Yuqi Jin

EC601 Product Design

Nov 19. 2023

# Societal Impact Paper

In our increasingly data-driven world, the proliferation of data-based systems powered by artificial intelligence and machine learning has significantly transformed numerous aspects of our lives. From decision-making in healthcare to recommendation systems, these systems heavily rely on vast amounts of data to make predictions, automate processes, and provide insights. However, data bias is an often critical issue permeating these systems. The systematic error in data results in inaccuracies or unfairness in outcomes and poses significant challenges to the integrity, ethics, and fairness of data-based systems. This bias can originate from a number of places, such as incomplete or unbalanced datasets, algorithms processing the data, and historical preconceptions ingrained in data collection techniques.

In the research article, A Survey on Bias and Fairness in Machine Learning, Mehrabi et al. (2022) mention that the algorithms are the main driving force trained based on the database containing the bias. The algorithms can learn from these data and train in the predictions. They also found that the algorithms can be more biased depending on the design choices. Then, they show an example of a web search engine about how biased algorithms affect real-world users' decisions. The idea shown in Figure 1, according to the popularity and user interests, the information with bias would be at the top of search results.

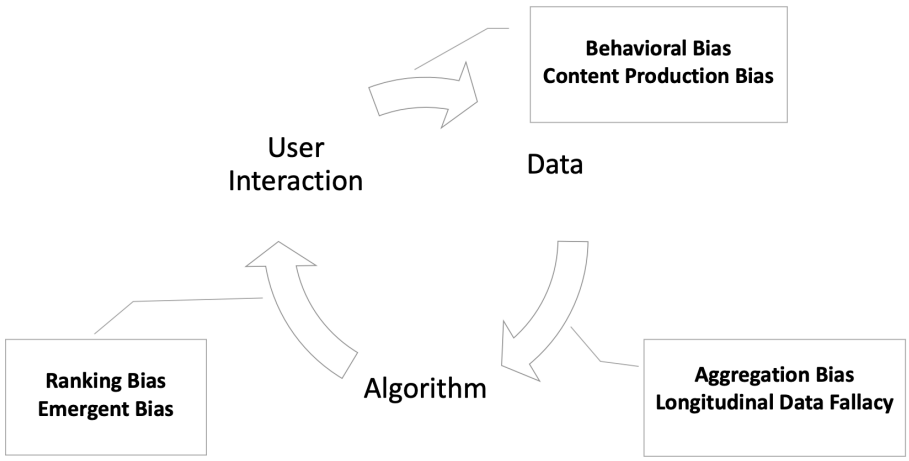


*Figure 1: Examples of bias definitions placed in the data, algorithm, and user interaction feedback loop*

In the research article, AI Fairness for People with Disabilities: Point of View. Trewin (n.d.) points out that fairness for people with disabilities is unique, with fairness with other protected attributes. They emphasize that "For systems that will make or influence decisions affecting human lives, a broad range of user stakeholders must be involved in development, including people with disabilities who can help developers to think through the possible implications of the technology."

Mehrabi et al. (2022) discuss the destructive consequences of some examples of data bias in medical applications. "In medical domains, there are many instances in which the data studied and used are skewed toward certain populations—which can have dangerous consequences for the underrepresented communities." They mention that many studies do not focus on covering all races to present the research. It emphasizes that the biased database and healthcare algorithms may not equally impact all patients. These studies likely provide additional insights into how biases in data collection and algorithmic decision-making influence healthcare outcomes, potentially disadvantaging certain groups.

The article Mitigating bias in machine learning for medicine extends the topics of how to mitigate the database bias in medicine. Vokinger et al. (2021) introduce how to mitigate the bias shown in Figure 2. During the data collection and preparation stage, bias can enter if the training data used to develop the ML system does not represent the population it is intended to serve. For instance, a skin disease recognition system may perform poorly when reviewing photographs of patients with darker skin tones if it was primarily trained on images of people with white skin. To mitigate this bias, transparency on the demographic representation in the training data is essential for reducing this bias. In order to prevent biased results, it is imperative to compile big, diverse datasets representing a range of patient populations. Under the model development, biases in the data may be reinforced. When taught in well-represented populations, machine learning models may outperform those trained in underrepresented ones. To address this, methods like adversarial de-biasing and oversampling, or de-biasing algorithms, compel models to consider underrepresented groups. However, these methods are still being developed, and additional studies are required to prove their reliability. For the model evaluation, it is significant for the model to generalize across various patient groups and must be assessed before authorization. Biases and mistakes can be exposed by evaluating the model's performance across subgroups—interpretability and explainability techniques aid in comprehending how the model makes its predictions.

Nevertheless, the precision or significance of the explanations obtained from these models may have limits. The last step is deployment (post-authorization), when the patient population in a clinical setting is unique from the training data. This phenomenon, known as "domain shift," might lead to bias during deployment. The ML system must be monitored after authorization to find biases resulting from variations in risk levels, patient subgroup performance

disparities, or demographic variances. Additionally, feedback loops, where outcomes influence practice and create new biases, must be identified and addressed.

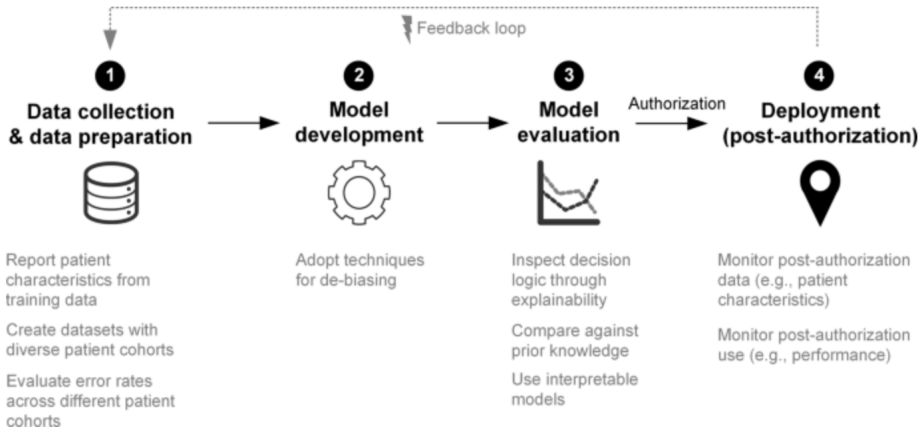


Diagram outlining proposed solutions on how to mitigate bias across the different development steps of ML-based systems for medical applications: (1) Data collection and data preparation, (2) Model development, (3) Model evaluation, and (4) Deployment.

*Figure 2: Strategies for mitigating bias across the different steps in machine learning systems development*

In the article Mitigating AI/ML Bias in Context, Vassilev et al. (2022) show the technical method of machine learning to produce more consistent, traceable, and repeatable decisions compared to humans. Also, they show how to make the outcomes less discriminatory. They present four scenarios, which are shown in Figure 3 and Figure 4.

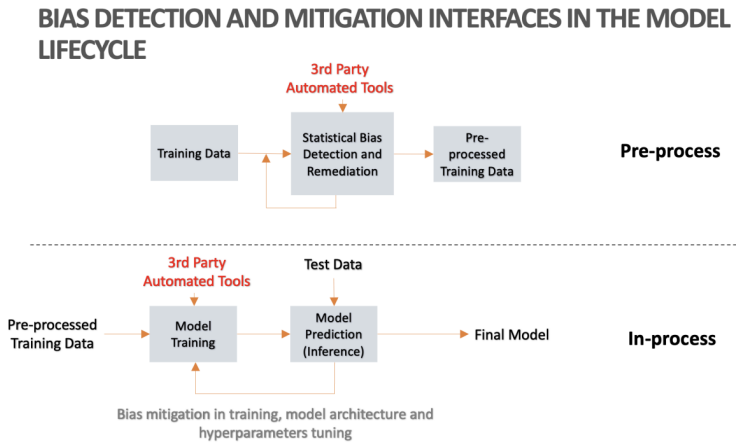


*Figure 3: pre-process and in-process Workflows for Scenarios 1 and 2*

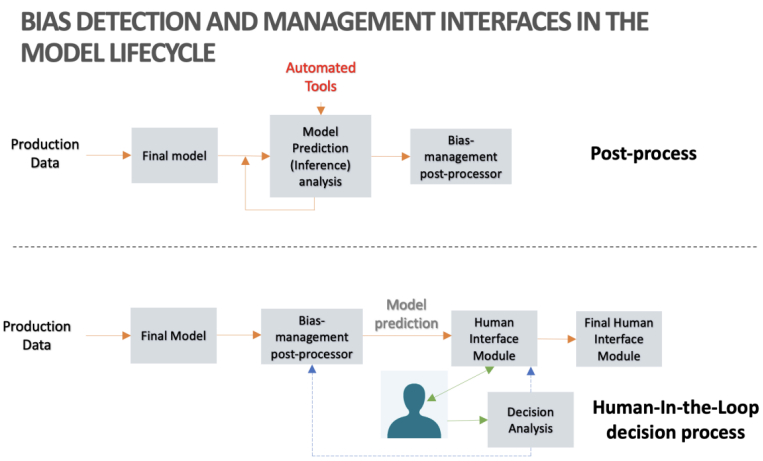


*Figure 4: Post-Process and HITL Decision Process Workflows*

Scenario 1 is called pre-process dataset analysis for detecting and managing bias. It has the functionality to identify the techniques according to the third-party tools to accomplish the mitigation. Scenario 2 is focused on the in-process. This scenario delineates strategies utilizing third-party automated tools and suggests best practices for making algorithmic adjustments during model training to mitigate bias. During the model training phase, potential alterations may involve modifications to the objective (cost) function or the imposition of new optimization constraints. Scenario 3 is primarily on post-process; it makes the learned model become an opaque system. During the post-process phase, the predictions would be called by a function. The last scenario is human-in-the-loop decision flow for identifying and managing cognitive bias. The model is produced from the previous three scenarios, which have undergone computational bias management and then put forth for a decision-making task within the credit underwriting domain. The primary aim is to scrutinize how a human engages with the AI system's output, specifically identifying any further biases that might emerge due to this interaction.

The data bias in the database system leads to a neglected social problem. Nowadays, machine learning and AI would worsen the problem and spread further in the real world. Mehrabi et al. (2022) explain how the bias in the data influences the users and algorithms. Then Mehrabi et al. (2022) and Trewin (n.d.) show the importance of the world's need to emphasize the dangers of data bias and the fairness of people with disabilities. Moreover, Vassilev et al. (2022) and Vokginer et al. (2021) present their projects on developing and deploying ML-based systems in medicine to detect and mitigate bias, more significantly to prevent health care inequality for particular groups.

# Reference

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. USC Information Sciences Institute. Retrieved from https://arxiv.org/pdf/1908.09635.pdf

Trewin, S. (n.d.). AI Fairness for People with Disabilities: Point of View. IBM Accessibility Research. Retrieved from https://arxiv.org/pdf/1811.10670.pdf

Vassilev, A., Booth, H., & Souppaya, M. (2022, November). Mitigating AI/ML Bias in Context: Establishing Practices for Testing, Evaluation, Verification, and Validation of AI Systems. Retrieved from https://www.nccoe.nist.gov/sites/default/files/2022-11/ai-bias-pd-final.pdf

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. Retrieved from https://www.nature.com/articles/s43856-021-00028-w#Fig1