Analysis of Various Explanation Methods on Sleep Score Regression

The Failure of Microsoft's Tay and What it Means for AI Governance

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
Benjamin Orndorff

Spring 2022

Technical Project Team Members
Benjamin Orndorff

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

Kent Wayland, Department of Engineering and Society

Daniel Graham, Department of Computer Science

Introduction

The past 10 years of AI development have been marked by the exponential growth of hype and progress in AI technology. Between 2020 and 2021 alone, Total AI investments jumped from $36 billion to $77.5 billion (Mehta et al., 2021). AI has also become more and more vital to many of our industries from AI used for diagnosis in the medical industry to targeted ads generated from AI used by many brands and businesses. The applications and use of AI have become very broad and it is only getting broader while researchers and businesses pour more money into improving the accuracy of these systems.

However, in recent years, there has been growing concern over the implementation of these AI systems as time and time again controversy has risen over issues with them. For example, in 2019, An algorithm used on more than 200 million people in US hospitals to determine which patients would likely need extra care was found to heavily favor white people over black people (Vartan, 2019). The double-edged sword of the amazing capabilities AI provides is that they are oftentimes complex black boxes that do not explain the rationale behind decisions which makes it hard for the systems to be trusted in critical situations like healthcare. This growing concern over the realistic application of AI has prompted increased investments into AI safety, a field of research looking into ways to improve understanding of the decisions produced by AI, and AI governance, establishing accountability to guide the creation and deployment of AI systems in an organization. In my technical research, I will apply AI safety methods to a sleep score regression model and analyze the benefits of each while in my STS research topic I will focus on AI governance policies in the context of the Microsoft Tay controversy. As AI has become more advanced and more integrated into our society there has been an increased need for AI to be made and applied ethically to prevent avoidable catastrophes from decisions made using uninterpretable and unregulated systems.

Analysis of Various Explanation Methods on Sleep Score Regression

Which methods of explanation are best suited for medical AI and what are the benefits and disadvantages of each?

Sleep is one of the most crucial aspects of living a healthy life but despite that 35.2% of adults in the US get less than the recommended 7 hours of sleep every night (CDC, 2017). This lack of sleep in the short term affects judgment and mood while in the long term it could bring major health issues. In the interest of improving sleep quality, many applications have been developed that analyze users' sleep to determine patterns within them and assign sleep scores. However, these generated sleep scores are not easily interpretable by humans as these apps do not explain why a score was assigned or what factors were most important in determining it. In my technical research, I will first develop a sleep score regression model to predict a user's sleep score given their information and then test multiple post-hoc explanation methods on it to create an explainable sleep score for users. My goal is to 1) improve the interpretability of sleep scores so that users can use them more effectively to improve sleep quality and 2) Analyze a variety of

explanation methods on this regression model to determine the advantages and disadvantages of each.

Methods

I will begin my research by collecting sleep data to be used to train the model. The data will be collected using the Sleep Cycle app and Fitbit app and will contain information on the date, nutrition, exercise, caffeine intake, stress level, sleep time, wake-up time, weather, location, and sleep score of the user. Using the data, I will develop a regression model to predict the sleep score for a night using the collected information. I will test a random forest model, a neural net, and a logistic regression model to find the model with the highest accuracy.

I will then modify the model to include interpretability in the form of post-hoc explanations from 3 different explanation methods: LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), and Anchors (Ribeiro et al., 2018). Each of these methods is a perturbation method which means they slightly modify the input to a prediction model and use the changes produced in the output to determine the importance of each aspect of the input. In this research of sleep scores, this may involve removing certain features of the input like caffeine intake or bed time and then viewing the changes in the sleep score prediction to assign importance values to each of the importance features. By including these explanation methods into the sleep score regression models the output will contain both the predicted sleep score as well as the importance values for each of the input features for that sleep score.

Using the 3 explainable models I have developed, I will test the human interpretability of each of them by testing with humans. For each model, I will have participants go through 3 user cases with the recorded sleep notes of the user and the model output of a sleep score with the model's explanation. Additionally, there will be a case with a non-explanation model to compare the results of explanation vs. non-explanation models. The participants will be asked to select the top 3 reasons why a user received the sleep score they received as well as if the user slept well or not. I will use these results to determine which explanation method is best suited for the task.

The Failure of Microsoft's Tay and What it Means for AI Governance

How do the many proposed regulations on AI deal with situations like Microsoft's Tay?

In 2016, Microsoft released an AI chatbot, Tay, onto Twitter to learn from user interactions on the platform. Tay was supposed to model and communicate like a teenage girl, but after being targeted by users on the platform, began posting offensive tweets leading to it being shut down a mere 16 hours after its release. According to Microsoft, they performed extensive stress-testing and implemented multiple filters before release, but in the end, Tay's system was not able to handle a coordinated attack on it causing the resulting unacceptable behavior to occur (Lee, 2016). Not only is there a concern about the reliability of Tay's AI system if it was able to be hijacked so quickly but also issues related to the effects of the offensive tweets Tay posted on individuals who have read them. Only a few months after the

incident, however, Microsoft released Zo, another AI chatbot similar to Tay, without any mention of the improvements or measures taken to prevent another Tay incident (Riordan, 2016).

Background

With the increasing use of AI by large companies like Microsoft, who are incentivized by the extreme efficiencies provided by their use, there has been growing concern among AI ethics researchers about the reckless implementation of these systems. One response to the growing unregulated use of AI by large companies was the growth of the AI governance field which looks into and proposes regulations on how organizations should implement and operate AI. There has been a growing push for AI governance as seen by the projected industry growth to $1,016 million by 2026 compared to $50 million in 2020 (MarketAndMarkets, 2021). With this increase, many new AI governance organizations like The Future of Humanity Institute's AI Governance Research group have been created to research methods of ethically developing and applying AI to advise investors, developers, and regulators of AI. This has also led to at least 60 governments since 2017 developing their own set of policies for AI research and use (*OECD AI's Live Repository of over 260 AI Strategies & Policies*, n.d.). With the many potential regulations proposed by both governments and AI governance groups, how do those frameworks deal with situations like Tay's?

In my STS research, I will be performing a case study of the Microsoft Tay situation examining how it fits within AI governance frameworks proposed by governments, research organizations, and corporations. I will select policies proposed by the European Union, The Future of Humanity Institute, and KPMG for a diverse perspective on AI governance.

Methods: Data Collection

My data collection will consist of literature reviews related to both Microsoft Tay and AI governance frameworks from the three organizations I mentioned above. For literature, on Microsoft Tay, I will be looking at primary sources such as the tweets Tay tweeted and the responses to the situation by media, Microsoft, and regulators, as well as secondary sources discussing the Tay situation in various frameworks. For AI governance literature I will be examining generalized frameworks proposed by 3 organizations: the European Union, The Future of Humanity Institute, and KPMG. Each organization represents a different view on AI governance from 3 different stake holder groups: governments, research groups, and large corporations. By analyzing AI governance from the perspective of these 3 organizations I will be able to view the differences and values each place on the multiple aspects of AI use such as development, application, and testing. Furthermore, With the many different incentives and demerits to AI use for each of these stakeholders, I will examine if the AI governance frameworks each provides can be reconciled with the others in order to achieve a single proposal that fits the needs of everyone.

Looking further into each framework, each organization has outlined their framework within the past 2 years. In May 2021, the European Union released draft regulations aimed specifically at the development and use of AI (*A European Approach to Artificial Intelligence* |

Shaping Europe's Digital Future, 2022). Since the draft was comprehensive and highly publicized, along with the original document I will also be using secondary sources that provide commentary and analysis of the regulations either from an AI governance perspective or governmental perspective. The Future of Humanity Institute has multiple published research papers on AI governance topics that analyze situations of AI use through the lens of AI governance making possible proposals with it (FHI). KPMG has also released a document, "The Shape of AI Governance to Come" (*The Shape of AI Governance to Come - KPMG Global*, 2021) with their laid-out framework along with the role of businesses within that framework.

Methods: Analysis

Using the understanding of the Microsoft Tay situation and AI governance frameworks I will examine how those organizations' AI governance frameworks would deal with the Microsoft Tay situation. Each framework should outline relevant actors, the responsibilities of each, and methods of prevention for the situation. By analyzing frameworks from policy makers, researchers, and corporations I will compare their views on the roles each organization should play when it comes to AI. Finally, I will use my analysis of the Tay situation to propose a generalized response to similar situations like Tay (i.e., large companies allowing AI to interact with people without thorough testing) from my own analysis of the literature.

Through this case study of Microsoft Tay and these organizations' frameworks I will have an understanding of the current state of the AI governance field and flaws within the frameworks I analyze. I also hope to show potential solutions to these flaws and propose solutions to similar issues like Tay's that policy makers will be able to utilize.

Overall Conclusion

The growing AI industry has been focused mainly on improving the accuracy and applications of AI but recently there has been growing concern over the ability to interpret the decisions AI makes. This concern has caused an increase in research on AI safety and AI governance to combat the potentially harmful effects of the application of AI. In my research, I have evaluated methods of AI safety in my analysis of a sleep score regression model and AI governance frameworks as I applied them to the Microsoft Tay controversy in hopes of understanding the current state of AI safety and AI governance research.

References

A European approach to Artificial intelligence | Shaping Europe's digital future. (2022, February 23). Digital-Strategy.ec.europa.eu. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

CDC. (2017). *CDC - Data and Statistics - Sleep and Sleep Disorders*. Centers for Disease Control and Prevention. https://www.cdc.gov/sleep/data_statistics.html

FHI, F. of H. I. -. (n.d.). *Future of Humanity Institute*. The Future of Humanity Institute. https://www.fhi.ox.ac.uk/ai-governance/#publications

Lee, P. (2016, March 25). *Learning from Tay's introduction - The Official Microsoft Blog*. The Official Microsoft Blog. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

MarketAndMarkets. (2021, March). *AI Governance Market Size, Share and Global Market Forecast to 2026 | MarketsandMarkets*. Www.marketsandmarkets.com. https://www.marketsandmarkets.com/Market-Reports/ai-governance-market-176187291.html

Mehta, B., Mousavizadeh, A., & Darrah, K. (2021, December 2). *AI Boom Time*. Tortoise. https://www.tortoisemedia.com/2021/12/02/ai-boom-time/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://ojs.aaai.org/index.php/AAAI/article/view/11491

Riordan, A. (2016, December 13). *Microsoft's AI vision, rooted in research, conversations*. Stories. https://news.microsoft.com/features/microsofts-ai-vision-rooted-in-research-conversations/

Starre Vartan. (2019, October 24). *Racial Bias Found in a Major Health Care Risk Algorithm*. Scientific American. https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

The shape of AI governance to come KPMG International home.kpmg/ShapeofAIGovernance. (2021). https://assets.kpmg/content/dam/kpmg/xx/pdf/2021/01/the-shape-of-ai-governance-to-come.pdf