Generating Custom Real-World Activity Data to Train an Artificial Intelligence Cloud Cybersecurity Model

(Technical Report)

Equity in Artificial Intelligence: Identifying Bias and its Causes in Intelligent Systems

(STS Research Paper)

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

Claire Williams

May 9, 2023

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

Alice Fox, Department of Engineering and Society

Rosanne Vrugtman, Computer Science

Overview

As technology continues to develop at an exponential rate, an emerging field is becoming increasingly prevalent in our lives and is poised to change society and the world as we know it: Artificial Intelligence (AI). While the possibilities of intelligent machines are endless, we must be cognizant of the inherent limitations and biases these machines have. In this project, I will discuss my firsthand experience with AI at my summer internship. I will also conduct a literature review to explore what it means for artificial intelligence to be biased and identify some of the root causes of this bias.

Problematization

AI is becoming increasingly prominent in our society, so it is important that we acknowledge AI's salient limitations. AI systems are regularly the subject of studies identifying systemic bias. A 2018 study found that facial recognition software is trained on overwhelmingly lighter-skinned data and that misclassification rates for minorities are orders of magnitude higher than for privileged groups (Buolamwini & Gebru, 2019). This apparent bias leads to inequities; In 2020, Robert Williams, a black man, was wrongfully arrested in front of his wife and two daughters due to inaccurate facial recognition (Burton-Harris & Mayor, 2020). In a more general example of the hidden effect of algorithms on society, AI-generated "risk assessment" scores used in the criminal justice system to predict the likelihood of a future crime have been shown to disproportionately mislabel black offenders as high-risk and white offenders as low-risk (Angwin et al., 2016). These inequities also exist beyond race and outside of the criminal justice system. In 2015, Amazon had to stop using an AI recruiting tool as it realized it discriminated against women by downgrading resumes with words like "women's" from "women's chess club captain"

(Dastin, 2018). AI is beginning to permeate all aspects of our society so, moving forward, we must design these systems as equitably as possible.

Guiding Question

How can we define and identify biases in artificial intelligence and where do these biases come from?

Projected Outcomes

In order to understand how to address bias in AI, we must first understand where this bias comes from. From there, we can determine how to improve new systems. We must also continue to be cognizant of bias, as there may be some sources of bias we have yet to identify. AI is liable to perpetuate existing inequalities, meaning the most disenfranchised members of society are the most vulnerable to this systemic bias. Therefore, we must actively understand this bias in order to address it and this project aims to investigate this topic.

Technical Project Description

My technical experience related to this topic is my previous internship at Amazon Web Services (AWS) in the summer of 2022. I worked as a software engineering intern for AWS Detective, a security service that helps AWS users monitor their cloud resources and activity for threats. AWS Detective takes in all of a user's activity data and organizes it into a graph model. Users can use this model to identify and investigate security threats and, as this tool evolves, it will hopefully continue to get smarter.

This experience taught me first-hand the huge potential of using AI to extract meaning from large amounts of data. I observed how the algorithm could extract relevant history that may

be related to an attack from large time periods worth of activity logs. I also witnessed how data-driven these algorithms are and how thoroughly data scientists have to test AI algorithms before they can be reliably used for an application. From this experience, I was able to see how the training data greatly affects the results of an AI and just how new these recent developments truly are. From observing this work firsthand, I understand the magnitude of the importance of ethically developing intelligent machines.

Preliminary Literature Review & Findings

As AI becomes more prolific, many analyses are being conducted on the results of these "intelligent" systems and there is resounding evidence of algorithmic biases being baked into many of these systems. However, the outlook is not all bleak. A 2017 study examined bail decisions and built a model to improve the predictive accuracy of these decisions. When using the model's decisions instead of the judge's and looking at the criminal reoffense rates, the model's decisions reduced the resultant crime rate and decreased the jailing rate while simultaneously reducing racial disparities (Kleinberg et al., 2017). The bias of this system was considered throughout the process, demonstrating that, with the proper attention, building an equitable system is an achievable goal.

While this study is promising, there are countless counterexamples of intelligent systems disproportionately affecting vulnerable groups. The National Institute of Standards and Technology (NIST) identifies three types of bias: Statistical/Computation Biases, Human Biases, and Systemic Biases (*There's More to AI Bias Than Biased Data, NIST Report Highlights | NIST*, 2022). Statistical and computation biases originate from the data sources the intelligence machines are trained on. For example, facial recognition programs perform significantly better

on light-skinned faces, and the data used to train them is also overwhelmingly light-skinned. One study examining these programs looked at programs trained with 79.6% and 86.2% fair-skinned data (Buolamwini & Gebru, 2019). Conversely, human biases stem from the humans building the systems. The technology industry is male-dominated and mostly White and Asian and this can cause inadvertent blind spots when these intelligent systems are being designed (Crawford, 2016). For example, during a federal investigation regarding a persistent wage gap, many Google employees, who develop the algorithms used in their AI, were found to support an "anti-diversity" manifesto (Noble, 2018).

Finally, systemic biases are not biases that an AI creates, but rather that it perpetuates. For example, in criminal justice models, overpoliced neighborhoods record more crime, which leads to greater policing (Silberg & Manyika, 2019). Also, in a 2017 study, researchers found that when using machine learning techniques to conceptualize language, the AI "learned" the human-like biases present in language. They used a common technique called word embedding where the AI learns words by learning associations between words. The AI found stronger associations between female names and family words or male names and careers (Narayanan et al., 2017). If deployed in the real world, the model would perpetuate outdated stereotypes. In another study on natural language processing, researchers built a machine that learned "doctor" as a masculine noun whereas "nurse" was feminine. These stereotypical word associations, and possibly many more harmful ones, are learned by the machine from a massive corpus of text data because these stereotypes are present in how we use our language. It is possible to manually decouple these associations, but this removes important context about the real world. Researchers attempted to debias these word embeddings, but found that it was not feasible as it

removed essential elements of the machine's understanding of language and, depending on the application, would strongly affect performance (Caliskan, 2021).

While we can mitigate these biases, we cannot entirely eliminate them. Therefore, it is exigent that we understand bias in AI and its root causes. Since many AI processes are highly theoretical and quantitative, some researchers have endeavored to mathematically prove that it is impossible to achieve true "fairness" in AI. In a 2016 study, AI researchers defined three types of fairness and found that they are mutually incompatible and in order to make a machine more "fair" by one metric, there was a tradeoff in another metric (Kleinberg et al., 2016). Based on these results, researchers stress that developers must be as transparent as possible about the assumptions their models are making (Friedler et al., 2016).

STS Project Proposal

Science and Technology Studies (STS) is an interdisciplinary field concerned with the interactions between society and technology. As technology becomes more and more ingrained in our day-to-day lives, these interactions become progressively more nuanced. AI has permeated most aspects of society, with facial recognition being used in law enforcement and security, medical learning algorithms being used for diagnosis, and smart home devices being used for daily life. As these intelligent machines become more prevalent, it becomes increasingly important that they are designed with foresight. These machines are poised to replace institutional systems so they must be designed fairly because once they are in place, they will be harder to modify. Additionally, since these machines are involved in such fundamental processes (search algorithm suggestions, hiring decisions, image and speech recognition e.g.), the potential damage bias can cause is exponential.

To approach this problem, we will explore two lenses: the machine learning engineers' data-driven approach and the views of marginalized people which these biases most affect. We will examine quantitative studies researchers have done on AI algorithms. Some of these studies use statistical analyses of the results of intelligent systems to uncover biases. Others discuss the mathematical framework behind these algorithms and where, when transforming the real world into numbers, biases inevitably emerge. These authors and their work are valuable because AI is a complex and rapidly evolving field that requires a lot of education to understand and those who currently work with AI are the most qualified to evaluate it.

On the other hand, this problem is not just technical, but largely social as well, and the group that will feel its effects most is the minorities these biases will affect. Therefore, we must view this problem through the lens of marginalized groups, including racial minorities, gender minorities, and other minority groups, as AI is broadly used and its applications can have a variety of effects. AI learns existing biases and perpetuates inequities and, if left unchecked, this system will continue to enforce itself. Therefore, we will explore real-world examples of this bias and hear from the perspective of the minorities that have been and will continue to be affected by this.

To investigate this topic, we will use the Social Construction of Technology (SCOT) framework. This framework focuses on how society affects the development of technology and vice versa. According to SCOT, technological development is influenced by many social factors. These technological developments in turn change society. This framework is well suited to address this since AI has applications across a broad range of fields and this is a flexible approach. Additionally, AI is a recently developing issue that is clearly shaping society so it is important that we look at these complex interactions.

In order to develop a comprehensive understanding of bias in AI, we will perform a literary analysis, reviewing a range of sources from studies by experts in the field of AI to articles from social workers with relevant experience. We will explore sources that use both quantitative and anecdotal evidence across some of the many fields AI is involved in, such as natural language processing machines, AI in criminal justice reform, and predictive algorithms.

Barriers & Boons

Throughout this research, we must be cognizant that AI is a relatively new and rapidly evolving field. This is a limitation of studying this topic, as the research being done is still relatively new, and therefore we must continue to consider new information as it arises. Other potential blindspots arise that may influence the scope and perspective of the literature review. For example, as a researcher, I may be less attuned to issues of discrimination against marginalized groups I do not personally belong to. Also, my personal cultural norms and values will influence the sources I will select for this research. To offset the impacts of these limitations, I will be critical and self-reflective throughout the entire process, especially when considering sources of information. I will also try to choose a wide range of authors to represent multiple perspectives.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias — ProPublica*. ProPublica. Retrieved March 17, 2023, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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 *Proceedings of Machine Learning Research* 81:77-91 Available from https://proceedings.mlr.press/v81/buolamwini18a.html.

Burton-Harris, V., & Mayor, P. (2020, June 24). *Wrongfully Arrested Because Face Recognition Can't Tell Black People Apart*. American Civil Liberties Union. Retrieved March 16, 2023, from https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart

Caliskan, A. (2021, May 10). *Detecting and mitigating bias in natural language processing*. Brookings. Retrieved March 16, 2023, from https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/

Crawford, K. (2016, June 25). *Opinion | Artificial Intelligence's White Guy Problem*. The New York Times. Retrieved March 17, 2023, from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. Retrieved March 17, 2023, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016, September 23). *[1609.07236] On the (im)possibility of fairness*. arXiv. Retrieved March 17, 2023, from https://arxiv.org/abs/1609.07236

Garvie, C., Bedoya, A., & Frankle, J. (2016, October 18). *Unregulated Police Face Recognition in America*. Perpetual Line Up. Retrieved March 15, 2023, from https://www.perpetuallineup.org/

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017, August 26). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. 10.1093/qje/qjx032

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016, September 19). *[1609.05807] Inherent Trade-Offs in the Fair Determination of Risk Scores*. arXiv. Retrieved March 15, 2023, from https://arxiv.org/abs/1609.05807

Narayanan, A., Bryson, J. J., & Caliskan, A. (2017, April 14). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. Science. 10.1126/science.aal4230

Noble, S. (2018). A Society, Searching. In *Algorithms of Oppression: How Search Engines Reinforce Racism* (pp. 15 - 63). NYU Press. https://safiyaunoble.com/wp-content/uploads/2020/09/Algorithms_Oppression_Introduction_Intro.pdf

Silberg, J., & Manyika, J. (2019, June 6). *Tackling bias in artificial intelligence (and in humans)*. McKinsey & Company. Retrieved March 16, 2023, from https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

There's More to AI Bias Than Biased Data, NIST Report Highlights | NIST. (2022, March 16). National Institute of Standards and Technology. Retrieved March 17, 2023, from https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights