Analysis of Systemic Bias Introduced by Machine Learning Applications

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

John Fishbein

Spring 2021

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

Advisor

Sean M. Ferguson, Department of Engineering and Society

Analysis of Systemic Bias Introduced by Machine Learning Applications

The advancements in machine learning in the past few decades have led to increased availability of powerful machine learning models. Using open-source software packages such as Scikit-learn, TensorFlow, or Keras, a developer with minimal experience can employ powerful machine learning models to any data and problem they choose. With this availability comes the opportunity for vast misuse. Both the data and the methods of analysis used in production-level machine learning need to be carefully understood. These powerful, hyper-technical tools have the potential to be misinterpreted by society in a myriad of critical ways.

In the subfield of machine learning, a big challenge is overcoming the bias inherent in the data that is used. On a high level, machine learning works by providing large amounts of data to a model so that an algorithm can learn the patterns and draw conclusions from it. The conclusions reached are highly dependent on the diversity and quantity of data used in the training process. If the given data contains bias, the model's predictions will as well. These issues of bias in the results of machine learning algorithms are a significant problem especially given the increase in their availability and use. This technology is powerful, and few experts fully understand it. Thus, this combination automatically makes machine learning dangerous. Data and the misuse of data is the heart of the problem. In the following paper, I will present these issues of systemic bias and discuss how we can better equip professionals to handle and avoid these problems, and also reach interventions at a myriad of other levels, such as in the math field with the advancement of transparency and in policy institutions.

At its core, machine learning is the intersection of computer science with statistics and applied to big data. Modern algorithms are used to create powerful models that can do a wide variety of things. Take, for example, the convolutional neural network. This machine learning

technique is most often used in the task of image recognition and is used in practice for anything from classifying food images to modern facial recognition software. With the help of software packages, anyone with a minimal coding background can create their own convolutional neural network in under 20 lines of code. The potential for this technology given the open access to it, is incredible. However, because of its widespread accessibility, those who use this technology without fully understanding what is going on "under the hood", can create a product with results that mislead the consumers.

Algorithmic Bias – What is the problem and how is it handled?

Algorithmic Bias is a major concern when it comes to modern machine learning. In today's world, machine learning / AI algorithms are used to make decisions that affect the day to day lives of people. Due to the hyper-technical nature of these algorithms, it can be difficult to evaluate precisely how the algorithm is working and whether the result is biased. At the highest level, machine learning works by "training" the algorithm using known data so that the model can in turn predict unknown data. {add overfitting discussion} Bias can arise from many different underlying issues both with the data used and the underlying model structure. One such example of model specifics within the black box that can produce algorithmic bias is model overfitting. In more complex machine learning models, overfitting is essentially when the model learns to match the training data near-perfectly but cannot generalize to new data. This, in effect, can cause the model to make biased predications. However, algorithmic biases most often occur when the "known" data that is used for training is flawed in some way, and then the resulting predictions become flawed. For example, certain facial recognition algorithms used in production have been shown to contain statistical bias. As discussed in Buolamwini's article in

Time Magazine, facial recognition software was created using a convolutional neural network (Buolamwini, 2019). However, this model exhibited a significantly higher error rate in identifying women versus men. A similar discrepancy was noted on the error rates of identifying minorities. After investigation, this flaw was due to the fact that the known data used to train the model consisted almost exclusively of men.

A similar case of algorithmic bias has been observed in advertisement recommendation software. A widely accepted use case of machine learning techniques is advertisement recommendation software. Companies like Google and Facebook use their user's data to produce predictions of what types of Ads a given user will respond positively to. Google will then sell these predictions to advertisers and other interested parties for the purpose of optimizing some metric related to Ad revenue. In 2015, Datta et. al clearly demonstrated that these mechanisms operate with algorithmic bias (Datta et al., 2015). Through several carefully executed experiments, they modified several aspects of google user profiles and observed the ads that were seen by these fake users. In their first study, they randomly created 1000 users and set 500 of them to have a "male" gender and the other "500" to have a female gender. Then, after recording the resulting ads that were seen, they found that the most commonly seen ad for the "male" group was an ad for an executive level career coaching service. They reported that 402 out of the 500 male users were shown this ad at least once, but only 60 out of the 500 female users received the same ad. Clearly, the advertisement recommendation service considers the gender feature as relevant in the advertisement recommendation, and in the case of this particular ad, the predictions contain algorithmic bias.

When a problem like this is seen in production, it is often times due to the developer's irresponsibility or ignorance. The developer needs to carefully consider the data being used and

thoroughly investigate the test results for bias. This problem needs to be fundamental to the curriculum of every introductory machine learning course because of its potential for dangerous consequences, politically and socially. Ruha Benjamin, in her article Race After Technology: Abolitionist Tools for the New Jim Code, presents the implications that follow from ”cultural coding [that is] embedded into the technical coding of software programs” (Benjamin, 2019). It is far-fetched to think that a prejudiced engineer is sitting somewhere and intentionally writing maliciously discriminatory code. The much more likely scenario is that biased data is used in the machine learning process and skewed results are present in production. Furthermore, since these models are visually nothing more than complex black boxes giving output, it is easy for the results to be misinterpreted by those who do not fully understand their inner workings – the algorithm and the data.

In the current landscape of machine learning, it can be very difficult to detect this algorithmic bias. In the case of Google Ad discrimination shown by Datta et. Al, it is clear that in that specific case, the Ads suggested by google were biased based on gender. However, it is possible to conceive of cases where it would be necessary to account for gender in the recommendation of Ads. Since machine learning models are general thought of as complex black boxes, it becomes difficult to distinguish between algorithmic bias and statistical inference. In order to analyze these predictions more deeply, there needs to be a way to quantify the relative influence a given piece of the data has on the prediction. This gives rise to the field of AI Explainability, which aims precisely to provide a deeper level of insight into how a “black-box” model makes its predictions. This is ultimately a way for these model biases and misinterpretations to be understood and detected.

This misinterpretation is multi-faceted and not always easy to detect. In her article, Whitman discusses the prevalence and use of predictive modeling in institutions— specifically schools/universities dedicated to higher education (Whitman, 2020). Her evaluation addressed the biased results that higher learning institutions are applying to nudge students into greater academic success. She talks about the approach of breaking down data for these models into two categories: "attributes" and "behaviors"—where attributes are inherently unchangeable, and behaviors are determined by what students have control over. In this discussion, she argues that behavior modifications being achieved using the predictive models are not necessarily the university's desired outcome. The correlation between a given student's actions and success is subject to change depending on how administrators define it; this directly influences the reliability of the results when used to assist students. This highlights the foundational concept of why there is a need for the creators of such predictive models to better understand their data.

In the last 25 years, incredible breakthroughs in the field of Artificial intelligence have created immense value for society. It is increasingly evident that systemic bias can be unintentionally introduced into influential technology through these machine learning applications. As engineers, we need to be able to successfully identify and prevent these biases from pervading the field of machine learning. Furthermore, we as a society need to move to a setting in which industry is held more accountable for the algorithmic bias that their enterprise grade models introduce. A shift is needed in how computer scientists and society view machine learning. It is a powerful tool, but data and the context in which it is applied must be carefully understood with respect to its implications in society. The machine learning black-box algorithms are no longer sufficient because of the explicit bias that they introduce, and thus we need to dive deeper.

Moving Forward – How can society combat these problems?

Systemic bias is present in today's world and has detrimental consequences on society as a whole. In the context of machine learning, it is vitally important that professionals are aware of and are capable of handling these issues when they arise. There needs to be a shift in how computer scientists and technical professionals view machine learning and an increased standard of accountability for those who employ said algorithms in the real world. Due to the incredible technological advancements seen in the field of machine learning over the past 20 years, the world has accepted the view that machine learning algorithms are powerful black boxes that, when given the right data, can predict just about anything. This viewpoint is dangerous. It has been demonstrated in countless cases that algorithmic bias can creep into these predictions, allowing discrimination and prejudice to spread throughout the world. This restrictive, black box view of machine learning is no longer sufficient. Society needs to recognize these dangers and adequately handle them so that the field of machine learning can grow responsibly and reach its amazing potential for the betterment of society rather than at its detriment.

In this line of logic, Joy Buolamwini, a prominent representative of the issue of algorithmic bias in machine learning, founded the Algorithmic Justice League: AJL. The AJL is a coalition working to promote awareness of and combat the discrimination introduced by various applications of machine learning. Buolamwini started her campaign into this issue when she encountered algorithmic bias in her work with facial recognition software. She identified cases of discrimination in widely accepted models and algorithms used in facial recognition applications, specifically disparities in identifying people of different races. This biased software is used across the globe for a multitude of different use cases, and thus is actively spreading

social injustice throughout the world. While this injustice is an indirect result of the algorithms, it must be addressed, nonetheless. The main goal of the AJL is to advocate for increased regulation of enterprise level machine learning applications to guard against this systemic bias.

Working with several other researchers within the AJL including UVA professor Vicente Ordonez, Buolamwini introduced a paper titled "Facial Recognition Technologies in the Wild: A Call for a Federal Office" that proposes a new regulatory branch of the federal government to be created specifically for this purpose (Buolamwini et al., 2020). The AJL's proposal is focused on the applications of Facial Recognition Technologies (FRTs). Specifically, the AJL's proposition discusses the establishment of ethical standards and principles with respect to the implementation of FRT and the development of better, more thorough testing standards before the deployment of new FRT (Buolamwini et al., 2020). To achieve this goal, they propose a framework of regulation centered around the concepts of "intended use" and "valid deployment". The intended use of a given FRT, as defined by the AJL, describes the context in which the proposed algorithm will be used. This intended use is then used to identify a risk category in which to consider the given technology. Then, based on this category, the FRT would be evaluated against testing standards to determine whether it is a valid deployment. This framework provides a method through which to systematically evaluate the efficacy of a given FRT and regulate its behavior (Buolamwini et al., 2020).

Efforts such as these are a great step towards this improved view of machine learning and have the potential to be generalized for other veins of machine learning applications. However, there still remains a big problem in terms of making this federal body a reality. It is very difficult to detect these biases when machine learning algorithms are considered to be black boxes. In

other words, how can we identify misuse and how can we establish generalized testing standards if we cannot quantify how a given piece of data effects the output of this black box software?

In parallel to Buolamwini's work in the field of Algorithmic Bias, a relatively large body of study has been produced with the aim of providing better tools to reason about and understand a given black box model. Over the past decade, this work has given rise to the field of AI Explainability. In the seminal paper of this field of Explainability, Carnegie Mellon University professor Anupam Datta and his colleagues introduce a method using which one can quantify the relative influence of a given input on the output of essentially any black-box machine learning framework (Datta et al., 2016). The scope and generality of these proven results offer a fundamental way to "open" the black box of a given model and gain a better understanding of its dependence on the data which it is given. This mathematical advancement is a necessary step towards a more insightful and effective way of identifying and regulating against algorithmic bias in predictive models.

In the few recent years since this publication has been released, many efforts have begun in order to implement these results and make the underlying technology more available to the machine learning community. Companies like Truera and Microsoft's InterpretML are actively making strides to enable this technology to be used. The use cases introduced by this new technology are vast and include but are not limited to providing model transparency, evaluating model fairness and bias, and analyzing the conceptual soundness of a model given the data used to create it (Kundu et al., 2021). This is all a very new field of study and innovation, but the underlying technology provides the groundwork to achieve the level of regulation that Buolamwini and the Algorithmic Justice League are advocating for. These innovations have the potential to form a quantitative basis for regulatory oversight (Kundu, 2021).

Strides in this vein are already in motion around the world, independent from the efforts put forth by the Algorithmic Justice League. Specifically, within the financial industry, this technology is starting to be used for regulatory purposes. In addition to the gender and racial biases shown in prior examples, there also exist comparable cases of algorithmic bias in the financial sector. Consider a machine learning model used to determine whether or not a given person should qualify for a loan based on personal information. This model determination can also be subject to algorithmic biases depending on the data used in training (Sen, 2016).

In cases of algorithmic bias in this area, there are emerging regulatory efforts with the aim of holding the financial industry accountable. Namely, the Veritas Consortium, led by the Monetary Authority of Singapore, is an initiative for Responsible AI including many major banks such as Citibank, Goldman Sachs, etc. ("MAS Partners Financial Industry to Create Framework for Responsible Use of AI," 2019). This initiative began very recently in 2019, but now in 2021 the model explainability company Truera, represented by Anupam Datta, was selected to spearhead the work aimed at accomplishing this goal. Here, we are seeing the beginning of necessary changes in regulatory bodies utilizing this powerful new technology. It is vital that we continue down this path and improve regulation using these new revolutionary quantification techniques.

Addressing the problem early on - Educating professionals against biases

Along with these regulatory efforts, an essential part of preventing this algorithmic bias from continuing to infiltrate machine learning applications is better preparing our professionals to handle these issues. Consider the current UVA undergraduate course: CS 4774 Machine Learning. Professor Nguyen teaches a well-designed course serving as an introduction to the

standard practices and use cases of modern machine learning techniques. Simply by examining the topics covered in the course, it is clear to see that he provides an overview of many of the basics of machine learning (Nguyen, 2021). However, after completing the course myself, only a very small portion of the material covers model evaluation techniques, and the majority of the focus is how to apply machine learning techniques to a given problem. Compare this instead to the equivalent introductory course within the Machine Learning Department at Carnegie Mellon: 10301 Introduction to Machine Learning. Largely, the course covers very similar material to the UVA course. However, clearly listed in the learning objectives of the course is that students should be able to analyze a given ML technique to identify the bias implicit in the algorithm (Mitchell & Gormley, 2021). This is an important component that sets introductory machine learning courses such as these apart.

Despite the efficacy of the examined machine learning courses, both still leave something to be desired for preparing professionals to identify and prevent the introduction of algorithmic biases. With the advancements in model explainability discussed above, these curriculums need to be modified to include a more rigorous introduction to the current best practices of model evaluation. Given that it is possible for an inexperienced developer to employ complex machine learning techniques with minimal experience, it is crucial that aspiring professionals learn the techniques to properly evaluate their work. The innovations in AI Explainability have been proven to provide powerful insights into model predictions and thus can be used by developers to better understand their models. Even though this technology is almost brand new and very much on the cutting edge of development, we must start introducing it into our introductory machine learning education programs.

Conclusion

Ultimately, there are clear next steps that must be taken in order to reduce the systemic bias introduced by machine learning applications. The technological innovations to provide the necessary level of insight into a model's predictions have been created and now these techniques need to be accepted by the field of machine learning in general. Regulatory bodies now have the tools at their disposal to create real and effective regulations over enterprise machine learning applications to protect against these systemic biases. Organizations like the Algorithmic Justice League have the technology needed to quantify cases of algorithmic bias so that they can be prevented. Furthermore, this technology needs to be introduced into machine learning curricula so that future professionals are capable of deeply evaluating their own work. Coupled together, these next steps can go to great lengths to reduce the systemic bias introduced from machine learning. With these steps, our society will be able to view machine learning in a much more mature light. No longer will machine learning algorithms have to be viewed as black boxes. This technology can be used to provide insightful conclusions about the innerworkings of any general machine learning technique and thus can be used to evaluate model performance at a general, widespread scale.

References

Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. 258.

Buolamwini, J. (2019, February 7). Artificial Intelligence Has a Problem With Gender and Racial Bias. Time. https://time.com/5520558/artificial-intelligence-racial-gender-bias/

Darby, S. C., Ewertz, M., McGale, P., Bennet, A. M., Blom-Goldman, U., Brønnum, D., Correa, C., Cutter, D., Gagliardi, G., Gigante, B., Jensen, M.-B., Nisbet, A., Peto, R., Rahimi, K., Taylor, C., & Hall, P. (2013). Risk of ischemic heart disease in women after radiotherapy for breast cancer. The New England Journal of Medicine, 368(11), 987–998. https://doi.org/10.1056/NEJMoa1209825

Kataria, T., Bisht, S. S., Gupta, D., Abhishek, A., Basu, T., Narang, K., Goyal, S., Shukla, P., Bansal, M., Grewal, H., Ahlawat, K., Banarjee, S., & Tayal, M. (2016). Quantification of coronary artery motion and internal risk volume from ECG gated radiotherapy planning scans. Radiotherapy and Oncology, 121(1), 59–63. https://doi.org/10.1016/j.radonc.2016.08.006

Murphy, H. (2017, October 9). Why Stanford Researchers Tried to Create a 'Gaydar' Machine—The New York Times. https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html

Spetz, J., Moslehi, J., & Sarosiek, K. (2018). Radiation-Induced Cardiovascular Toxicity: Mechanisms, Prevention, and Treatment. Current Treatment Options in Cardiovascular Medicine, 20(4), 31. https://doi.org/10.1007/s11936-018-0627-x

Wang, H., Bai, J., & Zhang, Y. (2008). A normalized thoracic coordinate system for atlas mapping in 3D CT images. Progress in Natural Science, 18(1), 111–117. https://doi.org/10.1016/j.pnsc.2007.08.004

Wennstig, A.-K., Garmo, H., Hållström, P., Nyström, P. W., Edlund, P., Blomqvist, C., Sund, M., & Nilsson, G. (2017). Inter-observer variation in delineating the coronary arteries as organs at risk. Radiotherapy and Oncology : Journal of the European Society for Therapeutic Radiology and Oncology. https://doi.org/10.1016/j.radonc.2016.11.007

Whitman, M. (2020). "We called that a behavior": The making of institutional data. Big Data & Society, 7(1), 2053951720932200. https://doi.org/10.1177/2053951720932200

Xue, J., Kubicek, G., Patel, A., Goldsmith, B., Asbell, S. O., & LaCouture, T. A. (2016). Validity of Current Stereotactic Body Radiation Therapy Dose Constraints for Aorta and Major Vessels. Seminars in Radiation Oncology, 26(2), 135–139. https://doi.org/10.1016/j.semradonc.2015.11.001

Sen, S. (2016, September 1). *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*. https://www.youtube.com/watch?v=Ln3JwVcHxJo&t=1331s

Buolamwini, J., Ordonez, V., & Learned-Miller, E. (2020). *Facial Recognition Technologies in the Wild: A Call for a Federal Office*.

Kundu, S. (2021, January 12). Trustworthy AI: Essential and within reach. *Truera*. https://truera.com/trustworthy-ai-essential-and-within-reach/

Kundu, S., Datta, A., & Gopinath, D. (2021, March 25). Machine learning explainability is just the beginning. *Truera*. https://truera.com/machine-learning-explainability-is-just-the-beginning/

MAS Partners Financial Industry to Create Framework for Responsible Use of AI. (2019, November 13). *Monetary Authority of Singapore*. https://www.mas.gov.sg/news/media-releases/2019/mas-partners-financial-industry-to-create-framework-for-responsible-use-of-ai

Mitchell, T., & Gormley, M. (2021). *Carnegie Mellon University Computer Science 10-301, Spring 2021*. Carnegie Mellon University, School of Computer Science. http://www.cs.cmu.edu/~mgormley/courses/10601/about.html

Nguyen, R. (2021). *University of Virginia Computer Science 4774, Spring 2021*. University of Virignia, Department of Computer Science. https://www.cs.virginia.edu/~nn4pj/teaching