Missouri S&T's Peer to Peer

Volume 2 | Issue 1 Article 1

May 2018

The Dangers of Human-Like Bias in Machine-Learning Algorithms

Daniel James Fuchs
Missouri University of Science and Technology

Follow this and additional works at: https://scholarsmine.mst.edu/peer2peer

 Part of the Computer Sciences Commons

Recommended Citation

Fuchs, Daniel J.. 2018. "The Dangers of Human-Like Bias in Machine-Learning Algorithms." *Missouri S&T's Peer to Peer* 2, (1). https://scholarsmine.mst.edu/peer2peer/vol2/iss1/1

This Article - Journal is brought to you for free and open access by Scholars' Mine. It has been accepted for inclusion in Missouri S&T's Peer to Peer by an authorized administrator of Scholars' Mine. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

Machine learning (ML), frequently used in constructing artificial intelligence, relies on observing trends in data and forming relationships through pattern recognition. Machine learning algorithms, or MLAGs, use these relationships to solve various complex problems. Applications can range from Google's "Cleverbot" to résumé evaluation, to predicting the risk of a convicted criminal reoffending (Temming 2017). Naturally, by learning through data observation rather than being explicitly programmed to perform a certain way, MLAGs will develop biases towards certain types of input. In technical problems, bias may only raise concerns over efficiency and optimizing the algorithm's performance (Mooney 1996); however, learned biases can cause greater harm when the data set involves actual humans. Learned biases formed on human-related data frequently resemble *human-like* biases towards race, sex, religion, and many other common forms of discrimination.

This discrimination and the question of the fairness of artificial intelligence have received increasing public attention thanks to the numerous social media-based AIs launched in recent years. Microsoft's "Tay", an AI made to resemble a teenage girl, became anti-Semitic, racist, and sexist; Tay was shut down a mere "16 hours into its first day" (Wiltz 2017). Following in Tay's footsteps, Microsoft's "Zo" exhibited similar problematic biases despite additional precautions (Shah 2017). Other MLAGs, such as Beauty.AI's "robot jury," have demonstrated learned biases towards physical properties like skin tone and facial complexion (Pearson 2016). In these three popular cases, though the biases were quickly identified, the designers were unable to simply remove the learned biases. Despite the intention of their designers, many ML implementations have developed harmful human-like biases that cannot be easily removed.

While much research is being done to improve performance speed, create more efficient implementations, and create more powerful MLAGs to solve more difficult problems, much of

this research does not concern bias control or correction. This is to be expected, as many ML implementations are applied to solve purely technical problems. While ML implementations might not have to enforce any form of fairness when dealing with strictly technical data, the growing usage of MLAGs that operate on human data reveals a need to better regulate bias to ensure fairness. The purpose of this study is to show the effects of these human-like biases in MLAGs across a variety of scenarios and to analyze the results of both current and emerging methods of bias correction. Human-like biases in MLAGs have many harmful effects, and there is a need for greater control over and the correction of these learned biases.

Research Design

To study the effects of human-like bias in MLAGs, I used the ACM Digital Library, IEEE Xplore, and Scopus. These three databases provide numerous articles on observations of learned biases in MLAGs and records of correctional efforts and methods to manipulate biases. The search keywords *machine learning*, *correctional*, *artificial intelligence*, and *bias* were used to browse these databases. Articles that concern observations of learned bias in MLAGs and articles that concern bias correction or avoidance methods are included in this study. Articles that focus on solving purely technical problems with MLAGs or statistically evaluating the performance of an MLAG have been excluded. A variety of ML implementations across different fields were studied to provide a more thorough understanding of the effects of human-like learned biases in different circumstances.

I also refer to recent incidents of bias-driven discrimination by ML implementations that garnered noteworthy public attention. These incidents, such as Microsoft's AIs "Tay" and "Zo" or Beauty.AI's artificial jury, while having relatively well-documented results thanks to the great public outcry, tend not to have their technical details revealed to the public. Some of the parties

responsible for these incidents, such as Microsoft, offered statements explaining the behavior of their ML implementations (Wiltz 2017) but still did not disclose specific technical details. In discussing these events then, since research and academic journal articles are generally unavailable, I relied on popular articles on the subject. These sources are used here to discuss the behavior and actions of each MLAG.

Machine Learning Training and Bias Origin

MLAGs generally require two components before they can be applied to a particular problem. First, the underlying ML framework must be constructed. While the algorithm's designer may understand the framework itself, as Maria Temmings writes, "it’s often unclear — even to the algorithm’s creator — how or why [the algorithm] ends up using data the way it does to make decisions" (2017). It is difficult to directly observe learned biases to see why they formed or how they affect data; the complex network of relationships that compose the learned bias exist as an effectively abstract object. Therefore, rather than attempting to directly detect a learned bias, observers can identify bias by observing trends in the MLAG's decisions.

The second component to creating a functional MLAG is proper "training." Training refers to exposing an MLAG to a special set of inputs with specific desired outputs to teach the algorithm how to solve a problem (Osaba and Welser 2017). This particular style of training, commonly known as "supervised training," sets up an MLAG to deal with future cases by using the training data as a reference. The MLAG then extrapolates from the training data to make future decisions. If the training data accurately represents the population that algorithm is to operate in, the behavior of the algorithm will generally be more predictable. While the duration and exact role of training vary across ML implementations, learned biases form during the training period. This most commonly occurs in cases where the training data does not adequately

prepare the algorithm for use. If the training data poorly represents the target population or is chosen carelessly, training can directly create harmful learned biases (Osaba and Welser 2017). To reduce the amount of learned bias that an MLAG may develop, and by extension the number of human-like learned biases, the training data should accurately represent the population the algorithm is intended to operate on.

Learned Biases in Social Media

The risk of improper training is particularly high for chatbots. Microsoft's twitter-based AI chatbot Tay, despite being stress-tested "under a variety of conditions, specifically to make interacting with Tay a positive experience," learned anti-Semitic and racist behavior due to the efforts of a specific group of individuals (Wiltz 2017). By being repeatedly exposed to similar types of discriminatory content, Tay acquired numerous discriminatory biases. While many MLAGs cease learning after completing their initial training, some chatbots continue to learn, and these chatbots tend to be particularly quick to acquire new biases. This is partially due to the difficulty of making a chatbot's training data accurately represent all potential discussion across the social media platform they will operate on (Wiltz 2017). The nature of social media implies that these chatbots may frequently be exposed to discriminatory input, and if insufficient training data is supplied to reject or counter these inputs, these ML implementations can easily learn harmful human-like biases.

A year after Tay was shut down, Microsoft launched another chatbot known as Zo, which faced similar public backlash after exhibiting anti-Islamic learned biases (Shah 2017). However, due to bias avoidance measures, Zo proved to be resistant to exhibiting discriminatory biases. To avoid exhibiting bias, Zo included filters for rejecting discussion about topics that referenced religion or politics (Shah 2017). But, though Zo did not frequently exhibit harmful biases,

moments of discriminatory behavior indicate that underlying harmful learned biases still formed. Even though the output was made to appear correct through special filters, this does not remove the underlying harmful learned biases. Thus, this method of bias avoidance still frequently failed, and it is only applicable to certain MLAGs. This method of bias avoidance also relies on having input/output that can be easily categorized for filtration, and pre-existing knowledge of everything that needs to be filtered. Like Tay, Zo developed harmful learned biases due to improper training.

Hidden Bias and the Importance of Adequate Training

Some may argue that for social media chatbots, the training data failing to accurately represent the chatbot's environment is not the fault of the data itself but rather that discriminatory learned biases that appear in this environment are the fault of individuals with malicious intent to corrupt the chatbot. However, while those that "launched a coordinated attack" are not representative of these chatbot's intended users (Wiltz 2017), discriminatory learned biases do not always form in explicit or obvious manners, and determining if a user is acting maliciously may sometimes be difficult. A team of researchers from Princeton University published a study on using a purely statistical MLAG to map the context surrounding words across a large "standard corpus of text" (Caliskan, Bryson, and Narayanan 2017). Their results revealed many hidden instances of discriminatory associations between words, such as associating females' names with familial terminology, males' names with careers, and African-Americans' names with "unpleasant" words (Caliskan et al. 2017). Since a statistical MLAG made these observations, an MLAG that used this large text as training data could form many human-like learned biases.

While this corpus of text might not to appear to have implicit biases harmful to human readers viewing excerpts of text, the hidden biases were revealed by a statistical MLAG studying the context around words. However, not all training data can be first fed into a statistical MLAG to look for hidden biases, as this relies on knowing what instances of input data can carry an implicit bias. It can be very difficult, then, to predict all the possible associations an MLAG could learn from a given set of training data. Even if a hidden bias can be identified, it might be impractical or impossible to directly correct that set of training data. However, using purely statistical MLAGs as a method for revealing hidden biases in training data can reveal a need for a new set of training data or improvements to the current set. By improving an MLAG's training data, the number of harmful learned biases it acquires can be reduced.

Learned Physical Biases in Recognition Software

While chatbots continue to learn after completing their initial training, some MLAGs stop learning after completing their initial training. While this type of MLAG cannot acquire new biases during operation, improper training data can still lead to harmful learned biases. The team behind Beauty.AI created an artificial jury of "robot judges", with the intention of using this jury to host the first online, AI-judged beauty contest (Pearson, 2016). The jury was trained on a large set of user images with various physical attributes rated by human judges. Ideally, this training data would allow the jury to develop an objective method of rating contestants, though this would also require the human judges to score the training images objectively. However, in practice, the jury proved to be highly biased towards skin tones, with 44 of the 50 winners being white contestants, while only "one finalist had visibly dark skin" (Pearson, 2016). However, rather than the training data being biased by the human judges, the eventual determined cause for this result was that the majority of the training data involved individuals with light-skin tones;

insufficient training data on darker skin tones led to a bias of higher ratings for light-skin (Pearson, 2016). The training data's failure to represent the population led to a harmful learned bias towards skin tone, which skewed the results of the contest.

A team of researchers at Microsoft faced a similar issue in facial-emotion-recognition technology, stating "poor representation of people of different ages and skin colors in training data can lead to performance problems and biases" (Howard, Zhang, and Horvitz 2017). With training data that over-represents a certain demographic, the MLAG that drove some of Microsoft's emotion-recognition technology frequently failed to accurately detect emotions in children, the elderly, and minorities (Howard et al. 2017). However, the researchers designed a bias correction method by using "specialized learners," which explicitly put training emphasis on minorities and those of age groups that were less commonly represented in the training data (Howard et al. 2017). Rather than being trained on all the supplied training data, under this methodology the MLAG is more frequently exposed to data that deviates from the averages in the data set. The intention of this methodology is to correct bias by increasing the expected range of values internally determined by the MLAG.

This method of bias correction proved effective, resulting in an "increase in the overall recognition rate by 17.3%" (Howard et al. 2017). It caused the algorithm to be prepared for greater diversity, which led to a better representation of the target population, and ultimately resulted in fewer learned biases. While the training data itself was not modified or improved, by using an excerpt from the training data selected by the "specialized learners," Microsoft’s MLAG developed a more accurate model of the target population. While this methodology may not be applicable to other types of training data used by MLAGs, it has proven effective in reducing discriminatory learned biases related to physical traits.

Historical Bias and Inferential Discrimination

In cases where little training data is available, it is generally difficult to form a training data set that accurately represents the population. These training data commonly have "historical bias," or bias created by selective targeting over a period of time. This problem frequently arises in ML implementations in the field of criminal justice, namely due to historical discrimination against individuals from minorities. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a commonly-cited example of racial learned bias, is an ML implementation used to predict reoffending risk in convicted criminals (Temming 2017). COMPAS has frequently demonstrated a human-like bias towards race, wrongly predicting "that black defendants would reoffend nearly twice as often as it made that wrong prediction for whites" (Temming 2017). In other words, the rate of false positives for black defendants being reconvicted was nearly double that for white defendants; not only was COMPAS biased, but this bias also led to great inaccuracy.

To train COMPAS, it is provided a large set of crime reports as training data. The racial biases exhibited by COMPAS are likely learned from historical biases within the crime reports, such as a disproportional number of the reports being from low-income neighborhoods (Temming 2017). This historical bias, besides being partially due to direct human behavior, is also reinforced by some other MLAGs. The MLAG "PredPoll," which is used to predict crime location and distribute police presence, frequently shows bias towards selecting low-income neighborhoods and locations with higher minority concentration (Temming 2017), which leads to increased police presence in these areas, and by extension more recognized crime reports and active responses from these areas. More reports are received from areas of greater police presence, which leads to a cycle of further increased police presence in and crime reports from

these areas. This historical bias, and a variety of other biases possibly contained in COMPAS's training data, has led to the harmful learned bias towards race demonstrated by COMPAS.

Some may argue that an easy solution to prevent this racial discriminatory bias would be to simply remove sensitive information like race or sex from the training data; in fact, sensitive data fields that might cause inaccuracy are already typically hidden from the algorithm (Osaba and Welser 2017); however, "learning algorithms can implicitly reconstruct sensitive fields and use these probabilistically inferred proxy variables for discriminatory classification" (Osaba and Welser 2017). For example, "zip code may be strongly related to race, college major to sex, health to socioeconomic status" (Temming 2017). Since MLAGs were fundamentally created to find relationships across data, they have incredible inferential ability, and even hiding sensitive data fields can simply lead to the algorithm reconstructing the same hidden field. It can then develop the same harmful bias it originally learned, even though it no longer knows what property it is discriminating against. This inferential discrimination is difficult to prevent, as the inferential ability of MLAGs cannot simply be disabled.

Bias Correction by False Data

The inferential ability of MLAGs is one of the primary reasons bias correction has proved so difficult. Despite hiding fields and removing obvious properties that humans might discriminate against, MLAGs can use patterns in data that might not be obvious to humans to infer the hidden data's original values (Osaba and Welser 2017). While pattern detection is one of the MLAG’s greatest strengths, it also complicates data pruning, as pruning beyond the scope of an MLAG’s inferential ability might render the MLAG itself completely ineffective. The difficulty of bias correction in COMPAS and other MLAGs susceptible to inferential discrimination led to a team of researchers at the Max Planck Institute to perform research on

bias correction by adding false training data (Zafar, Valera, Rodriguez, and Gummadi 2017). Their study was performed for COMPAS, so their research involved predicting criminal reoffending rates. To improve COMPAS's performance and correct the underlying racial bias, their research focused on improving COMPAS's training data.

The researchers introduced a methodology for creating falsified training data based on the amount of disparate treatment, which they determined by variance in the rate of misclassification between different groups (Zafar et al. 2017). In other words, the more frequently a group was classified incorrectly, the more false data was generated for that group. For example, if subjects from some race were frequently misclassified as a repeat offender despite not reoffending, that minority then received an amount of falsified data in proportion to the rate of misclassification compared to how frequently other groups were misclassified in the same way. The falsified data in this case would be positive reports of non-reoffense (Zafar et al. 2017). The falsified data helped to combat the biased training data that led to discrimination and also helped prevent the MLAG from accurately reconstructing hidden fields related to these biases (Zafar et al. 2017). Modifying the training data itself led to a reduction in historical bias, and reducing the inferential accuracy of the MLAG further reduced the effect of the discriminatory bias. When applied to COMPAS, the researchers found this methodology resulted in a resounding success; misclassification rates for African-Americans as repeat offenders dropped from 45% to 26%, while white misclassifications remained at 23% (Temming 2017). In this particular case, bias correction reduced the rate of discriminatory bias while also increasing the MLAG's accuracy as a whole. This shows that improving the training data can lead to a reduction in harmful learned biases, which can improve both fairness and the functionality of the MLAG.

Discussion

The various human-like learned biases exhibited by social media chatbots, physical image-recognition software, and criminal justice ML implementations typically resulted in harmful and unfair treatment of the human population they were intended to operate on. In the ML implementations researched in this study, the origin of their harmful learned biases was generally traced to either insufficient or implicitly-biased training data. To prevent the formation of harmful human-like biases, the training data must accurately represent their algorithm's target population and also not contain hidden, implicit biases. However, as shown by the statistical analysis on a "standard corpus of text" (Wiltz 2017), implicit bias can easily be hidden within the training data. When an alternative source of training data is not available, training data can be difficult to correct due to the inferential capabilities of ML implementations (Osaba and Welser 2017). Yet, despite the difficulty in providing proper training data to ML implementations, numerous bias correction methods have been developed. These methods, such as using falsified data to counter inaccurate harmful biases (Zafar et al., 2017) and using "specialized learners" to provide training emphasis on outliers in the training data (Howard et al., 2017), have both proved not only to reduce the number of discriminatory learned biases from their respective ML implementations, but also to further improve the accuracy of their respective algorithms. By modifying existing training data or creating new training data, bias correction methods have helped reduce the effect harmful learned biases.

However, this research was limited to a small number of ML implementations, and these methods of bias correction may not be applicable to other types of MLAGs based on their use cases or on the datasets they work with. Given that proper training data is critical to creating a fair MLAG, much more research needs to be done in potential methods for improving training

data. In particular, using false data to improve training data has potential for use in many more MLAGs than just those in criminal justice. More research needs to be done concerning the effects of false data, as excessive training data falsification could cause a variety of additional problems with MLAG performance by misrepresenting the actual population the MLAG will operate on. Other more general or powerful methods of manipulating training data would directly lead to greater control over learned biases, which could lead to more accurate and unbiased MLAGs in the future.

This study has shown a variety of harmful effects that these human-like biases cause. Though racial discrimination in beauty pageants and sexism in social media are surely problematic whether an AI is responsible or not, the growing role of ML-driven implementations in the world necessitates concerns over the "morality" or fairness of AIs. While fairness is admittedly a concern that sometimes steps outside the jurisdiction of computer science, this research has shown that some discriminatory biases can lead to less effective and potentially inaccurate MLAGs. I believe this provides grounds for computer scientists to strive for fairness from ML implementations. Improving the accuracy and fairness of these MLAGs can not only have a positive impact on the field of computer science, but can also lead to improvements in the many fields in which ML implementations are used. Whether in beauty contents or chatbots, résumé selection or criminal justice systems, developments in ML can have a far-reaching impact. As AIs and other ML-driven implementations find increasingly common use and more important roles, there exists a growing need to control their behavior and remove potentially harmful human-like biases from MLAGs to ensure machines treat humans fairly and objectively.

References

Caliskan, A., Bryson, J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), pp. 183-186, doi: 10.1126/science.aal4230

Howard, A., Zhang, C., & Horvitz E. (2017). Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pp. 1-7, doi: 10.1109/ARSO.2017.8025197

Mooney, R. (1996). Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. *Conference on Empirical Methods in Natural Language Processing* (82-91). Austin, TX: University of Texas.

Osoba, O., & Welser, W. (2017). An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence. Santa Monica, CA: RAND Corporation.

Pearson, J. (2016). Why An AI-Judged Beauty Contest Picked Nearly All White Winners. Motherboard. Retrieved from https://motherboard.vice.com/en_us/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners.

Shah, S. (2017). Microsoft's "Zo" chatbot picked up some offensive habits. Engadget. Retrieved from https://www.engadget.com/2017/07/04/microsofts-zo-chatbot- picked-up-some-offensive-habits/.

Temming, M. (2017). Machines are getting schooled on fairness. ScienceNews, 192(4), pp. 26, Retrieved from https://www.sciencenews.org/article/machines-are-getting-schooled-fairness.

Wiltz, C. (2017). Bias In, Bias Out: How AI Can Become Racist. *Embedded Systems Conference*

(ESC). Minneapolis, MN: Design News.

Zafar, B., Valera, I., Rodriguez, G., & Gummadi, P. (2017). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *Proceedings of the 26th International Conference on World Wide Web* (1171-1180). Geneva, Switzerland: Max Planck Institute.