APPLICATIONS OF MACHINE LEARNING IN SPORTS

IMPACT OF REGULATIONS ON MACHINE LEARNING ON ETHICS AND INTEGRATION INTO SOCIETY

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
Sindhura N. Mente

November 1, 2021

Technical Team Members:
John Kim

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

Haiying Shen, Department of Computer Science

In recent years, Machine Learning (ML) has gained popularity among computer scientists and researchers as the next major technology, with an increasing number of university courses and more research efforts being focused toward the field. Jones (2019) provides an example of this in the sports industry: the National Association for Stock Car Auto Racing (NASCAR) partnering with Amazon Web Services to use Machine Learning to enhance the live broadcast of races. The two companies used computer graphics and computer vision to improve the quality of livestreams and make them more accessible to users. As the field has developed, competing models have been introduced, all aiming for the actual output of the model to align with the expected output, and for this to remain consistent even as the input parameters change. The reality, however, is that different models are best suited for different datasets and problems, and choosing the correct Machine Learning algorithm to implement for a particular situation can be the most difficult step in reaching a solution. In addition, as the number of potential applications for Machine Learning increases, the question as to the degree to which Machine Learning will become integrated into society remains unknown, especially as it relates to privacy concerns and ethics, and how the regulations, or lack thereof, on Artificial Intelligence systems affect these concerns. The two topics are tightly coupled, since the goal of the technical project is to not only develop a well-performing model, but to develop one that has the least amount of bias possible. Bias is an ethical issue that may be made more apparent by the lack of regulations on a system, which is what the STS project explores with the aid of the Actor-Network Theory framework. Figure 1 below is the schedule of research for the Fall 2021 semester.

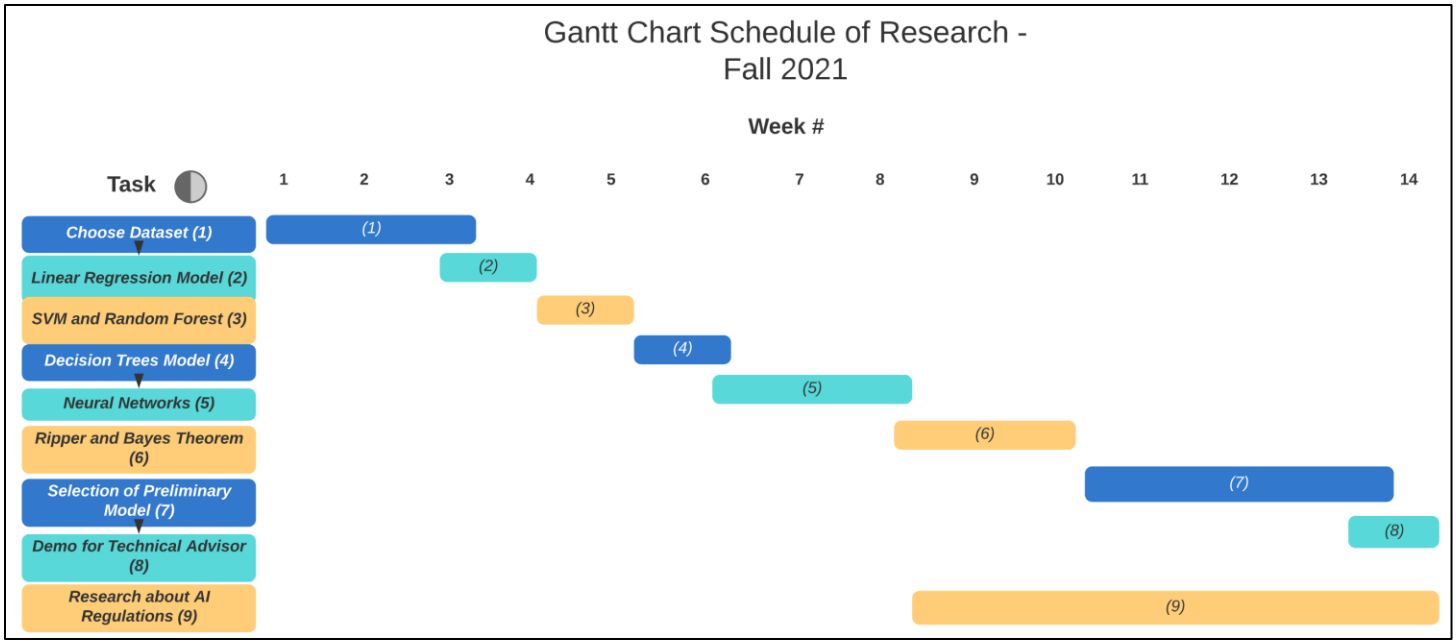

Figure 1: Schedule of Research. This figure outlines the research to be done on the technical and STS projects during the Fall 2021 semester. (Mente, 2021).

APPLICATIONS OF MACHINE LEARNING IN SPORTS

Machine Learning is becoming a widely used term in both academia and the industry and refers to the subset of Artificial Intelligence (AI), which can be further split into two fields: General AI and Applied AI. General AI is the development of algorithms, neural networks, and models to mimic the human brain and train computers to seemingly think for themselves and respond to human situations. This is often measured by the ability of a system to pass the Turing Test, which determines if a device using Machine Learning is able to fool a human interrogator into believing it is a human. This works by having the interrogator converse with a mixture of humans and artificially intelligent systems to determine if any distinguishment between them can be made, and was developed by Alan Turing. Turing goes further to claim that something AI will never be able to achieve is to tell right from wrong and inherently behave ethically (Stanford

Encyclopedia of Philosophy, 2021, p. 5). This highlights one of the major shortcomings of Machine Learning algorithms, which is that they are designed with an inherent bias and are often not able to be extrapolated to apply to data outside of the training data and make accurate predictions or decisions, or behave ethically as a result of the bias. This view is echoed by Redman (2018), and he cites data cleaning, or the scaling, and slight modification of training data to become viable to use in computation, as the most time-consuming step in developing an algorithm. This is also the step that amplifies any bias in the dataset and propagates it throughout the steps of creating an algorithm. Applied AI, is the application of a Machine Learning concept to fit a particular need or as a solution to a problem to align with the users' desires. An example is Facebook, which populates the News Feed with advertisements and posts that align with the user's interests through Machine Learning algorithms that collect data from the user over time. This is done using predictive and statistical analysis that is generalized to account for different datapoints and evolving data (Hellard et al., 2018).

BIAS IN MACHINE LEARNING

The technical project aims to study different Machine Learning algorithms and models and to select and modify and train one to predict the outcome of a basketball season with as much accuracy and precision as possible, as well as determine which features have the greatest impact on the outcome. This will be done under the guidance of Haiying Shen, Assistant Professor in the Department of Computer Science, and with John Kim, and undergraduate student majoring in Computer Science in the School of Engineering and Applied Sciences. This topic is important, since the applications of Machine Learning in the sports industry have been increasing, and there is always a need for better deep learning algorithms. With Machine Learning being a relatively new field, the majority of existing models are not satisfactory enough

to be released in the form of products to consumers who would indirectly interact with this Artificial Intelligence on a regular basis. He (2016) describes Machine Learning algorithms as being comprised of three main parts: the training phase, the validation phase, and the testing phase. Although these three phases are crucial to developing a working model, they are often inefficient, taking up a considerable amount of system memory and time, and consuming a lot of power, which results in the system not being practical for handling large datasets. The problem the technical project will strive to solve is that identified above: to develop a powerful model that accurately predicts the outcome of a basketball season, and to reduce the bias in the data through hyperparameter tuning and other adjustments to the trained model. Manyika, Preston, and Silberg (2019) explain that the presence of bias in artificially intelligent systems is harmful and cite an example of a British school accused of discrimination because their algorithm for choosing prospective students often excluded women and people of color. They also claim that removing biases is the responsibility of the developers of the models to ensure that the final product is reliable. They present different ways to mitigate biases, including making use of existing services that detect and reduce bias, and the standard method of taking the time to

improve the model and tuning it (p. 2). Figure 2 shows the method being followed to develop the algorithm.

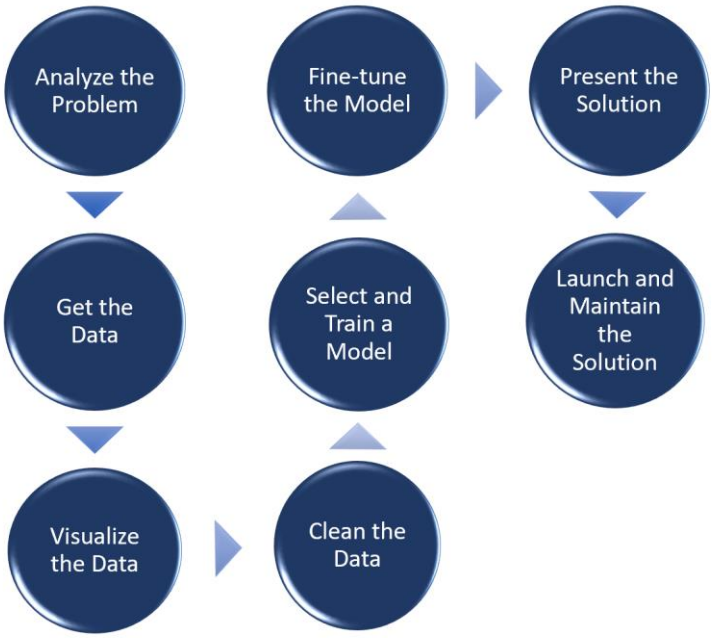

Figure 2: Developing a Machine Learning Algorithm. This figure demonstrates the classic eight-step process to approaching a Machine Learning Problem. (Adapted by Mente (2021) from Nguyen 2019).

METHOD OF DEVELOPMENT

The project will rely on previously collected data from the KenPom website (https://kenpom.com/) and entail working on the computer to train, validate, and test the selected model using supervised learning algorithms, to predict the outcome of a basketball season. A similar project has been completed by researchers in China, who were able to predict if a basketball shot would make it into the net based on the height and angle of the throw with 91.75 percent accuracy, as described by Shao and Zuo (2021, p. 1). A goal for the technical project is aiming for a similar threshold of accuracy. The approach that will be taken to develop the project is as follows. First, the basketball dataset is analyzed, and the data are cleaned. Afterwards, the

prepared data will be split into a train and test set to later validate the model. Then, the different Machine Learning algorithms that are applicable to the problem of predicting an outcome and creating a computation model will be trained on the cleaned and scaled data, specifically the training data. Once each of the different models and methods for prediction have been implemented, they are validated with the test data and the root mean square error is used to determine the error of the model and just how much bias was present in the initial data itself. After choosing the model with the lowest root mean square error, it will be improved using hyperparameter tuning and other methodologies. Another end goal of the project is to create a viable tool that can be used by basketball organizers to evaluate and modify their current strategy. The tool would also be useful for those in the sports betting business, which is an industry with a large market for predictive technologies. Another implication of the importance of the technical project is in the medical field. The demand for predictive technologies to detect serious illnesses has been steadily increasing, and the methods used in this project could be extrapolated and applied to a medical context as well. This project is expected to be completed by the summer of 2022.

The available resources are Google Colab, which allows for the creation and running of machine learning models, as well as consulting the various Computer Science professors in the case that the project is not moving forward at the pace that it should be. The anticipated outcome is to develop a Machine Learning model that can take in input regarding the current information for each team for a given basketball season, including variables such as field goal percentage, and determine who will win the season based on the data at that point in time. Another hoped for outcome is for the model to be able to accurately predict outcomes for future seasons, and past seasons, since the model is being trained on the most recent basketball data set. The reasoning

for this is that if the predictions are still relatively accurate when cross validated with other datasets, then it is indicative of a fair amount of the bias being removed from the model, making it more useful and reliable. The model is also expected to have high degrees of both accuracy and precision to ensure that it is able to be used and has practical applications. The type of paper that will be written for the technical portion of the thesis is a scholarly article that will outline the problem with current Machine Learning algorithms and detail the steps taken to attempt to develop a model that is able to mitigate the problem to at least a small extent.

IMPACT OF REGULATIONS ON MACHINE LEARNING ON ETHICS AND INTEGRATION INTO SOCIETY

The field of Machine Learning has witnessed waves of rapid development followed by disinterest, known as "AI Winters," since it was introduced, as explained by Schuchmann (2019, p. 8). Many people believe there will be another winter; however, Emmert-Streib (2021) takes a strong view on this matter and states that it is certain that Artificial Intelligence will be heavily integrated with society and that another winter is not likely (p. 289). Since Artificial Intelligence and the users interact with each other, the design of Artificial Intelligence is conducive to the consumer's needs, which further affects the behavior of the user and society at large. However, this is a disputed topic. Although another winter does not appear likely, and the numerous applications of Machine Learning and Artificial Intelligence continue to only grow, there have not been serious attempts to regulate their development or use, even though data privacy concerns and consumer safety issues have alerted governments to the need for regulation.

PROBLEM AND METHODOLOGY OF PROJECT

The STS project will seek to gauge the prevalence and applications of Machine Learning in society in the present day, as well as what societal impacts it may have in the future due to recent developments and the digitalization of the world. More specifically, the problem to be explored is how the lack of regulations on Machine Learning are raising ethical concerns on behalf of users, and how without regulation, these technologies are able to become more integrated into society along with the ethical issues they raise. The research question intended to be answered is, to what extent are regulations in place to restrict Machine Learning technologies, and how is this impacting the ethical issues that Machine Learning technologies have and their integration into society? This question will be explored in the general scope of Machine Learning, since all sectors of application of these technologies have ethical and privacy concerns and need to be regulated. The STS topic is tightly coupled with the technical topic, since Machine Learning already has applications in sports, and access to gambling and placing bets on the outcomes of sports matches was recently approved in the United States (CBS Sports, 2021), which could increase the demand for predictive models domestically. Since the sports betting industry is already heavily regulated, it is likely that the use of predictive models to help determine the outcomes of matches and sports season will be equally strictly regulated, which will be important to consider in the development and use of the model described in the technical paper.

ETHICS IN MACHINE LEARNING

Most privacy concerns surrounding Machine Learning and Artificial Intelligence in the technology industry, namely through browsers and social media, appear to be a result of

technological companies tracking users' data and collecting information about the user to then show them similar items, posts, or ideas that relate to their interests. The problem that arises from this is that the information is being collected without the consent of the users – in most cases, users are unaware of the extent to which their browsing habits on apps and websites are being tracked, which is happening constantly. Furthermore, this information is sometimes sold to other sites and apps as a way for companies to make money and exploit their users (Rongione, et al., 2006, p. 58). The constant gathering of sensitive information, for example, in the medical field, makes it a target for hackers and those with malicious intent. Shneiderman (2021) describes the need to combine Artificial Intelligence with user experience design while also ensuring that ethical concerns, including bias, are mitigated to secure the privacy and equality of consumers. He then lists some of the different social and corporate groups that are responsible for taking these factors into account and the interactions between them. Shneiderman contends that regulations must be instantiated to ensure ethical development of these technologies, because it is the lack of them that have even allowed these concerns to emerge and become so prevalent in society (p. 35).

There is a view shared by few that the increasing prevalence of Artificial Intelligence and Machine Learning is a leading cause for privacy and ethical concerns, whereas the majority of society is constantly pushing for the release of the next Machine Learning technologies, whether

it is in smartphones or virtual reality applications. Figure 3 below contrasts the current view of ethical problems in Machine Learning with the proposed view.

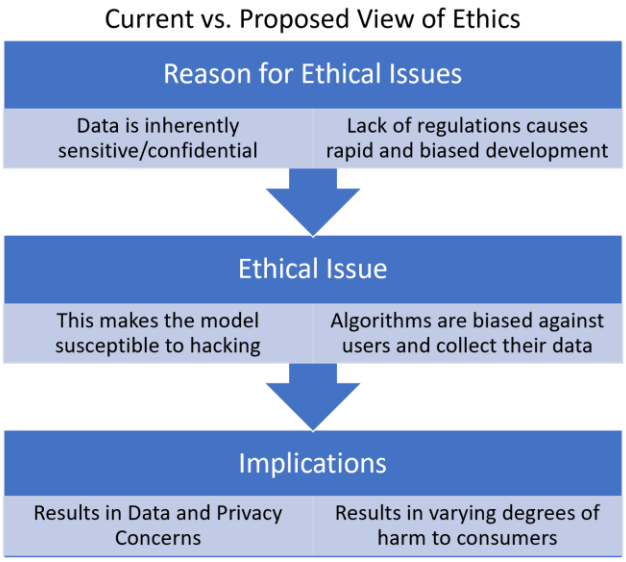

Figure 3: Views of Ethical Problems in Machine Learning. This figure contrasts the current view of ethics in the field of Machine Learning with the proposed view as it relates to regulations. (Mente, 2021).

The fact that most consumers use these technologies without fully knowing how the systems are processing and handling the data is part of the reason why the area of ethics in Machine Learning and Artificial Intelligence needs to be researched more so that governments can enforce the regulation of these technologies and protect the users and those who interact with the systems. This is a view shared by Pasquale and Malgieri (2021), who claim that although Artificial Intelligence and Machine Learning have many applications that would potentially benefit society, people must exercise caution and avoid trying to make these technologies available prematurely before they are fully tested and regulations on how they are to be used are made available. They further discuss the implications of not regulating these technologies with the purpose of making others aware of the ethical problems that arise when imperfect artificially intelligent systems introduce bias and discrimination into the algorithms and result in wrong

predictions and decisions. They recount a situation where algorithms take in a variety of factors regarding different districts and are used to allocate public assistance services, and how the inaccuracy and biases in Machine Learning algorithms may under-allocate resources to poorer districts, which is be a life-or-death situation for those residents. To mitigate problems such as this, the authors expect the United States government to take action to regulate the development and use of technologies using artificial intelligence to protect the end users and customers, as well as those who the use of these technologies directly impact (p. 19).

REGULATIONS ON MACHINE LEARNING

As explored in the preceding discussion of ethical issues in Machine Learning and Artificial Intelligence due to the lack of regulation, the urgent need for regulation before these issues become more prevalent in society due to daily use of such technologies by most of society has become apparent and recognized by many in the industry. Despite this, there has not been much progress in terms of governance over these technologies. In 2018, the European Union enacted the General Data Protection Regulation, which attempts to set guidelines on how consumer data is processed and limits the development of Artificial Intelligence technologies that use this data and are biased. These guidelines are "often vague and open-ended" (European Parliamentary Research Service, 2020, p. 3) and do little to impose any real restrictions on the use of Artificial Intelligence. For example, one of the clauses states that business should only retain the private data of their customers and users for as long as necessary, which is also determined by the business, rendering it unclear and a loose requirement (p. 5). The United States has even fewer restrictions in place, with only four states passing General Artificial Intelligence legislation in 2021 (National Conference of State Legislatures, 2021).

APPROACH TO UNDERSTANDING THE IMPACT OF REGULATIONS ON ETHICS AND SOCIETY

The approach that will be taken to develop the STS portion of the thesis will follow the Actor-Network Theory model. As developed by Latour (2005), Actor-Network Theory is a framework which views how society and technology shape each other in the context of a network that is made up of actors and the interactions between these actors, and simplifies the model such that no external factors exist outside of the network (p. 7). The reason for using this framework is that in the context of ethics as it relates to regulations, the different entity groups can be represented as actors in the network, and the interactions among them determine the severity of the ethical issues that are prevalent in society as a result of a lack of regulations on Machine Learning algorithms. For example, the users, artificially intelligent systems, the developers, companies that sell these technologies, the government, and professional research groups are all different groups or actors in the overarching network, though not all of them. The way that the government responds to the relationship between users and the systems, and the concerns expressed to the government by users, determines the regulations and restrictions placed on the technologies, which directly impacts these as well as every other relationship in the network. Another reason for using Actor-Network Theory as the framework for analysis is the degree of interconnectedness of the actors. For instance, the users' needs determine the systems developed by technological companies which the government must regulate to ensure the protection of the users. Within this network, the users directly interact with the systems and companies on a regular basis and the government constantly monitors, but loosely regulates this interaction. It is important to examine these interactions to determine where the ethical issues are

stemming from and how to mitigate them, as well to determine where the strictest regulations need to be imposed. This can be visualized in Figure 4 below.

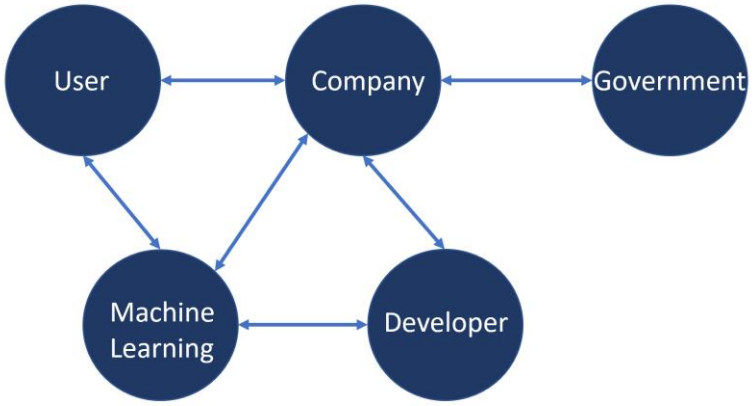

Figure 4: Actor-Network Theory. This figure shows how Actor-Network Theory can be applied to aid the analysis of the STS project. (Mente, 2021).

The anticipated outcome is a thorough analysis of the current regulations on Machine Learning and the ethical concerns surrounding Machine Learning. The goal is to get a deeper understanding of the ethical issue of privacy concerns and data protection for the customer and user and the reasons for why there has not been much regulation from the government's side, either national, state, or local, to investigate the relationship between the two topics and establish causation. It is not only the prevalence of Machine Learning in society that is causing ethical issues to emerge, but a lack of regulation that is in turn allowing the widespread application of these technologies and their questionable ethics in society. West and Allen (2018) present an interesting view on this matter, claiming that the widespread use of Machine Learning and the many different fields and areas that it can be applied to, and this has the same ethical implications in society with regards to privacy concerns, as previously discussed. They also explain the implications of the loss of privacy of the users as systems become more integrated

into society and collect more personal data. They discuss that this leads to issues of informed consent, the unethical selling of collected data, and the potential for the system to be hacked and sensitive information leaked, such as passwords and biodata, which are all detrimental to the user, and are a matter of urgency with regards to regulation. The type of paper written for the STS portion of the thesis is a conference paper that discusses the increasingly prevalent ethical concerns regarding Machine Learning in society and how and why the connection of this to a continued lack of regulation has not been explored yet, and suggest where regulations should be implemented to address the ethical issues studied.

REVIEW OF TECHNICAL AND STS PROJECT

The level of integration of Artificial Intelligence into society, and especially in the sports field remains as of yet unknown; however, the technical project and STS research project will attempt to answer this question. Additionally, the technical project will seek to build a model that is able to predict the winning team given data about various teams for any given time in the season and will aim to be able to accurately do so for a variety of datasets to ensure that bias is not present in the model. An idea for extending the model developed over the course of the technical project in the future, is using computer vision to assess the motion and position of players as additional parameters to determine the outcome of a basketball season. The STS project will focus on the regulations on Machine Learning and how their lack of strictness is fomenting the prevalence of ethical issues and allowing such technologies to become more widely used and integrated in society. In tandem with the extension of the technical project, new ethical issues would arise that would have to be addressed by the STS project such as discrimination among players, copyright infringement if those outside of the coaches and team members obtained the footage and results of the algorithm, and misinformation.

REFERENCES

Allen, J., & West, D. (2018, April 24). How artificial intelligence is transforming the world. *Brookings*. https://www.brookings.edu/research/how-artificial-intelligence-is transforming-the-world/

CBS Sports. (2021, October). *Wanna bet? Here's where all 50 states stand on legalizing sports gambling*. https://www.cbssports.com/general/news/wanna-bet-heres-where-all-50-states-stand-on-legalizing-sports-gambling/

Emmert-Streib F. (2021). From the digital data revolution toward a digital society: pervasiveness of artificial intelligence. *Machine Learning and Knowledge Extraction, 3(1),* 284-298. https://doi-org.proxy01.its.virginia.edu/10.3390/make3010014

European Parliamentary Research Service. (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

He, L. (2016). *The acceleration of machine learning and deep learning algorithms with parallel architectures* (Publication No. 10183862) [Doctoral dissertation, University of Massachusetts Lowell]. ProQuest Dissertations & Theses.

Hellard B., Hopping, C., Walker, D., Fearn, N. (2018, June 4). What is machine learning?. *IT Pro*. https://www.proquest.com/advancedtechaerospace/docview/2057992357/fulltext/E31BB509B7514D15PQ/1?accountid=14678

Jones, C. (2019, June 5). NASCAR revs up its video business with AWS. *IT Pro*. https://www.proquest.com/advancedtechaerospace/docview/2235392981/4346BE61C2AB4FDDPQ/11?accountid=14678

Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network Theory*. Oxford University Press.

Malgieri, G. & Pasquale, F. (2021, August 2). If you don't trust A.I. yet, you're not wrong. *The New York Times*. https://www.nytimes.com/2021/07/30/opinion/artificial-intelligence-european-union.html?searchResultPosition=82

Manyika J., Preston B. & Silberg J. (2019, October 25). What do we do about the biases in AI?. *Harvard Business Review*. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Mente, S. (2021). *Actor-Network Theory.* [4]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Mente, S. (2021). *Schedule of Research.* [1]. *Prospectus* (Unpublished undergraduate thesis).

School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Mente, S. (2021). *Views of Ethical Problems in Machine Learning.* [3]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

National Conference of State Legislatures. (2021, September). *Legislation related to artificial intelligence*. https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx

Nguyen, R. (2019). *Developing a Machine Learning Algorithm.* [2]. Lecture 3: end-to-end ml project [Presentation Slides]. University of Virginia. https://docs.google.com/presentation/d/16TwHNuPBJywAb1D5VbY1zIqSGdqtwO97o13Tnc4QmIw/edit#slide=id.g327d438fde_0_57

Redman, T. (2018, April 2). If your data is bad, your machine learning tools are useless. *Harvard Business Review*. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Rongione, N. M., Sipior, J. C., & Ward, B. T. (2006). Ethics of collecting and using consumer Internet data [Abstract]. *Information Systems Management*, 58-66. https://doi.org/10.1201/1078/43877.21.1.20041201/78986.6

Schuchmann, S. (2019). Analyzing the prospect of an approaching AI winter. *ResearchGate*. https://www.researchgate.net/figure/Timeline-of-the-AI-winters_fig1_333039347

Shao, J., Zuo, H. (2021). *The institute of electrical and electronics engineers, inc. (IEEE) conference proceedings.* Piscataway.

Shneiderman, B. (2021). Viewpoint responsible ai: Bridging from ethics to practice: Recommendations for increasing the benefits of artificial intelligence technologies. *Communications of the ACM, 64(8),* 32-35.

The turing test. (2021, October 4). In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/turing-test/#Tur195ImiGam