International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

An Insight and Resolve into Algorithmic Bias in Artificial Intelligence Systems

Prateek Bajaj

SAP Labs India, Bangalore, India
ptkbajaj[at]gmail.com
ORCID: 0000-0003-2105-860X

Abstract: ***With the growing usage of Artificial Intelligence and Machine Learning in businesses and decision making, there has been a debate churning on the topic of bias and fairness in technology today. In this paper, the authors talk about bias in algorithms and artificial intelligence systems, understanding the underlying cause for the same, looking deeper into the existing research, and deducing an elaborate alternative process to train and practice AI allowing for fair, unbiased decision making.***

Keywords: AI, ML, ethics, bias, fairness, algorithm, decision

Declarations: Not Applicable

1. Introduction

There has been an increase in the use of Artificial Intelligence in the recent times, including some sensitive areas such as hiring, healthcare, justice, as suggested by Silberg, et al. [1]. But the use of AI and computer algorithms in the art of decision making isn't new. Silberg's research [1] suggests that in 1988, the UK Commission of Racial Equality found a medical school in Britain guilty of bias against women applicants and non-European names. The underlying problem, after extensive discussions and investigations resulted in the argument that these decisions were made matching the data fed into systems consisting of human decisions. More so, the same algorithm had an accuracy of more than 90 percent, resulting in a highly accurate system trying to mimic its training data.

AI being the front runner for decision making today, affecting not just individuals but businesses, economies, and industries suggests the need for understanding the process of decision making in AI, and taking the right steps towards developing processes that help not just dilute the bias in the current systems, but completely erase bias from all future AI and decision making systems.

Before we dig deeper into analyzing the causes and effects of bias in AI, let us understand the concept of AI and algorithms for decision making first. Friedman, et al. [2] puts it in a manner that signifies to the researchers even today that AI is an elaborate process applied to 'narrow inference tasks' where large volumes of data are present, to find associations.Panch et al. [4] has further emphasized on the fact that there is no 'general purpose' replacement for human intelligence available in AI today. Such inferences take the shape of determining the importance of human involvement and interference in training artificially intelligent algorithms. The result suggests that AI isn't a smart all-intelligent system that starts discriminating and originating bias on its own. These prejudices stem from every particle of data fed into these algorithms training them.

2. Defining Bias

Green, et al. [5] in their research on the subject of fairness in machine learning suggest a variety of definitions associated with bias and fairness. This research deduces that bias can be defined as any form of preference, discrimination, for or against an individual, a group, or characteristics. What needs to be emphasized on is the fact that bias isn't just towards or against people.

Decisions differing in ethical judgement for certain sensitive attributes or protected characteristics, as suggested by Silberg et al. [1] may also result in bias. What is also interesting to note however, is that the absence of unwanted bias may not always be sufficient to call a system fair. There is no one-size that fits all scenarios for deployments of such systems.

The first inference we make towards resolving the issue of bias in AI is the importance of understanding the social contexts into which these systems are being deployed.

3. The Causes Resulting in AI and Algorithmic Bias

Different researches conclude with a plethora of reasons causing algorithmic bias. From the most common problem of a lack of clear standards of 'fairness' to insufficient contextual specificity. The following lists down the most relevant reasons associated with bias in artificial intelligence systems and algorithms:

1. Data as the source of bias: Extensive research conducted on determining the underlying factors responsible for bias in systems points out to data as the cause for creating bias, as backed by Angwin's [7] research. Kleinberg [6] in their research on the subject of discrimination in algorithms points the relevance of creating a distinction and separating the AI model into two algorithms, i.e. the trainer and the screener. The trainer may be biased due to the underlying data associated with the training process, and the screener that is responsible for making predictions based on the

trainer. Such a system, Kleinberg proposes results in the proof for the hypothesis of data as the source of bias.

Balukbasi's [8] research points out on the example of word embeddings in an NLP (Natural Language Processing) algorithm trained on news articles and reports exhibiting gender stereotypes found in society. But the actual concern here is that even when algorithms are programmed to exclude protected characteristics directly, there have been multiple studies proving the indirect encoding of biasby other variables. For instance, a hiring algorithm may consider a prospect's credit history as a factor for determining viabilityor may favor words more commonly found on certain applications, as indicated by Hao [9].
Another example of data bias could be oversampling certain groups. Data generated by users could also be a responsible source of bias in algorithms.

2. Clear Standard of Fairness: Systemic bias has been long prevailing in the society, and it has been difficult to suggest a single standard and definition of fairness. A very limited few in the world are responsible for training algorithms catering to masses, with the added burden of a clear lack of standards for fairness. Most of the existing standards too are qualitative, not quantitative, causing multiple interpretations. The evaluators looking into such bias hold the same qualitative standards too, subjecting algorithms to implicit bias. Panch, et al. [3] identifies this concept of standards of fairness in their research.

3. Lack of Contextual Specificity: Each AI system and algorithm created in the world varies in design, objectives and the diversity of people they are designed for. As Panch et al. [4] recommends in their research, most systems created cater to consumers from different cultures, backgrounds, beliefs, preferences, and more. There are no such generic data catering to all such groups.

4. Measuring and Mitigating Bias in Algorithms

The research conducted till now helps the readers deduce the concepts of bias and fairness, their causes, and the implications associated with them. Next, the authors formulate a process handling multiple outcomes to measure the reasons behind bias in certain algorithms and models.

A look into human decision making: But before analyzing the same, what is more important is to understand ways in which these models and algorithms are trained. Largely, all models today are trained by data created by and involving human beings. However, human beings, the key carriers of these biases are not always objective while making decisions. Pligt et al. [10] deduces in their study on decision making that this process (of decision making) is a very subjective one, and more often than not, it is not easy to trace back the reasoning behind every decision made. Studies such as those conducted by Pligtclearly demonstrate that human decision making is much more complicated than just a standard objective set of questions resulting in an outcome. The human brain factors in a huge gamut of emotions, past experiences, knowledge, intrinsic biases, and their outcomes from previous experiences to make decisions. Most models handling artificial intelligence today do not yet factor in the same amount of resources while making decisions, which has been elaborated hereafter. Researches such as those conducted by Lone et al. [11] identify major differences between cognitive processes of humans and artificial intelligent systems, clearly indicating a wider lack of emotional intelligence in AI. As Lone et al. suggest, emotional intelligence and experiences can only be instilled in humans, not installed. In AI systems however, such parameters can only be installed for now, resulting in a huge flaw in such systems to conduct decision making with a lack of high emotional intelligence. Moreover, Pomerol, et al. [12] observe in their research on the topic of decision making in humans and artificial intelligence about two factors responsible for decision making, i.e. 'diagnosis of the situations' and a 'look-ahead reasoning'. The research, although distinctively identifies AI's relationship with the first aspect, i.e. diagnosis of situations, it demonstrates AI's lack in the second aspect, i.e. look-ahead reasoning, which has its main concepts revolving around preferences and uncertainty. This has resulted in AI models to be objective in making decisions, making it easier for researchers to drill down on reasoning behind each decision.

If there is a streamlined process used to understand the reasoning behind decision making in AI models, it would be easier to understand the underlying biases induced in those decisions, and in turn, mitigate those biases.

The concept of reverse engineering: Seong Joon, et al. [13] presented in their research multiple ways of reverse-engineering black-box neural networks. Multiple researches on the subjectindicates that repeated queries to machine learning models (at the very basic level) can result in effective reproduction of a machine learning-trained AI model. These procedures are commonly known as 'model extraction attacks'. It is possible to infer hyperparameters of a deep neural network by observing responses of a sequence of queries sent to an AI model. This is possible and has been carried out for various models, as indicated by Florian et al. [14] in their research on deducing information about machine learning models. The research simulated a number of machine learning models, including logistic regression models, support vector machines, and decision trees. The resulting conclusions from the rigorous analyses and attacks on the aforementioned models deduced information such as confidence values, scores, and predictions, which forms the basis for reverse engineering artificial intelligence models to fetch several parameters responsible for a prediction in a model.

Seong et al. [13] in their research concluded on building on the concept of metamodels, which are basically models trained to predict attributes of other models. Metamodels, as derived by Seong, work on the concept of submitting *n* query inputs to an actual machine learning model *a*, takes the corresponding outputs from the actual ML model *a*, and uses it as an input to return predicted model attributes as its output. Consider the example of MNIST digit classifiers, which can be representative of a generic learning model technique, using an MNIST dataset and training the conceptual metamodels on the same. Seong et al [13] in their research suggest three major methodologies to infer

metamodels out of this digit classifier: *Kennen-O, Kennen-I, Kennen-IO*. Each of these works on the concept of feeding queries to the actual machine learning models and using the outputs from those queries to determine the deciding attributes behind these models.

Taking such research forward, any AI model can be reverse engineered to some extent, to fetch the hyperparameters associated with their decision making. Following is an example implementationused to derive decision-making attributes in AI, helping in understanding how decisions are made in AI models:

Consider an AI model used for sending ad-suggestions to prospective customers of a company. With the help of *Membership Inference attacks methodology*, as inferred by Alteniese et al. [15], this ad-suggesting AI model can be reverse-engineered to provide information about whether a particular data sample has been included in the training of this model. Using this concept, a data sample of a particular raceor gender corresponding to select products (more stereotypically used by that race/gender), can be fed into this system to determine whether a biased dataset or a data sample has been used.

This deduction is largely different from any other research conducted in this field of bias, as the research here suggests practical applications of machine learning models and reverse engineering, as well as refined processes to understand what an AI model uses to reach at results, which may or may not be biased.

Case for a dummy AI model:Consider a dummy AI model that helps select candidates for a particular position in a company. It considers criteria such as skills, age, experience, education, gender, race, etc. for each role. One of the leading factors responsible for helping this model make decisions is the historic data used to train this model. An AI model is as good as the data used to train it with. Hypothesize now that this model is used for a role largely dominated by male candidates, resulting in data that is skewed towards more male candidates. Moreover, consider this position to be largely preferred by candidates who may not hold a doctorate. Multiple permutations of education, genders, age groups, and skills can be created and sent as requests to this model with response of each request being stored. Within a very few requests, this reverse engineering process would help the researchers determine the bias, not just towards the male gender, but also against candidates with doctorates.
Now, this bias may or may not have been induced by the creators of this algorithm, or by the actual hiring teams, but just due to the lack of enough diverse data, biases (which may not seem unfair at the first go) may propagate further to the decision making models.

Reverse engineering such models to understand the underlying data is one way to measure bias. And there are several other ways that can be suggested to mitigate such bias. Next, we include some socio-technical solutions/suggestions that can prove to be effective in the most generic issues:

1) Awareness towards context of AI: What is extremely important for this problem is to anticipate the domains prone to bias, especially in cases when systems have been known to create bias with skewed data.
2) Proactive training of AI models: Most important of all the solutions would include establishing processes to test for and mitigate bias at early levels of training and testing.
3) Diversifying the teams building such AI models and algorithms, hence resulting in reducing human bias in the whole group of people building AI models.
4) Algorithms shouldn't be left to make decisions completely without any human intervention just yet. More humans and machines working together can result in a more balanced approach towards decision making.

Such suggestions can provide a pathway towards a less biased AI and algorithmic universe, while considering not just technical, but also the aforementioned socio-technical solutions to fight the bias in AI that persists in the society.

5. Conclusion

With the current advancements in the field of AI and Machine learning, there is a dramatic need for extensive research on the concepts of bias and other human emotions while dealing with AI models, especially due to the expansive usage of AI and algorithms in society today. This research talked about the definitions of fairness and bias, understood the core reasons behind bias in AI and some of its implications, focused on filtering data to reduce bias, and proposed technical as well as socio-technical solutions to the problem. Such research can directly impact the society and people at a large scale, from hiring algorithms, to ad-suggesting systems, to healthcare, and justice.

References

[1] Silberg, Jake and Manyika, James. "Notes from the AI frontier: Tackling bias in AI (and in humans). June 2019.
[2] Batya Friedman and Helen Nissembaum, "Bias in computer systems," ACM Transactions on Information Systems, July 1996, Volume 14, Number 3, pp. 330–347.
[3] Panch T, Szolovits P, Atun R. "Artificial intelligence, machine learning and health systems." J Glob Health. 2018 Dec.
[4] TrishanPanch, Heather Mattie and Rifat Atun. "Artificial intelligence and algorithmic bias: implications for health systems"
[5] Ben Green and Lily Hu. "The myth in the methodology: Towards a recontextualization of fairness in machine learning". 35th International Conference on Machine Learning, Stockholm, Stockholm, Sweden, July 10–15, 2018; Meredith Whittaker et al., AI Now Report 2018, AI Now Institute, New York University, December 2018
[6] Jon Kleinberg et al., Discrimination in the age of algorithms, SSRN, February 2019
[7] Julia Angwin et al., "Machine Bias," ProPublica, May 2016
[8] TolgaBolukbasi et al., "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," Proceedings of the 30th

International Neural Information Processing Systems, pp. 4356–4364; Google Developers Blog, "Text embedding models contain bias. Here's why that matters," blog entry by Ben Packer, Yoni Halpern, Mario Guajardo-Céspedes& Margaret Mitchell, April 13, 2018.

[9] Karen Hao, "This is how AI bias really happens—and why it's so hard to fix," MIT Technology Review, February 4, 2019.

[10] J. van der Pligt, "Psychology of Decision Making", International Encyclopedia of the Social & Behavioral Sciences 2001, Pages 3309-3315, 2001.

[11] Zahoor Ahmad Lone, Dr Shah Alam, "Emotional Intelligence; A Flaw In Robots", International Journal Of Technology Enhancements And Emerging Engineering Research, Vol 1, Issue 4, 2013.

[12] Jean-CharlesPomerol, "Artificial intelligence and human decision making", Vol. 99, Issue 1, European Journal of Operational Research.

[13] Seong Joon Oh, et al., "Towards Reverse-Engineering Black-Box Neural Networks", ICLR 2018.

[14] Florian Tramèr,Fan Zhang,Michael K. Reiter, Thomas Ristenpart, "Stealing Machine Learning Models via Prediction APIs", 2016.

[15] Giuseppe Ateniese, Giovanni Felici, Liugi V. Mancini, Angelo Spognardi, Antonio Villani, and Domenico Vitali. "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers." IJSN, 2015.