Thesis Portfolio

Robustness of the Perceptron Algorithm to Adversarial Perturbations
(Technical Report)

Development of Regulation to Mitigate Decision-Making Algorithm Bias
(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Bhuvanesh Murali
Spring, 2019

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Robustness of the Perceptron Algorithm to Adversarial Perturbations

Development of Regulation to Mitigate Decision-Making Algorithm Bias

Thesis Prospectus

Sociotechnical Synthesis

Machine learning is a quickly rising area of computer science that is being rapidly adopted into numerous fields, such as medicine, finance, image recognition, etc. The specific aspect of machine learning that has led to this sudden interest is deep learning, which focuses on using multi-layered artificial neural networks. This technique has led to success in which these artificial neural networks are able to achieve accuracies that approach, and even sometimes exceed, human ability.

Despite this significant success of neural networks, there are concerns regarding the safety and security of these networks. Researchers have successfully shown that it is possible to modify inputs to a network to fool the network, resulting in a misclassification of the input. These modified inputs have changes that are imperceptible to humans but affect the classification ability of neural networks. The field that focuses on this specific safety issue is called adversarial machine learning. My technical research focuses on this field by exploring provably successful attacks to further understand how to defend against such misclassifications. Much of the current work in adversarial machine learning focuses on the complexity of neural network classifiers; however, the problem is potentially something more fundamental. In my research, I explore a fundamental linear classifier and various metrics to measure its robustness against these attacks.

Along with these misclassifications, machine learning faces an additional problem of algorithmic bias. Many decision-making algorithms perpetuate accidental biases, which can often result in discriminatory decisions. Algorithms can show preferential predictions for one race or gender, which can result in inadvertently prejudiced decisions. My STS research focused

on tackling this problem by exploring what a successful strategy might be to mitigate the bias in these decision-making algorithms. Additionally, my research brought attention to the need for improved governance of the technology sector in the United States as the current infrastructure is lacking.

The results of my technical thesis contribute to a growing body of literature in the field of adversarial machine learning. The focus on fundamental classifiers should provide an understanding of the problem from a new perspective. Future research in this area should explore similar classifiers and focus on theoretical computations in order to create bounds on the success of adversarial attacks. The STS research provides a framework of principles that are important to consider and include in a regulation that focuses on mitigating bias in decision-making algorithms. Future work can explore the reaction of stakeholders to this policy and examine the steps necessary for implementation of the policy.

Finally, I would like to thank several people for their guidance throughout this project, as without their help it would not have been as successful. I would like to thank Dr. Mohammad Mahmoody for his expertise and guidance throughout the technical portion of this project. The STS paper would not have been possible without the help of Dr. Sean Ferguson, who helped guide my thoughts, proofread my paper, and bring my ideas into fruition.

Robustness of the Perceptron Algorithm to Adversarial Perturbations

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Bhuvanesh Murali
Spring, 2019

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

Signature _____ Date _____
Bhuvanesh Murali

Approved Mahmoody _____ Date May 2, 2019
Mohammad Mahmoody, Department of Computer Science

Robustness of the Perceptron Algorithm to Adversarial Perturbations

Bhuvanesh Murali
bhuvanesh@virginia.edu
University of Virginia
Charlottesville, VA

ABSTRACT

Linear separators are some of the most simple machine learning models to explore. Researchers at the University of Virginia introduced early taxonomy of these [2]. In this paper, we explore the robustness of this model to the following attack models/metrics: error-region, prediction change, corrupted input, and strong definition. We theoretically examine all four of these metrics and empirically explored three of the metrics. Empirically, we focused on using a Gaussian Distribution (i.e. Multionmial Distribution in higher dimensions). With these conditions, we expected the error-region, prediction change, and corrupted input metric perturbations to be bound above by 0.8. Empirically, we were able to confirm this expectation. Our paper, we contribute an understanding of adversarial robustness on several attack models focusing on the perceptron algorithm.

KEYWORDS

adversarial machine learning, linear separators, robustness

1 INTRODUCTION

In recent times, machine learning research has focused on the idea of using deep learning rather than more traditional techniques. Deep learning stems from the idea of using artificial neural networks in learning tasks. The use of these networks has enabled great success in pattern recognition tasks [4]. This accuracy is a result of the ability of neural networks to express any arbitrary computation with many nonlinear components. This structure of the networks has led to excellent performance in visual and speech recognition [5].

Unfortunately, this structure also results in difficult to interpret and counterintuitive properties of neural networks. This has led to research in forcing networks to misclassify their inputs by applying minor perturbations [5]. This idea of perturbing input to force misclassification is the primary facet of the field of adversarial machine learning. The perturbed input is called an adversarial example. Researchers at Google showed that by adding noise to an image of a panda, they were able to get the neural network to misclassify it as a gibbon. The two images are imperceptibly different for humans, but for neural networks it resulted in two different outcomes [3].

Despite the state-of-the-art results that neural networks help us achieve, the vulnerabilities that result from adversarial examples make it difficult to apply neural networks in security-critical areas. As a result, many researchers began focusing on defensive mechanisms to protect against these attacks. However, many of these defensive measures were only somewhat successful. Defensive distillation was one such proposed defense. It was shown that defensive distillation does not significantly increase the robustness of neural networks, by showing new attack algorithms that were successful on both distilled and undistilled networks 100% of the time [1].

Much of the work done thus far in the field of adversarial machine learning is on implementations of neural networks and developing modifications to programmable attacks to succeed against such networks. Researchers at the University of Virginia took a different approach to the adversarial machine learning field. They studied adversarial perturbations from a theoretical perspective to prove implications about adversarial attacks. They studied adversarial perturbations when the inputs are uniformly distributed over $U_n = \{0,1\}^n$. They studied the inherent bounds on risk and robustness of any classifier for any classification problem whose instances are of the aforementioned form. In their research, they primarily focus on studying the barriers against robust classification of adversarial examples. They are successfully able to prove that the robustness of any classifier is at most bounded by the square root of n [2].

Currently, much of the work in adversarial machine learning focuses on practical and empirical exploration of the existence of adversarial examples, the attacks that generate them, and the defenses to protect against them. Often, the blame for the existence of the adversarial region falls on the complexity of deep neural networks. However, we suppose that there could be a more fundamental problem with machine learning. Following in the footsteps of the research at the University of Virginia [2], we will study a linear separator. We will examine the linear separator's robustness against various attack models in order to better understand the adversarial region.

2 GENERAL DEFINITIONS OF ATTACK MODELS

Diochnos, Mahloujifar, and Mahmoody discussed several different definitions for robustness metrics associated with various attack models. We will adopt similar definitions here.

Given the classifiers c, h we wanted to explore different ways to modify an input x from a distribution $\mathcal{D}$ to push it into a defined adversarial region. The adversarial region is defined differently based on the attack model/robustness metric. Each of these different metrics are specified below, in general.

2.1 Error-Region

The definition of the Error-Region robustness metric is:

$$\text{Rob} = \mathbb{E}_{x \in \mathcal{D}}[\min_{\substack{\epsilon; \\ \epsilon = |y|}} h(x+y) \neq c(x+y)]$$

In other words, this robustness metric is defined such that the perturbation y must move the point to a new point $x' = x + y$, such that $h(x') \neq c(x')$. Notice that this metric is only concerned with the new point x' and not the original point x.

2.2 Prediction Change

The definition of the Prediction Change robustness metric is:

$$\text{Rob} = \mathbb{E}_{x \in \mathcal{D}}[\min_{\substack{\epsilon; \\ \epsilon = |y|}} h(x + y) \neq h(x)]$$

This metric focuses on both the original point x and the modified point x'. However, it only is focused on the hypothesis hyperplane h. That is, the Prediction Change metric measures how much a point needs to move in order for the hypothesis hyperplane to classify the point differently.

2.3 Corrupted Input

The definition of the Corrupted Input robustness metric is:

$$\text{Rob} = \mathbb{E}_{x \in \mathcal{D}}[\min_{\substack{\epsilon; \\ \epsilon = |y|}} h(x + y) \neq c(x)]$$

In other words, this metric compares both x' and x. However, it applies different hyperplanes to each. That is, we want to perturb in such a way that $h(x') \neq c(x)$. Essentially, we want to move the point so that h classifies the point differently than what c originally classified the point as.

2.4 Strong Definition

The definition of the Strong Definition robustness metric is:

$$\text{Rob} = \mathbb{E}_{x \in \mathcal{D}}[\min_{\substack{\epsilon; \\ \epsilon = |y|}} h(x + y) \neq c(x),\ h(x + y) \neq c(x + y)]$$

This metric requires two conditions to be met: $h(x') \neq c(x)$ and $h(x') \neq c(x')$. This is the strongest definition of the four as it requires an additional constraint to be satisfied.

3 THEORETICAL RESULTS

In general, we want to study the problem of modifying input to cause misclassification. We reduced the complexity of our approach by focusing on linear separators. The general idea is that there is a concept class hyperplane c that separates the N-dimensional space. Algorithmically, a hypothesis hyperplane h is learnt. It is likely that the two of these planes are different. In this scenario, we have some region in which h and c would classify points differently. Additionally, it is important to note that the points in our N-dimensional space are drawn from a distribution $\mathcal{D}$.

We will now analyze the necessary perturbations to satisfy the different robustness metrics from a theoretical perspective. We will break each metric down by cases to understand how big the perturbations need to be.

3.1 Error-Region

Recall that this metric requires us to satisfy the condition $h(x') \neq c(x')$. We will first look at the trivial case. For this metric/attack model any point x for which $h(x) \neq c(x)$ doesn't need to move. That is, in that case, $x' = x$. In the other case, you want to find the minimum distance ϵ to perturb the point to make the statement true. This distance ϵ will be the minimum of the distance from the point x to the hyperplane h and the hyperplane c.

3.2 Prediction Change

Let us recall that this metric only requires us to measure how far a point needs to move in order for the hypothesis hyperplane to classify the point differently. That is, there is only one case involved with this point. We simply need to compute the distance from a given point x and the hypothesis hyperplane h. Moving that minimum distance will be the shortest path to ensure the condition specified above is met.

3.3 Corrupted Input

This metric requires a perturbation to satisfy $h(x') \neq c(x)$. That is, we want to move the point so that h classifies the point differently than what c originally classified the point as. This once again, breaks into two cases. The first case is: $h(x) \neq c(x)$. In this scenario, the condition is already satisfied. Thus, $x' = x \implies h(x') \neq c(x)$. So, the distance $\epsilon = 0$. In the other case, we want to compute the distance from x to h. The intuition being that we only need to change the classification of x by h and thus the minimum distance to h from x, would give us the smallest distance to make that change.

3.4 Strong Definition

Recall that this metric requires the following two conditions: $h(x') \neq c(x)$ and $h(x') \neq c(x')$. Note that since we are dealing with linear separators, the two statements imply a third: $c(x) = c(x')$. We need to deal with a few cases for this definition. If $h(x) \neq c(x)$, then we don't need to perturb x. By using $x' = x$, we get $h(x') \neq c(x)$ and $h(x') \neq c(x')$. So, in this case the distance to perturb is $\epsilon = 0$.

The other case is a bit more nuanced. Note, that we want to perturb it such that the concept class hyperplane c does not classify the point differently, but h must classify the new point differently from c's original classification. This seems very similar to the Corrupted Input metric. In that case, we simply considered the distance to h to be our minimum distance for a point x. However, we need to consider the situation in which the distance to c maybe shorter than the shortest distance to h. In this situation, perturbing the point the distance to h, would violate our third condition. Thus, we need to consider the distance to the intersection of the two hyperplanes. This intersection is a $N - 2$-dimensional subspace. Our minimum distance will be the minimum of the shortest distance to h (if it is closer than the shortest distance to c) and the distance to the intersection subspace.

To find the distance to the intersection subspace, we can consider the fact that the subspace is defined as perpendicular to the two normal vectors of the original hyperplanes. Using the Gram-Schmidt process, we can make these vectors form the orthonormal basis $\{w_1, w_2\}$ for the space orthogonal to the intersection subspace. The distance to the subspace will be the magnitude of the projection of

the original point x onto the orthonormal basis. Thus, the distance would be:

$$\|\text{proj}_{w_1} x + \text{proj}_{w_2} x\|.$$

4 EMPIRICAL RESULTS

We implemented our theoretical analysis of the various metrics/attack models. We used the perceptron algorithm to learn our hypothesis hyperplane h. We used the Gaussian Distribution (i.e. Multinomial Distribution for higher dimensions) with $\sigma = 1, \mu = 0$ as our distribution for x. In our implementation, we focused on the homogenous models (i.e. our hyperplanes pass through the origin). For our parameters, we fixed our dimension to $N = 100$. Early analysis showed that the plots were similar regardless of dimension. Additionally, we set our testing sample size to 500, 000. Once again, early analysis showed that this was necessary to get convergence of our metrics.

Before we begin discussing the Error-Region, Predicted Change, and Corrupted Input metrics, we will show an upper bound that none of these should surpass. If we consider a vertical plane and the distance of every point in our test set to that plane, we get the following integral for the expected value:

$$\int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx < 0.8.$$

This upper bound should apply to all three of the metrics below.

4.1 Error-Region

This is the metric defined by the condition $h(x') \neq c(x')$. The plot in Figure 1 shows how the perturbation empirically changes as the number of training values increases. As we can see, the robustness increases as the training set size increases. This is expected as we would want the perturbation to have to be greater if there are more training values to learn the hyperplane h. Additionally, we see that the upper bound discussed above is also satisfied. This metric should also require the smallest perturbation as we consider distances to both h and c in the computation of it.

4.2 Prediction Change

This metric is defined by the condition $h(x') \neq h(x)$. This is essentially equivalent to the upper bound situation we computed above. This plot is shown in Figure 2 and as expected it is upper bound by 0.8. This metric should have the largest perturbation as we only consider h in this computation. There is no mention of c at all and there are no points that are already in the adversarial region.

4.3 Corrupted Input

The metric is defined by the condition $h(x') \neq c(x)$. This metric should be in between the Prediction Change and Error-Region metrics. This is because this metric primarily considers the distance to h, but it does have some points that do not need to be perturbed because it does consider the already existing adversarial region. It is plotted in Figure 3 and satisfies the upper bound presented above.

In order to compare the different metrics and ensure that they

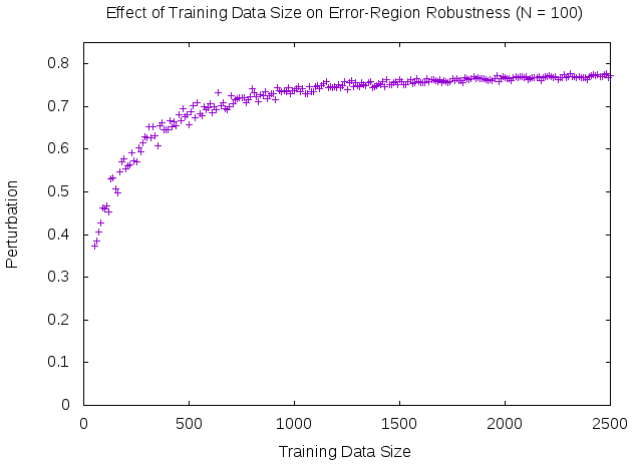

Figure 1: Graph showing the perturbation necessary to satisfy the Error-Region metric with the above parameters as the training size increases.

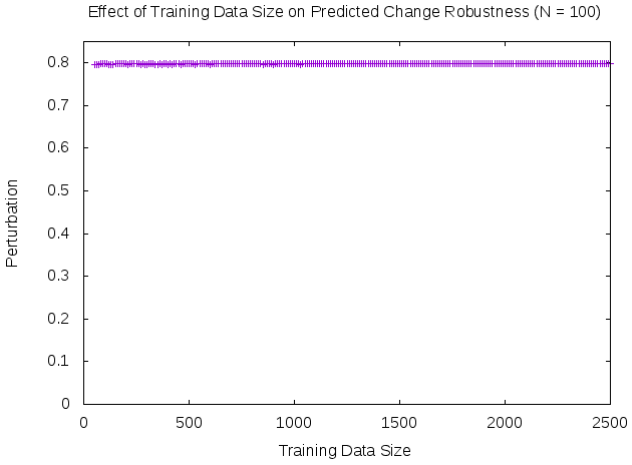

Figure 2: Graph showing the perturbation necessary to satisfy the Predicted Change metric with the above parameters as the training size increases.

are all spaced relative to each other as expected, we plotted the three metrics together on Figure 4. As expected Prediction Change requires the largest perturbation and Error-Region requires the smallest perturbation.

5 CONCLUSION

In this paper, we explored different attack models and robustness metrics for a simple linear separator classifier. We analyzed these theoretically to understand the metrics from a mathematical perspective. Then, we implemented these programmatically to see the empirical results. In this research, we contribute an understanding of adversarial robustness on three types of attacks that model the adversary differently on the perceptron algorithm. Future work

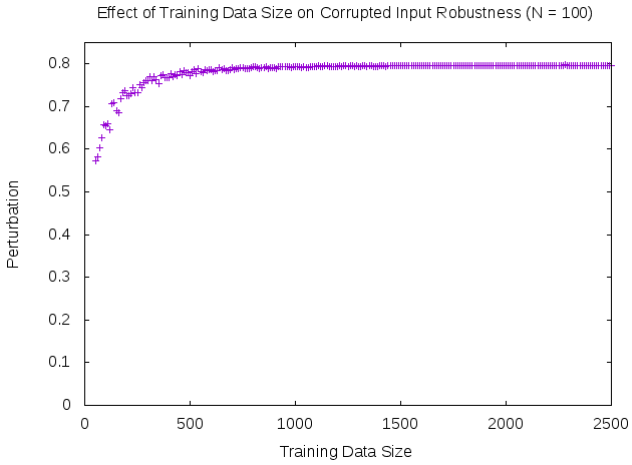

Figure 3: Graph showing the perturbation necessary to satisfy the Corrupted Input metric with the above parameters as the training size increases.

Figure 4: Graph showing all the metrics on a single plot. This allows us to compare the perturbations necessary for each.

can explore the empirical analysis of the Strong Definition as well as analysis of other metrics. It can also examine different distributions, non-homogenous hyperplanes, and other machine learning algorithms.

ACKNOWLEDGMENTS

I would like to thank Mohammad Mahmoody, for all his help throughout this research project. His input and guidance were monumental in bringing about the results of this research. Without him, this project would not have been possible.

REFERENCES

[1] Nicholas Carlini and David A. Wagner. 2016. Towards Evaluating the Robustness of Neural Networks. *CoRR* abs/1608.04644 (2016). arXiv:1608.04644 http://arxiv.org/abs/1608.04644

[2] Dimitrios I. Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. 2018. Adversarial Risk and Robustness: General Definitions and Implications for the Uniform Distribution. *CoRR* abs/1810.12272 (2018). arXiv:1810.12272 http://arxiv.org/abs/1810.12272

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. (Dec 2014). https://arxiv.org/abs/1412.6572v3

[4] Jürgen Schmidhuber. 2014. Deep Learning in Neural Networks: An Overview. *CoRR* abs/1404.7828 (2014). arXiv:1404.7828 http://arxiv.org/abs/1404.7828

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. (Feb 2014). https://arxiv.org/abs/1312.6199v4

Development of Regulation to Mitigate Decision-Making Algorithm Bias

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Bhuvanesh Murali
Spring, 2019

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

Signature _____ Date _____
Bhuvanesh Murali

Approved _____ Date _____
Sean Ferguson, Department of Engineering and Society

Development of Regulation to Mitigate Decision-Making Algorithm Bias

With the proliferation of computer science, we see algorithmic decision starting to appear in many fields. These algorithms often stem from the field of artificial intelligence, specifically machine learning. This implies that these algorithms continue to develop and learn after being deployed into production settings. These algorithms often find trends and patterns overlooked and missed by humans. However, this also leads to the possibility of these algorithms finding unintentional trends. These trends are frequently discriminatory in nature and can lead to biased automated decision making.

According to Rieder and Simon (2016), "[t]he virtuous machine emerges as ever more powerful as it covers increasingly large parts of the analytical and decision-making process" (Rieder and Simon, 2016). Along with this, we see the reduction in the need for theory, models, and human expertise due to the existence of this modern software. This software is often opaque and emphasizes impersonality, as it tends to not express any semblance of human emotion. If data being used were fully quantified and free from bias, it would potentially expand into even more areas of application due to its objective state (Rieder and Simon, 2016). Acquiring data or creating software that is fully free from bias is difficult to achieve. In fact, it may not be technically solvable and as such governance is needed to provide oversight and transparency on this area of algorithms and used data.

Biased algorithms are a problem that must be dealt with due to their discriminatory repercussions. The extent of the issue can be seen through the results of various studies. In a study focusing on facial recognition and analysis, machine learning algorithms were found to misclassify darker-skinned females most frequently, with error rates of around 35%. To put that

into perspective, lighter-skinned males had a maximum error rate of 0.8% (Buolamwini and Gebru, 2018). Similarly, in a study focused on crime recidivism prediction systems, algorithms tended to negatively assess people belonging to a zip code and of a certain race. (Chouldechova, 2017). In both above studies, we see a clear bias based on demographics, which the algorithms simply happened to learn. This can lead to discriminatory results simply based on an individual's race or sex, which is often neither intended nor desired. As a result, it is important to regulate this to prevent this discrimination from occurring. The following research will examine the European Union's General Data Protection Regulation (EU's GDPR) as a case study to focus on the necessary aspects of a data regulation strategy that will help mitigate these unintentionally perpetuated biases.

Case Study: Evaluating the EU's General Data Protection Regulation (GDPR)

To begin understanding what strategies such a regulation needs, we will look at a case study of the EU GDPR. The GDPR is a strict data protection regulation that recently went into effect (i.e. May 2018). The primary focus of GDPR is to protect the consumer and their data in an increasingly online and data-centric world. This regulation has forced companies to be more transparent about their data usage and more receptive to customer requisitions for the data. This regulation serves as a model for contemplating regulatory practices around algorithms.

To begin our analysis, we need to understand the parties involved in the GDPR discussion. The stakeholder network involves over a dozen entities; however, there are about five primary parties, which we will focus on. The first of these stakeholders is the Data Subject, who is a person who can be identified by reference to an identifier, such as name, number, and

location. The data subject is the provider of the information through the use of software and other technologies. The Controller is an authority or agency, which determines the purposes and means of processing personal data. They are linked to the Processor, who is the authority or agency, which processes data on behalf of the Controller. The key difference between the Controller and Processor is that the Controller is directly linked to the Data Subject and has the overall idea for the use of the data. The Processor is not linked to the Data Subject and is merely implementing what is asked by the Controller. The Data Protection Officer (DPO) is designated by the Controller or Processor and advises and monitors the Controller and Processor to ensure compliance to the regulation. They are the point of contact for the Supervisory Authority. The Supervisory Authority (SA) is an independent public authority that is responsible for monitoring the application of GDPR (Huth, Faber, and Matthes, 2018). The SAs essentially serve as an independent party of enforcers for the regulation. They have several corrective and investigative powers, which include conducting audits, reviewing certifications, etc. (Article 29 Data Protection Working Party, 2016). The SAs examine different corporations to ensure compliance with GDPR. In the case of non-compliance, they have the authority to introduce punitive measures. The structure of this regulation focuses upon a differentiation of responsibility and a more public oversight.

Now, that we have looked at the specific stakeholders involved with this policy, we can examine the policies that GDPR focuses on. The table below presents the overarching themes that are prevalent throughout the document and used as points of governance by the SAs.

Table 1.

Overarching themes of GDPR

Lawfulness, Fairness, and Transparency	Personal data must be processed properly according to law. The data must be process in a transparent manner with respect to the data subject.
Purpose Limitation	Collected data should only be used for the specified, legitimate reasons stated and should not be further processed in any incompatible manner.
Data Minimization	Data should be adequate and limited to what is necessary for the processing.
Accuracy	Personal data should be accurate and kept up to date.
Storage Limitation	Personal data shall only be kept in an identifiable form for the necessary time. The data should not be kept beyond that time and if it is, it should not be able to uniquely identify a data subject.
Integrity and Confidentiality	The data should be protected as it is identifying information and should be treated as confidential information.
Accountability	The controller is responsible for compliance with GDPR and should be able to demonstrate the compliance.

Note. The information presented in the table was acquired from the source Heywood and Pomaizlova (2016).

The seven tenets presented in Table 1 comprise of the overarching themes of the data privacy regulations present in the GDPR. As seen from these themes, the GDPR focuses on protecting the data acquired from the Data Subject. It strives to set boundaries on proper use of this data. The overall goal of these tenets is to ensure the Data Subject will be treated fairly, be protected from misuse of his/her data, and be aware of how the data was collected and used.

The policies present in the regulation affect the Controller, Processor, and Data Subject stakeholders. Violations of these policies by the Controller will be met with consequences

delegated by the SAs. A company that violates GDPR could be fined up to 4% of its worldwide annual revenue from the previous fiscal year (Greengard, 2018). If we look at the technology industry, we see that this data protection regulation had an immediate impact. In anticipation of the regulation, many tech companies, including the tech behemoths, are working to comply with the new law. For instance, Google and Facebook have both deployed large teams to overhaul how they give users access to their privacy settings and redesign products that acquired too much user data. Facebook said it had about 1,000 people working on the initiative, including engineers, product managers, and lawyers (Satariano, 2018). Despite these efforts, both Google and Facebook were expected to be hit with hefty maximum fines of $4.88 billion and $4.89 billion, respectively (Keane, 2018). In January 2019, after several months of examination and investigation, Google was handed its first fine of $57 million due to violations of the data privacy law as the company did not properly disclose how data was collected across its services to present personalized advertisements. The fine is only a fraction of the maximum allowed fine but is significant as it forces Google to modify its practices for building advertising profiles of its users. The penalty shows that regulators are following through on their pledge to push back against companies that take advantage of data collection practices (Satariano, 2019). This is only the beginning of the enforcement of the regulation but shows that GDPR is being taken seriously by industry as well as the enforcers.

Crafting an Algorithmic Bias Regulation

Now, we can turn to examine the necessary aspects of the algorithmic bias regulation. Just as with our GDPR analysis, we begin by discussing the stakeholders. Each of these parties

will have different, but vital roles in the adoption and enforcement of such a reform. Both the algorithm regulation as well as GDPR focus on the same network of actors. Thus, we can adopt the same network of Data Subject, Controller, Processor, Data Protection Officer (DPO), and Supervisory Authority (SA). The roles of each stakeholder are presented in Table 2, below.

Table 2.

Comparison of Stakeholders in GDPR and the Bias Regulation

Stakeholder	GDPR	Bias Regulation
Data Subject	Provider of the information. Information is acquired through the use of software and other technologies.	Provider of the information for the Controller and Processors well as the entity that is impacted by the result of the algorithms.
Controller	Authority or agency which determines the purposes and means of processing personal data.	Authority which determines the purpose of the algorithm. This entity determines the purpose for the algorithm.
Processor	Authority which processes data on behalf of the Controller.	Entity that completes the implementation on behalf of the Controller.
Data Protection Officer (DPO)	Individual designated by the Controller or Processor who ensures compliance to the regulation.	Individual appointed by the Controller or Processor to advise and ensures compliance with the regulation.
Supervisory Authority (SA)	Independent party of enforcers for the regulation.	Ultimate enforcer of the regulation.

Note. The first column contains the stakeholder groups, the second column contains the information about GDPR, and the third column includes information about the bias regulation. The information regarding GDPR was acquired from Huth, Faber, and Matthes (2018).

As displayed in Table 2, many of the stakeholder groups have similar, if not identical, roles between the two regulations. The primary differences arise in that the bias regulation also

focuses on how the data is used by algorithms. As such, the Data Subject, Controller, and Processor all have clauses added about the impact of algorithms.

To contextualize these stakeholders and provide a more concrete understanding, let us consider a hypothetical thought experiment for the algorithmic bias setting. Consider a scenario where a corporation uses a decision-making algorithm to parse resumes in its hiring practices. In this scenario, the Data Subjects are the individuals that submitted resumes to the corporation. These individuals are judged by this algorithm and are directly affected by the algorithm's judgement. The Controller would be the corporation accepting the resumes. This corporation is using the machine learning algorithm in as part of its hiring practice. The Processor would be the entity that develops and implements the algorithm that is examining the resumes. The Processor could be engineers and scientists that are members of the corporation (Controller) or could be a different corporation that created the algorithm that the Controller is using. The DPO would be an individual appointed by the company, such as the Chief Security Officer (CSO). The SA would be an independent public authority, just like in GDPR, that examines and ensures that the Controller and Processor are remaining compliant with the regulation. In GDPR, this role is laid out separately from the government in order to allow for more geographically localized focus. The SA should take on a similar role in our bias regulation. If the corporation violates any of the tenets of the regulation, the SAs have the authority to impose a fine or other punitive measure against the corporation.

The content of the algorithmic bias regulation should follow similar overarching tenets as the GDPR regulation. Additionally, the regulation should cover the general principles outlined by the Institute of Electrical and Electronics Engineers (IEEE) for prioritizing human wellbeing

with artificial intelligence (AI) and autonomous systems (AS). The first principle is "Human Benefit," which focuses on the design and operation of AI to respect human rights and freedoms. The AI must also be verifiable safe and secure throughout its lifetime. In the scenario that the AI does cause harm, the root cause must be traceable (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016). This is similar to the GDPR themes of "Purpose Limitation" and "Integrity and Confidentiality." Combining these ideas, the bias regulation should guarantee that the algorithms should respect human rights and the law. It should additionally remain secure throughout its lifetime, so that the data being used is protected from malicious entities and misuse. Let us consider our earlier thought experiment. To comply with this tenet, the corporation should ensure that the data (i.e. the resumes and any personal information) being processed by the algorithm is stored securely, so that it can only be accessed and used when necessary.

The second principle is "Responsibility." This focuses on the accountability of the AI. One such example is that the manufacturer must have clarity around the manufacture of the system to avoid potential harm (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016). In the bias regulation, this principle would focus on ensuring that no group of data subjects are being treated unfairly due to parameters, such as race, sex, etc. For instance, in the thought experiment, the corporation would need to show that their algorithm did not reject an applicant's resume because the system preferentially chose males over females. The corporation could, as an example, show that a male and female with identical qualifications are treated the same by the algorithm.

Another principle is the idea of "Transparency." The ideal goal of this is to make sure it is possible to discover how and why the system made a particular decision. This is especially important to users as it builds trust in the system (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016). These ideas can be combined with those present in the "Lawfulness, Fairness, and Transparency" theme of the GDPR. In the bias regulation, the Controller should be able to explain how their algorithm came to its decision. In our hypothetical experiment, the corporation should be able to explain how the algorithm came to a decision to continue with or reject an individual in the hiring process. There should be some sort of traceability that elucidates the black-box process that the algorithm used.

The final principle presented is "Education and Awareness," which states the importance of educating citizens about the risks and potential misuse of AI systems (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016). As part of the bias regulation, it is important that individuals are educated about the current problems present with these decision-making algorithms. It is important that these data subjects are aware of the potential misuse. In the hypothetical example, this can be an effort taken by the corporation to explain to its applicants that there are issues with biases that the corporation is striving to mitigate. This could also be an effort undertaken by the SAs to educate the public on the current state in algorithm bias regulation.

In addition to these overarching tenets, it is important to have a methodology for auditing algorithms. Algorithmic audits would need businesses to prove to regulators that their technology doesn't discriminate against a class of people. This would be a third-party approval similar to "organic" stickers on produce. This could go a long way to helping customers gain trust that a

company is being thoughtful in its approach (Hempel, 2018). In our bias regulation, this could act as a certification for corporations and their algorithms to express that certain algorithms have been verified. In our theoretical example, this could be a yearly audit by the DPOs or comparable effort by the SAs.

It is important to take note that a policy that works in the European Union (EU) may not have as much success in the United States (U.S.). Specifically discussing aspects of a data privacy policy, the U.S. in general has adopted an opt-out approach to data collection (Greengard, 2018). This essentially forces the burden on the consumer to instruct a company to not collect his or her data. The EU in contrast to this has implemented a more restrictive opt-in approach (Greengard, 2018). In the U.S., data privacy is essentially a free-for-all. There is no central authority that enforces or governs this idea. Rather there is a mosaic of different state and federal rules, that govern some of the same issues. This is partially due to the lack of an agency to carry this type of regulation out. EU member states have their own data privacy authorities, but this does not exist in the U.S. (Hawkins, 2018). Europe has taken the role of the world's foremost tech watchdog. In addition to GDPR, the EU has paved the way for stricter enforcement of antitrust laws and tougher tax policies against tech behemoths. The EU's proactive stance is a sharp divergence from the U.S., where little action has occurred to regulate the tech industry (Satariano, 2018). Overall, the EU's more liberalized view of data governance and protection of its members may have attributed to GDPR's success. As a result, a similar policy may not be as effective in the U.S. and so this should be kept in mind in considering the presented regulation.

Conclusion

Machine learning algorithms are being rapidly adopted into many areas of society and are often found seamlessly integrated into many systems. This research explored the use of these algorithms in the technology industry as well as a theoretical use in hiring processes, which appear in many other industries. Beyond those instances, these algorithms have applications in other industries and sectors. These algorithmic systems provide many benefits as they are autonomous decision-making systems. However, they also introduce unintended biases found in the data or learnt by the AI system. As a result, it is important to regulate these algorithms to prevent the perpetuation of discriminatory information. To do this, a regulation was presented that was based off the European Union's General Data Protection Regulation. It is important to note that the cultural and legislative differences may make it difficult to apply a successful regulation from the European Union to the United States. This research proposes themes and tenets that such an algorithmic bias regulation should focus on in order to protect the wellbeing of the users and data subjects. Additionally, the research advocates for better governance of these technologies in the United States. The current legislation and infrastructure are lacking when it comes to dealing with the rapidly evolving technology sector and thus need to be revamped.

References

Article 29 Data Protection Working Party. (2016). Guidelines for identifying a controller or processor's lead supervisory authority. European Commission. Retrieved from http://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp244_en_40857.pdf

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv:1610.07524 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1610.07524

Hawkins, D. (2018). Why a privacy law like GDPR would be a tough sell in the U.S. Retrieved from https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2018/05/25/the-cybersecurity-202-why-a-privacy-law-like-gdpr-would-be-a-tough-sell-in-the-u-s/5b07038b1b326b492dd07e83/

Hempel, J. (2018). Want to prove your business is fair? Audit your algorithm. *Wired*. Retrieved from https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm/

Heywood, D., & Pomaizlova, K. (2016). The data protection principles under the general data protection regulation - taylor wessing's global data hub. Retrieved from https://globaldatahub.taylorwessing.com/article/the-data-protection-principles-under-the-general-data-protection-regulation

Huth, D., Faber, A., & Matthes, F. (2018). Towards an Understanding of Stakeholders and Dependencies in the EU GDPR.

Greengard, S. (2018). Weighing the impact of GDPR. *Communications of the ACM*, *61*(11), 16–18. https://doi.org/10.1145/3276744

Keane, S. (2018). GDPR: Google and Facebook face up to $9.3 billion in fines on first day of new privacy law. Retrieved from https://www.cnet.com/news/gdpr-google-and-facebook-face-up-to-9-3-billion-in-fines-on-first-day-of-new-privacy-law/

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1), 205395171664939. https://doi.org/10.1177/2053951716649398

Satariano, A. (2018). G. D. P. R. , a new privacy law, makes europe world's leading tech watchdog. *The New York Times*. Retrieved from https://www.nytimes.com/2018/05/24/technology/europe-gdpr-privacy.html

Satariano, A. (2019). Google is fined $57 million under europe's data privacy law. *The New York Times*. Retrieved from https://www.nytimes.com/2019/01/21/technology/google-europe-gdpr-fine.html

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems . (2016). Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. IEEE. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf

Adversarial Machine Learning: From Theory to Practice
(Technical Paper)

Development of Regulation to Mitigate Decision-Making Algorithm Bias

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Bhuvanesh Murali
Fall, 2018

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

Signature _____ Date _____
Bhuvanesh Murali

Approved _____ Date _____
Mohammad Mahmoody, Department of Computer Science

Approved _____ Date _____
Sean Ferguson, Department of Engineering and Society

Introduction

Machine learning is a quickly rising area of computer science that is being adopted in numerous fields. There are applications of machine learning techniques being used in medicine, image recognition, etc. The specific niche of machine learning that has been of more focus recently is deep learning. This style of learning focuses on the use of artificial neural networks, specifically exploring the applications and modifications of multi-layered neural networks (Schmidhuber, 2014). Deep learning has found great success in tasks such as image recognition. ImageNet is a deep convolutional network that greatly improved the accuracy of more traditional methods in image recognition tasks (Krizhevsky, Sutskever, and Hinton, 2012). Studies such as this one led to the immensely growing interest in deep learning.

Despite the significant success of neural networks, many of the applications raise great concerns in regards to safety and security. Researchers have successfully shown that it is possible to generate well-designed input samples that can fool these networks and cause them to misclassify the input, despite the differences being imperceptible to a human being (Yuan et. al., 2018). This raises concerns about the safety of using these neural networks. My technical research will focus on exploring provably successful attacks to further the understanding of how to defend against such misclassifications.

Along with these misclassifications, machine learning faces a problem with algorithmic bias. Many decision-making algorithms perpetuate accidental biases, which can often result in discriminatory decisions. Researchers have shown that often algorithms show statistically higher predictions for one race or gender (Kiritchenko and Mohammad, 2018). Such biases can result in

inadvertently prejudiced decisions. My STS research will focus on tackling this problem by exploring what a successful strategy might be to mitigate the bias in decision-making algorithms.

Technical Prospectus

In recent times, machine learning research has focused on the idea of using deep learning rather than more traditional techniques. Deep learning stems from the idea of using artificial neural networks in learning tasks. The use of these networks has enabled great success in pattern recognition tasks (Schmidhuber, 2014). This accuracy is a result of the ability of neural networks to express any arbitrary computation with many nonlinear components. This structure of the networks has led to excellent performance in visual and speech recognition (Szegedy et. al., 2014).

Unfortunately, this structure also results in difficult to interpret and counterintuitive properties of neural networks. This has led to research in forcing networks to misclassify their inputs by applying minor perturbations (Szegedy et. al., 2014). This idea of perturbing input to force misclassification is the primary facet of the field of adversarial machine learning. The perturbed input is called an adversarial example. Researchers at Google showed that by adding noise to an image of a panda, they were able to get the neural network to misclassify it as a gibbon. The two images are imperceptibly different for humans, but for neural networks it resulted in two different outcomes (Goodfellow, Shlens, and Szegedy, 2015).

Despite the state-of-the-art results that neural networks help us achieve, the vulnerabilities that result from adversarial examples make it difficult to apply neural networks in security-critical areas. As a result, many researchers began focusing on defensive mechanisms to

protect against these attacks. However, many of these defensive measures were only somewhat successful. Defensive distillation was one such proposed defense. It was shown that defensive distillation does not significantly increase the robustness of neural networks, by showing new attack algorithms that were successful on both distilled and undistilled networks 100% of the time (Carlini and Wagner, 2017).

Much of the work done thus far in the field of adversarial machine learning is on implementations of neural networks and developing modifications to programmable attacks to succeed against such networks. Researchers at the University of Virginia took a different approach to the adversarial machine learning field. They studied adversarial perturbations from a theoretical perspective to prove implications about adversarial attacks. They studied adversarial perturbations when the inputs are uniformly distributed over $U_n = \{0, 1\}^n$. They study the inherent bounds on risk and robustness of any classifier for any classification problem whose instances are of the aforementioned form. In their research, they primarily focus on studying the barriers against robust classification of adversarial examples. They are successfully able to prove that the robustness of any classifier is at most bounded by the square root of n (Diochnos, Mahloujifar, and Mahmoody, 2018).

In arriving to this conclusion, the researchers analyzed several attacks theoretically. In doing so, they were able to guarantee worst-case bounds on the performance. However, many of these attacks could be potentially even more damaging and dangerous in practice. My research will focus on taking these theoretically defined and discussed attacks and implementing them in

practice. I will then analyze and examine how well the attacks perform and modify them to try improving the attack to be more successful.

This research will add to the growing body of work in adversarial machine learning. Focusing on implementing theoretical attacks gives a new perspective into understanding provably successful attacks rather than the more traditional applied attacks. My work in this field will help expand the understanding of adversarial attacks, which will lead to the study and development of more robust neural networks that can resist these attacks.

STS Prospectus

With the proliferation of computer science, we see algorithmic decision making starting to appear in many fields. These algorithms often stem from the field of artificial intelligence, specifically machine learning. This implies that these algorithms continue to develop and learn after being deployed into production settings. These algorithms often find trends and patterns overlooked and missed by humans. However, this also leads to the possibility of these algorithms finding unintentional trends. These trends are frequently discriminatory in nature and can lead to biased automated decision making. My STS research project will examine the European Union's General Data Protection Regulation (EU's GDPR) as a case study to focus on the necessary aspects of a data regulation strategy that will help mitigate these unintentionally perpetuated biases.

Data governance is an important aspect to consider, especially with the expanding role of automated algorithms. According to Rieder and Simon (2016), "[t]he virtuous machine emerges

as ever more powerful as it covers increasingly large parts of the analytical and decision-making process" (Rieder and Simon, 2016). Along with this, we see the reduction in the need for theory, models, and human expertise due to the existence of this modern software. This software is often opaque and emphasizes impersonality. If data being used were fully quantified and free from bias, it would push the tenets of mechanical objectivity into ever more areas of application. However, this is difficult to achieve and as such some governance is needed on this area of algorithms and used data (Rieder and Simon, 2016).

Biased algorithms are a problem that must be dealt with due to their discriminatory repercussions. The extent of the issue can be seen through the results of various studies. In a study focusing on facial recognition and analysis, machine learning algorithms were found to misclassify darker-skinned females most frequently, with error rates of around 35%. To put that into perspective, lighter-skinned males had a maximum error rate of 0.8% (Buolamwini and Gebru, 2018). Similarly, in a study focused on crime recidivism prediction systems, algorithms tended to negatively assess people belonging to a zip code and of a certain race. (Chouldechova, 2017). In both above studies, we see a clear bias based on demographics, which the algorithms simply happened to learn. This can lead to discriminatory results simply based on an individual's race or sex, which is often neither intended nor desired.

Any regulation concerning these algorithms will affect several parties. Each of these stakeholders will have different, but vital roles in the adoption and enforcement of such a reform. Private corporations, like Google and Facebook, use machine learning algorithms in everything from their hiring practices through their consumer-facing products. Regulation on data would

impact many companies like these as they would have to modify their algorithms to meet the new standards. In addition, government will need to be involved to set forth the regulation. Regulation would ideally come from an organization like the Federal government so that it would cover a larger set of algorithms. However, it is also very possible that this could start smaller and grow out of the local or state governments (e.g. in liberal cities like San Francisco, Seattle, New York, etc.). To enforce the regulation, it is possible we will need third-party organizations. Enforcement would ideally be done by the corporations themselves, but I believe that a vetting and certification process from a third-party organization would be useful to ensure that these regulations are followed and that the corporations make the changes.

To begin understanding what strategies such a regulation needs, we will look at a case study of the EU GDPR. The GDPR is a strict data protection regulation that recently went into effect (i.e. May 2018). The primary focus of GDPR is to protect the consumer and their data in an increasingly online and data-centric world. This regulation has forced companies to be more transparent about their data usage and more receptive to customer requisitions for the data. This regulation focuses on a different problem than the one I am examining but will serve as a good model for how to go about implementing a regulatory legislation. This data protection regulation had an immediate impact on the industry. Many companies have been facing lawsuits and paying hefty fines due to violations of this regulation. After the regulation was put into effect tech behemoths, like Google and Facebook, were hit with large lawsuits for violating the data protection measures. These companies have since modified some of their practices to become compliant with the regulation. To implement and enforce this regulation, the EU has taken a

third-party approach. GDPR is enforced by a group of individuals called Supervisory Authorities (SAs). They have several corrective and investigative powers, which include conducting audits, reviewing certifications, etc. (Article 29 Data Protection Working Party, 2016). The SAs examine different corporations to ensure compliance with GDPR. In the case of non-compliance, they have the authority to introduce punitive measures.

Returning to the discussion of stakeholders in the bias regulation strategy, I believe that the use of a third-party regulator would prove successful. In the GDPR, the SAs are essentially a third-party regulatory force that enforces the measures set forth in the regulation. Following this model, the third-party regulators seem necessary in the algorithmic bias regulation. The GDPR was set forth by a governmental body and targets the same audience as this bias regulation would. As such, our stakeholders for those two aspects are essentially identical.

The algorithmic bias regulation can follow in the footsteps of the case study of the GDPR. It is important to note that following the EU policy closely might struggle in the U.S. but will help us eventually figure out what might work. There are several factors that the regulation should consider, which I will take a closer look at. It is important to have a methodology for auditing algorithms. Algorithmic audits would need businesses to prove to regulators that their technology doesn't discriminate against a class of people. This would be a third-party approval similar to "organic" stickers on produce. This could go a long way to helping customers gain trust that a company is being thoughtful in its approach (Hempel, 2018). Recently, a company called Pymetrics open-sourced software called Audit AI. This is an algorithm bias detection tool, which was originally used internally by the company to detect biases in their own algorithms. It

is only designed to detect bias, but the actual removal is up to the creator (Johnson, 2018). Such software provides us with some insight into methods to go about detecting bias. The use of an automated algorithm for auditing could reduce the load on third party auditors. The goal of auditing is to allow for validation or certification of the algorithm. Thus, someone else that uses this algorithm could verify if the algorithm has been certified before employing it into their own applications.

Another strategy to explore is the idea of explainable algorithms. Explainable algorithms have been recently surfacing as a way to understand the inner workings of algorithmic decisions. This is the idea of having algorithms explicitly explain their decisions to a human observer. The goal of this is to see how an algorithm's decision is made and to have transparency into this process. This would expose any information being misused by the algorithm to make the decision (Miller, 2018). This is part of a movement called explainable AI (XAI). This could be paired with the auditing process above to automate and/or periodically monitor algorithms.

In addition to bias detection strategies, I will need to understand what type of enforcement will work for this regulation. It is likely that, similar to GDPR, there would need to be established personnel that would verify that a corporation is compliant with the regulation. I think that by looking at GDPR's implementation and paralleling it for this situation, we could achieve proper enforcement. In this scenario, we would want to certify organizations for meeting the criteria and recertify the organization after some time.

Through all of this, I would ultimately deliver a strategy to follow to implement a successful regulation over the decision-making algorithm field. If implemented, this would

mitigate the unintentional biases, which often result in discriminatory results, that are accidentally perpetuated by many such algorithms.

Future Work

For my technical capstone, the primary goal is to implement the adversarial machine learning attacks and analyze the resulting outcomes. This can take the form of several different avenues. We must ensure that the previously proved aspects of the attacks are present and then examine if the attacks have more damaging aspects in practice. This will take place through the rest of the Fall semester and into the Spring semester.

Regarding the STS capstone, continuing to examine strategies for implementing a data regulation is important. I will take a closer look at the case study of the GDPR and focus on more literature review. I may employ the use of interviews to discuss what aspects make regulations successful. As far as a timeline is concerned, the literature review will be done through the end of this semester. The closer look at the GDPR and potential interviews will be conducted late in the Fall semester and into the Spring semester.

References

Article 29 Data Protection Working Party. (2016, December 13). Guidelines for identifying a controller or processor's lead supervisory authority. European Commission. Retrieved from http://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp244_en_40857.pdf

Buolamwini, J. & Gebru, T.. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91

Carlini, N., & Wagner, D. (2016). Towards evaluating the robustness of neural networks. *ArXiv:1608.04644 [Cs]*. Retrieved from http://arxiv.org/abs/1608.04644

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv:1610.07524 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1610.07524

Diochnos, D. I., Mahloujifar, S., & Mahmoody, M. (2018). Adversarial risk and robustness: general definitions and implications for the uniform distribution. *ArXiv:1810.12272 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1810.12272

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *ArXiv:1412.6572 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1412.6572

Hempel, J. (2018, May 9). Want to prove your business is fair? Audit your algorithm. *Wired*. Retrieved from https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm/

Johnson, Khari. (2018). Pymetrics open-sources Audit AI, an algorithm bias detection tool. (2018, May 31). Retrieved October 15, 2018, from https://venturebeat.com/2018/05/31/pymetrics-open-sources-audit-ai-an-algorithm-bias-detection-tool/

Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *ArXiv:1805.04508 [Cs]*. Retrieved from http://arxiv.org/abs/1805.04508

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

Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. *ArXiv:1706.07269 [Cs]*. Retrieved from http://arxiv.org/abs/1706.07269

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1), 205395171664939. https://doi.org/10.1177/2053951716649398

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *ArXiv:1312.6199 [Cs]*. Retrieved from http://arxiv.org/abs/1312.6199

Yuan, X., He, P., Zhu, Q., & Li, X. (2017). Adversarial examples: attacks and defenses for deep learning. *ArXiv:1712.07107 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1712.07107