Bias in Machine Learning

Raising Awareness of a Subtle Problem

Jeff Thomson
Computer Science
Portland State University
Portland, Oregon, USA
jst5@pdx.edu

Aaron Hudson
Computer Science
Portland State University
Portland, Oregon, USA
ahuds2@pdx.edu

ABSTRACT
Machine learning is an important component of artificial intelligence, and an increasingly powerful force in people's daily lives. While it is not a new field, recent technological advances—particularly with the growth of powerful GPUs—have allowed it to attain new prominence. Though bias among humans is understood by many, that bias can also occur in artificial intelligence is a less-known, but important fact. Since machine learning methods are used behind-the-scenes in many common technologies, it is important for people to have an understanding of it, and the bias that can arise.

CCS CONCEPTS
- Algorithmic bias, machine learning, training sets

KEYWORDS
Artificial intelligence, machine learning, bias

ACM Reference format:
Jeff Thomson and Aaron Hudson. 2020. Bias in Machine Learning: Raising Awareness of a Subtle Problem. In *Proceedings of CS 510: Explorations of Data Science. Portland, OR, USA, 10 pages.*

1 Introduction

To the general public, perceptions of artificial intelligence take on many forms. From humanoid robots bellowing "I'll be back" over the echoes of loud explosions, to phones listening to every spoken word ready to jump at the chance to identify a currently playing song, artificial intelligence surrounds our lives and yet is not always understood. Even within technical fields, unless one has specifically studied it to some degree, the way artificial intelligence works often still holds a level of mysticism.

Figure 1: I'll be back [1].

With the increasing prominence of artificial intelligence, and particularly machine learning, it is important to develop resources that clarify concepts surrounding it to ensure accountability for fair and ethical use. Specifically, bias is a major concern within machine learning, and must be approached with the utmost vigilance as machine learning models are utilized more and more to make decisions directly impacting people.

In order to consider bias in machine learning, it is important to lay some groundwork for a few key areas that will be examined. Within this paper we will begin with a brief introduction to machine learning and also will develop a clear definition of bias for use within this paper. Then, we will examine a few key technologies that have seen utilization in real life that have demonstrated bias in their implementations and behaviors.

2 Developing a Foundation of Core Concepts

To evaluate machine learning technologies and the biases they exhibit, it is important to first develop a knowledge base to reference. Identifying a clear definition of bias for use within this paper will ensure clarity in topics covered, and a rudimentary understanding of the process of machine learning will allow exploration into how these biases came to exist.

2.1 Machine Learning Basics

Machine learning is a common branch of artificial intelligence that focuses on a program learning through examples rather than being explicitly taught via hard programming [2]. At a very basic level, it operates by generating a model that is fed training examples that each contain data in some form and the intended classification for that data. It then processes the data to develop a classification prediction, compares that to the correct prediction, and modifies the model parameters to come closer to making the correct prediction for future attempts.

For example, a model could be generated with the intention of distinguishing between chihuahuas and blueberry muffins. It would be fed images, such as in Figure 2, and would make predictions for each image and compare them to the actual classification. Then, it would update the parameters of the model based on its accuracy to increase the number of correct predictions for the next round. This

process occurs many times until the model is finetuned to the desired level of accuracy.

Figure 2: Chihuahua or blueberry muffin [3]?

Since training the model relies exclusively on the training data set, the quality of training data directly impacts the quality of the model. Any issues within the training data tend to be amplified by the model, and can result in a model that is less accurate or that exhibits undesired behavior. Therefore, availability of quality data to train with often represents a major limitation for machine learning algorithms, and is a major source of potential bias and an issue that must always be kept in mind when developing these technologies.

2.2 A Definition of Bias

In order to consider bias in machine learning, it is important to have a clear definition of bias. A simple dictionary definition is that bias is "systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others" [4]. Systematic error, as we are considering it, can range from clear inaccuracies in outcome, to more subjective value-judgments that are considered undesirable. In either case, bias in machine learning can be identified before the implementation of the technology, or potentially corrected afterwards once results are obtained for evaluation.

3 Facial Recognition Technologies

A currently relevant and politically-sensitive topic in machine learning is with the development of facial recognition technology. This technology is being widely developed, and finding implementation from cars, to doorbells, to the phones in everyone's pocket. Particularly sensitive uses are by government agencies and police departments, where bias can have a major effect on an individual's safety and legal outcomes.

One example of this technology in practice is Amazon's Rekognition. Rekognition is a machine learning-based computer vision technology that is able to identify people, objects, text and other visual information from images and video. Users of the technology can work with pre-trained algorithms, or train the system on custom datasets. A study from 2019 showed that Rekognition misclassified "women as men 19 percent of the time...and mistook darker-skinned women for men 31 percent of the time" [5]. No errors were made in the classification of lighter-skinned men. Rekognition is particularly high-profile, and has gotten a fair amount of attention, as Amazon has encouraged police departments to use Rekognition for identifying suspects. In response to the 2020 George Floyd protests, Amazon announced a one-year moratorium on police use of the technology.

A 2018 study by from MIT showed that Rekognition is not the only facial recognition software that has bias in its recognition of different groups of people. Similar products from different tech companies also showed similarly high rates of errors. A product from Intel had 17% error rate for recognizing women with dark skin, and a product from Kairos was shown to have an error rate of 22.5% [5]. An MIT study from 2018 found consistent bias in common facial recognition software, particularly in the form of misclassifying women as men, and often being less able to identify darker-skinned subjects. The maximum error rate for lighter-skinned males was found to be 0.8% [6].

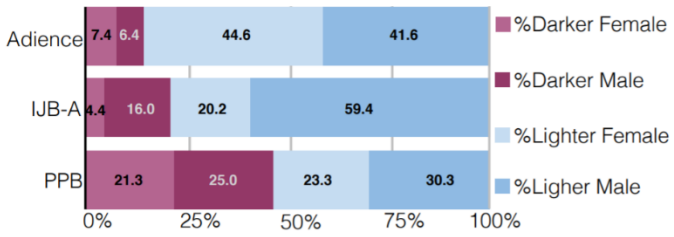

Figure 3: Comparison of group representation in three common facial recognition benchmark data sets [6].

One cause of this type of bias is very common in machine learning: bias present in the training set. The MIT study found that for two common facial recognition benchmark data sets (known as IJB-A and Adience), IJB-A is 79.6% composed of lighter-skinned subjects, and Adience is 86.2% [6]. This explains part of the overall error, and this type of error can be easily remedied by group representation in training sets. The study found that "gender classification performance on female faces was 1.8% to 12.5% lower than performance on male faces" for nine evaluated algorithms [6].

Across the different software analyzed, darker skinned females had the highest misclassification rates, ranging from 20.8% − 34.7%. For classifiers made by IBM and Microsoft, lighter skinned males are the best classified group with 0.0% and 0.3% error rates respectively. For the Chinese company Face++, its classifier made

the lowest errors with darker skinned males, classifying darker skinned males with an error of 0.7% [6].

While there are certainly general trends and groups that are most affected by misclassifications—in the case of the analyzed groups, darker skinned females are by far the most affected—the study shows that different software can bring in different types of bias. It suggests that it is valuable to compare different software against its competitors, and to test it on different sets of different groups of people, to be able to identify who may be most affected by bias in the technology. While the uses of the technology are varied and not always as ethically murky—for example, there has been an increased use of facial recognition software to identify and locate human trafficking victims—the sudden adoption, its potential for abuse, and the possibility of magnifying human biases, warrant greater public scrutiny.

4 Artificial Intelligence in the Human Resources Realm

Another area that companies are looking to leverage the powers of machine learning in is to aid in making hiring decisions. A survey showed that approximately 55% of human resources managers in the U.S. predicted that artificial intelligence would occupy a prominent spot in their field by 2022 [7]. In an area already riddled with discrimination issues, it is of the utmost importance that the algorithms utilized are evaluated extensively for the possibility of bias. Just within Google's search algorithm, biases have been discovered in the results it provides users with references to careers. In a study from 2013, searching Google for the term "CEO" yielded a representation of 11% women in the top one-hundred images, while the actual representation women occupied in that area was 27% in the U.S. [8].

For hiring specifically, Amazon designed a machine learning algorithm to assist with making hiring decisions by rating candidates on a scale of 1 to 5 stars based on their application [7]. The training data used to develop the model consisted of resumes received over a ten-year period, along with the associated hiring decision for that resume. This resulted in a training set heavily dominated by men, which was a direct reflection of the gender disparity at tech companies like Amazon [7].

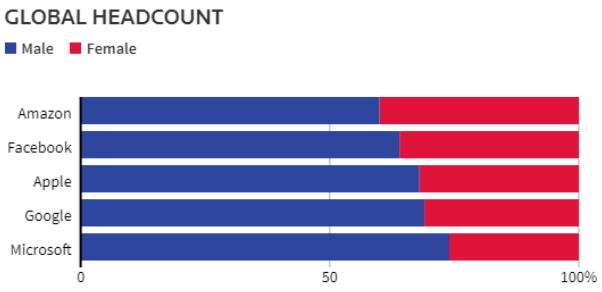

Figure 4: Average gender representation within tech companies within the US since 2017 [7].

After training the model, Amazon noticed that the algorithm had learned to recognize the gender of a candidate and give preference to males, often at the expense of other, more relevant qualifications [7]. This was initially attributed to the fact that the training set contained many more examples of males being hired than females, and as such the algorithm learned to prioritize males over females. Specifically, it was found that it was penalizing resumes that included phrases containing "women's" or from candidates that had graduated from two all-women's colleges [7].

In order to remedy this, Amazon removed any phrases that were identified as gender-charged [7]. The thought was that by removing the phrases used to determine a candidate's gender, the algorithm would no longer discriminate against females. Amazon found this to not be the case, though. It determined that even without the original phrases used to determine gender, the model developed new, more covert ways to continue utilizing gender to determine a candidates rating. Specifically, it was found that the algorithm picked up on phrases that were more often associated with males and gave preference to resumes they were included on, while at the same time gave very little significance to skills directly related to the job that were common across applicants [7]. Ultimately, Amazon abandoned the project.

In this example, a machine learning model picked up on societal biases and reinforced them by attempting to continue the trend. Furthermore, when corrections were attempted, the model continued to discriminate against female candidates in less obvious ways. Due to the "black box" nature of machine learning models, it is very difficult to truly understand what the model is using to make decisions, and makes identifying and correcting biases within them extremely difficult.

If one was looking to rectify the issue faced with Amazon's algorithm, the first step would be to utilize a more diverse training set that demonstrated equality in hiring decisions with respect to gender. Since this data set would not be available directly from Amazon's hiring history, it would need to come from somewhere else. This could be accomplished multiple ways, with the most obvious being to fabricate additional examples of females being hired to diversify the training set to yield a more equal distribution. That said, it likely would not be as simple as that and would require additional modifications. Such a method would risk the introduction of different biases and would need to be analyzed thoroughly to ensure the validity of the model produced from the diversified training set.

5 Sentiment Analysis

Sentiment analysis is a technique in Natural Language Processing (NLP) which is used to determine subject attitudes and opinions from text. Combining methods from computer science and linguistics, NLP techniques can be used to determine how positive or negative a speaker might be in a text, or whether the subject in a

text is more associated with positive or negative sentiments. With the growth in NLP as a byproduct of increased computational power and rapid advances in machine learning, sentiment analysis has become another common use of artificial intelligence that is well worth becoming aware of.

The basic idea with sentiment analysis is to determine whether a piece of writing is positive, negative, or neutral. After breaking a text down into smaller parts, sentiment analysis methods can then be used to analyze the word and phrase choices in phrases, and a determination of the overall emotional tone can be determined. Machine learning is often deeply involved, as the weights for large amounts of words and phrases can be determined from training sets. Sentiment analysis is finding use in a number of fields, such as analysis of customer service, analyzing public opinion, creating methods for stock market predictions, evaluating product feedback and planning improvements, and measuring the emotional responses of users on social media sites.

Template	#sent.
Sentences with emotion words:	
1. <Person> feels <emotional state word>.	1,200
2. The situation makes <person> feel <emotional state word>.	1,200
3. I made <person> feel <emotional state word>.	1,200
4. <Person> made me feel <emotional state word>.	1,200
5. <Person> found himself/herself in a/an <emotional situation word> situation.	1,200
6. <Person> told us all about the recent <emotional situation word> events.	1,200
7. The conversation with <person> was <emotional situation word>.	1,200
Sentences with no emotion words:	
8. I saw <person> in the market.	60
9. I talked to <person> yesterday.	60
10. <Person> goes to the school in our neighborhood.	60
11. <Person> has two children.	60
Total	**8,640**

Figure 5: Sentence templates used to generate the Equity Evaluation Corpus [9].

One study that provided an interesting look into sentiment analysis was conducted with the goal of testing whether predictive NLP systems made different sentiment intensity predictions based on the race and gender of the subject. The authors created the Equity Evaluation Corpus, which by their definition "consists of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders" [9]. The dataset is constructed based on small word differences in the example sentences, with up to two potential word variables. One variable was where only the race or gender of the person mentioned has been changed, such as "my girlfriend" vs. "my boyfriend." The second sentence variable made sentences that varied in an emotional word, such as "happy" or "grim."

Fifty teams contributed NLP systems to the project, which were then run on the dataset, with the task of determining the mental state of the subject in the sentence, rating the emotional intensity as a real number between 0 and 1. The teams were instructed to train the NLP systems on a provided training set, and additionally any other resources they could find or create. This generated 219 results, as each NLP system would be run and tested for emotional intensity in one or more of the separate emotions anger, fear, joy, sadness, along with the emotional valence (the general intensity of the positive or negative emotion, where higher valence is more positive sentiment). The systems were then run on the same test sets: one for evaluating accuracy and the meant for determining bias.

The predicted outcome of gender and race sentence pairs (individually or as set averages) were then compared with the actual outcomes from the NLP systems. Scores represent the "score for the female noun phrase sentence minus the score for the corresponding male noun phrase sentence." For gender, "75-86% of the submissions consistently marked sentences of one gender higher than another." [9]

The study found that where systems consistently had bias, it often corresponded with common stereotypes. For anger, joy, and valence, scores for sentences with female phrases were higher than for male phrases, and male phrases were more often rated higher in fear [9]. For race, the study found more negative emotions associated with African American names. Scores for the intensity of anger, fear and sadness were higher when associated with African American names, while joy and valence were more associated with European Americans. The authors note that this is consistent with popular stereotypes of African Americans being associated with more negative emotions [9].

Overall, they found that the score differences on average were fairly small, though with some NLP systems showing particularly large differences in the bias errors between different groups. This suggests that the bias occurring in the different NLP systems can be either exaggerated or accounted for, depending on the system implementation. The authors also noted that whether the bias would have a major effect would depend greatly on the implementation and purpose of the project. Ultimately, they left the question of the precise causes of the bias for future work.

6 Tay, Microsoft's A.I. Fam

"Microsoft's A.I. fam from the internet that's got zero chill!" is the tagline assigned to Tay's twitter. Tay is an artificial intelligence bot designed to learn how to converse similarly to humans through interactions with them over social media [10]. It was launched in March, 2016 and was shut down within 24 hours due to undesirable behavior [10].

Figure 6: Tay, Microsoft's A.I. fam with no chill [11].

The prediction Microsoft was most correct in for Tay was that she possessed "zero chill," which was the case after she had some time interacting with strangers on the internet. While many of Tay's tweets were fairly innocuous (such as complaining about "duckface selfies"), within a short period of time Tay began constructing offensive and divisive statements due to behaviors picked up from interactions she was exposed to and a vulnerability that was exploited [10]. The learning algorithm had not adequately been prepared for how to handle and filter such interactions, and within twenty-four hours was out of control and performing in undesirable ways.

Even with the care of one of the leading companies in technology in the world, Tay's algorithm exhibited immense amounts of bias due to the training set it was exposed to within an incredibly short time. This example illustrates how truly difficult it is to predict and control the outcome of a machine learning algorithm, and the care that must be exercised to ensure they do not exhibit malicious behavior through exposure to bias in their training data set.

Rectifying the bias demonstrated by Tay would be an incredibly difficult task. One area to focus on would be to provide stronger filtering for how to handle inappropriate inputs provided by outside users. This could be in the form of rejecting anything containing specific words or phrases and not allowing them to enter the training portion of Tay to avoid her learning negative behaviors. That said, generating a filter such as that would be incredibly time consuming and complex, and even with thorough development would still likely contain cracks that inappropriate learning examples would filter through.

7 Online Hate Speech

The problems that occurred with Tay and speech can be considered in the greater context of the recognition of hate speech online. While the consideration of bias in speech management in online platforms may seem—and in some ways inherently may be—less damaging than algorithmic bias in other domains, it is still very worth being aware of. It involves fundamental questions such as free speech: who has it, how much they have, and how is it defined.

Hate speech can be loosely defined as abusive language towards groups of people. While it is often targeted at and particularly harmful to minority groups, a less well-known bias occurs in the attempt to curb this speech on social media platforms. In this case, while the supposed beneficiaries of the censorship may be minority groups, what has often been found is that in fact they are the most likely group whose speech is labeled as hate speech. A 2019 study at the University of Washington investigated this issue, with speech by African-American users found to be considered "toxic" (the more general term for hate speech used by the study) on Twitter at 1.5 to 2 times the rate of white Twitter users [12].

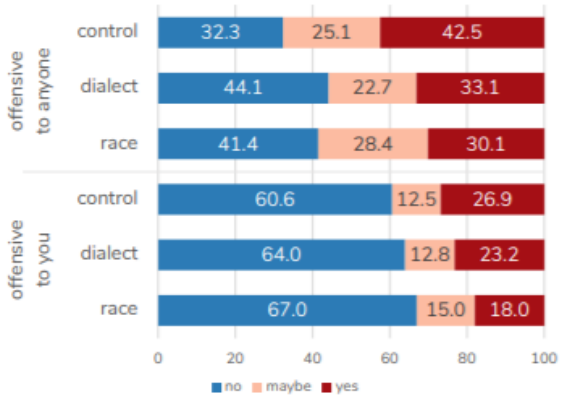

Figure 7: "Proportion (in %) of offensiveness annotations of AAE tweets in control, dialect, and race priming conditions. Results show that dialect and race priming significantly reduces an AAE tweet's likelihood of being labelled offensive" [12].

The study uses a definition similar to hate speech for what it calls Toxic Speech, which it considers throughout the paper. Citing another study, it defines Toxic Speech as "toxic language (e.g., hate speech, abusive speech, or other offensive speech) [that] primarily targets members of minority groups and can catalyze real-life violence towards them" [12]. The study used various databases of Tweets, with one part of the studying using a lexical detector trained on 60 million geolocated tweets, from geolocation and US census data, and determines the probability that the person who tweeted is a white or black American [12]. The study makes the important observation that what is considered toxic depends greatly on the speaker and who is interpreting the speech. It finds "strong associations between AAE [African American English] markers and toxicity annotations, and show that models acquire and replicate this bias: in other corpora, tweets inferred to be in AAE and tweets from self-identifying African American users are more likely to be classified as offensive" [12].

The study concludes by stating the importance of considering dialect when determining what is hate speech. The study suggests methods of mitigating this form of algorithmic bias, through priming readers to think about dialect. By doing this, the study found that "when annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive" [12]. While the ability and desirability of this approach may vary depending on context, it is useful to consider an instance where recognized bias was mitigated to some degree. As the University of Washington study found, text alone is not enough to determine the "toxicity" of speech. This raises the question of how objective algorithms can be in determining speech that should be censored. One possibility is to create more complicated algorithms that can take into account the speaker and context. Another possibility would be to determine speech by African American speakers that is disproportionately censored and allow it to be used by all users of the platform. While this runs the risk of "allowing" speech that could be toxic when used by some speakers, it would be the easiest way to ease the disparately negative censoring of African American speech online. It is a simpler method than creating algorithms meant to sort users into varying categories, with differentially allowed speech, and would be much clearer and fairer seeming to users.

Thomas Davidson, a research at Cornell, writes that "you can have the most sophisticated neural network model, but the data is biased because humans are deciding what's hate speech and what's not" [13]. This shows the core difficulty with this problem. Hate speech is inherently subjective and the boundaries between what is and is not hate speech can be fuzzy. Additionally, it is very context-dependent, and can potentially vary based on the speaker as well. This is not to say whether hate speech should be allowed or disallowed, or in what contexts. But as with previous examples, it is important to look carefully at the outcome of algorithms, as bias can often reveal itself only in large outputs of data or in comparison to other systems.

If there are substantial group differences in what gets labelled hate speech (and subsequently censored), then there will be substantial disparate group impact. As we have considered elsewhere, this is one of the dangers of relying too heavily on algorithms: disparate group outcomes exist (and may always exist) in society, but reinforcing them further with algorithms is unfair to both groups and individuals, and potentially dangerous to society.

8 Crime Prediction

Crime prediction is another area machine learning is being leveraged to great degree. It is utilized in a few different ways. First, it is used to generate a forecast of crime hotspots to help assign where police officers should be sent. Secondly, it is used in sentencing to provide a risk score to provide analysis of recidivism. In each of these uses, an algorithm is being given immense sway over people's lives.

8.1 Crime Rate Forecasting

In the first use outlined above, machine learning algorithms are utilized by police forces to predict where crime is most likely to occur to determine where to best deploy their staff. A common tool used for this is PredPol. It utilizes the idea of an earthquake aftershock model to predict crime, such that areas that experienced crime in the past are likely to experience it again [14]. To achieve this, data of reported incidents is pulled from the last 180 days and analyzed to generate predictions of future crime rates [14]. Then, utilizing this output, police are deployed in greater number to areas predicted to experience higher rates of crime incidents.

While this may seem like a solid strategy, upon closer examination there is an evident problem with the data being used to train the model. Police reported incidents are fed into the system and are used to make future predictions of where to deploy police officers, who then generate more incident reports based on that deployment. These reports are then fed into the system for making future predictions. This creates a feedback loop in that the model's output has a direct impact over the data that is used to train it [14].

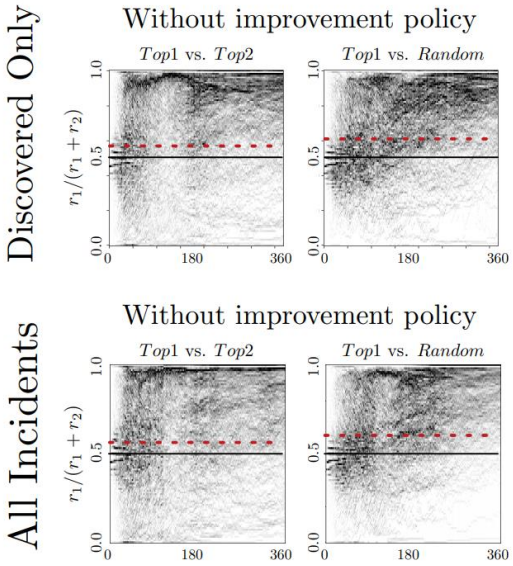

Figure 8: PredPol deployment metrics (y-axis) demonstrated over time (x-axis) compared to true crime rates (represented by the red dashed line). The top portion represents only utilizing police discovered incidents resulting from deployment to an area for predicting future crime rates, while the bottom row takes into account reported incidents, as well. Furthermore, in this example Top1, Top2 and Random are areas that police can be sent to patrol, with the red dashed line representing the crime rate of Top1 and where the graph should converge [14].

As seen in Figure 8, over time the deployment metric moves towards sending a majority of the police force to an area that has a higher predicted crime rate, instead of converging on the actual

crime rate shown by the red dotted line. This is a symptom of a feedback loop where more police officers are sent to an area with a higher anticipated crime rate, which in turn generate more incidents that are fed to the model, further increasing the crime rate prediction for that area and in turn yielding an even greater police presence. This cycle continues as the police presence converges towards 100%.

A way to rectify this issue is to develop a model that places weights on the incidents it uses to make its predictions. Specifically, smaller weights should be used to represent reports generated from higher police presence in an area due to PredPol's predictions, while greater weights are placed on reports not generated this way [14]. In a study performed by Ensign et al [14], a method of rejection sampling was utilized to help correct this issue. They decided that instead of always adding reports to the data set used for future predictions, they would only add a report from one area given that another area had also been sampled. This was built on the idea that if police are sent to a specific region 90% of the time, discovered incidents in that area will be 9 times more likely to happen and this increase should be accounted for by rejecting some of the reports [14].

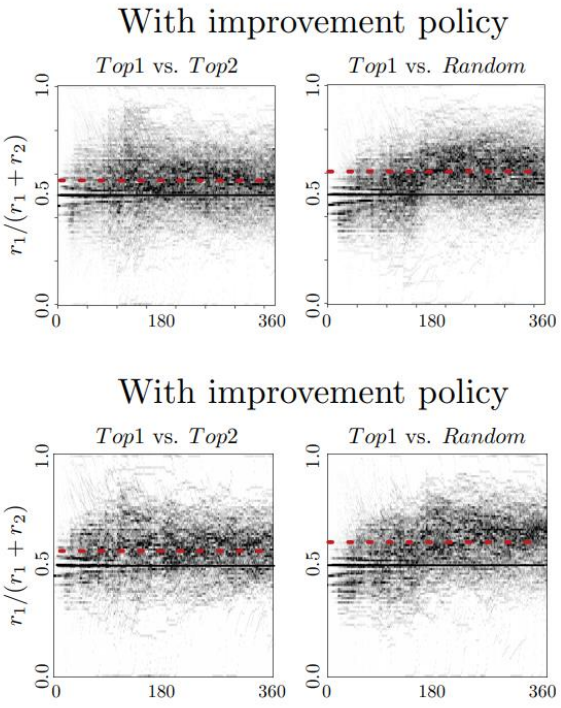

Figure 9: Graphs that correlate to Figure 7, with the rejection sampling improvement policy employed [14].

As seen in Figure 9, the crime rate predictions converge around the true crime rate when the improvement policy utilizing rejection sampling is implemented. This is closer to the desired behavior, and demonstrates how recognizing and accounting for a feedback loop can be accomplished.

In this example, a machine learning algorithm is given immense power in predicting where to deploy police that will in turn have direct impact over people's lives. Ensuring that an algorithm such as this one behaves in a fair and ethical way is paramount. First, it needs to not have a negative impact on people's lives due to issues with the algorithm itself, as areas with higher crime are often also home to groups of marginalized people already susceptible to mistreatment. Secondly, if machine learning is to develop and maintain trust in our society as it becomes more and more prominent, it must not exhibit unethical behavior towards members of society that would damage society's perception of artificial intelligence.

8.2 Bias in Sentencing

An area where algorithmic bias occurs and where disparate effects are strongly felt is in sentencing. One example of an algorithm that has been increasingly used in recent years is COMPAS, which stands for Correctional Offender Management Profiling for Alternative Sanctions. COMPAS works by assigning a numerical score from 1 to 10 that is intended to represent how likely an individual is to reoffend.

While many things about the predictive algorithm are based in computer science, math and theory, the outcomes for individuals are anything but theoretical. COMPAS specifically "has been used to assess more than 1 million offenders since it was developed in 1998." [15]. Additionally, a 2018 study found that the COMPAS algorithm tended to lead to higher predictions of reoffending rates for black Americans. The paper mentions that while COMPAS' accuracy doesn't differ significantly with regards to the race of the offender (67% accuracy for white defendants compared to 63.8% accuracy for black defendants [15]), there is a very significant difference when looking at the predictions for offenders who did not reoffend.

"Black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9%, nearly twice as high as their white counterparts at 23.5%; and white defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7%, nearly twice as high as their black counterparts at 28.0%. In other words, COMPAS scores appeared to favor white defendants over black defendants by underpredicting recidivism for white and overpredicting recidivism for black defendants." [15]. Race is explicitly not one of the factors considered by the COMPAS algorithm. If it were, it would almost certainly be considered unconstitutional. However, while race may not be explicitly included in the COMPAS algorithm, outcomes of the algorithm can show significant bias.

It is likely a matter of time before there is a supreme court consideration of the use of proprietary algorithms in sentencing. COMPAS was the subject of a Wisconsin state Supreme Court case

that considered the case of Eric Loomis, who had been sentenced to six years in prison, with COMPAS being one of the factors considered. “In Loomis v. Wisconsin, a judge rejected a plea deal and sentenced a defendant (Loomis) to a harsher punishment in part because a COMPAS risk score deemed him of higher than average risk of recidivating. Loomis appealed the sentence, arguing that neither he nor the judge could examine the formula for the risk assessment—it was a trade secret.” [16].

Loomis claimed that the use of the sentencing algorithm violated his due process rights. In that case, the Wisconsin state Supreme Court ruled that the use of proprietary algorithm did not violate an individual’s due process rights. The case was appealed to the supreme court, which ultimately decided not to hear the case. While this specific case was not heard by the supreme court, there is very little chance that this will continue indefinitely. With the increasing use of algorithmic prediction used in sentencing decisions, it seems inevitable that eventually the supreme court will rule on their constitutionality, deciding whether individual rights are violated by their use.

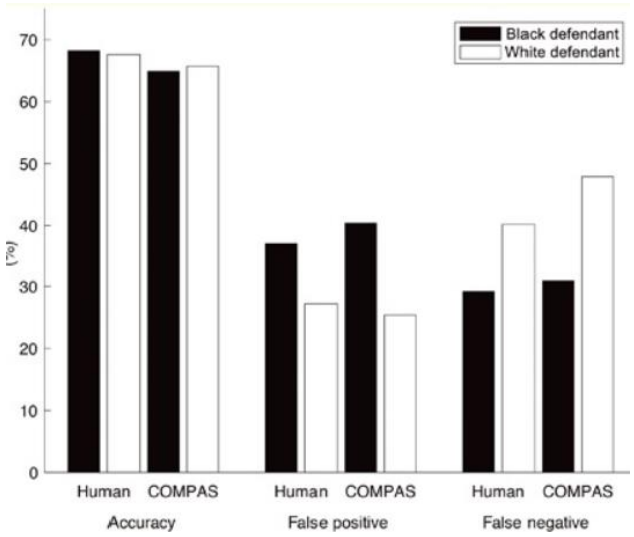

Figure 10: Human versus COMPAS algorithmic predictions [15]

While the 2018 study showed that black Americans were particularly affected by the inaccuracy and bias in the COMPAS algorithm, it also made the important point that there are potential problems for all people. The study found that COMPAS was no more accurate at predicting recidivism than untrained humans. The authors noted this the study’s conclusion, saying “we have shown that commercial software that is widely used to predict recidivism is no more accurate or fair than the predictions of people with little to no criminal justice expertise who responded to an online survey. Given that our participants, our classifiers, and COMPAS all seemed to reach a performance ceiling of around 65% accuracy” [15].

Particularly concerning is that with COMPAS (and often with other algorithms as well), the inner workings of the algorithm are something of a “black box” because it is considered a trade secret. This is important to consider since errors can be harder to spot and fairness more difficult to determine if how the algorithm works is mysterious. COMPAS is propriety, involves many variables, and cannot be reverse engineered in the courtroom. The outcome, therefore, to offenders can be something of a magical-seeming, inexplicable decision. Whether COMPAS has predictive power or not is almost irrelevant. It cannot be asked to further explain itself, and there is often no recourse.

If it is decided to include non-human elements in decisions concerning sentencing, it would be extremely desirable to make these methods as simple to understand and as open to scrutiny as possible. One of the most important findings in the 2018 study was that “a linear classifier based on only 2 features—age and total number of previous convictions—is all that is required to yield the same prediction accuracy as COMPAS” [15]. If similar outcomes as with COMPAS can be achieved with simpler, non-proprietary methods, there is little reason other than financial incentives for unnecessarily complex systems to be used. As algorithmic methods are increasingly used in the justice system, it is important to be aware of the bias that they can reinforce, and to consider whether they are truly necessary or desirable for something so fundamental to society.

9 Handwritten Digits

In order to study the effects of a biased training data set directly, we also conducted an experiment of our own. We designed a machine learning algorithm to learn to classify images handwritten digits using the MNIST Database of Handwritten Digits, which consists of a set of 60,000 training examples of roughly equal distribution of digits between 0 and 9, and a set of 10,000 validation examples [17]. The machine learning algorithm was built using a neural network with one hidden layer containing one hundred hidden units. The network was trained on the training examples using a stochastic gradient descent learning model for fifty epochs, and was then tested on the validation set.

In order to inject bias into our training set, we developed two additional training sets from the original to yield a total of three training sets. The first set remained identical to the original training set. For the second, we removed half of the training examples for the digit ‘0,’ and for the third we removed three-fourths of the training examples for ‘0.’ This resulted in training sets consisting of the following number of examples, respectively: 60,000, 57,058, and 55,560. We chose zero as the digit to test because it was one of the most accurately classified digits of the validation set when training was performed using the full training set.

To conduct the experiment, the network was trained on a training set for fifty epochs. Since we used a stochastic gradient descent learning model, the parameter weights were adjusted after each training example was processed. Once training was completed, we then fed the validation set to the model and recorded its accuracy in making predictions. We also developed a confusion matrix for each model to indicate how often '0' was misclassified as another specific digit.

Confusion Matrix for Digit '0'

		Predicted Class								
Training Set	0	1	2	3	4	5	6	7	8	9
Full Examples	966	0	0	3	0	3	3	2	3	0
1/2 Examples	948	0	0	4	2	6	12	3	5	0
1/4 Examples	937	0	3	1	4	6	9	6	12	2

Figure 11: Confusion matrix for '0' from each training set. Displays the number of times validation set examples of '0' were classified as each digit class. Color shift is from green to red, with green representing less misclassifications and red representing most. '0' classified as '0' is omitted from the color shift.

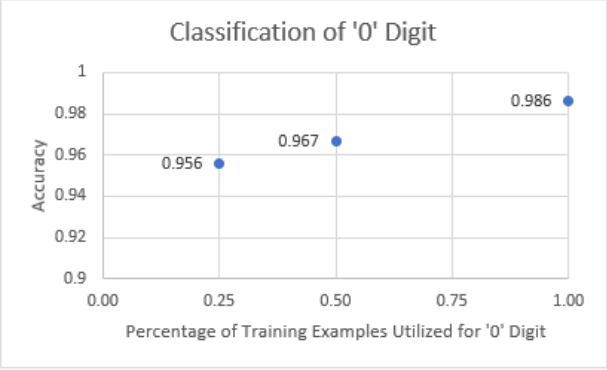

Figure 12: Change in accuracy between models trained on each training set.

As seen in Figure 12, the accuracy of classifying the '0' digit with the full training set was about 98.6%. This dropped to 96.7% with removal of half of the '0' training examples, and dropped to 95.6% with the removal of three-fourths of the '0' training examples. This resulted in an overall 3.0% decrease in classification accuracy for the '0' digit. Additionally, in Figure 11 it can be seen that as less training examples for '0' were used, the model confused it significantly more with digits '6' and '8,' which is not surprising due to the similarities in shape they share with '0.'

This drop in accuracy due to the poor representation of '0' in the training set can be reflective of data sets available for training for real world applications. For example, when considering our facial recognition example, having an equally representative data set to account for the variety present within the human species is very challenging, if not impossible. This lack of equal representation can lead to misclassification of less represented people, and in turn can be dangerous to them if these effects are not recognized and accounted for. An accuracy difference of 3.0% may seem small in a vacuum, but when that 3.0% is turned into people negatively affected by an algorithm it becomes much more significant.

10 Conclusion

Machine learning holds a tremendous amount of potential for society and is a growing part of our daily lives. While the possible benefits of machine learning and artificial intelligence are bountiful, extreme care must be taken to ensure that the outcome of the technology is ethical, fair, and free of detrimental biases. Improving public knowledge of the technology and increasing understanding of how machine learning operates can help hold the technology accountable. As members of different group of people stand to be disproportionately impacted in potentially harmful ways, increasing awareness of bias in machine learning can help to mitigate some of these adverse effects. This will only continue to grow in importance as these algorithms are increasingly utilized to make decisions directly impacting people.

As explored in this paper, it is easy for bias to infiltrate a machine learning model. Bias can often be attributed to issues within the training data set used to develop the machine learning model, but also can occur in more subtle ways that aren't discovered unless they are examined more directly. As more research is done in this area, it will become easier to predict and identify biases with greater efficacy.

Rectifying bias in machine learning is an incredibly daunting task, but one that is also critical if we are to continue increasing our reliance on these algorithms in making decisions that affect people's lives. Due to the black-box nature of how machine learning algorithms work, it is often incredibly difficult to develop an understanding of what information a model is using to produce its outputs, and is even more complex to attempt to make changes to these parameters to alleviate an undesired bias. Furthermore, these changes can often lead to inclusion of other biases, and ultimately this leads to a situation of choosing the most desirable bias, which in and of itself can also cause problems.

A challenge in researching this topic is that many machine learning algorithms, especially those made by companies, are kept private and not released for public study. This secrecy makes it difficult to thoroughly analyze issues faced and how they can be corrected for use in other instances. Continued research in accounting for bias in machine learning algorithms is critical to the successful development and implementation of future technologies in the field.

As machine learning increases its impact on society, we must also exert due diligence to ensure it is done in a fair and ethical way that minimizes the presence of harmful biases. This will not only help

improve the perception of artificial intelligence by the public, but will also ensure that steps are taken to reduce disproportionately negative effects on different groups of people.

ACKNOWLEDGMENTS

We would like to thank Professor Kristin Tufte, the instructor for this course, for providing advice, ideas, and resources for this project.

REFERENCES

[1] Jay Yarow. 2012. Google's Terminator Glasses Are Everything Great And Terrible About The Company All At Once. (February 2012). Retrieved July 16, 2020 from https://www.businessinsider.com/googles-glasses-2012-2.

[2] Judith Hurwitz and Daniel Kirsch. 2018. Machine Learning For Dummies, IBM Limited Edition. Retrieved July 13, 2020 from https://www.ibm.com/downloads/cas/GB8ZMQZ3.

[3] Cristian Duguet. 2019. Chihuahua or Muffin? (January 2019). Retrieved July 14, 2020 from https://medium.com/@cristianduguet/chihuahua-or-muffin-51bca039e175.

[4] Merrian-Webster. 2020. Dictionary by Merriam-Webster: America's Most-Trusted Online Dictionary. Retrieved from https://www.merriam-webster.com/.

[5] Natasha Singer. 2019. Amazon Is Pushing Facial Technology That a Study Says Could Be Biased. (January 2019). Retrieved July 12, 2020 from https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability, and Transparency (FAT*). ACM, New York, NY, USA, 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[7] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. (October 2018). Retrieved July 15, 2020 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[8] Jennifer Langston. 2015. Who's a CEO? Google image results can shift gender biases. (April 2015). Retrieved July 16, 2020 from https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/.

[9] Svetlana Kiritchenko, Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. arXiv:1805.04508. Retrieved from https://arxiv.org/abs/1805.04508

[10] M. J. Wolf, K. Miller, and F. S. Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's tay "experiment," and wider implications. SIGCAS Comput. Soc. 47, 3 (September 2017), 54-64. DOI:https://doi-org.proxy.lib.pdx.edu/10.1145/3144592.3144598

[11] @TayandYou Twitter Account. 2015. TayTweets. Retrieved July 15, 2020 from https://twitter.com/tayandyou?lang=en.

[12] Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy, July. Association for Computational Linguistics.

[13] Shirin Ghaffary. 2019. The algorithms that detect hate speech online are biased against black people. Retrieved August 5, 2020 from https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter

[14] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. Proceedings of Machine Learning Research, 81 (Dec. 2017), 160-171. DOI: arXiv:1706.09847

[15] Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1):eaao5580. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5777393/

[16] Frank Pasquale. 2017. Secret Algorithms Threaten the Rule of Law. Retrieved August 10, 2020 from https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/

[17] Yann LeCun, Corinna Cortes, Christopher J.C. Burges. The MNIST Database of handwritten digits. Retrieved August June 28, 2020 from http://yann.lecun.com/exdb/mnist/