Programming Ethics and Reducing Bias into Machine Learning systems

(Technical Project)

Exploring Ethics Within Machine Learning Systems

(STS Project)

A Thesis Prospectus

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

Akira Durham

November 8, 2024

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.

Introduction

The rise of software development and use of that software in everyday life has led to an increase in technology dependency for many research fields and businesses. One application of this technology dependency would be in machine learning systems, propagating themselves in regular use through various products such as Google's search engine, Youtube's recommendations, Instagram's feed, and many more products. These services have a wide-spread reach, with billions of active users every day. As a result of these services, it would not be an overstatement to say our world runs on machine learning.

Machine learning systems are built on top of huge amounts of data, requiring consistent data collection from online sources to further improve and develop models. These models utilize data to generate guesses on behavioral or statistical trends, giving companies the power to guess what a user will type next, what users want to see on social media, and how best to keep users active, to name a few use cases. Data collection comes from users interacting with websites and links, allowing companies to pull their data through cookies as well as logging what users are doing over time to track behaviors and trends.

Once these systems are online, the models generally stay online without receiving new data to learn from, essentially locking them into the past, which can lead to outdated data, difficulties in fixing past errors, or improving the model on discovery of bias (Koshiyama et al., 2022, p.42). As a result of this, models can reflect inherent bias and discrimination that stems from the data, the developers, or the overall creation process (Franks, 2020). Due to these practices, discrimination can be difficult to detect both for users and for the developers themselves when deploying these models (Dubber et al., 2020, pg. 201). This can be seen in several case studies, such as Amazon's discriminatory hiring model, Google's over

representation in Gemini, and Apple's face detection not recognizing people of color as well as people with a whiter complexion, to name a few real-world examples.

To combat these challenges, my technical project is focused on decreasing discrimination and inherent bias within datasets and machine learning models. The project aims to achieve this by reviewing variance between the dataset and real-world representation, and ensuring a relative level of equality across represented groups, as well as reviewing the output of the model to ensure there is not a discriminatory process within. Similarly, my STS project seeks to understand how to improve trust for black-box models and improve accessibility by improving representation for underrepresented groups. Through these two projects, I aim to improve machine learning understanding and improve the underlying ethicality of models.

Technical Topic

Recent research suggests that a key point in the persistence of discrimination within machine learning models comes from biased data, such as mislabelled data, unrepresentative data, and human evaluation error to name a few (Srinivasan, 2021, p.44). Among these data-based errors, certain developmental decisions can disproportionately prioritize certain groups. Due to a lack of governmental regulations and differing company standards, there is a lackluster amount of ethical assurance behind these new models, where the goal is unclear to users. Tackling these challenges is key to developing a more equitable machine learning model. My approach addresses these through three main stages: a pre-model focus on data analysis, a development phase where we equip the model with tools to detect biases, and a post-model process where we test the model's performance for fairness, with respect to the goal of the model.

The data analysis step is key in discovering inherent biases within the data itself because this bias will propagate itself through the model and its output. To tackle this we can analyze the data for a variety of biases, outlined in the article by Srinivasan (2021) alongside finding inherent patterns within the data. By performing this analysis, we can locate biases in the data and decide if this bias would be acceptable given the purpose of the model, thereby placing accountability on developers making these decisions.

The next step would be during the development phase of the model, where one can balance out the biases found within the data. To address these biases programmatically, we can remove or randomize indicators within the data that allow for discrimination, increase representation of underrepresented groups through copying of existing data and slightly altering the averages within the group, as well as adjust the weighting of these values to increase the representation in the output. While performing these actions can help to alleviate underrepresentation, caution is necessary to not break existing relationships within the data, increase bias in another direction, overfitting the data due to the augmentation, and understand if underrepresentation is an integral part of the data. The data analysis and development phases can be combined and repeated, providing developers with a more robust, representative dataset.

Testing the model's output is essential to validate the ethicality and ensure balanced representation. To accomplish this task, we can compare these results to similar models, measure qualitative metrics for representation in the output and compare to real-world distributions, as well as stress test the model with biased data, to detect how the model handles disproportionate data sets and better understand the decisions the model is making. Ensuring a full suite of model testing will assist the developers in tuning as well as increasing transparency for users by displaying this anit-discrimination process.

Compiling these three steps into one framework allows developers to ensure they account for discrimination within their models throughout the entire model building pipeline. By holding themselves and the model accountable, developers become more active in breaking down discrimination and improving model transparency. The main costs to integrate this framework into existing development processes is time, developer cost, and potential cloud virtual machine cost when retraining the model, although much of this cost is alleviated if done during initial development rather than importing this framework onto an older model. The deliverable in this case is this framework and how to approach model creation in order to reduce discrimination.

STS Topic

Improving understanding of machine learning model black boxes can improve representation, educate the public, and hold both the models as well as their creators accountable for discriminatory actions ("STS Concepts", 2021). Viewed through the lens of enhancing public understanding of machine learning models, it is clear that machine ethics plays a key role in preventing discrimination and embedding human values into AI systems (Anderson & Anderson, 2011, p.53).

People interact with machine learning models in all facets of life and in every field, showcasing the need for further understanding of the concepts behind them as well as how prevalent they are in today's world. Machine learning models process data and can transform it into lists of numbers that represent objects in the data. They then detect trends and similarities across this data, giving it the ability to recognize patterns, predict those patterns, and make decisions based on these patterns. For example, a machine learning model can learn the patterns

of a “good” resume from some input data labeled good and bad, giving the model the ability to classify resumes and be applied to resume screening in the hiring process for a company. Understanding how a machine learning model works is key to removing the “black box” implementation of these models. This information allows one to walk through the entire process from through, and understand why the model produces certain outputs (Koshitama et al., 2022, pg.49). Creating transparency in these models builds trust and encourages users to critically think about machine learning systems. As users gain additional insight into how these models operate, from the initial data collection to understanding why a model would give a particular output, they can identify potential biases.

Simply understanding these models is not enough to combat the discrimination. The focus should be on holding developers and the model responsible for discriminatory actions, which allows for accountability in these flawed systems and increases the likelihood of more ethical development. An article from Carnegie Mellon outlines how Amazon had developed a machine learning algorithm to improve hiring processes, but the model was trained on current employee resumes, resulting in an overrepresentation of white men (Iiriondo, 2018). In this example, we can see the effects of a biased data set in how the model was discriminating against women and people of color, whereas if the developers had spent the time to analyze the data further, they would have come to the conclusion that this data was not representative of their hiring practices. IBM describes other situations such as “healthcare, online advertising, and predictive policing tools”, where bias within the training data leads to discrimination that can negatively affect people through unfair enforcement of laws, enforcing gender stereotypes, and even bodily harm (IBM Data and AI, 2024).

Through additional research, I want to explore non-discrimination practices, improving public knowledge, and building trust with people in machine learning models, as these are key to holding developers accountable while improving model effectiveness. The main challenges with this approach include clearly explaining complex models, balancing transparency with privacy, as well as a lack of current regulations (Dubber et al., 2020, pg. 201). A combination of governmental regulations, internal ethical frameworks, and public perception typically constrains companies to current ethical standards. However, due to a lack of government regulation in machine learning development and a widespread public confusion on how these systems work, machine learning companies are largely self-regulated, which leaves much to be desired for the standardization of ethics as well as ensuring equity. In a study by Joy Buolamwini, she demonstrates the discriminatory biases of popular machine learning systems powered by large tech companies such as Google and Microsoft, proving that these companies cannot be relied on to sufficiently guarantee ethicality within their models (Buolamwini, 2018). Developing stronger regulations and improving discrimination auditing from a government regulation perspective can increase company integrity and improve public perception. From this exploration, the deliverable will be a report detailing methods for companies to improve model transparency without compromising data integrity, while including consideration of the models' societal impact throughout the development process.

Conclusion

A systemic approach to reduce machine learning bias and constructing checkpoints during the development cycle creates a healthier, more ethical framework for creating machine learning models. By discussing methods to combat bias from within datasets and exploring

processes to reduce models' negative impact, we can educate developers in creating ethical models. Implementing the methods discussed in this paper allows developers to design more generalizable models and better understand how to reduce underrepresentation and overrepresentation. Users will benefit from these processes by having a more transparent view of how machine learning models work, understanding how discrimination can become a part of the process, and utilizing equitable models in day to day life.

References

Anderson, M., & Anderson, S. L. (Eds.) (2011). Machine Ethics. Cambridge University Press.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Dubber, M. D., Pasquale, F., Das, S., & Oxford Handbooks Online Law (2020). The Oxford Handbook of Ethics of AI. New York, NY: Oxford University Press.

Fiske, A., Tigard, D., Müller, R., Haddadin, S., Buyx, A., & McLennan, S. (2020). Embedded Ethics Could Help Implement the Pipeline Model Framework for Machine Learning Healthcare Applications. *The American Journal of Bioethics*, *20*(11), 32–35. https://doi.org/10.1080/15265161.2020.1820101

Franks, B., & O'Reilly Online Learning: Academic/Public Library Edition (2020). 97 Things About Ethics Everyone in Data Science Should Know: Collective Wisdom From the Experts. Beijing: O'Reilly.

Hollis, C., & Christy, S. (2024, October 16). *Ai, Machine Learning & Big Data laws 2024: USA*. GLI. https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/usa/

IBM Data and AI. (2024, August 21). *Ai Bias Examples*. IBM. https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples

Iriondo, R. (2018, October 11). *Amazon scraps secret AI recruiting engine that showed biases against women - machine learning - CMU - Carnegie Mellon University*. Machine

Learning | Carnegie Mellon University. https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html

Koshiyama, A., Kazim, E., & Treleaven, P. (2022, April 1). Algorithm Auditing: Managing the Legal, Ethical, and Technological Risks of Artificial Intelligence, Machine Learning, and Associated Algorithms. Computer, 55(4), 40 - 50.

Londono, L., Valeria Hurtado, J., Hertz, N., Kellmeyer, P., Voeneky, S., & Valada, A. (2024, April 1). Fairness and Bias in Robot Learning. Proceedings of the IEEE, Proc. IEEE, 112(4), 305 - 330.

Srinivasan, R., & Chander, A. (2021, August 1). Biases in AI Systems. Communications of the ACM, 64(8), 44 - 49.

STS Concepts. Sci-Tech Asia. (2021, September 8). https://scitechasia.org/sts-resources/sts-concepts/

Vollmer, S., Mateen, B., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., ... Hemingway, H. (2020, March 16). Machine learning and artificial intelligence research for patient benefit : 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ: British Medical Journal, 368, 1 - 0.