MACHINE LEARNING IN CYBER SECURITY

PROPAGATING RACIAL BIAS THROUGH MACHINE LEARNING ALGORITHMS

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Callie Hartzog

November 1, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Catherine Baritaud, Department of Engineering and Society

Daniel G. Graham, Roseanne Vrugtman, Department of Computer Science

Cybersecurity has been a concern of companies and software engineers as long as private information has been stored digitally. As societies around the world progress towards more technological driven practices and lifestyles, the concern over the safety and security of sensitive data increases. A by-product of the upsurge in use of technology is the emergence of cyber criminals. Often called hackers, cyber criminals gain illegal access to data and often use this data for profit. Social security numbers, banking information, passwords, and other susceptible personal data is all stored online and at risk of being stolen by hackers. In the past two years, cyber-attacks have increased by 400 percent and are unlikely to fall moving forward (Riley, 2021). Good cybersecurity practices are essential to protecting data and thwarting the attacks of hackers. Rather than having software engineers manually inspect technology to assess potential vulnerabilities, machine learning techniques can be used to streamline cybersecurity and improve the rate at which threats to security are detected.

The technical research will focus on current practices in cyber security with regard to the use of machine learning algorithms. The research will then look into further areas of research in this field and areas of cyber security that could benefit the most from machine learning. The coupled STS research will delve into racial bias that arises in machine learning algorithms and measures that can be taken to reduce this bias. The primary motivation of this research is to analyze both the benefits and faults of a technological future more centered around machine learning. The technical and STS reports will be completed during the spring of 2022.

MACHINE LEARNING IN CYBER SECURITY

A huge vulnerability in current technology is found in Internet of Things (IoT) devices. IoT devices are often common household appliances such as doorbell cameras, smart

thermostats, and wireless gaming devices. Their direct access to the internet allows for a variety of attacks from hackers if they are not secured properly (Sivanathan et al., 2020, p. 1). The most common practice for securing IoT devices is the use of firewalls, which protects the device from external access but fails to protect it against any internal attacks. There has been a more recent movement within the cyber security community to secure IoT devices using intrusion detection systems that are implemented using machine learning algorithms (Smys et al., 2020, p. 1).

The use of recurrent neural networks (RNNs) to train computers to recognize cybersecurity threats in programs has become another popular way to secure various devices and computer systems. Recurrent neural networks are a type of neural network that allows the computer to retain information it has previously learned, as seen in Figure 1 below. A computer is trained to recognize patterns, classify images, or sort data based on given parameters. The computer then can use its knowledge to apply the same practice towards previously unseen data (Zahangir et al., 2019, p. 28).

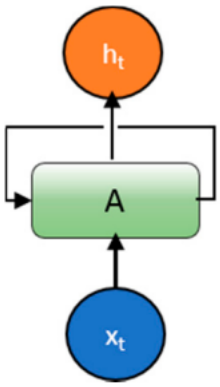

Figure 1: A basic recurrent neural network: Input is processed then fed back into the machine, allowing predictions to be made from a retained memory of previous results (Zahangir et al., 2019).

There a several different sets of data that can be used to train algorithms to detect security threats. When using a data-set of 22 different attack types to train a recurrent neural network, Xin

et al. (2018) initially received a test accuracy of 83.28% (p. 12). This is a good base line for using RNNs to identify security threats but through optimizations the accuracy can increase.

SHORTCOMINGS OF MACHINE LEARNING ALGORITHMS

Both a benefit and deterrent to the use of machine learning is the number of different types of algorithms that can be implemented. Depending on the device or system that is being secured, the machine learning algorithm that yields the most accurate risk detection results may vary. The variety of machine learning algorithms that can be applied provides a wide array of options that may best fit securing a certain device, but also requires more testing to determine which algorithm is the most accurate for each situation. Alqahtani et al. (2020) used seven different algorithms [Decision Tree (DT), Random Forest (RF), Random Tree (RT), Decision Table (DTb), Artificial Neural Network (ANN), Naive Bayes (NB), and Bayesian Network (BN)] and trained them on the same data set to implement various intrusion detection systems and compare the resulting accuracies (p. 2). In Figure 2 on page 4, all the accuracies were either 90% or above but they differed widely within the 90-100% range. Considering the algorithms are being used to secure systems that potentially contain very sensitive and important data, achieving an accuracy closest to 100% in detecting security risks is extremely important.

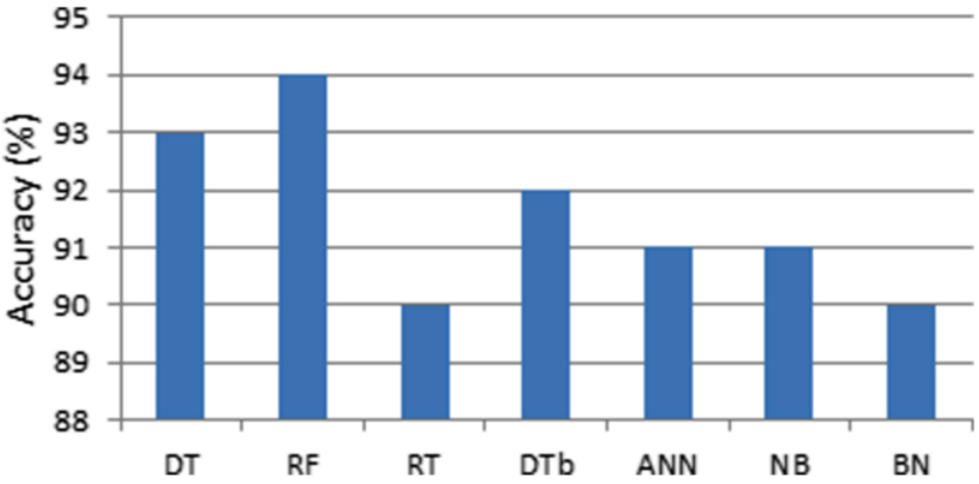

Figure 2: Accuracy results of different machine learning algorithms: The accuracy in detecting risks in an intrusion detection system is widely dependent on the algorithm implemented to create the system (Alqahtani et al., 2020).

Another issue with machine learning is that the accuracy of an algorithm is only as good as the data it is trained on. There are many pre-existing data sets that contain features useful to training machine learning algorithms to detect cyber security threats, such as the KDD Cup 99 and ADFA data sets. The KDD Cup 99 dataset contains a wide variety of attack types, which is useful for creating a blanket protection system that can identify many different kinds of attacks. However, datasets like the ADFA data set are more targeted. The ADFA data set focuses on detecting intrusions from the host level, such as an internal attack originating from an individual gaining access to admin controls of the system. In addition to these data sets, algorithms can be trained by collecting data directly from a computer system. This makes the algorithm more aligned with the specific system at hand as it analyzes incoming and outgoing traffic to the system, then learns to form typical use patterns that will be the basis to predict if abnormal use occurs (Yavanoglu & Aydos, 2017, p. 5-6). Over the years these data sets have been improved through the addition of new features and attack patterns. The quickest and most efficient way to

improve the accuracy of these algorithms is to improve the data sets they are being trained on. This can be achieved by researching in depth the typical attacks systems face from hackers and adding data on these attacks to the data bases so they are as robust as possible.

PROPAGATING RACIAL BIAS THROUGH MACHINE LEARNING ALGORITHMS

Machine learning is quickly becoming perceived as the more rational implementation of artificially intelligent systems. It is the backbone of many autonomous systems, such as self-driving cars, factory robots, and language translators. Machine learning is built upon the existence of data. In order to train computer systems, they first need to analyze a data set and extrapolate meaning from the data. Collecting data, however, is typically done with some human involvement which introduces bias into the data. Often once the data is collected and analyzed by hand, it mainly represents the white majority and fails to include minorities. Google's Automated Retinal Disease Assessment tool is machine learning-based and data driven, but fails to work properly in less technologically advanced countries such as India. When testing their program in India, Google failed to accommodate for lower resolution images, which are a common occurrence in countries whose technology is not as developed as the United States. Through leaving out this social group in the original test data, Google has created a program that now fails to identify signs of vision loss due to diabetes in a country where upwards of 60 million people suffer from diabetes (Abrams, 2019).

Bias is not only present due to the failure of considering technological disparities across communities, but can also arise directly from the data set used to train a model. Hardesty (2018) reported that the typical error of detecting the gender of an individual was only 0.8% for light-skinned males, but rose above 34% for darker-skinned females. Bias in this case arose from the

fact that the data set was 83% white (Hardesty, 2018). Facial recognition and prediction methods for things such as crime can inherently introduce racial profiling if programs are not trained on a widely representative data set that includes minorities (Budds, 2017).

SOCIAL CONSTRUCTION OF MACHINE LEARNING

Racial bias is prevalent in many machine learning algorithms. This project will apply the Social Construction of Technology (SCOT) theory to machine learning to analyze the impact of programs on various social groups, and further, the social groups that these programs fail to consider and incorporate into their product (Pinch & Bijker, 1987). Machine learning programs already impact many different social groups, whether intentional or not. Figure 3 on page 6 shows some various examples of machine learning being used in modern society. This is just a sample of the impact of machine learning, and it will only grow to impact social groups in more ways as machine learning becomes more ingrained into society.

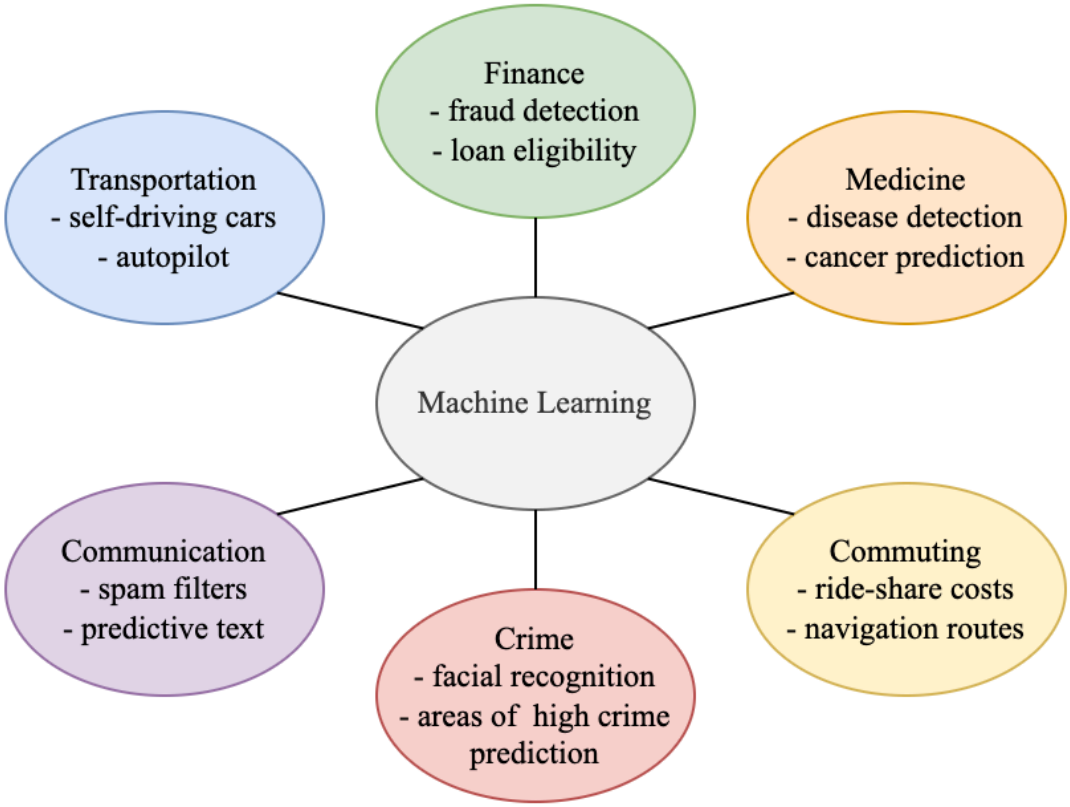

Figure 3: Examples of machine learning: Different common applications of machine learning in modern society (Hartzog, 2021).

In a typical SCOT model, the engineer creating a piece of technology interacts with different social groups and received feedback in order to improve the technology. The current SCOT model for machine learning can be seen in Figure 4 on page 8. Rather than following a traditional SCOT model, the engineer only choses to interact with very specific social groups, typically the white majority (Metz, 2021). However, they still extend their technology to other minority social groups but do not receive feedback from them. No tradeoffs are made to accommodate the minority groups. The previously mentioned case studies of racial bias in machine learning demonstrate this.

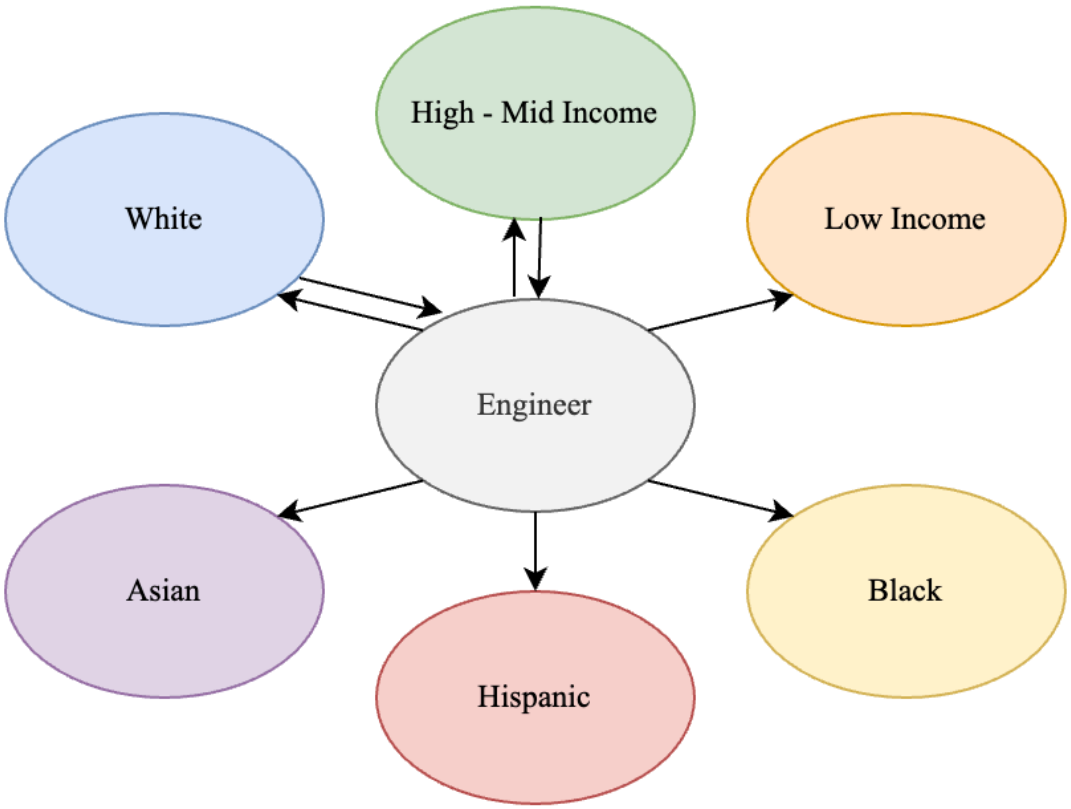

Figure 4: SCOT model: Current model of tradeoffs between the group of engineers that implement a machine learning program and the social groups they interact with (Hartzog, 2021).

An unintended consequence of some of these programs and the flawed SCOT model is that they create immoral predictions. In the United States, there are laws against which characteristics can and cannot be used for decision making. This is most commonly seen in anti-discrimination laws where employers cannot discriminate against employees or potential hires based on attributes such as race, gender, religion, disability status, etc. (Hentze & Tyus, 2021). However, machine learning algorithms do not have these restrictions. This can cause their results to reflect discriminatory tendencies against a specific population if data is collected in a way that only reflects a certain population. While a majority of this issue stems from bias in the data used to train algorithms, the algorithms themselves can cause bias. Machine learning algorithms

operate using various weights and parameters, which help them determine what data is important and what is not. If the parameters are adjusted to improve performance of the program by making it produce a more general approximation, it is possible that the complexity of the program may be compromised which could remove important patterns found in a minority of the data (Veale & Binns, 2017, p. 2-3). When these programs are being applied towards humans, this tends to directly reflect racial bias appearing in the results as the data that concerns minority social groups is ignored. Figure 5 below depicts different forms of bias that can arise from machine learning in finance. Finance is a field where the use of machine learning algorithms is very common. The data that the algorithms use to make predictions in finance often reflect racial bias.

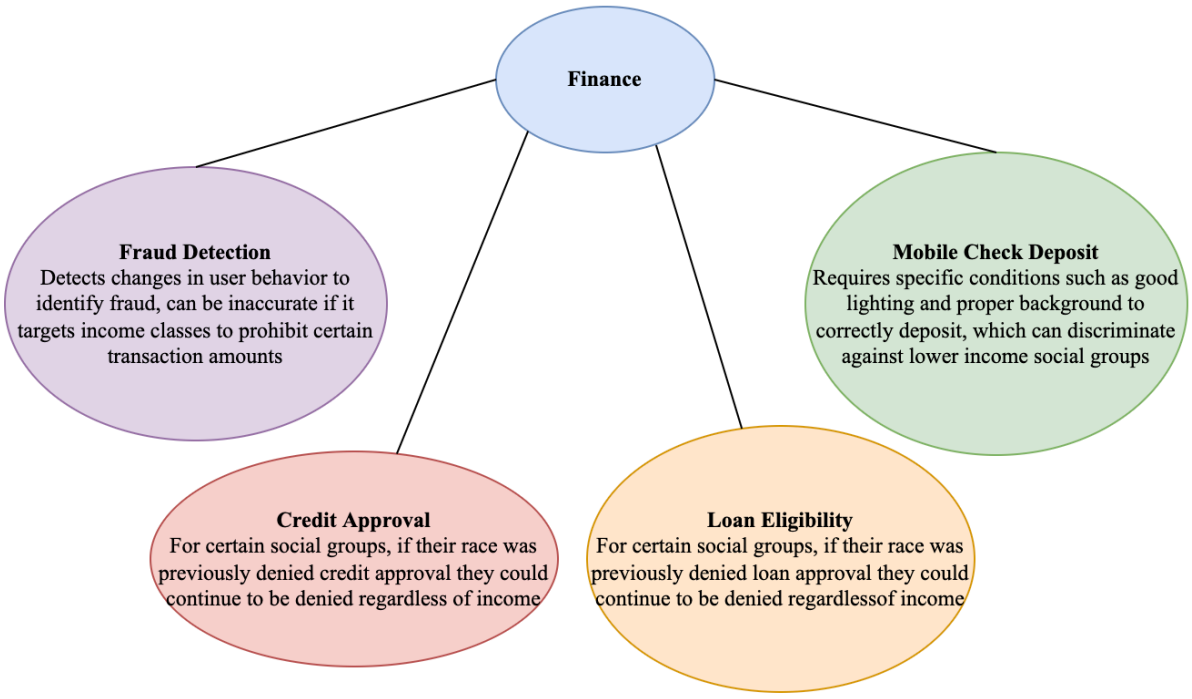

Figure 5: Examples of bias in finance: Different areas of finance use machine learning algorithms and can produce biased results if the data they are trained on is biased (Hartzog, 2021).

Bias can never be eliminated in machine learning; it is essential for the algorithms to learn trends and patterns. Removing all bias would prohibit the algorithm from knowing what it should be learning, and what data is irrelevant (Jaton, 2021, p. 3). Since bias can never be removed completely, it is more vital to focus on ensuring that the data an algorithm does receive is accurate in its portrayal. To achieve this goal, for any machine learning programs that either involve data on humans or whose function would impact various social groups, data collection standards should be put in place to ensure all data is fair and moral in its representation.

THE FUTURE OF MACHINE LEARNING

Machine learning is a promising field as society looks towards automation as a replacement for many jobs and practices. One of the topics gaining interest is the use of deep learning and neural networks to detect vulnerabilities in software and improve the current cybersecurity practices in place. An often overlooked issue with machine learning is the presence of racial bias in data used to train algorithms. Minority communities are regularly either forgotten or ignored when collecting data for machine learning. This can further racial bias already present in a society when these programs are used to make generalizations about a population they do not represent.

REFERENCES

Abrams, C. (2019, January 26). Google's effort to prevent blindness shows AI challenges. *The Wall Street Journal.* https://on.wsj.com/39tnDLo

Alqahtani, H., Sarker, I.H., Kalim, A., Minhaz Hossain, S.M., Ikhlaq, S., & Hossain S. (2020) Cyber intrusion detection using machine learning classification techniques. *Computing Science, Communication and Security, 1235*(1), 121-131. https://doi.org/g3mh

Budds, D. (2017, July 25). Biased AI is a threat to civil liberties. The ACLU has a plan to fix it. *Fast Company.* https://bit.ly/3zAAeae

Hardesty, L. (2018, February 11). Study finds gender and skin-type bias in commercial artificial-intelligence systems. *MIT News.* https://bit.ly/3lVW4R4

Hartzog, C. (2021). *Examples of machine learning.* [Figure 3]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Hartzog, C. (2021). *SCOT model.* [Figure 4]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Hartzog, C. (2021). *Examples of bias in finance.* [Figure 5]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Hentze, I., & Tyus, R. (2021, August). *Discrimination and harassment in the workplace*. National Conference of State Legislatures. https://www.ncsl.org/research/labor-and-employment/employment-discrimination.aspx

Jaton, F. (2021). Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society, 8*(1), 1-15. https://doi.org/g3mn

Metz, C. (2021, March 15). Who is making sure the A.I. machines aren't racist?. *The New York Times.* https://nyti.ms/3jztsfP

Pinch, T. J. & Bijker, W. (1987). The Social construction of facts and artifacts. *In The Social construction of technological systems: New directions in the sociology and history of technology.* Cambridge, MA: MIT Press.

Riley, T. (2021, February 22). The Cybersecurity 202: Cybercrime skyrocketed as workplaces went virtual in 2020, new report finds. *The Washington Post*. https://wapo.st/3EFGNw5

Sivanathan, A., Gharakheili, H. H., & Sivaraman, V. (2020). Managing IoT cyber-security using programmable telemetry and machine learning. *IEEE Transactions on Network and Service Management, 17*(1)*,* 60-74. https://doi.org/gwv5

Smys, S., Basar, A., & Wang, H. (2020) Hybrid intrusion detection system for Internet of Things (IoT). *Journal of ISMAC*, *2*(4), 190-199. https://doi.org/gksp9n

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 1-17. https://doi.org/gdcfnz

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018) Machine learning and deep learning methods for cybersecurity. *IEEE Access, 6*(1)*,* 35365-35381. https://doi.org/gf9fq3

Yavanoglu, O., & Aydos, M. (2017). A review on cyber security datasets for machine learning algorithms. 2017 *IEEE International Conference on Big Data (Big Data),* 2186-2193. https://doi.org/gmt5j4

Zahangir, M. A., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics, 8*(3)*,* 292. https://doi.org/gfw52f