Shashwat Vidhu Sher

MLA 373

Professor Kochenderfer

7 June 2023

Human Bias in Training and Testing Data for AI Systems

AI systems depend highly on good training and testing data (both quantity and quality) to drive machine learning processes (Schwartz et. al. 15). As is commonly known in academia and industry, human bias can creep into the data collection phase and is a major challenge for the researchers involved in training and testing AI systems. This paper will highlight problems like selection bias, insensitivity to social inequities, and exclusion bias that plague data sets and suggest ways to overcome them.

Selection bias is perhaps the most common of the data collection issues and is defined as sampling data from a narrow or undiversified group. Using a specific group to collect data can severely impact the training of an AI system, and the system is likely to fare poorly when exposed to data from a larger set. For example, when trained on a limited set of past applicants from a specific industry, an AI system for candidate screening would give sub-optimal results when exposed to a more extensive data set comprising applicants from diverse fields. This behavior results from heuristics learned from the training data that do not apply to the test data (Larkin). Homogenous and small data sets are not the only problems, as even in large and diversified data sets, there could be issues related to historical inequities. (Pullum).

Racial, gender, and age-specific stereotypes often create problems in training data, even if collected over a larger population. Historical misrepresentation can affect data collected from public records and make an AI system wrongly classify a particular group with negative connotations. For example, it is often seen that in “online ad-targeting, African-American-identifying names tended to result in more ads featuring the word 'arrest' (Silberg and Manyika)." Researchers might not be sensitive to such representations, so a subconscious bias toward a particular group can affect the collected data. In addition to training data, human bias in the form of exclusion can also affect testing data.

Exclusion bias is the erasure of outliers and adversarial data inputs from test data. Such practice leads to misplaced confidence in an AI system’s predictive capabilities. Removing data points deviating too wildly from standard measurements is an accepted practice, as such inputs often result from recording or reporting errors. However, sometimes in marginalized populations, such outliers are not undesirable noise but relevant points on which a system should be tested (Schwartz et. al. 19). Also, as shown by Szegedy et al., not testing an AI system against "Adversarial examples," i.e. inputs that deviate from the standard, can result in incorrect predictions when the system faces real-world data (Szegedy et. al. 10). Being cognizant of exclusion bias and understanding which data to remove and which to keep is essential to test set creation. Though the problem of human bias may seem daunting, methods and processes exist to overcome it, as discussed next.

Analyzing the data set and its periodic update effectively reduces human bias. Undoubtedly, analyzing a large amount of data is no mean feat, but some processes can make this task manageable. For example, checking the data on certain parameters like age, gender, and race can help identify some obvious biases afflicting the data. Documenting the steps taken to clean the data can then inform the next round of data collection. Also, updating the data set regularly can help prevent problems like “population drift—where the nature of the underlying population changes over time (Hand and Khan).” In this way, any selection bias that gets baked into the AI system due to the previous data set can become less accentuated when the model is retrained with the latest data.

In addition to analysis and updates, there is a push toward data transparency and audit to combat human bias. For example, involving research groups from social sciences to look into data collected from varied communities can help identify social biases. By publishing the methods used in selecting the sample space for data collection and being transparent on the reasons for exclusions of specific data points, AI system designers can benefit from recommendations for improvement (Silberg and Manyika). A feedback channel and partnership with external groups can help AI researchers prepare better quality training and test data.

Complete elimination of human bias from AI data sets may never be possible, but it is a goal worth striving for. Sensitivity to such biases amongst researchers and adopting processes that can reduce them is the way forward in creating AI systems that can serve their purpose satisfactorily without perpetuating prejudices.

Works Cited

Hand, David J, and Shakeel Khan. “Validating and Verifying AI Systems.” *Patterns (New York, N.Y.)*vol. 1,3 100037. 12 Jun. 2020, doi:10.1016/j.patter.2020.100037. Accessed 16 Apr. 2023.

Larkin, Zoe. “AI Bias - What Is It and How to Avoid It?” *levity.ai,* 16 Nov. 2022, [levity.ai/blog/ai-bias-how-to-avoid](https://levity.ai/blog/ai-bias-how-to-avoid). Accessed 16 Apr. 2023.

Pullum, Laura. “Verification and Validation of Systems in Which AI is a Key Element.” *sebokwiki.org*, 31 Oct. 2022, [sebokwiki.org/wiki/ Verification\_and\_Validation \_of\_Systems \_in\_Which\_AI\_is\_a\_Key\_Element](https://sebokwiki.org/wiki/Verification_and_Validation_of_Systems_in_Which_AI_is_a_Key_Element). Accessed 16 Apr. 2023.

Schwartz, Reva, et al. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. National Institute of Standards and Technology, 15 Mar. 2022, doi:[10.6028/NIST.SP.1270](https://doi.org/10.6028/NIST.SP.1270). Accessed 16 Apr. 2023.

Silberg, Jake and James Manyika. “Tackling bias in artificial intelligence (and in humans).” *mckinsey.com*, 6 Jun. 2019, [www.mckinsey.com/featured-insights/artificial- intelligence/tackling-bias-in-artificial-intelligence-and-in-humans](http://www.mckinsey.com/featured-insights/artificial-%09intelligence/tackling-bias-in-artificial-intelligence-and-in-humans). Accessed 16 Apr. 2023.

Szegedy, Christian, et al. *Intriguing Properties of Neural Networks*. arXiv.org, 19 Feb. 2014, [arxiv.org/abs/1312.6199](http://arxiv.org/abs/1312.6199). Accessed 16 Apr. 2023.