Countering Bias in Personalized Rankings

From Data Engineering to Algorithm Development

Team Details

Bishnu Kashyap
(M20AIE227)

Sri Rama Sasi Teja Bhattiprolu

(M20AIE309)

Introduction:

Machine learning algorithms have penetrated every aspect of our lives. Algorithms make movie recommendations, suggest products to buy, and who to date. They are increasingly used in high-stakes scenarios such as loans and hiring decisions. There are clear benefits to algorithmic decision-making, unlike people, machines do not become tired or bored, and can take into account orders of magnitude more factors than people can. However, like people, algorithms are vulnerable to biases that render their decisions “unfair”. In the context of decision-making, fairness is the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics.

Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people. A canonical example comes from a tool used by courts in the United States to make pretrial detention and release decisions. The software, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), measures the risk of a person to recommit another crime. Judges use COMPAS to decide whether to release an offender, or to keep him or her in prison. An investigation into the software found a bias against African-Americans

COMPAS is more likely to have higher false positive rates for African-American offenders than Caucasian offenders in falsely predicting them to be at a higher risk of recommitting a crime or recidivism. Similar findings have been made in other areas, such as an AI system that judges beauty pageant winners but was biased against darker-skinned contestants or facial recognition software in digital cameras that overpredicts Asians as blinking. These biased predictions stem from the hidden or neglected biases in data or algorithms.

In this survey we identify two potential sources of unfairness in machine learning outcomes— those that arise from biases in the data and those that arise from the algorithms. We review research investigating how biases in data skew what is learned by machine learning algorithms, and nuances in the way the algorithms themselves work to prevent them from making fair decisions—even when the data is unbiased. Furthermore, we observe that biased algorithmic outcomes might impact user experience, thus generating a feedback loop between data, algorithms and users that can perpetuate and even amplify existing sources of bias.

BIAS IN DATA, ALGORITHMS, AND USER EXPERIENCES:

Most AI systems and algorithms are data driven and require data upon which to be trained. Thus, data is tightly coupled to the functionality of these algorithms and systems. In the cases where the underlying training data contains biases, the algorithms trained on them will learn these biases and reflect them into their predictions. As a result, existing biases in data can affect the algorithms using the data, producing biased outcomes. The outcomes of these biased algorithms can then be fed into real-world systems and affect users' decisions, which will result in more biased data for training future algorithms.

For example, imagine a web search engine that puts specific results at the top of its list. Users tend to interact most with the top results and pay little attention to those further down the list. The interactions of users with items will then be collected by the web search engine, and the data will be used to make future decisions on how information should be presented based on popularity and user interest. As a result, results at the top will become more and more popular, not because of the nature of the result but due to the biased interaction and placement of results by these algorithms. The loop capturing this feedback between biases in data, algorithms, and user

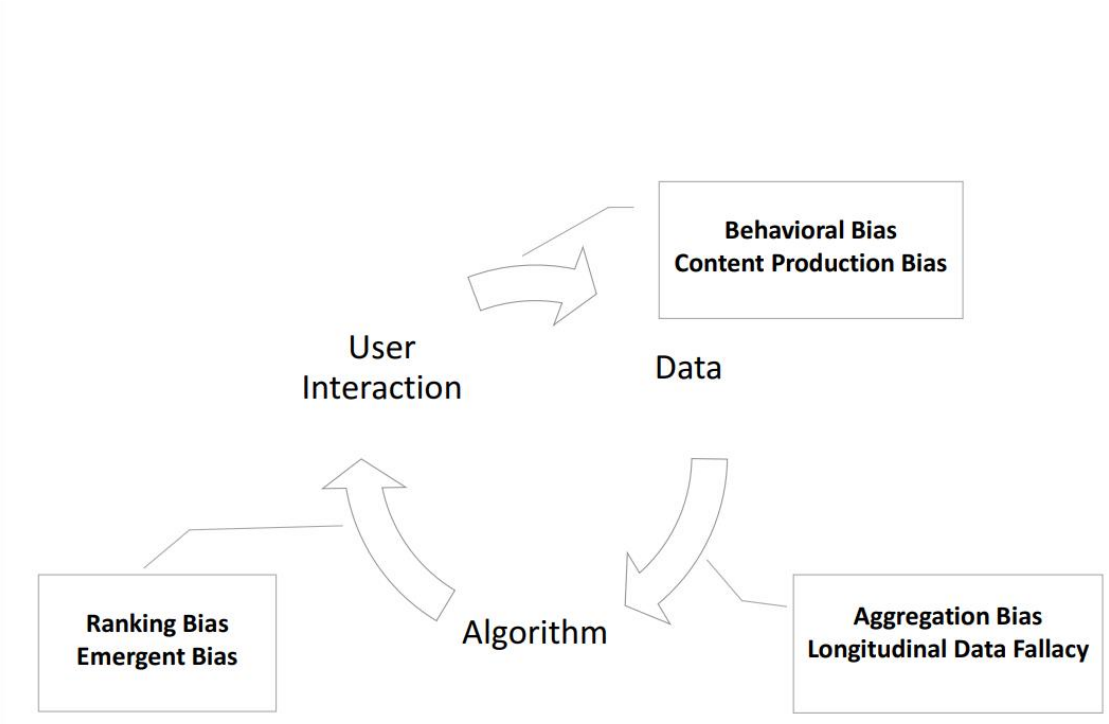


Fig. 1. Examples of bias definitions placed in the data, algorithm, and user interaction feedback loop.

Types of Bias

Bias can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks. here we prepare a complete list of different types of biases with their corresponding definitions that exist in different cycles from data origins to its collection and its processing. Here we will reiterate the most important sources of bias introduced in these two papers and also add in some work from other existing research papers. Additionally, we will introduce a different categorization of these definitions in the paper according to the data, algorithm, and user interaction loop.

Measurement Bias:- Measurement, or reporting, bias arises from how we choose, utilize, and measure particular features.

An example of this type of bias was observed in the recidivism risk prediction tool COMPAS, where prior arrests and friend/family arrests were used as proxy variables to measure level of “riskiness” or “crime”—-which on its own can be viewed as mismeasured proxies. This is partly due to the fact that minority communities are controlled and policed more frequently, so they have higher arrest rates. However, one should not conclude that because people coming from minority groups have higher arrest rates therefore they are more dangerous as there is a difference in how these groups are assessed and controlled

Omitted Variable Bias:- Omitted variable bias occurs when one or more important variables are left out of the model.

An example for this case would be when someone designs a model to predict, with relatively high accuracy, the annual percentage rate at which customers will stop subscribing to a service, but soon observes that the majority of users are cancelling their subscription without receiving any warning from the designed model. Now imagine that the reason for cancelling the subscriptions is appearance of a new strong competitor in the market which offers the same solution, but for half the price. The appearance of the competitor was something that the model was not ready for; therefore, it is considered to be an omitted variable

Representation Bias:- Representation bias arises from how we sample from a population during data collection process . Non-representative samples lack the diversity of the population, with missing subgroups and other anomalies. Lack of geographical diversity in datasets like ImageNet (as shown in results in demonstrable bias towards Western cultures.

Aggregation Bias:- Aggregation bias (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population. An example of this type of bias can be seen in clinical aid tools. Consider diabetes patients who have apparent morbidity differences across ethnicities and genders. Specifically, HbA1c levels, that are widely used to diagnose and monitor

diabetes, differ in complex ways across genders and ethnicities. Therefore, a model that ignores individual differences will likely not be well-suited for all ethnic and gender groups in the population. This is true even when they are represented equally in the training data. Any general assumptions about subgroups within the population can result in aggregation bias.

Discrimination

Similar to bias, discrimination is also a source of unfairness. Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally, while bias can be considered as a source for unfairness that is due to the data collection, sampling, and measurement. Although bias can also be seen as a source of unfairness that is due to human prejudice and stereotyping, in the algorithmic fairness literature it is more intuitive to categorize them as such according to the existing research in these areas. In this survey, we mainly focus on concepts that are relevant to algorithmic fairness issues. contain more broad information on discrimination theory that involve more multidisciplinary concepts from legal theory, economics, and social sciences which can be referenced by the interested readers.

METHODS FOR FAIR MACHINE LEARNING

There have been numerous attempts to address bias in artificial intelligence in order to achieve fairness; these stem from domains of AI. In this section we will enumerate different domains of AI, and the work that has been produced by each community to combat bias and unfairness in their methods. It provides an overview of the different areas that we focus upon in this survey. While this section is largely domain-specific, it can be useful to take a cross-domain view. Generally, methods that target biases in the algorithms fall under three categories:

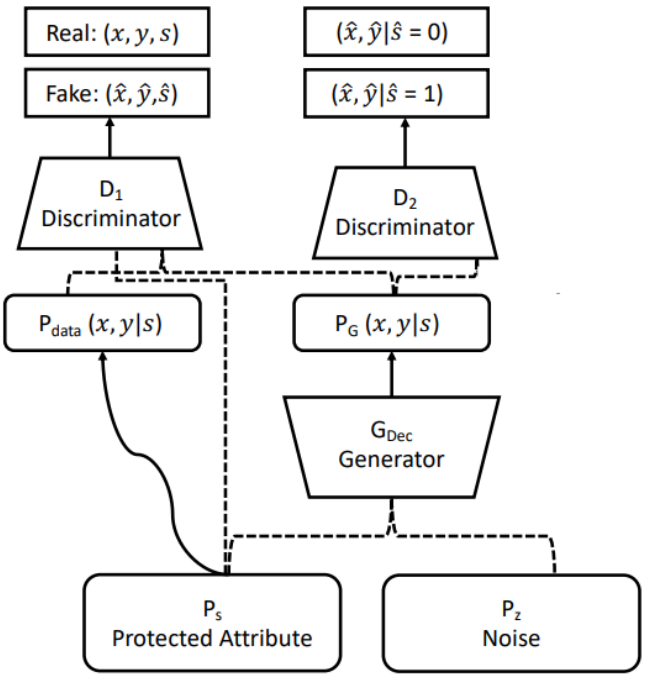
(1) Pre-processing. Pre-processing techniques try to transform the data so that the underlying discrimination is removed. If the algorithm is allowed to modify the training data, then pre-processing can be used.

(2) In-processing. In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used during the training of a model— either by incorporating changes into the objective function or imposing a constraint.

(3) Post-processing. Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase.

Algorithm	Reference	Pre-Processing	In-Processing	Post-Processing
Community detection	[104]	✓		
Word embedding	[23]	✓		
Optimized pre-processing	[27]	✓		
Data pre-processing	[76]	✓		
Classification	[159]		✓	
Regression	[14]		✓	
Classification	[78]		✓	
Classification	[155]		✓	
Adversarial learning	[90]		✓	
Classification	[63]			✓
Word embedding	[20]			✓
Classification	[125]			✓
Classification	[102]			✓

Algorithms categorized into their appropriate groups based on being pre-processing, in processing, or post-processing.



Structure of FairGAN as proposed

Comparison of Different Mitigation Algorithms

The field of algorithmic fairness is a relatively new area of research and work still needs to be done for its improvement. there are already papers that propose fair AI algorithms and bias mitigation techniques and compare different mitigation algorithms using different benchmark datasets in the fairness domain. For instance, a propose a geometric solution to learn fair representations that removes correlation between protected and unprotected features. The proposed approach can control the trade-off between fairness and accuracy via an adjustable parameter. In this work, authors evaluate the performance of their approach on different benchmark datasets, such as COMPAS, Adult and German, and compare them against various different approaches for fair learning algorithms considering fairness and accuracy measures. In addition, IBM's AI Fairness 360 (AIF360) toolkit has implemented many of the current fair learning algorithms and has demonstrated some of the results as demos which can be utilized by interested users to compare different methods with regards to different fairness measures.

OUR WORK

In this work we have taxonomized and summarized the current state of research into algorithmic biases and fairness—with a particular focus on machine learning. Even in this area alone, the research is broad. Subareas, from natural language processing, to representation learning, to community detection, have all seen efforts to make their methodologies fairer. Nevertheless, every area has not received the same amount of attention from the research community. Figure 7 provides an overview of what has been done in different areas to address fairness—categorized by the fairness definition type and domain. Some areas (e.g., community detection at the subgroup level) have received no attention in the literature, and could be fertile future research areas.

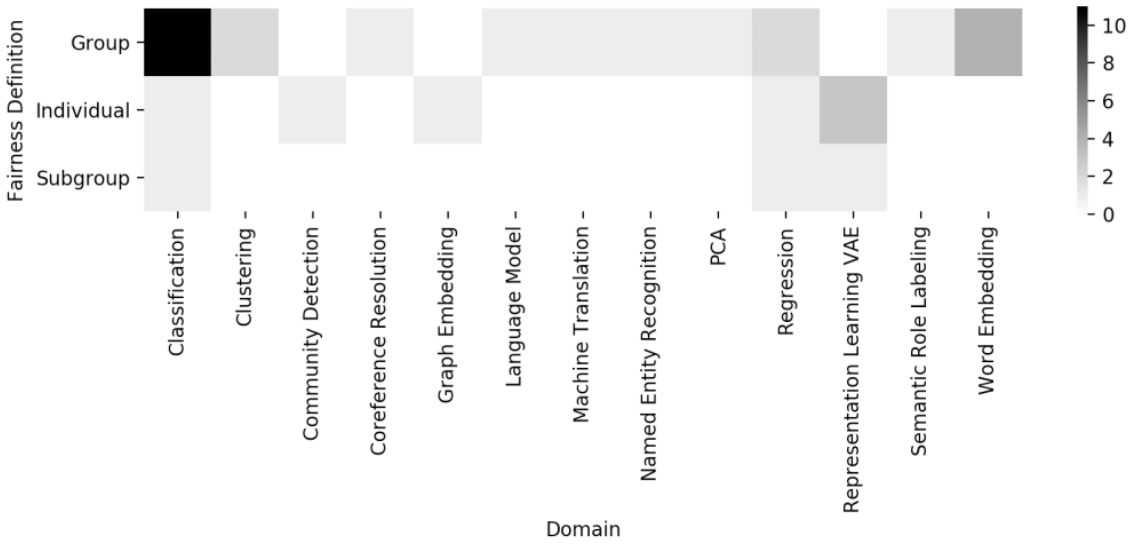


Fig. 7. Heatmap depicting distribution of previous work in fairness, grouped by domain and fairness definition.

CONCLUSION

we introduced problems that can adversely affect AI systems in terms of bias and unfairness. The issues were viewed primarily from two dimensions: data and algorithms. We illustrated problems that demonstrate why fairness is an important issue. We further showed examples of the potential real-world harm that unfairness can have on society—such as applications in judicial systems, face recognition, and promoting algorithms. We then went over the definitions of fairness and bias that have been proposed by researchers. To further stimulate the interest of readers, we provided some of the work done in different areas in terms of addressing the biases that may affect AI systems and different methods and domains in AI, such as general machine learning, deep learning and natural language processing. We then further subdivided the fields into a more fine-grained analysis of each subdomain and the work being done to address fairness constraints in each. The hope is to expand the horizons of the readers to think deeply while working on a system or a method to ensure that it has a low likelihood of causing potential harm or bias toward a particular group. With the expansion of AI use in our world, it is important that researchers take this issue seriously and expand their knowledge in this field. In this survey we categorized and created a taxonomy of what has been done so far to address different issues in different domains regarding the fairness issue. Other possible future work and directions can be taken to address the existing problems and biases in AI that we discussed in the previous sections.