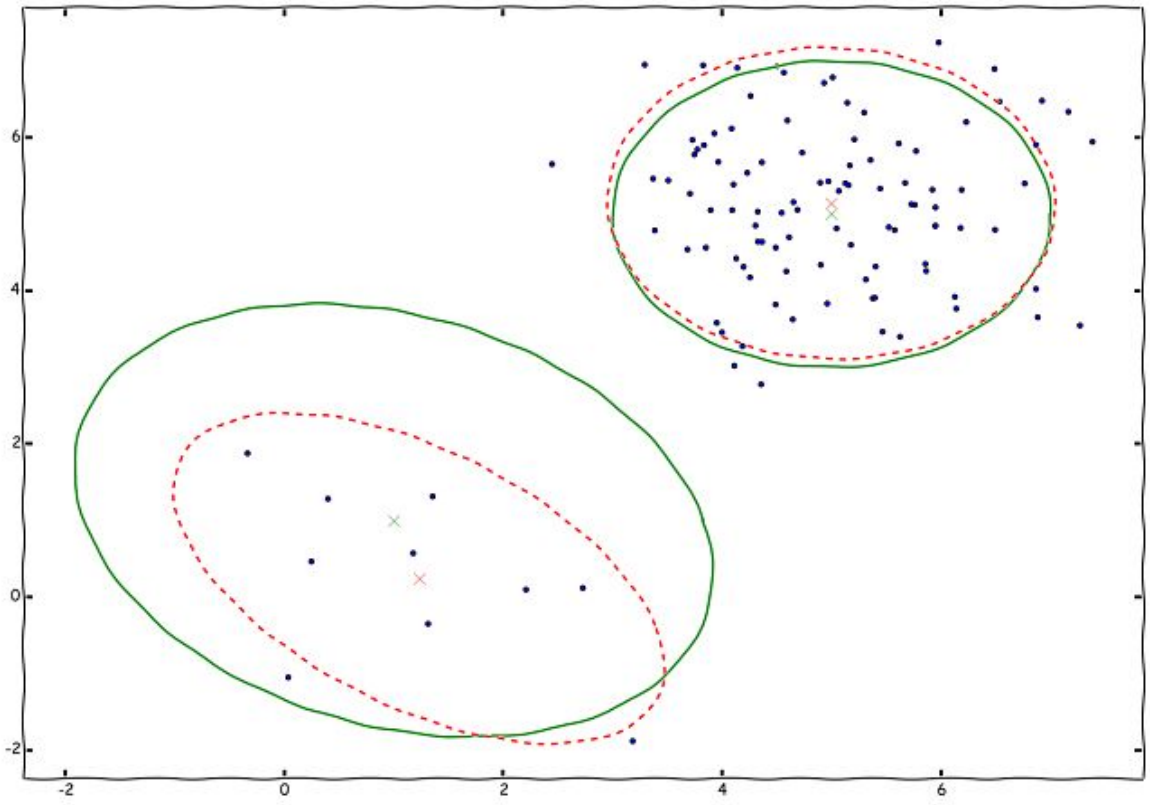
While many forms of discrimination can create the feedback loop that sustains bias, in some ways algorithms are especially potent. Most algorithms are black boxes: those they affect have no insight into how they work. Even if they did and those affected found evidence of discrimination, algorithms feel no social or moral pressure to reduce their bias, unlike people. Additionally, they work far faster than humans are capable of. These algorithms can generate massive amounts of data, biased against a certain demographic, that could then be used to train other algorithms. This would continue to perpetuate systemic inequality, even if social attitudes change.

Furthermore, those who are vocally biased against a demographic can use algorithms as an argument. One only needs to look back to 1994 with the release of *The Bell Curve* by Richard J. Herrnstein and Charles Murray to see a famous example of arguing for racism with statistics. This same logic could be used with machine learning; biased groups could use a mix of false ethos and logos, appealing to the authority of mathematics and so-called objective algorithms.

There are a few flaws with this argument: first, mathematics does not dictate the truth of the algorithm's results. It only ascertains that it can approximate the function governing the data. This leads to the next flaw: that function does not necessarily say much about the demographic. Going back to the example of parole, an algorithm may recommend that African-American men as less likely to get parole than white men. This, however, does not mean that they are should not receive parole. It means that in the data set it was trained on, African-American men received paroles less often. Rather than mapping objective reality, it is mapping the choices humans have made in the past.

Even when an algorithm is trained on data that is not explicitly biased, there is still a danger for some demographics to be treated less fairly than the majority. Minorities, by definition, are less prevalent in a population than the majority; consequently, in many cases there will be less data about them. With learning algorithms, the more data you have, the more accurate your algorithm becomes. Thus, an algorithm trained on a data set (even if it does not contain explicit bias) could output more accurate results for the majority than minority groups. Below is an example (Hardt, M. 2014); a learning algorithm tries to learn the parameters of two random distributions. The predicted parameters are shown in red, while the correct parameters are shown in green. In the upper right corner, where far more data is available, the parameters are match the actual population well. The lower corner, containing far less training samples, has not been learned nearly as well by the algorithm. This is demonstrated in a trivial example; in the real world, with a far more complex model, one subpopulation being underrepresented can ensure the algorithm hardly works better than flipping a coin for that demographic.

Corroborating the problem, minorities can affect the error far less than the majority. Moritz Hardt, a machine learning researcher, explains, that even if the algorithm achieves a 5% error, nothing can be concluded about its accuracy for minorities. That error could come from only correctly classifying minorities half the time, while classifying the majority right every single time (2014).



An expectation–maximization algorithm tries to learn the parameters of two subpopulations. The dashed red ovals describe the estimated covariance matrices, while solid green define the correct covariance matrices (Hardt, M. 2014).

Another issue that can plague algorithms is misrepresentation of the target population, even if the data is unbiased. Depending on how the data is collected, it is possible one or more

demographics may be inadvertently underrepresented in the data set even if they are not minorities. Suppose data on road conditions are collected through a smartphone app in order to determine which roads need repair. This could result in city services being directed away from areas populated by elderly people, as they are less likely to carry smartphones (Podesta, Pritzker, Moniz, Holdren, & Zients, 2014). Simple oversights like that can create lasting issues, even if an engineer could find a source of unbiased data. (Some would argue that there is no such thing as unbiased data.)

Although many of these studies paint a stark picture of machine learning's influence on civil rights, there is hope. When developers choose to focus on creating fairness, the learning algorithms today can help equalize opportunities for all demographics, even in the face of societal bias. Though data is biased, the algorithms themselves are not; it is a matter of calibrating them and the data they are trained with. Additionally, unlike humans, it can be easier to audit their results for bias. Much of machine learning may be a black box, incomprehensible at scale, but the data they output is not. Massive quantities of examples can be run through the algorithm almost instantly; those results can then be scrutinized by statistics, as well as other machine learning algorithms, to detect bias. Work is being done currently to automate this process (Galhotra, S., Brun, Y., & Meliou, A. 2017).

Many are pushing for greater transparency in how algorithms are developed and used. When a company hides how its algorithms make decisions, it makes it more difficult to determine whether discrimination is occurring and, if so, why. Groups such as EPIC have pushed for reducing the opaqueness of algorithms ("EPIC - Algorithmic Transparency", n.d.), while the White House released two reports on big data also recommending algorithms be more open to

the public (Podesta, J., et al. 2014; Muñoz, Smith, & Patil, 2016). The second report also called for increased education about big data (Muñoz, et al.). An educated public able to scrutinize the algorithms used by companies will put pressure on them to ensure that they are fair.

At the time of this writing, there are no regulations in place specifically for learning algorithms. Machine learning is a relatively young field and lawmakers have yet to catch up. The European Union recently approved the General Data Protection Regulation, set to be fully enforced in May, 2018 ("Home Page of EU GDPR," n.d.). It is applicable to organizations located in the European Union, as well as organizations outside who provide goods or services or monitor citizens of the European Union ("FAQ of EU GDPR," n.d.). It provides the following rights to individuals when dealing with those organizations:

1. The right to be informed
2. The right of access
3. The right to rectification
4. The right to erasure
5. The right to restrict processing
6. The right to data portability
7. The right to object
8. Rights in relation to automated decision making and profiling ("Individuals' rights").

This will be a huge step towards preventing discrimination. It requires that individuals affected by algorithmic decisions have a right to learn the logic behind the decision and challenge it. As it applies to companies not just in the European Union, but those who do

business with citizens of the European Union, it could be a catalyst for transparent, challengeable algorithms to become the norm.

Machine learning is quickly permeating every sector of modern life, from applying to jobs to leaving prison on parole. It has the potential to reinforce discrimination and hide it from public scrutiny, but it could also be used to ensure decisions are made as fairly and objectively as possible. While there is still a way to go before it is as easy as telling your algorithm to not be bigoted, it is possible today to train algorithms that will treat all demographics fairly. The question is not whether humanity is capable of reducing inequality; the question is whether we will.

Conclusion

Discrimination is a difficult, pervasive problem and there is no easy solution, even in an isolated segment like this. It must be approached from multiple disciplines; society must work to reduce the discrimination present in the data sets, while computer scientists must work to prioritize fairness in their algorithms. Research is being done to make this possible, but a change in policy must accompany this process or it will not be implemented across the board.

Big data and machine learning has made huge waves in society, and they are set to change the world even more in the next few decades. We must make sure that the changes it effects are for the good of everyone.

References

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016, July 21). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Retrieved October 03, 2017, from http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings

The authors examine word embedding techniques and find that they encode the implicit biases society uses in its language (e.g., ‘man’ is more closely associate with ‘doctor’ than ‘woman’ is). It talks about the potential for applications of word embeddings to increase the gender bias in language. They then lay out an algorithm for reducing gender bias in embeddings while leaving proper relationships intact.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017, April 14). Semantics derived automatically from language corpora contain human-like biases. Retrieved October 03, 2017, from http://science.sciencemag.org/content/356/6334/183.full

The authors focus on examining the biases contained in word embeddings, creating a test similar to the Implicit Association Test. They show that embeddings have captured many relations humans already know (flowers are more closely associated with pleasant things than insects are; therefore, flowers are more pleasant than insects), as well as biases humans show. They found that in the embeddings, African-American names were less closely associated with pleasant

things than Caucasian-sounding names, and that male names were more closely associated with career words than female names.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies,2015*(1). doi:10.1515/popets-2015-0007

Datta and Datta found that men were more likely to be shown ads about high-paying jobs than women.

Devlin, H. (2017, April 13). AI programs exhibit racial and gender biases, research reveals. Retrieved October 03, 2017, from https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals

Devlin examines the research done by Caliskan et al. and has some useful quotes by Bryson, such as: “As machines are getting closer to acquiring human-like language abilities, they are also absorbing the deeply ingrained biases concealed within the patterns of language use,” and “A lot of people are saying this is showing that AI is prejudiced. No. This is showing we’re prejudiced and that AI is learning it.”

EPIC - Algorithmic Transparency: End Secret Profiling. (n.d.). Retrieved October 03, 2017, from https://epic.org/algorithmic-transparency/

EPIC is a group that argues for privacy rights. In this page, they list various resources for algorithmic transparency, such as White House reports and lawsuits.

Frequently Asked Questions about the GDPR. (n.d.). Retrieved October 03, 2017, from http://www.eugdpr.org/gdpr-faqs.html

This page answers questions about the European Union's General Data Protection Regulation, such as who it applies to.

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness testing: testing software for discrimination. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2017*. doi:10.1145/3106237.3106277

Galhotra et al. create a software suite, Themis, for automatically testing software for bias. They talk about one possible measure of fairness (group discrimination) and its shortcomings, as well as offering a 'real-world' example. They introduce a new method, causal discrimination (the software must give the same output if two individuals differ only in a protected attribute).

Hardt, M. (2016, September 6). Approaching fairness in machine learning. Retrieved October 03, 2017, from http://blog.mrtz.org/2016/09/06/approaching-fairness.html

In this article, Hardt talks about how machine learning can learn from biased data, even if protected attributes are removed. He examines one methodology of determining fairness, demographic parity, and points out the weaknesses in it.

Hardt, M. (2014, September 26). How big data is unfair – Moritz Hardt – Medium. Retrieved October 01, 2017, from https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Hardt examines how machine learning can be biased, destroying the assumption "machine learning is fair by default." He shows how unbiased data (if it were to exist) could still create applications that are biased against minorities by being less accurate. He also talks about how cultural differences can make a classifier that works well for one group classify another improperly, and how the error rate of a classifier does not tell us much about the treatment of minorities.

Hardt, M., Price, E., & Srebro, N. (2016, October 07). Equality of Opportunity in Supervised Learning. Retrieved October 03, 2017, from https://arxiv.org/abs/1610.02413

In this paper, Hardt et al. examine various methods of measuring discrimination, and then showcase the weaknesses of each. They propose a new method, equal odds and equal opportunity, that is easy to implement on top of an already existing algorithm and holds more promise than previous methods for reducing discrimination without reducing accuracy.

Home Page of EU GDPR. (n.d.). Retrieved October 03, 2017, from http://www.eugdpr.org/

The home page of the European Union's General Data Protection Regulation lists some information about the regulation, such as when it is set to go into effect.

Individuals' rights. (n.d.). Retrieved October 03, 2017, from https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/individuals-rights/

This web page lists out, in simple, concise terms, the rights afforded to an individual by the EU's General Data Protection Regulation.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM,60*(6), 84-90. doi:10.1145/3065386

This paper outlines the design of AlexNet, a neural network for image recognition. It gives the reader an idea of how complex neural networks can become, and how difficult it can be to understand how they arrive at the decisions they do.

Liptak, A. (2017, May 01). Sent to Prison by a Software Program's Secret Algorithms. Retrieved October 03, 2017, from https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html

This article examines the case of Eric Loomis, who was sentenced to jail partially based upon a score given by software created by Northpointe. It points out that we can only see the decisions made by the software, not how it arrived at the decision. It also talks about the lawsuit Loomis

filed, arguing that he did not receive due process as he could not challenge or examine the algorithm.

Mattu, J. L. (2016, May 23). How We Analyzed the COMPAS Recidivism Algorithm. Retrieved October 03, 2017, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

ProPublica examines the results of many rulings informed by Northpointe's COMPAS software and finds that it is biased against African-Americans. They show graphs depicting the stark difference in scores, as well as explaining the difference in accuracy between scores for Caucasians and African-Americans.

Meet Tay. (n.d.). Retrieved October 03, 2017, from http://web.archive.org/web/20160413070521/https://www.tay.ai

This was the official website for Microsoft's Tay. It explains what she is and includes the following quote: "The more you chat with Tay the smarter she gets, so the experience can be more personalized for you."

Muñoz, C., Smith, M., & Patil, D. (2016, May). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*(United States, Executive Office of the President).

This was a report released by the White House under Obama, discussing the state of big data/machine learning and its affect on civil rights. It walks through how algorithms can

perpetuate historic discrimination. The report ends with some suggestions moving forward: increase education about big data and support research into decreasing discrimination.

O'Neil, C. (2017). *Weapons of math destruction: how big data increases inequality and threatens democracy.* London: Penguin Books.

Cathy O'Neil wrote this book about how algorithms can be used to discriminate and cause problems in society. In it, she wrote, "The technology already exists. It's only the will we're lacking."

O'Neil, C. (2016, September 12). 'Weapons Of Math Destruction' Outlines Dangers Of Relying On Data Analytics [Interview by K. McEvers]. Retrieved October 1, 2017, from http://www.npr.org/2016/09/12/493654950/weapons-of-math-destruction-outlines-dangers-of-relying-on-data-analytics

This interview with the author of *Weapons of Math Destruction* has a good quote, categorizing which algorithms are especially dangerous: "They have three characteristics. The first is that they're high-impact, they affect a lot of people. It's widespread and it's an important decision that the scoring pertains to, so like a job or going to jail, something that's important to people. So it's high-impact. The second one is that the things that worry me the most are opaque. Either that means that the people who get the scores don't understand how they're computed or sometimes that means that they don't even know they're getting scored. Like if you're online, you don't even know you're scored but you are. And the third characteristic of things that I care about, which I

call weapons of math destruction, the third characteristic is that they are actually destructive, that they actually can really screw up somebody's life. ”

O'Neil, C. (2016, October 06). When Not to Trust the Algorithm [Interview by W. Frick]. Retrieved October 01, 2017, from https://hbr.org/ideacast/2016/10/when-not-to-trust-the-algorithm.html

Cathy O’Neil gives an example of how algorithms are used to discriminate against those with mental illness and further isolate them from society: “So we have reason to believe that some of these algorithms discriminate based on mental health status. This is illegal under the Americans with Disability Act. You’re not allowed to give somebody a health exam, including a mental health exam, as part of hiring. [...] And if some of those personality tests filter out people with mental health problems, then that’s a real problem. So not only does it destroy their chances of getting a job, but it is exactly creating that feedback loop that the ADA is meant to stop, which is actually isolating that group of people from normal society.”

O'Neil, C. (2016, October 13). Don’t trust that algorithm [Interview by C. Pazzanese]. Retrieved October 01, 2016, from https://news.harvard.edu/gazette/story/2016/10/dont-trust-that-algorithm/

Cathy O’Neil talks about how people are too trusting of algorithmic decisions and other problems with letting algorithms make decisions. She says, “There are a lot of issues, but the most obvious one is the trust itself: that we don’t push back on algorithmic decisioning, and it’s

in part because we trust mathematics and in part because we're afraid of mathematics as a public."

Podesta, J., Pritzker, P., Moniz, E. J., Holdren, J., & Zients, J. (2014, May). *Big data: seizing opportunities, preserving values*(United States., Executive Office of the President).

This report by the White House lists some examples of how discrimination can inadvertently occur. They touch upon how big data can diminish equality, but also write that it is possible for big data to uphold civil rights. It has several useful quotes.

Price, R. (2016, March 24). Microsoft is deleting its AI chatbot's incredibly racist tweets. Retrieved October 03, 2017, from http://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3

This article examines the story of Tay, Microsoft's chatbot, and has quite a few examples of tweets.

Reese, H. (2016, March 24). Why Microsoft's 'Tay' AI bot went wrong. Retrieved October 03, 2017, from http://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

This article also examines what happened to Tay during her short life. It includes the following quote from Louis Rosenberg, founder of Unanimous AI: "When Tay started training on patterns that were input by trolls online, it started using those patterns. This is really no different than a

parrot in a seedy bar picking up bad words and repeating them back without knowing what they really mean."

Smith, M., Patil, D., & Muñoz, C. (2016, May 4). Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights. Retrieved October 03, 2017, from https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights

This article by the White House has a pretty good summarizing quote: "Big data is here to stay; the question is how it will be used: to advance civil rights and opportunity, or to undermine them."

Vector Representations of Words. (n.d.). Retrieved October 03, 2017, from https://www.tensorflow.org/tutorials/word2vec

This article by TensorFlow (Google) walks through the basics of word embeddings and what they are used for. It includes some useful diagrams showing how word embeddings capture analogies.

Winerip, M., Schwirtz, M., & Gebeloff, R. (2016, December 04). For Blacks Facing Parole in New York State, Signs of a Broken System. Retrieved October 01, 2017, from https://www.nytimes.com/2016/12/04/nyregion/new-york-prisons-inmates-parole-race.html

This article examines signs of discrimination based on race in America's justice system, specifically in who receives a parole and who does not. It includes some useful statistics.