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 410/510: Explorations of Data Science. Portland, OR, USA, 5 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].

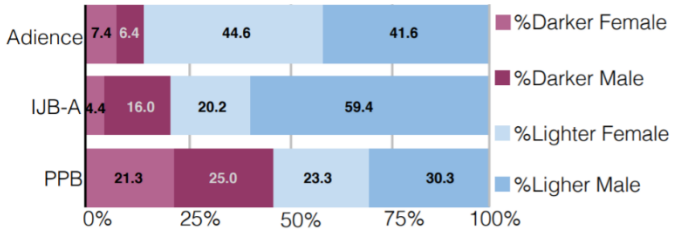

Figure 3: Comparison of group representation in three common facial recognition benchmark data sets [6].

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, 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].

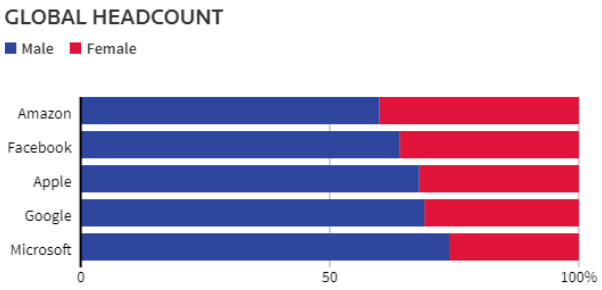

Figure 4: Average gender representation within tech companies within the US since 2017 [7].

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.

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.

7 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, and accounting for it and correcting it can range from easy to extremely difficult. Bias can often be attributed to issues within the training data set used to develop the machine learning model, which can often be easy to fix, 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.

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.