HARRISBURG
UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Project Proposal

AN INVESTIGATION INTO THE EFFECTS AND ETHICAL IMPLICATIONS OF BIASES ON MACHINE LEARNING ALGORITHMS

Matthew R Sabo

12/13/2023

ABSTRACT

It is an indisputable fact that in the past decade, the use of Artificial Intelligence has seen a rapid increase in personal and professional settings. As Artificial Intelligence becomes more prevalent in society, more concerns are raised about their ethical use. One of the most pressing ethical issues in the field of Artificial Intelligence concerns the effects of bias and ways to minimize the amount of bias in Artificial Intelligence. This paper is intended to be a compilation of research done on the current state of ethics in the field of artificial intelligence, the effects that artificial intelligence has on humans, and methods that have been developed to mitigate bias in machine learning algorithms. The paper will also define the methodology that will be followed for testing the bias-mitigation methods in Project II.

ABOUT THE AUTHOR

Matthew Sabo is a third year Computer and Information Science Major at Harrisburg University of Science and Technology. He has experience programming in Python, Java, JavaScript, and C. He is currently striving to get into Harrisburg University of Science and Technology's 5-year master program for Computer and Information Science.

KEYWORDS

- Bias

  Bias has many different definitions depending on the field of discussion. It is defined technically as a preference towards a certain attribute which is more geared towards things like algorithmic bias. It is defined in social fields as prejudice in favor or against one thing (Szczekocka et al., 2022).

- Machine Learning Algorithm

  A Branch of Artificial Intelligence that focuses on the development of AI Models using data and algorithms to imitate the way that humans learn.

- Artificial Intelligence

  Computer Systems that can perform tasks that require human intelligence. Some of these tasks consists of recognizing speech, making decisions, and identifying patterns

- AI Model

  Refers to the program that is the basis of an Artificial Intelligence System.

1 Introduction

Bias is a very important issue to be aware of when developing an AI Model. There have been methods developed to try to develop ways to minimize the amount of bias that exists in the models that they use to ensure fairness in their models. However, since there are numerous types of bias this makes it difficult to have one solution for multiple types of bias. This report aims to compile research on the current state of ethics in the field of artificial intelligence, the effect that AI has on humans, and the current methods that have been developed to mitigate bias in machine learning algorithms.

1.1 Background

When attempting to create a solution for mitigating bias in machine learning algorithms, it is important to understand the types and categories that biases fall into. Knowledge of these categories allows for better understanding and implementation of the methods and procedures that are used to reduce biases in machine learning algorithms. There are four categories that biases fall into, the first being Human Cognitive Bias. This category is defined by human bias that might impact the design and application of a system. Examples of human cognitive bias are automation bias which is the bias towards recommendations made by an AI system, confirmation bias which is humans favoring predictions of AI systems that confirm existing beliefs or hypotheses, and societal bias which is bias that originates from society at large (Piecerak et al, 2022). The second category of bias is data bias which is bias in the datasets that AI models are trained from. Examples of data bias are sampling bias which occurs when data is not collected randomly from a group and selection bias which is when a datasets samples are chosen that are not reflective of the real-world distribution (Piecerak et al, 2022). The third category of bias is machine learning model architecture bias which contains model interaction bias and informativeness bias. The fourth category of bias consists of all the other biases that do not fall within the previously stated categories (Piecerak et al, 2022). By knowing the categories of biases that occur in machine learning algorithms, it allows for a better understanding of the ethical implications of bias and methods that have been developed to mitigate bias.

2 Project Topic Statement and Justification

This section will go over the project statement, the justification for this project, the goals of the project, and a list of the project's deliverables.

2.1 Project Statement

This project will be an investigation into how biases, conscious and unconscious, affect the development of artificial intelligence and the ethical implications that may arise from the previously mentioned biases. The project will also test the effectiveness of the current methods that have been developed to mitigate bias in machine learning algorithms.

2.2 Justification

In 2014, Amazon created an AI hiring model that used machine learning algorithms to review job applicants' resumes. This AI model would allow Amazon to automate the resume review process and be able to hire employees at a faster pace. However, this project was discontinued in 2018 due to the model having a significant bias toward women. This bias would cause the algorithm to penalize resumes that included gendered terms like "woman" and penalized resumes of applicants that were graduates of all-women's colleges (Dastin, 2018). With the increased use of AI in professional and personal environments it is imperative to understand the effects of bias on AI and machine learning algorithms since it is possible for those biases to have detrimental impacts on people's lives.

2.3 Deliverables

The following deliverables will be presented at the completion of this project.

- Project I Poster

  An academic poster that presents the work done on this project and the projected work to be completed for Project II.

- Project I Final Presentation

  A formal presentation given to peers that presents the work that has been completed on this project and discusses the future steps that will be completed in Project II.

- Project I Final Paper

  A paper that compiles the current research done on the effects of bias in machine learning algorithms and methods that have been developed to reduce bias in machine learning algorithms. It also defines the next steps that will be taken to complete the research that will be completed in Project II.

3 Literature Review

In preparation for this project, research was done on the current state of ethics in the field of artificial intelligence, the effects that AI have on humans, and previously developed methods to mitigate certain types of bias.

3.1 Current State of Ethics

A paper in the 2022 IEEE International Conference on Big Data entitled *Standardization on Bias in Artificial Intelligence as Industry Support* defined "Responsible AI" as a list of principles that an AI system should be to be considered trustworthy (Piecerak et al, 2022). However, the principles are difficult to define, with different companies and organizations having different principles that should be followed to ensure that their artificial intelligence is "responsible". For example, research that has been done on this subject proposes that the principles of "Responsible AI" are as follows: ethics, transparency, regulation and control, socioeconomic impact, design, and responsibility. IBM considers the principles as: explainability, fairness, robustness, transparency, and privacy, and Microsoft's principles of Responsible AI are defined as: fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. Even though the principles of each organization differ slightly, the consistent principles surround ethics and the principle of fairness in an AI system (Piercerak et al, 2022).

The paper *Building Ethics into Artificial Intelligence* talks about how the industry can and should standardize the inclusion of ethics into AI models. It also explains that one of the starting points of ethics in artificial intelligence is the Belmont Report. The paper explains that in the Belmont Report, there should be principles that are followed to ensure that ethics are sustained in Human-

AI interactions. The principles include three main requirements. The first principle being: a person’s personal autonomy should not be violated or in other words, a person should be able to maintain their free will when interacting with an AI. The second principle outlined is that any benefits that are brought about by the AI should outweigh the potential risks of the AIs use. The third and final principle outlined in the Belmont Report is that the benefits and risks of the AI should be distributed fairly among its users, or in other words, the AI should not discriminate against someone’s personal background (Lesser et al, 2018). The paper then pushes the idea that ethics should be taught as the AI curricula to ensure that ethics are built into every artificial intelligence that is developed.

3.2 Human-AI Interactions

In the report *Exploring the Relationships between Artificial Intelligence Transparency, Sources of Bias, and Types of Rationality*, the authors research the impact that AI systems have on the human decision-making process. The participants in this study were split into two groups. One group would be in an environment that would be led by humans and the other would be led by an SML-based AI where the participants would be asked “simple, no accountability, low risk decision making situations”(Mäkinen & Valtonen, 2022). When interviewing the participants after the testing, the authors concluded that for the participants in the AI group there seems to be an over-reliance on the information and recommendations made by the AI for one participant without an information technology background. Meanwhile, for one participant that had a negative opinion on artificial intelligence, there seemed to be authoritative stigmatism when given information and recommendations by the AI. The authors concluded that AI is seen as authoritative in high and low stake decision making environments and that the presence of an AI system does not inherently suppress choice but can suppress the rationality creation step in the decision-making process (Mäkinen & Valtonen, 2022).

3.3 Methods Developed to Mitigate Bias

During research, multiple methods to mitigate bias in machine algorithms were found.

3.3.1 Gender Masking

An article in the 2023 Multimedia Tools and Applications journal titled *Evaluating and mitigating gender bias in machine learning based resume filtering* provides a method to reduce gender bias in AI models called Gender Masking. This method works by extracting or masking specific gendered terms like he/she, man/woman, and boy/girl from a resume before finding the similarity in the job requirements(Kaur et al, 2023). The authors found that by implementing Gender Masking to preprocessing datasets substantially reduced the amount of gender bias in the models and significantly increased the model's ability to recognize bias in (Kuar et al, 2023). Now with the understanding of how this research was conducted and the results from the research, it is possible to test the effectiveness of the method during Project II.

3.3.2 Reject Inference

Reject Inference is the concept of including the rejection data into the dataset to avoid selection bias (Annas, Benyacoub, & Ouzineb, 2022). Selection bias is commonly found in credit scoring models because many of these models are trained on datasets that do not include rejection data. Because of this, it can unintentionally cause discrimination on loan or credit card applications (Annas, Benyacoub, & Ouzineb, 2022). In the paper from the Journal of the Operational

Research Society entitled *A Monte Carlo simulation framework for reject inference*, the authors study three main reject inference techniques that are widely used. The first technique is called simple augmentation, which builds a model on accepted applicants and applies this model to rejected applicants (Anderson et al, 2022). The next technique is parceling which is a hybrid method that combines proportional assignment and augmentation. Proportional assignment is the random portioning of rejects into good and bad classifications based on the number of good and bad applicants. The final reject inference technique is fuzzy augmentation which separates the rejected applicants into a partial good and bad which is generated in the form of a weight variable (Anderson et al, 2022). The methodology that is used in this paper is not applicable to this project but the research on the methods used for Reject Inference will be used to test this method.

3.3.3 Feature Attribution Explanation (fae)

The final method researched to mitigate bias is called Feature Attribution Explanation (fae). This method is defined by its ability to detect whether a data point contributes too much to the output of a model. The paper *Feature Attribution Explanation to Detect Harmful Dataset Shift* discusses the three categories that a dataset can shift and how to detect a harmful shift. These three categories are: Covariate Shifts which is defined by the changes and differences in the distribution of input variables between training and test data, Label Shifts which is defined by the changes and differences in the distribution of target variables between training and test data, and Concept Shifts which is defined by the changes in the relationship between the input variables and class variables(Huang, Wang, & Yao, 2023). The article tests two different FAE methods, Gradient Saliency and Grad*Input, which are both found to be effective in detecting a harmful dataset shift. This method will be used in the research that will be completed in Project II.

4 Research Design and Methodology

The following section gives an overview of the design and methods for the proposed research that will be done during Project II.

4.1 Overview

The research that will be done in Project II will consist of training multiple AI models to test the effectiveness of the methods previously mentioned in the Literature Review. This will also be used to investigate how bias affects the output of an AI model.

4.2 Environment

The AI models will be developed using the Python programming language. The models will be developed using open-source libraries from TensorFlow which will provide functionalities like model tracking, performance monitoring, and model retraining. The hardware that the AI models will be trained on is the Harrisburg University of Science and Technology's HPC Computer.

4.3 Procurement of Training Datasets

The methods of obtaining the datasets needed for the models will differ depending on the bias reduction method being tested.

4.3.1 Gender Masking Dataset

For testing the Gender Masking method, the dataset that was used in the paper *Evaluating and Mitigating Bias in Machine Learning Algorithms* will be utilized for this project. This dataset is a publicly available dataset that consists of a set of resumes. This data will be used to evaluate the effectiveness of gender masking as a bias reduction method. The dataset was taken from https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset.

4.3.2 Reject Inference Dataset

When researching the procurement of the training dataset for the researched methods, the original plan for reject inference was to use the dataset that was used in the paper *A Monte Carlo Simulation Framework for Reject Inference*. However, the dataset that was used is currently unavailable to the public. The current plan is to research publicly available datasets that would work for this project. The author has contacted the authors of the paper and is waiting for a response to see if the simulated dataset that was developed would be able to be used.

4.3.3 Feature Attribution Dataset

For testing the Feature Attribution method, the plan is to utilize one of the datasets that were used in the paper *Feature Attribution Explanation to Detect Harmful Dataset Shift*. The dataset that was chosen was MNIST since it is publicly available. The dataset is taken from https://www.kaggle.com/datasets/hojjatk/mnist-dataset.

5 Conclusion

To conclude, this paper first introduced the topic being discussed and the justification behind this topic. After listing the deliverables for this project, this paper then compiled the current research that has been done on the current state of ethics in the field of artificial intelligence, which discussed concepts like the "Responsible AI" (Piecerak et al, 2022) and discussed the industry standardizing the inclusion of ethics in AI models by referencing the principles outlined in the Belmont Report (Lesser et al, 2018). Then this paper discussed some ethical concerns that may arise from the use of AI especially in AI lead environments where it was found that AI can suppress the rational creation step of the decision-making process (Mäkinen & Valtonen, 2022). Which is concerning since with the popularity of applications like Chat-GPT, it can raise ethical concerns around the use of AI. The paper then discussed some bias-reduction methods that have been developed to mitigate the effects and presence of bias in machine learning algorithms.

After compiling the research this paper then discussed the research that will be conducted during Project II. The methodology gave an overview of the research and the goal of the research. The methodology explained how the author will procure the datasets needed to train the models which for the gender masking and feature attribution explanation methods use the same publicly available datasets that were used in the papers that those methods came from. However, for Reject Inference, the dataset that was used in the original paper is no longer available, so the author is currently waiting for a response from one of the authors of the academic article *A Monte Carlo Simulation Framework for Reject Inference*.

It is an indisputable fact that in the past decade, the use of Artificial Intelligence has seen a rapid increase in personal and professional settings. With this increased use, it is imperative that the effects of bias on machine learning algorithms are known and the ethical implications that come from that bias. This paper addresses these concerns by compiling the current research on this topic and defines a methodology to test current bias-reduction methods.

6 References

Anderson B., Grim II P. A., Hardin J. M., & Newman M. A. (2023). A Monte Carlo simulation framework for reject inference. *Journal of the Operational Research Society*, *74*(4), 1133–1149. https://doi.org/10.1080/01605682.2022.2057819

Benyacoub B., El Annas M., & Ouzineb M. (2023). Semi-supervised adapted HMMs for P2P credit scoring systems with reject inference. *Computational Statistics*, *38*(1), 149–169. https://doi.org/10.1007/s00180-022-01220-9

Dastin, J. (2018, October 10). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://reuters.com/article/idUSKCN1MK0AG/.

Gagandeep, Kaur J., Kaur S., Mathur S., Mathur S., Nayyar A., & Singh S. P. (2023). Evaluating and mitigating gender bias in machine learning based resume filtering. *Multimedia Tools and Applications: An International Journal*, 1–21. https://doi.org/10.1007/s11042-023-16552-x

Huang C., Wang Z., & Yao X. (2023). *Feature Attribution Explanation to Detect Harmful Dataset Shift*. https://doi.org/10.1109/ijcnn54540.2023.10191221

Huang C., Mao B., Yao X., & Zhang Z. (2023). An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819. https://doi.org/10.1109/tai.2022.3194503

Lesser V. R., Leung C., Miao C., Shen Z., Yang Q, & Yu H. (2018). Building Ethics into Artificial Intelligence. *ArXiv*. /abs/1812.02953 https://doi.org/10.48550/arXiv.1812.02953

S. J. Mäkinen and L. Valtonen. (2022). "Exploring the Relationships between Artificial Intelligence Transparency, Sources of Bias, and Types of Rationality," 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Kuala Lumpur, Malaysia, 2022, pp. 1296-1300, doi: 10.1109/IEEM55944.2022.9989994.

Pieczerak J., Szczekocka E., & Tarnec, C. (2022). Standardization on Bias in Artificial Intelligence as Industry Support. *2022 IEEE International Conference on Big Data (Big Data), Big Data (Big Data), 2022 IEEE International Conference On*, 5090–5099. https://doi.org/10.1109/BigData55660.2022.10020735