Towards Semantic Search in Building Metadata
(Technical Paper)

The Impact of Human Cognitive Biases on Rule-Based Machine Learning Models in Enabling Racial and Gender Biases
(STS 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
Aishwarya Gavili

December 7, 2021

Technical Team Members:
Aishwarya Gavili

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

Hannah Rogers, Department of Engineering and Society

Hongning Wang, Department of Computer Science

Introduction

"Garbage in, garbage out" (GIGO) is an expression that has been coined in computing to describe the phenomenon of poor quality inputs always producing poor quality outputs (McMahon, 2021). Despite its peculiarity, this expression's underlying concept is prevalent in machine learning biases where flawed data used to train machine learning models is oftentimes the reason behind flawed results. It should be noted that these biases entail systematic prejudice towards certain groups of people, including women and racial minorities. More recently, this issue has escalated to the point where large corporations such as Amazon have had to scrap existing machine learning algorithms in their recruitment tools and other technologies due to the lack of gender-neutrality they contained (Dastin, J., 2018). Nevertheless, it is important to understand that ultimately, a human controls the parameters of a model, the subsets of data chosen, how data is labeled, and how a model is interpreted. Therefore, a model and its outcomes are heavily exposed to human cognitive biases. This is the inspiration behind the STS portion of this prospectus exploring the influence of human cognitive biases in allowing machine learning models to enable racial and gender inequity.

The technical portion of this prospectus revolves around semantic search in building metadata that is used for a search engine API and front-end web application called UVAMonitor. Modern search engines such as Google encounter ambiguous web searches daily, which makes it imperative to understand the context of a query. Semantic search enables this by trying to understand the searcher's intent and query context to generate the most accurate search results (Pecánek, 2020). Consequently, the technical research for this project further expands on this and aims to provide a better understanding of the use of databases in a production environment and its intersection with NLP queries in the creation of effective search engines.

STS Topic

With the widespread digitization of data, machine learning has dominated different industries including healthcare, advertising, retail, government institutions, and more. In fact, in 2020 it was stated that the global machine learning market is projected to grow from $7.3B-$30.6B by 2024 (Columbus, 2020), highlighting the rapid dominance of the market. More specifically, the need to automate and perform classification tasks within machine learning has led to an increase of the use of facial recognition systems, recommender systems, and text analysis technologies over the past decade. This is especially evident with widespread use of face ID authentication in smart devices, Netflix's $1 billion recommendation engine (McAlone, 2016), and Google's search engine; all of which account for close to 300 million users worldwide (Georgiev, 2021; Haslam, 2021). However, being ruled-based at their core, these machine learning models and applications are greatly exposed to and influenced by human cognitive biases before and after deployment through algorithmic bias and interpretive bias.

Furthermore, these algorithmic and interpretive biases have tendencies to enable racial and gender biases in the systems they are incorporated in. The implicit and explicit motives behind doing so boil down to the entity utilizing the model. These motives range from higher model performance to monetary or political incentives, which will be further explored in the analysis below. Consequently, if not addressed and fixed, the risk that bias will be ingrained in countries' infrastructures will be high.

For this paper, it is first important to define the scope of rule based machine learning models. Rule based machine learning revolves around models and their outputs being made up of rules in the form of an if/then statement. Most commonly, learning classifier systems or LCS algorithms and decision tree models employ if/then rules to link independent variable states to dependent variable statements (ex. a class or action) (Urbanowicz & Browne, 2015). And these rules are also the foundation of facial recognition systems, text analysis, and recommender systems.

Additionally, it is important to understand the different connotations of bias employed in this paper. Bias can be defined as prejudice in favor of or against one thing (HarperCollins Publishers, 2021). Similarly, cognitive bias can be defined as ‘systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments.” (Kliegr et al., 2021) Interpretive biases are essentially cognitive biases that are employed in interpreting scenarios or events. Algorithmic biases, on the other hand, are systematic errors in a computer program that create unfair outcomes in its output (Wikipedia contributors, 2021).

Prior to the deployment of a model, algorithmic bias can be incorporated by using incomplete and under representative training data. This is especially evident in recruitment tools where job search algorithms and resume filtering algorithms learn word associations present in training data, which are highly biased when it comes to feminist linguistics (Leavy et al., 2020). More specifically, phrases like “female lawyer” or “female engineer” are seen as anomalies to societal norms and appear less in structured and unstructured data, decreasing their chance of being learned by algorithms (Lee, 2019). Similarly, the lack of gender and racial diversity in facial recognition training data sets has resulted in commercially available facial recognition technologies to cater poorly to darker-skinned complexions. In fact, most facial recognition training datasets are estimated to be more than 75% male and more than 80% white (Lee, 2019). As a result, the rate of error in identifying women, more so women of color, is extremely high (2020).

Following the deployment of a model, the interpretation of its results greatly influences how it is tuned and utilized for future use. The interpretability of a model can be measured by the likeliness the user accepts the model as an explanation for a prediction. With regards to model acceptance, humans tend to be very insensitive to model results by overestimating the strength of evidence but underestimating the weight of evidence (Kliegr et al., 2021). In other words, interpretive biases are weaved into the evaluation of models. For instance, in a widely used medicare prediction algorithm, black patients are assigned the same risk score as white patients even if they are more sick because of how health costs are used as a proxy for health needs (Obermeyer et al., 2019). More specifically, the algorithm falsely concludes that black patients are healthier than equally sick white patients due to lower healthcare expenditures of black patients in comparison to white patients; which is simply attributed to their historically lower incomes and inability to afford healthcare. The issue of interpretability lies in the fact that using health costs as the main predictive feature yielded high predictive accuracies, but failed to account for the fact that those predictive accuracies were the wrong evaluation metrics in the context of the situation (Obermeyer et al., 2019). Similarly, in Amazon's same day Prime delivery system, Amazon used features such as warehouse proximity and prime member quantity to train its models as they boosted its profits (Lee, 2019). In this case, profit was considered a viable evaluation metric due to the strength of evidence it provided with regards to the amount of revenue generated. However, the weight of evidence was not considered with how using such features excluded lower income neighborhoods populated with minorities.

Ultimately, the purpose of this STS research is to answer how human cognitive biases affect rule-based machine learning models with regards to enabling racial and gender biases. And it will be done by invoking the SCOT STS framework, which reinforces the idea that technology doesn't determine human action and rather human action shapes technology (Pinch, T. J., & Bijker, W. E., 1984). Moreover, from the examples above, it is clear that the constraints of a model along with the mined data fed into a model are set by humans. Similarly, the tuning of models and the interpretation of their results boil down to humans as well. Evidently, technology design can have varying results depending on the social contexts of when the technology was developed, which is a pillar of the SCOT framework (Pinch, T. J., & Bijker, W. E., 1984). This pillar of SCOT will further drive the analysis of which social groups or stakeholders contribute the most to the final say of decisions that dictate a model's parameters and evaluation metrics. And understanding the social context of when a machine learning model was developed will facilitate the identification of its "social peak" for which it is deemed acceptable for deployment. Additionally, by doing so, the implications of algorithmic formalism and algorithmic realism in relation to racial/gender biases will further be explored (Green & Viljoen, 2020). Consequently, the research conducted will delve more into the intricacies of how different kinds of cognitive biases influence model decisions, how the socio-economic status of stakeholders affects how a model is constructed and evaluated, and how to debias models when it comes to enabling racial and gender inclinations.

Technical Topic

Semantic search is an information retrieval process that focuses on the meaning behind queries rather than keyword matching when retrieving search results (Pecánek, 2020). Consequently, it is a culmination of natural language processing (NLP), user context, query stream context, and entity recognition. In the context of UVAMonitor, the system that utilizes semantic search for building metadata, semantic search allows users to search queries related to specific sensors without having the queries be equivalent to pre-determined keywords. An example of this would be the search engine yielding the same results for the following queries: *"temp > 20"* and *"temperature > 20"*.

Consequently, the technical research and project will explore the effects of the system undergoing the process of refactoring event handling and enhancing query parsing for the search bar, which fall under the categories of entity recognition and NLP in semantic search. This could possibly involve adjusting the search bar's functionalities and cleaning up the back-end query pipeline. Similarly, the project will delve into outlining the relationship between natural language processing and SQL queries. Particularly, the goal is for the NLP to SQL translation to handle more than one keyword and operator to allow joins between tables. By doing so, the search bar can ingest more complex queries.

The implications of the technical project are vital because they are a step towards computers being able to manipulate information on human behalf. More specifically, if the enhanced query parsing and optimized NLP to SQL query translation is extended to real world search engines, eventually, the need for humans to label data streams will be eradicated, leaving computers to manipulate and learn that information on their own.

The project will be completed in an individual setting with close supervision by the technical advisor and the alumni who worked on the system prior to graduating. This will entail frequent communication through Zoom meetings and Slack messaging. More specifically, an understanding of existing implementations of the system is important to allow for the additional optimizations the technical project will add on.

Conclusion

The STS research conducted will provide greater insight on how algorithmic and interpretive biases, which are influenced by human cognitive biases, impact rule-based machine learning models' racial and gender inclinations. Accordingly, the analysis will be done using the SCOT STS framework to further understand how human action determines technology use within machine learning and how different stakeholders impact the decision process of creating and interpreting machine learning models. The capstone technical project will explore the relationship between natural language processing and SQL queries in semantic search. Consequently, it will provide greater insight into the extent to which computers can manipulate and learn human queries, optimizing search engines as a whole. Through both analyses, the goal is to understand the interaction between humans and the data they create, mine, study, and utilize in modern technologies.

References

Barysevich, A. (2021, August 3). *Semantic Search: What It Is & Why It Matters for SEO Today*. Search Engine Journal. https://www.searchenginejournal.com/semantic-search-seo/264037/

Columbus, L. (2020, January 23). *Roundup Of Machine Learning Forecasts And Market Estimates, 2020*. Forbes. https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/?sh=3df976e5c020

Dastin, J. (2018, October 11). *Amazon scraps secret AI recruiting tool that showed bias against women*. U.S. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Feast, J. F. (2020, October 8). *4 Ways to Address Gender Bias in AI*. Harvard Business Review. https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2019). On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, *109*(4), 853–898. https://doi.org/10.1007/s10994-019-05856-5

Georgiev, D. (2021, November 19). *111+ Revealing Google Statistics and Facts to Know in 2020*. Review42. https://review42.com/resources/google-statistics-and-facts/

Green, B., & Viljoen, S. (2020). Algorithmic realism. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Published. https://doi.org/10.1145/3351095.3372840

HarperCollins Publishers Ltd. (2021, December 6). *Bias definition and meaning | Collins English Dictionary*. Collins Dictionaries. https://www.collinsdictionary.com/us/dictionary/english/bias

Haslam, A. H. (2021). *Netflix Review and Prices*. U.S. News & World Report. https://www.usnews.com/360-reviews/technology/streaming-services/netflix

Kliegr, T., Bahník, T., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*, 103458. https://doi.org/10.1016/j.artint.2021.103458

Leavy, S. (2020, May 14). *Mitigating Gender Bias in Machine Learning Data Sets*. ArXiv.Org. https://arxiv.org/abs/2005.06898

Lee, N. P. T. R. (2019, October 25). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Brookings. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

McAlone, N. M. (2016, June 15). *Why Netflix thinks its personalized recommendation engine is worth $1 billion per year*. Business Insider. https://www.businessinsider.com/netflix-recommendation-engine-worth-1-billion-per-year-2016-6?international=true&r=US&IR=T

McMahon, M. M. (2021, February 25). *What is Garbage in Garbage out?* EasyTechJunkie. https://www.easytechjunkie.com/what-is-garbage-in-garbage-out.htm

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pecánek, M. (2020, August 12). *What Is Semantic Search? How It Impacts SEO*. SEO Blog by Ahrefs. https://ahrefs.com/blog/semantic-search/#why-is-semantic-search-important.

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

S. (2020, October 26). *Racial Discrimination in Face Recognition Technology*. Science in the News. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Unconscious Bias | diversity.ucsf.edu. (2021). Unconscious Bias. https://diversity.ucsf.edu/resources/unconscious-bias

Urbanowicz, R., & Browne, W. (2015). Introducing Rule-based Machine Learning. *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*. Published. https://doi.org/10.1145/2739482.2756590

Wikipedia contributors. (2021, November 14). *Algorithmic bias*. Wikipedia. https://en.wikipedia.org/wiki/Algorithmic_bias