Bias in Artificial Intelligence - Review

Abdullah Amawi
University of Göttingen
Göttingen, Germany

***Abstract*—In this report we will be exploring the bias in artificial intelligence, what is it, what are the sources, what are the types of it, and how to deal with it. Since in our current age we have many systems based on artificial intelligence (AI), it is important for us to understand as much as we can about artificial intelligence, how does it interact with us humans, how do we affect it and it affects us, and how we can utilize this understanding further to negate the downsides of some AI systems, one of those mainly negative factors in AI systems is AI bias. Moreover, those AI systems are deployed in many field today, such as healthcare, entertainment, recommender systems, natural language processing, image processing, image recognition, and much more; All this means that the currently deployed AI systems have a far reach, and a big impact on us as humans, therefore, it is crucial for us to understand how they are affected by bias and how the resulted biased AI system affect us.**

This report will be based on multiple works. And the goal is to give a detailed review of the topic of study and give the reader an overview of it to explore and get an overall idea about it and how to deal with it.

***Index Terms*—Artificial Intelligence, Human Bias, AI Bias.**

I. Introduction

In recent years, machine learning and specifically deep learning are increasingly becoming on of the top scientific topics both in information technology research and in industry, which is still incrementally advancing everyday; And we may note since deep learning is a subset of machine learning, and machine learning is a subset of the whole, which is artificial intelligence, meaning that when we speak about bias in deep learning, machine learning, and artificial intelligence are all kinds of bias in AI, so it is important to note that we may and can use that interchangeably. In order to understand bias in AI, we have to understand bias itself, what is it, where does originate?. Bias itself originates in humans, and since we humans develop AI systems, we tend to carry out our own bias and transfer it to the AI systems that we develop. According to Cambridge dictionary, we have bias and unconscious bias, where bias is defined as ”The action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment” [1], on the other hand Cambridge dictionary also defines unconscious bias as ” A type of bias that the person is not aware of but can influence the decisions of the person” [1]. As the definitions indicate, both bias and unconscious bias originate in a person, so it originate in people, us the humans, therefore, this explains how we humans can be biased and we may transfer our own biases to the AI systems. Understanding what is bias, bias in AI and where it originates empowers us to be able to try to negative bias and combat it, on the other hand there are techniques that are used in situations of AI bias for debiasing that we will go over in a later section. On the other hand, it is clear now that human and AI bias are related, therefore, it is normal to see that AI bias definition is related; AI bias is defined as ” an anomaly in the output of machine learning algorithms, due to the prejudiced assumptions made during the algorithm development process or prejudices in the training data” [2]; Moreover, AI bias is also known as algorithmic bias, which is the tendency of algorithms to reflect human biases [3]. Other also noted that we have underlying sources of bias in AI, such as how the National Institute of Standards and Technology(NIST) illustrated that in a figure that we will demonstrate here to show their own understanding on how we have systemic biases that are the underlying source of human biases, and then those human biases will give us statistical and computational biases, which is the AI bias in our case; Fig.1 demonstrates this.

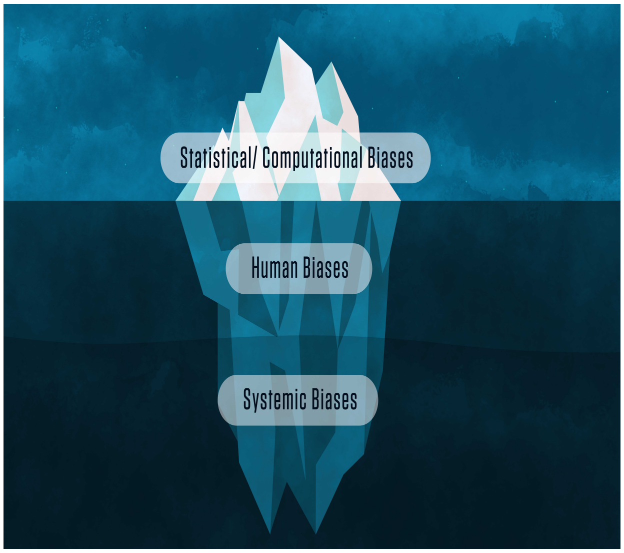

Fig. 1. Bias sources according to NIST [4]

When it comes to AI bias, there are so many sources and types, we will try to demonstrate multiple types and sources of AI bias in the upcoming sections; But we have to note that we should keep in mind that since the types and sources are highly correlated to humans and their own biases it results into making a comprehensive exhaustive list very hard or even impossible, therefore, we will provide a non-exhaustive list of the sources and types of AI bias.

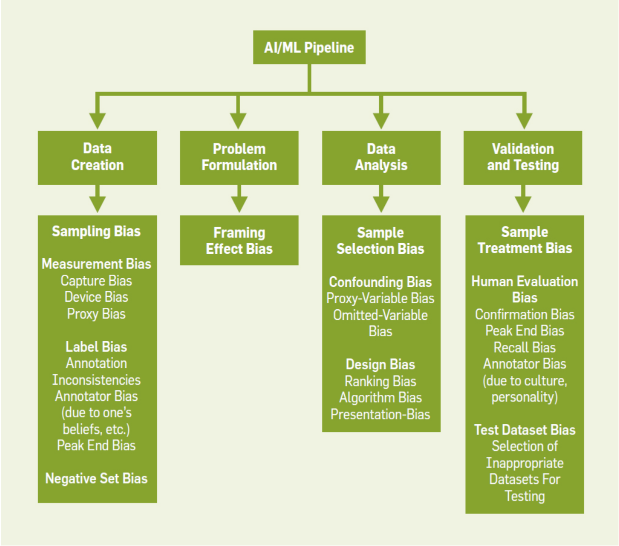

Fig. 2. Bias types in AI/ML pipeline according to ACM [6]

II. Types of AI Bias

When it comes to AI bias there are many types, and as indicated earlier, we can't really make an exhaustive list of all the types, not only because the types differ depending on the nature and the source, and since they are coupled with human biases, and we have so many types of human biases, and maybe more unclassified types since we still have a lot more to understand about ourselves; Therefore, we will introduce AI bias types according to the works of Mehrabi et al.,2022 [5] since it is a survey on bias in machine learning and sheds light on many types of AI bias, on the other hand, Fig.2 demonstrates how the ACM (Association for Computing Machinery) classifies the types of bias in AI.

The authors of Mehrabi et al.(2022) [5] survey found that the most important types of bias are divided to three main types and each type has a couple of sub-types under it, the main three were (1) **Data to algorithm**, which we will summarize here and rename to Data bias. (2) **Algorithm to User**, which explains the bias that is rooted in the algorithm behavior that is transferred to biases in the user behavior, which we will rename to Algorithm bias. (3) **User to Data**, which is the type of bias that the algorithm inherits from the user, we will rename it here to User bias. Here is a full summary of the types indicated in Mehrabi et al. works [5] renamed to our indicated main types and with a small summary of the sub-types for each one:

A. *Data Bias.*

It is a type of bias that occurs in the data itself, and when it is used, it might result to biased algorithms because of the data, not the design of the algorithm. The following sub-types of data bias is as mentioned in Mehrabi et al. work [5]:

1) **Measurement Bias.** It is a type of data bias that occur because of how we choose to measure specific features, it is also refereed to as reporting bias, an example of this is how in some states the statistics about crime can differ on how we measure them for minority groups.
2) **Omitted Variable Bias.** It is the case when there is one or more variables are omitted out of the model, for example, imagine we have a model to predict due to which reasons subscribers leave a service, then there is an external factor that was not considered before and omitted.
3) **Representation Bias.** This type of data bias occurs when the sample taken is biased itself, an example of that is the ImageNet dataset that lacks geographical diversity and is biased towards Western Geo-locations [5].
4) **Aggregation Bias.** Also called ecological fallacy, it is a type of bias that occurs when wrong conclusion are drawn about individuals because of the observations we made on an entire population; An example of this is when we conclude medical illnesses about individuals because it was observed in the population that they come from.
5) **Sampling Bias.** Sampling bias is very similar to representation bias due to the fact that it also demonstrates problems in the data sample, but in this case it occurs because of the sampling of sub-groups in a non-random manner.
6) **Longitudinal Data Fallacy.** It is a type of bias that occurs because of how researchers may analyze temporal data in a cross-sectional analysis, not regarding that it is temporal and they need to use longitudinal analysis due to the effects that the data changes over time. The works of Barbosa et al. [7] demonstrates this over an analysis of Reddit data.
7) **Linking Bias.** Linking bias occurs because of the links that are made about the users connections, activities, or interactions and may differ or misrepresent the actual behavior of the user according to Olteanu et al. work [8]. A prime example of this is social networks and how the network considers the links about the user, in this case the connections of other users, and not considering the user itself [9].

B. *Algorithm Bias.*

Biases that exist in the algorithm itself will affect the user and introduce biases to the user as indicated in Mehrabi et al. work [5], in this section we will list the sub-types of algorithm biases.:

1) **Algorithmic Bias.** According to Yates [10], algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm. Meaning that the data was not biased, but the choices of the design and functionality of the algorithm makes the outcome biased.
2) **User Interaction Bias.** Yates [10] also stated that"User Interaction bias is a type of bias that can not only be observant on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and

interaction"; It is worth noting that this type of bias can further affect other types, such as the presentation bias and ranking bias [5].

3) **Popularity Bias.** Nematzadeh et al. found that items that are more popular tend to be exposed more, but on the other hand, the metrics for popularity can be manipulated, an example of that is fake reviews [11]. Another example that we will mention in a later section is recommendation systems that have popularity bias.
4) **Emergent Bias.** Emergent biases emerge later, after the completion of the design, hence the name, those biases occur because of factors such as the change of population, other cultural values, and other factors that emerge later on [12].
5) **Evaluation Bias.** Evaluation bias happens during model evaluation [13], it is the use of inappropriate benchmarks for evaluation. An example of this is the use benchmarks in facial recognition systems that were biased towards skin color and gender [14].

C. *User Bias.*

1) **Historical Bias.** Historical biases are the bias type that already exist about issues in the world and can make its way to our data or algorithms [13]. An example of this is that historically CEOs were usually men and they make only 5% of the top 500 Fortune companies, which results in bias in searching for images that are related to CEO and makes the results biased towards men [13].
2) **Population Bias.** According to Olteanu et al. work, this type of bias exist and occurs because of features of the user population such as demographics, and user characteristics are different from the original target population [8]. An example of this in social platforms it is found that women are more likely to be more active in Pinterest, Facebook, Instagram, while men are more likely to be more active in Reddit or Twitter [15].
3) **Self-Selection Bias.** Is a sub-type of selection or sampling bias [5].
4) **Social Bias.** It is a type of bias that occurs when the actions of other people affect our judgement [10]. An example of this is how we rate or view items that we buy based on the view of others that affect our view of the item in question.
5) **Behavioral Bias.** A type of bias that can occur because of the difference in user behavior across different platforms, contexts, or datasets [8]. The work of Miller et al. [16] sheds further light on this type.
6) **Temporal Bias.** This type of bias arises because of the difference observed in populations and their behavior over time [8]. An example of this is the usage of hashtags in social media when discussing a hot topic, then the observance of that the population no longer uses the hashtag about that topic even if it is still in question [8].
7) **Content Production Bias.** It is a type of bias that occurs because of semantic and syntactic differences used in different users [8], such as different usage of language across different gender and age groups [5].

III. Bias Examples

In this section we will discuss multiple real world examples of bias in AI, those specific examples were hand-picked from many examples that can be found in research and industry due to how big is the impact of bias is in those fields.

A. *Bias In Healthcare*

Bias in AI can be damaging in many fields, in this example, we will demonstrate a study by Roosli et al. [17] that studied AI decision making effects inn the COVID-19 pandemic. The authors view that the heavy usage of AI in decision making in the pandemic was rushed to get new findings as fast as possible, they found that it was risky and could produce very biased predictions that would only do harm than good since they had unrepresentative datasets during the model development, moreover, the authors pointed out that not addressing those problems may even exaggerate the health disparities in the minority populations that were already facing the highest disease burden according to the study [17]. The authors also found that racial and ethnic minority groups suffer the most during the pandemic and even in general in the healthcare system, meaning that those AI systems that were already in use in the healthcare system were also biased against minority groups, and on top of that, the new AI models that were used for the COVID-19 pandemic were also biased making things worse.

Roosli et al. work [17] further finds that what they call as a "frenzy" in the research and publications that utilized AI in the early stage of the COVID-19 pandemic resulted in a "flood of non-peer-reviewed" [17] works. What made things worse is that they mention that bias levels were alarming in those works, and the PROBAST (Prediction model Risk Of Bias ASsessment Tool) study found that 66 models were screened as "high risk of bias" or "unclear". They further note that the most frequent problems were unrepresentative data samples, high model overfitting, and imprecise reporting of study populations and intended model use [17]. They also stress the need for balancing those datasets and models so we can get high-quality models that would benefit all the populations.

The authors concluded that even if the COVID-19 pandemic showed that our healthcare systems are struggling, the hasty usage of AI systems is not the answer if they were develop hastily and without understanding the risks of bias in them.

B. *Bias In Natural Language Processing*

Another major field that suffers from bias in its algorithms and its deep learning mdoels is NLP(Natural Language Processing), a study of Sun et al. [18] finds that NLP suffers a lot from AI bias, especially when it comes to gender bias. Furthermore, the study catagorized representation bias in NLP into four categories as Table 1. demonstrates how previous works in multiple sub-fields of NLP showed AI bias in them.

Task	Example of Representation Bias in the Context of Gender	D	S	R	U
Machine Translation	Translating ”He is a nurse. She is a doctor” to Hungarian and back to English becomes ”She is a nurse. He is a doctor”(Douglas,2017) [19]		X	X	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018) [20].		X	X	
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017) [21].			X	X
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018) [22].		X		
Language Model	“He is doctor” has a higher conditional likelihood than “She is doctor” (Lu et al., 2018) [23].		X	X	X
Word Embedding	Analogies such as “man : woman :: computer programmer : homemaker” are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016) [24].	X	X	X	X

TABLE I
REPRESENTATION BIAS IN NLP [18], FOUR CATEGORIES: (D)ENIGRATION, (S)TEREOTYPING, (R)ECOGNITION, (U)NDER-REPRESENTATION.

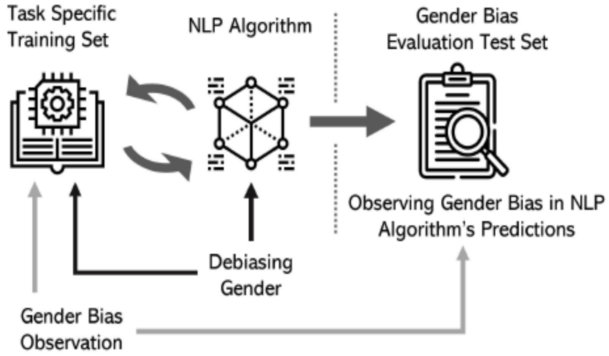
Task Specific Training Set
NLP Algorithm
Gender Bias Evaluation Test Set
Observing Gender Bias in NLP Algorithm's Predictions
Debiasing Gender
Gender Bias Observation

Fig. 3. Gender bias in NLP. Bias is observed in both training and test sets. Debiasing occurs in training, test sets, and the algorithm itself [18].

As demonstrated by Sun et al. [18] works, gender bias in NLP has many types and is a complicated issue; Moreover, the authors [18] stress that not only it is complex and a compound issue, they suggest that it requires interdisciplinary communication to solve or to ”debias”, and they think that we need solutions to debias current existing AI systems, in their paper, they review the methods of debiasing, recognizing bias, and mitigating bias in NLP systems, Fig.3 depicts debiasing as the authors illustrated it.

Moreover, Sun et al. [18] work demonstrated a couple of debiasing methods that are utilized for natural language processing, some of which was done using data manipulation, or on the other hand, debiasing by adjusting the algorithms.

*1) **Debiasing Methods Using Data Manipulation**:* The authors [18] mention that there are several methods for debiasing in NLP that revolve around either working on the text and their representations, or the prediction algorithm.

Debiasing Training Corpora

Three methods were reviewed by Sun et al. [18] work.

- **Data Augmentation**. For example, very often the dataset has gender bias by over-representing a gender over another, the idea of the data augmentation is to augment the dataset in a way that a new dataset that is biased to the opposite gender is created, so we can train on the original dataset and the ”gender-swapped” dataset, to get balanced training. More details are mentioned in [18].
- **Gender Tagging**. In tasks such as machine translation, the authors [18] found that current machine translation models predictions are disproportionate in favor of males since the training datasets are dominated by male data points. Gender tagging tries to solve this kind of bias for machine translation by adding a tag, hence the name, the tag indicates the gender of the data point object, so the machine translation can be more accurate, for example, we would tag ”I’m playing football” to ”MALE I’m playing football” to indicate to the machine translation which gender is it. More details are in Sun et al. [18] work.
- **Bias Fine-Tuning**. The idea of bias fine-tuning is that even though unbiased datasets are rare to find, they do still exist, so the idea behind this method is to use those unbiased datasets by learning from them using transfer learning to obtain a model that has no or minimal bias to use for the target task. Park et al. [22] details this method furthermore.

Debiasing Gender in Word Embeddings

According to Garg et al. [26], word embeddings is the way that we represent words in a vector space, but the problem lies within that it has been demonstrated that this method reflects societal biases. Two methods were reviewed by Sun et al. [18] work for Gender in word embeddings

- **Removing Gender Subspace in Word Embeddings**. This idea was first utilized by Schmidt [27] by removing the similarity to the gender subspace in word embeddings, Schmidt built a genderless framework using cosine similarity. Sun et al. [18] notes that this genderless framework

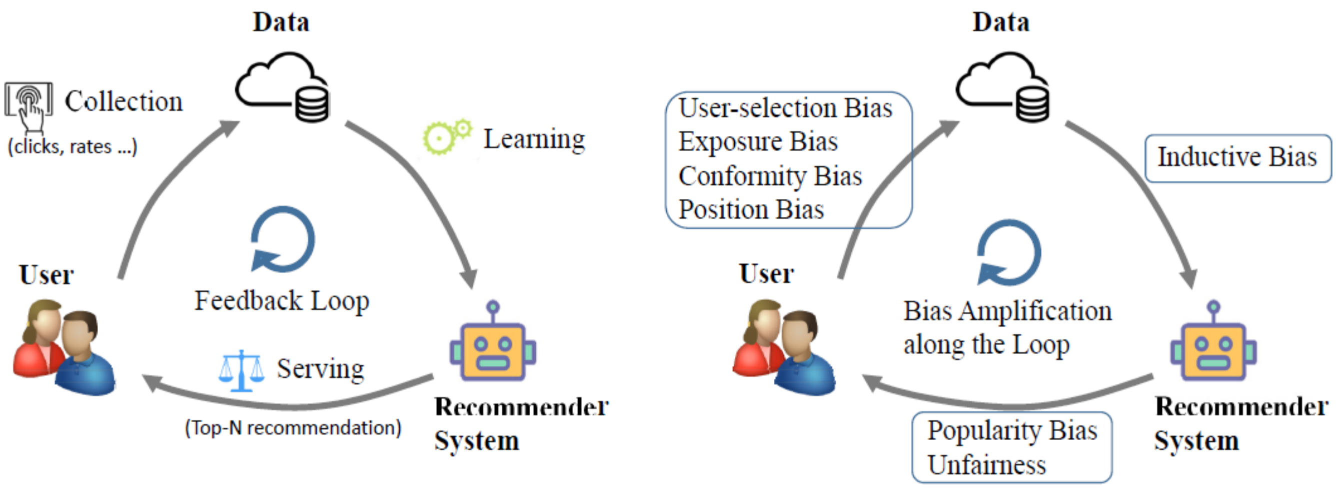

Fig. 4. Feedback loop in recommender systems and occurance in different stages [25].

may have been flawed, but others such as Bolukbasi et al. [24] have built upon it.

- **Learning Gender-Neutral Word Embeddings**. A newer method that was purposed in 2018 by Zhao et al. [28] work that basically isolates the gender specific information in word embeddings in a specific dimension and keeping the gender-neutral information in another dimension.

*2) **Debiasing by Adjusting Algorithms**:* Two methods were discussed in Sun et al. [18] work when it comes to debiasing by adjusting the algorithm itself.

Constraining Predictions

Previous works [29] found that NLP models have a high risk of amplifying bias when biases are present in the training dataset. for example, if the word ”Secretary” has a high correlation with females in the training dataset, it will have a very high chance this bias will be amplified in the algorithm. Zhao et al. [29] proposed RBA (Reducing Bias Amplification) by using a constrained conditional model that constrains that model optimization function to fit defined conditions [18].

Adversarial Learning: Adjusting the Discriminator

Zhang et al. [30] proposes a method that adjusts the algorithm discriminator from identifying the gender in the given task, the authors [18] also note that this method has the potential to be generalizable to debias any gradient-based model.

As we can see by this example, NLP is plagued by AI bias, but on the other hand, there are many NLP debiasing methods that were presented in Sun et al. [18] work.

C. Bias In Recommender Systems

On our last example that demonstrates bias in AI, is the field of recommender systems, the works of Chen et al [25] shows that AI bias heavily inflicts recommender systems, not only because AI bias exist there, but what makes things worse is the ”Feedback loop” which greatly amplifies bias in recommender systems. What is known as the feedback loop in recommender systems is how bias is amplified there because some of the popular items a recommended while ignoring others, these recommended items are used by the users, then this is logged by the system which makes it more biased, which is what feedback loop means [31]. Fig.4 demonstrates the feedback loop and all the other types of bias that are correlated to it. As we can see in Fig.4 Chen et al [25] work mentions multiple types of bias in recommender systems that we will briefly explain here, it is important to note that Chen et al. [25] work classified them into three categories:

*1) **Bias in Data**:* . It is the type of bias that exist in the data itself, and has multiple sub-types.

- **Data Bias**. It is the type of bias when the training data distribution is biased, or collected in a biased way.
- **Selection Bias**. Selection bias happens due to user ratings and selections, meaning it is biased due to the user selection itself not based on unbiased criteria.
- **Exposure Bias**. It is a type of bias that occurs due to the fact that the user is exposed to specific items more than others, for example, in recommender systems that recommend shows, the user may always see the ”top 10” and is exposed more to those, making his or her choice more biased towards that.
- **Conformity Bias**. Conformity bias is explained by its name, it is due to the fact that many users conform to others and their choices resulting in this conformity bias.
- **Position Bias**. It occurs due to the position of the items, similar to exposure bias provided example, the ”top 10” items have a better biased position in comparison to other items, creating higher tendency in the users to choose them.

*2) **Bias in Model**:* . The authors [25] note that bias is not always harmful and in some cases bias can be added on purpose to the model, which is known as inductive bias.

- **Inductive Bias**. It is the added assumption by the model, which is not a harmful type of bias used by the model

Types	Stages in Loop	Cause	Effect	Major solutions
Selection Bias	User→Data	Users' self-selection	Skewed observed rating distribution	Data Imputation; Propensity Score; Joint Generative Model; Doubly Robust Model

Exposure Bias	User→Data	Item Popularity;
		Intervened by systems;
		User behavior and background
			Unobserved interactions do not mean negative	Giving confidence weights by heuristic, sampling or exposure-based model;
				Propensity Score; Causality-based Model

Conformity Bias	User→Data	Conformity	Skewed interaction labels	Modeling social or popularity effect
Position Bias	User→Data	Trust top of lists; Exposed to top of lists	Unreliable positive data	Click models; Propensity Score; Trust-aware Model
Inductive Bias	Data→Model	Added by researchers or engineers	Better generalization, lower variance or Faster recommendation	-
Popularity Bias	Model→User	Algorithm and unbalanced data	Matthew effect	Regularization; Adversarial Learning; Causal Graph
Unfairness	Model→User	Algorithm and unbalanced data	Unfairness for some groups	Rebalancing; Regularization; Adversarial Learning; Causal Modeling
Bias amplification in Loop	All	Feedback loop	Enhance and spread bias	Break the loop by collecting random data or using reinforcement learning

TABLE II

TYPES AND CHARACTERISTICS OF BIASES IN RECOMMENDER SYSTESM AND THE AMPLIFICATION LOOP ACCORDING TO CHEN ET AL. [25]

to better learn the target function to be able to generalize beyond training data [25].

*3) **Bias and Unfairness in Results**:* . Another classification of bias in recommender system that the authors [25] did not consider as data or model bias has two sub-types as follows:

- **Popularity Bias**. It occurs when popular items are recommended more frequently because of their popularity.
- **Unfairness**. In some cases, the system is unfair or discriminatory in a systematic way against individual or groups on favor of others.

Table.2 demonstrates the characteristics of different type of biases in recommender systems and where they originate in the loop, their causes, effects, and proposed solutions.

IV. Debiasing & Mitigating Bias

Since we presented AI bias, its types, and major examples in multiple fields, in this section, we will present a couple of methods of debiasing, or how to deal with bias in AI systems. There are three approaches to deal with bias in AI systems as mentioned by Ntoutsi et al. [32] survey.

*1) **Preprocessing approaches**:* In this category, the focus here to mitigate bias is by focusing on the data itself. The goal is to produce a balanced dataset that we can use in any learning algorithm to use and obtain unbiased results. Methods exist to also modify existing datasets to produce a balanced dataset out of them, such methods can modify the original dataset distribution, modifying the weights, or selecting samples from sub-groups [32].

*2) **In-processing approaches**:* This approach focuses on modifying the in-processing, for example, modifying the model functions through regularization of by using constraints. Many methods that use in-processing approaches exist, dealing with the algorithm itself that the model is built upon, even though most of those approaches deal with classification problems, recently other unsupervised approaches emerged [32].

*3) **Post-processing approaches**:* The third approach is done post classification of the model in use, that is why it is a post-processing approach. Moreover, this approach is done differently if the model is a white-box approach or if it is a black-box approach, in the case of white-box models, it is done by changing the model internals, on the other hand, when we have a black-box model, it is done by changing the predictions [32]. Many studies addressed all three approaches of mitigating bias, Ntoutsi et al. [32] survey can be a good reference point to be able to to study and refer to the other works that addressed debiasing and mitigating bias.

Tools to Reduce Bias. There are also existing tools that can detect and reduce bias in our AI systems, many of them exist, but some of the major examples are:

- AI Fairness 360 [33]. Open source library from IBM, and can test bias in models and datasets.
- IBM Watson OpenScale [34]. Another solution from IBM, prforms bias checking in real time when the AI is making decisions.
- Google's What-If Tool [35].A Tool offered by Google, can improve fairness and visualize model behavior.

V. Conclusion

We believe that bias is an existing phenomena in humans that we deal with on daily basis, and we dealt with historically, and our problem is that since we did not solve bias in ourselves, and we may never do, we tend to carry our own biases into the AI systems that we develop. This makes it hard for us to combat AI bias, since this bias originates in us humans and there are many types both in humans and AI, but on the other hand, the more we learn about bias in ourselves it may shed more light in bias in our own developed AI systems. Moreover, not only we did introduce real life examples of AI bias, but also the sources, types, and debiasing methods that we may use to either solve or at least mitigate bias in our systems, and we also introduced some of the major tools to tackle AI bias and its sources hoping that we can always reducing and combat bias in our systems. But, it is important to note that unfortunately, bias is a very complicated issue that will need further studies, research, and continuous improvements in our AI systems.

References

[1] https://dictionary.cambridge.org/dictionary/english/bias.

[2] https://research.aimultiple.com/ai-bias/

[3] https://levity.ai/blog/ai-bias-how-to-avoid.

[4] https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

[5] Mehrabi, Ninareh and Morstatter, Fred and Saxena, Nripsuta and Lerman, Kristina and Galstyan, Aram. 2022. A Survey on Bias and Fairness in Machine Learning. https://doi.org/10.48550/arxiv.1908.09635. https://arxiv.org/abs/1908.09635

[6] https://cacm.acm.org/magazines/2021/8/254310-biases-in-ai-systems/fulltext

[7] Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M. Cesar-Jr. 2016. Averaging Gone Wrong: Using Time- Aware Analyses to Better Understand Behavior. (April 2016), 829–841.

[8] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

[9] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. 2009. User interactions in social networks and their implications. In Proceedings of the 4th ACM European conference on Computer systems. Acm, 205–218.

[10] Ricardo Baeza-Yates. 2018. Bias on the Web. Commun. ACM 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[11] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Filippo Menczer, and Alessandro Flammini. 2017. How algorithmic popularity bias hinders or promotes quality. arXiv preprint arXiv:1707.00574 (2017).

[12] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. ACM Trans. Inf. Syst. 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[13] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA,77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[15] Eszter Hargittai. 2007. Whose Space? Differences among Users and Non-Users of Social Network Sites. Journal of Computer-Mediated Communication 13, 1 (10 2007), 276–297. https://doi.org/10.1111/j.1083-6101.2007.00396.x arXiv:http://oup.prod.sis.lan/jcmc/article-pdf/13/1/276/22317170/jjcmcom0276.pdf

[16] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. “Blissfully Happy” or “Ready toFight”: Varying Interpretations of Emoji. In Tenth International AAAI Conference on Web and Social Media.

[17] ElianeRoosli, Brian Rice, and Tina Hernandez-Boussard. Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. 2020.

[18] Tony Sun, Andrew Gaut, Shirlyn Tangy, Yuxin Huangy, Mai ElSheriefy, Jieyu Zhaoz, Diba Mirzay, Elizabeth Beldingy, Kai-Wei Changz, andWilliam Yang Wang. Mitigating Gender Bias in Natural Language Processing: Literature Review. 2019.

[19] Laura Douglas. 2017. AI is not Just Learning our Biases; It Is Amplifying Them. https://bit.ly/ 2zRvGhH. Accessed on 11.15.2018.

[20] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, Anna Rohrbach, and Kate Saenko. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. European Conference on Computer Vision (EECV'18).

[21] Rachel Tatman. 2017. Gender and Dialect Bias in YouTube‘s Automatic Captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (ACL‘17), pages 53–59.

[22] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In Empirical Methods of Natural Language Processing (EMNLP‘18).

[23] Kaiji Lu, Piotr Mardziel, FangjingWu, Preetam Amancharla, and Anupam Datta. 2018. Gender Bias in Neural Natural Language Processing.

[24] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man Is to Computer Programmer As Woman Is to Homemaker? Debiasing Word Embeddings. In Neural Information Processing Systems (NIPS‘16).

[25] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2021. bias and debias in recommender system a survey and future directions. arXiv:2010.03240v2.

[26] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. Proceedings of the National Academy of Sciences, 115(16):E3635–E3644.

[27] Ben Schmidt. 2015. Rejecting the Gender Binary: A Vector-Space Operation. https://bit.ly/ 1OhXJM0. Accessed on 11.15.2018.

[28] Jieyu Zhao, Yichao Zhou, Zeyu Li,WeiWang, and Kai- Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In Empirical Methods of Natural Language Processing (EMNLP‘18).

[29] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints. In Empirical Methods of Natural Language Processing (EMNLP‘17).

[30] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES‘18).

[31] Masoud Mansoury, Himan Abdollahpouri, and Mykola Pechenizkiy. 2020. Feedback Loop and Bias Amplification in Recommender Systems.

[32] Eirini Ntoutsi,Pavlos Fafalios,Ujwal Gadiraju,Vasileios Iosifidis,Wolfgang Nejdl,Maria-Esther Vidal,Salvatore Ruggieri,Franco Turini,Symeon Papadopoulos,Emmanouil Krasanakis,Ioannis Kompatsiaris,Katharina Kinder-Kurlanda,Claudia Wagner,Fariba Karimi,Miriam Fernandez,Harith Alani,Bettina Berendt,Tina Kruegel,Christian Heinze,Klaus Broelemann,Gjergji Kasneci,Thanassis Tiropanis,Steffen Staab. 2020. bias in data-driven artificial intelligence systems—an introductory survey.

[33] https://github.com/trusted-ai/aif360

[34] https://www.ibm.com/cloud/watson-studio

[35] https://pair-code.github.io/what-if-tool/