International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 1978- 1985
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0270

Harmonizing Intelligence: A Holistic Approach to Bias Mitigation in Artificial Intelligence (AI)

Isha Mishra[1], Vedika Kashyap[2], Nancy Yadav[3], Dr. Ritu Pahwa[4]

[1,2,3]UG, Computer Science Engineering (AI&ML), Dronacharya College of Engineering, Gurugram, Haryana, India

[4]Associate Professor, Computer Science Engineering (AI&ML), Dronacharya College of Engineering, Gurugram, Haryana, India

Emails: *ishamishra0421@gmail.com[1], vedikakashyap33@gmail.com[2], nancyrao22222@gmail.com[3], ritu.pahwa@ggnindia.dronacharya.info[4]*

Abstract

Artificial intelligence (AI) is transforming the way we interact with data, leading to a growing concern about bias. This study aims to address this issue by developing intelligent algorithms that can identify and prevent new biases in AI systems. The strategy involves combining innovative machine-learning techniques, ethical considerations, and interdisciplinary perspectives to address bias at various stages, including data collection, model training, and decision-making processes. The proposed strategy uses robust model evaluation techniques, adaptive learning strategies, and fairness-aware machine learning algorithms to ensure AI systems function fairly across diverse demographic groups. The paper also highlights the importance of diverse and representative datasets and the inclusion of underrepresented groups in training. The goal is to develop AI models that reduce prejudice while maintaining moral norms, promoting user acceptance and trust. Empirical evaluations and case studies demonstrate the effectiveness of this approach, contributing to the ongoing conversation about bias reduction in AI.

Keywords: *Artificial intelligence, Bias, Machine-learning techniques, Ethical considerations, Interdisciplinary perspectives, Robust model evaluation techniques, Empirical evaluation.*

1. Introduction

A lot of companies have decided to invest in AI due to its potential for improvements and efficiencies [1] [2]. Nevertheless, the threat of inadvertent bias and its potential for damage can adversely affect a business's standing. There's good reason to exercise caution when dealing with this kind of risk and the associated lawsuit exposure. Algorithmic bias not only entails these dangers but also has the potential to result in subpar performance from an application, which can mean lost opportunities. When bias in lending procedures, for exampleunjustly discriminates against someone, it can result in loss of financial benefit as well as loss when prejudice gives preference to some individuals over others. The problem is exacerbated by the fact that a large gives preference to some individuals over others.in The problem is exacerbated by the fact that a large number of AI systems are not "explainable" in a way that allows deployers to assert that bias and of the other types of errors are not present [3] [4]. When making important business decisions, managers sometimes struggle to accept conclusions that lack justification since training data may contain subtle traces of bias and other poor decision-making habits [5] [6]. With so much at risk, it's necessary to develop a set of best practices to assist those deploying AI systems in avoiding these potential dangers. We offer a framework to identify, measure, and reduce bias sources as a first step in this direction. We start by reviewing the possible causes and effects of bias in AI systems. Next, we outline procedures that businesses can use to control bias in order to enhance AI results and give regulators, customers, and employees confidence in the findings. system may be biased as early as when the employer chooses a specific algorithmic objective from a potentially vague target. For instance, the true objective of an Advertising.

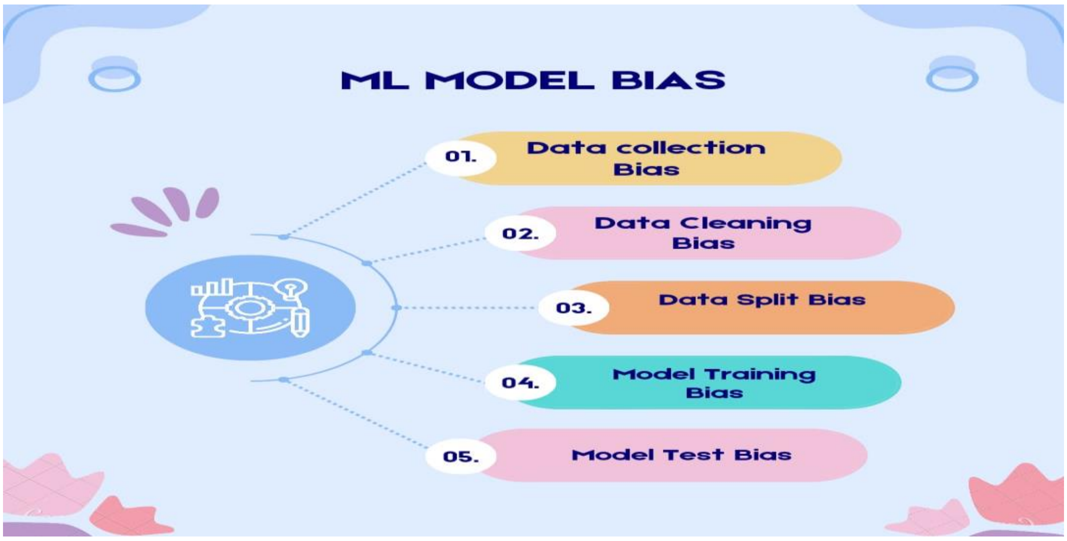

Figure 1 **A Guide to Different Bias Mitigation Techniques in Machine Learning**

2. Major Challenges

A variety of factors, including the method and input properties selected as well as the undiscovered correlations in the training set, can introduce bias into AI systems. We outline three broad categories of bias: those resulting from sample distribution, those connected to integrating business objectives into AI implementations, and utilized in training (including historical impacts) as well as those found in specific input samples.

2.1 Problems with Goal Representation

An AI system may be biased as early as when the employer chooses a specific algorithmic objective from a potentially vague target. For instance, the true objective of an Advertising targeting potential clients who are most likely to buy their product may be the focus of a firm. Companies need to determine which hypothesis, input properties, training labels, and reinforcement criteria will best achieve this aim because there is no simple method to convert this into an AI implementation [8]. For instance, a business that sells video games can assume that young males are the target market for their product and target consumers who fit this description, such as guys between the ages of 15 and 25.

Objectives of Proxy: In the aforementioned scenario, as it is impossible to pinpoint a customer's precise chance of making a purchase, the company may decide to focus on choosing certain individuals. Having characteristics akin to those of clients who have previously bought a comparable product. When attempting to promote new product features or break into new markets, this might not be the best option. The algorithm will always be subject to historical bias if the target selection is based solely on historical data without taking appropriate context into account. For instance, consumers from some areas might not have bought the goods because it wasn't previously advertised there. Similar to this, modifications to the product's attributes, cost, or outside fashions could make it appealing to buyers who didn't buy the prior iteration.

Selection of Features: The selection of which qualities to include may have been the most obvious instance of prejudice on the part of the mapping creators. For instance, a university's admissions process might consider recommendations from letters of recommendation, class rank, GPA, and results on standardized tests. Though the final objective might be the same (for example, forecasting college achievement), the characteristics used can lead to radically different conclusions. It is possible for ostensibly neutral input features to have unnoticed biases [5, 9]. Bias resulting from features that are left out but could positively affect some people's predictions if included is even harder to measure. It could be more difficult to translate extra information, such firsthand observations from a recommendation letter, into measurable and clearly defined attributes.

Data on Substitutes: Large amounts of mathematical data are required for AI models. These attributes are restricted to those that developers can readily obtain at scale, as training sets have to be huge and quantitatively represented. This could imply that a credit score is used by a job screening agency as a stand-in for a quality like "reliability" instead of letters of recommendation [18][19]. Such mathematical reductions could cause loss of information, which would cause the problem mapping to become distorted. Since surrogate data act as stand-ins for restricted input—such as zip codes for race, magazine subscriptions for gender or race, and purchasing patterns for health conditions—bias may be introduced.

2.2 Problems with Data Sets

In addition to datasets that pose problems during the mapping process, there can be concerns with training or production datasets. Making training sets is a laborious process.Preening a sizable data set is usually required [10], and for supervised Getting labels is a part of learning. In deep learning systems, this can mean making sure that uncommon cases are disproportionately overrepresented in order to provide the model with enough training opportunities to encounter such cases. Due to its size, complexity, and sometimes urgency, the process of creating a training data set may make up most of the effort required for AI systems and is often the source of problems [11]. Furthermore, the AI system is vulnerable due to the manipulable nature of training datasets [13].

Extraordinary Cases: The ability of AI systems to successfully generalize reactions to a variety of stimuli is a significant benefit. This can, however, work against the system if it encounters a class for which it was not designed. For instance, a neural network that has been trained to identify texts as German or English would continue to respond in place of expressing uncertainty when it came across a French sentence. These issues could lead to "silent" or "hidden" mispredictions, which could subsequently proliferate to further damage to the business application.

Inconsistent Data Sets: The model is unlikely to function properly if the data used in training and production are very different. To elaborate on the above point, commercial facial The accuracy of recognition algorithms, which are primarily trained on fair-skinned participants, varies greatly depending on the population: 34.7% of women with darker complexion and 0.8% of males with lighter skin are reported in [19]. Even if the model was originally trained on a dataset that corresponds to its intended purpose, the production data may change over time due to a number of variables, including seasonal fluctuations or external trigger events. Any such modification might have unanticipated consequences brought forth by mismatched data sets.

Data that has been manipulated: Training data can be manipulated to skew the results, as evidenced by the brief existence of the chatbot Tay, which quickly mimicked the hate speech of its Twitter followers [12, 31]. Programs created with small, public data sets are especially vulnerable to attacks of this kind. Analogously, data poisoning is acknowledged as a security problem in AI [13].

Unlearned Cases: Even highly trained models are not infallible; in fact, a high degree of accuracy would probably be the consequence of overfitting the data, indicating a poor likelihood of the model's ability to generalize to new cases. Because of this, even highly trained models will underperform on some classes of samples. Research has indicated that facial recognition datasets with inadequate representation of different ethnic groups might lead to trained models exhibiting significantly varied racial accuracy [14].

Ungeneralizable Characteristics: Model makers may decide to employ well-prepped subsets of their anticipated production data sets for training in order to get around the practical difficulties involved in creating sizable, labeled training sets. This might result in the training set's unique qualities being given more weight than those that apply to bigger datasets. For example, [15] shows how the training set's word distributions drive text classifiers to emphasize irrelevant phrases like "POST" in their classifications. Using standard newsgroup training sets, these classifiers were trained to classify articles as either "Christian" or "atheist".

Irrelevant Correlations: Predictions may be off if irrelevant input features and the outcome have correlations in the training data. For instance, using photos of dogs without snow and wolves in the snow, Ribeiro et al. trained a classifier to distinguish between dogs and wolves. Occasionally, following training, the model thinks that a dog in a snowy environment is a wolf [15].The distribution of meaningless correlations may not be exclusive to the training set, but rather occur in real-world data, notwithstanding their nongeneralizable characteristics. It could be that wolves are more likely than dogs to be found in the snow. However, the projection would be incorrect if the feature had any bearing on it because wolves are wolves no matter where they reside, including Grandmother's house.

Problems with the Use of Historical Data: Artificial intelligence systems have to learn from the past. Sadly, this entails picking up on the prejudices held by others there [5] and possibly passing on chances brought about by shifting surroundings [9].

2.3 Limitations of Special Sampling

Problems that can be identified by looking at the data from a single sample are categorized as solitary sample problems. The issue could be exclusive to that sample or widespread across all of the samples. When the data sets contain sensitive personal information that prevents the complete set from being made public, this classification is crucial. visible, but where people could be able to examine their own personal information.

False Information: To make sure the model learns properly. training data is frequently carefully selected. Regretfully, real-world data is rarely so pure. It could be tainted or lacking. Data that has been manually entered can be wrongly [16]. Inaccurate sources may be included in data that is automatically gathered [17].

Outdated Information: It's possible that the data used for production input and training is outdated. This might be especially true when big "dictionaries" are stored in the cache for quick access. As an illustration, Credit reports could be quickly accessed by downloading and storing them locally from an outside source. Unfortunately, since changing the dataset could reset the baseline for ongoing training experiments, developers could be reluctant to do so. sure that uncommon cases are disproportionately overrepresented in order.

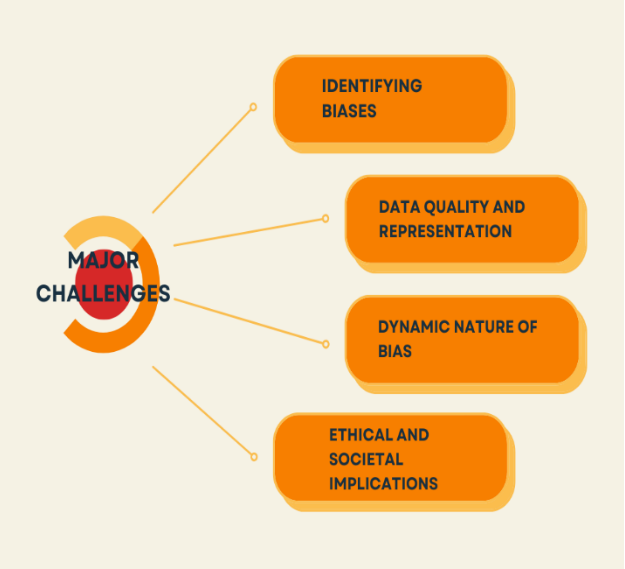
MAJOR CHALLENGES
IDENTIFYING BIASES
DATA QUALITY AND REPRESENTATION
DYNAMIC NATURE OF BIAS
ETHICAL AND SOCIETAL IMPLICATIONS

Figure 2 The Image Illustrates the Major Challenges in Bias Mitigation in AI

3. Managing Bias

It might be intimidating to face how to address the multitude of recognized possible sources of bias (and more are being found as the area develops). Given the multitude of problems that lead to bias creeping into the system, we can't expect one solution to address them all. Rather, we suggest combining evaluations, monitoring, data review, controlled experiments, business procedures, and quantitative assessments. We create certain ground rules for the procedures we wish to incorporate before going into detail about the aforementioned stages.The first requirement is that any procedure used to assess an AI system for bias must be feasible for people who aren't the main developers. This is significant because it could be necessary for non-technical management to understand the system or for an auditor or regulator to assess it. It's possible that even the primary developer is unaware of the inner workings of the model because complicated models are becoming more widely available from outside sources [16, 17]. As such, we highlight that none of our procedures necessitate an understanding of the inner workings of the model; instead, we simply take into account the input and output data when assessing the AI pipeline in its whole, feature engineering apart. Transparency in the data used as input is also crucial. The only method to confirm that the data is correct and free of inaccurate or protected info is to perform this. The individual to whom the data refers should have access to the information, even when it is private [18]. Methods for data versioning, data cataloging, tracking, and governance are also critical to ensure that the exact dataset used to train a particular model can always be accessed and examined. We organize the procedures into phases based on when they are most likely to be used throughout the AI system's deployment lifecycle. Nonetheless, as the phases are probably going to overlap and repeat throughout the course of these activities, this is mostly an organizing strategy for planning utilization.

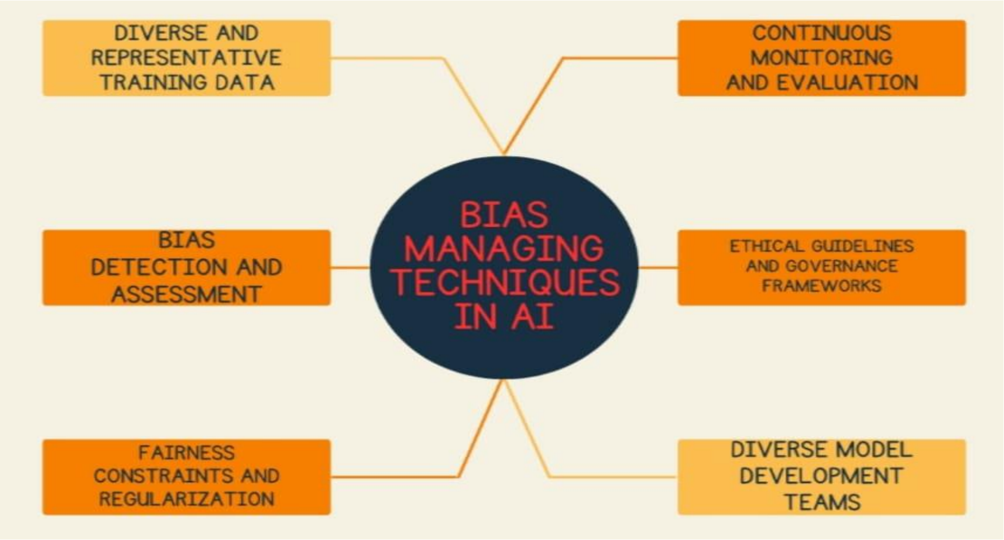
DIVERSE AND REPRESENTATIVE TRAINING DATA
CONTINUOUS MONITORING AND EVALUATION
BIAS DETECTION AND ASSESSMENT
BIAS MANAGING TECHNIQUES IN AI
ETHICAL GUIDELINES AND GOVERNANCE FRAMEWORKS
FAIRNESS CONSTRAINTS AND REGULARIZATION
DIVERSE MODEL DEVELOPMENT TEAMS

Figure 3 The Imaging Depicts Various Strategies and Technologies to Manage Bias in Artificial Intelligence (AI) System

4. Strategies to Mitigate Bias

Concerns among healthcare professionals are growing that AI systems may reveal, perpetuate, or even amplify prejudice. Variations in a subgroup's performance on a prediction task are commonly used to define bias [6, 7]. For instance, there may be a performance difference in an AI system that predicts a patient's future risk of breast cancer, making it more probable for black patients to receive the wrong classification of "low risk." Furthermore, because patient populations, treatment modalities, and prescription regimens may differ in the USA, an algorithm based on hospital data from German patients may not work well there. Healthcare systems have already observed incidents similar to this one. There could be a wide range of causes behind this disparity in performance. During the many stages of developing an AI model, such as data preparation creation, assessment, and use of the model in clinical environments [9]. This specific example suggests that the algorithm might have been mostly trained on White patient data, or Black patient health records might be harder to obtain. Moreover, the training of a model to predict risk may be impacted by underlying societal disparities in healthcare spending and access [6] [10]. Whatever the reason, the effect of an algorithm that disproportionately assigns false negatives would be to increase the number of undiagnosed/untreated cancer cases and decrease the number of follow-up scans, hence exacerbating health disparities for already marginalized communities. and gathering, bias might arise.

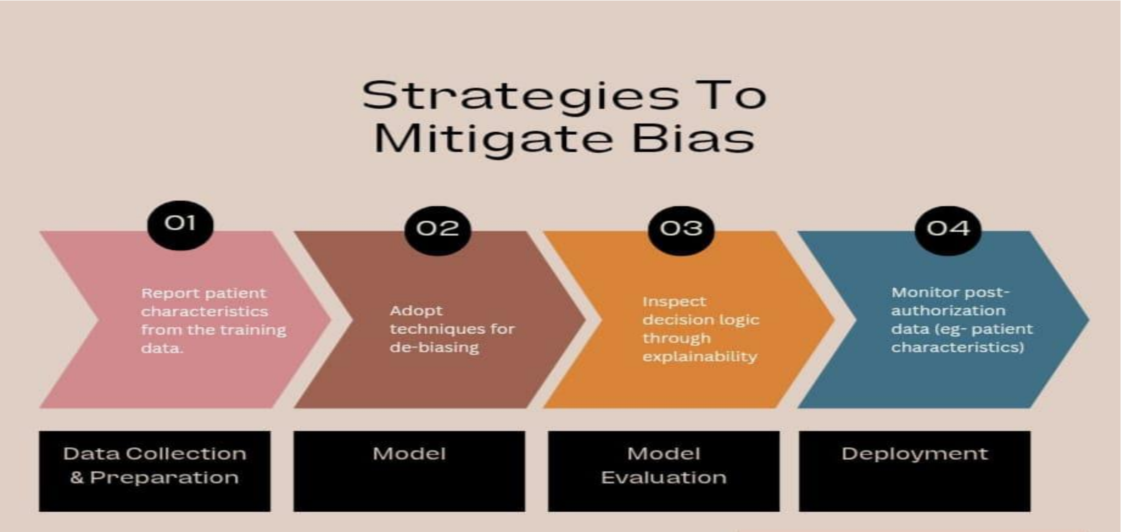
Strategies To Mitigate Bias
01
Report patient characteristics from the training data.
02
Adopt techniques for de-biasing
03
Inspect decision logic through explainability
04
Monitor post-authorization data (eg- patient characteristics)
Data Collection & Preparation
Model
Model Evaluation
Deployment

Figure 4 **The Image Aims to Show the Strategies to Mitigate Bias at Various Levels**

5. A Vision to The Future: Troubles with Automated Learning Ai Models

With the advancement of AI technology, bias in AI algorithms has to be examined further and mitigated. What level of bias is appropriate for an AI-created program is an issue that will surely come up in the future [20]. This is akin to asking what accuracy threshold a given AI system [20] can function at, supposing that it is hard to construct bias-free systems before implementation, even though prior groups have suggested that any performance gap is symptomatic of algorithmic bias [21]. The data and population that an AI system is trained on and then deployed to might also affect performance disparity. The datasets that AI algorithms are trained on within the categories of algorithm types itself now exhibit a great deal of variety [22] [23]. It is still unclear whether AI algorithms should be more localized and used more narrowly, or more generalizable and trained on bigger and more varied datasets to be applied to wider populations. In any event, in order for these issues to be researched and discussed in the upcoming years, AI models will need to be visible and explainable [24]. Another issue for the future is the capacity of AI algorithms to be changed or adjusted, much like how Kiyasseh et al. [20] added TWIX to their pre-existing SAIS algorithm. AI algorithms can be either adaptive or locked, meaning that once trained, the model will always provide the same result given the same input [25]. In this case, the AI model may be updated continuously rather than becoming outdated quickly as it learns from new data over time. However, if the incoming data are biased, there is a chance that continuous learning can exacerbate the bias already present or introduce new ones [26]. Thus, the key to implementing AI will be developing methods for ongoing bias reduction and regular bias detection.

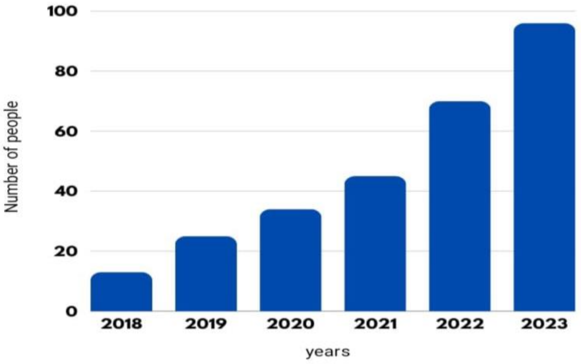

Figure 5 **The Image Illustrates the Increasing Rate of Bias in AI Over Time**

Conclusion

In the upcoming years, there will be a greater and greater incorporation of AI into medical technology and healthcare systems. Bias avoidance will be essential to the usability and integration of AI models. Kiyasseh & al. provide a novel method of mitigating prejudice using their TWIX technology. Bias reduction is being pushed at every stage of technology development, from model creation and overtraining to deployment and execution. Healthcare facilities, regulatory bodies, and inventors will all need to provide checks and balances on this endeavor.

References

[1]. Witten, H., Frank, E., and Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann Publishers Inc., San Francisco, CA.

[2]. McKinsey Global Institute (2017). “Artificial Intelligence: The Next Digital Frontier?”

[3]. Goodman, B. and Flaxman, S. (2016). “European Union regulations on algorithmic decision-making and a 'right to explanation'”. ICML Workshop on Human Interpretability in Machine Learning, New York, NY.

[4]. Kahn, J. (2018). “Artificial Intelligence Has Some Explaining to Do”. Bloomberg Businessweek, 12 Dec 2018.

[5]. Bolukbasi. T., Chang, K., Zou, J., Saligrama, A., and Kalai, A. (2016). “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings”. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain.

[6]. Angwin, J., Larson, J, Mattu, S., and Kirchner, L. (2016). “Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks”. ProPublica, 23 May 2016.

[7]. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. (2018). “Runaway Feedback Loops in Predictive Policing”. Proceedings of Machine Learning Research.

[8]. Hao, K. (2019). “This is How AI Bias Really Happens — and Why It’s so Hard to Fix”. MIT Technology Review, 4 Feb 2019.

[9]. Dastin, J. (2018). “Amazon scraps secret AI recruiting tool that showed bias against women”. Reuters Business News, 10 Oct 2018.

[10]. Del Balso, M. and Hermann, J. (2017). “Meet Michelangelo: Uber's Machine Learning Platform”. Uber Engineering, 5 Sep 2017.

[11]. Baylor. D. et al. (2017). “TFX: A TensorFlow-Based Production-Scale Machine Learning Platform”. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada.

[12]. Vincent, J. (2016). “Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day”. The Verge, 24 Mar 2016.

[13]. Bursztein, E. (2018). “Attacks against machine learning - an overview”.

[14]. Buolamwini, J. (2019). “Aritificial Intelligence Has a Problem with Gender and Racial Bias”. TIME, 7 Feb 2019.

[15]. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). “’Why Should I Trust you?’: Explaining the Predictions of Any Classifier”. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA.

[16]. Bier, D. (2017). “E-Verify Has Delayed or Cost Half a Million Jobs for Legal Workers”. CATO Institute, 16 May 2017.

[17]. Suhartono, H., Levin, A., and Johnsson, J. (2018). “Why Did Lion Air Flight 610 Crash? New Report Details Struggle”. Bloomberg, 27 Nov 2018.

[18]. 0'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown. New York, NY.

[19]. Buolamwini, J. and Gebru, T. (2018). “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY.

[20]. Kiyasseh, D. et al. Human visual explanations mitigate bias in AI-based assessment of surgeon skills. NPJ Digital Med. 6, 54 (2023).

[21]. Townson, S. Manage AI Bias Instead of Trying to Eliminate It. https:// (MIT Sloan

Management Review, 2023).

[22]. Gubatan, J. et al. Artificial intelligence applications in inflammatory bowel disease: emerging technologies and future directions. World J. Gastroenterol. 27, 1920–1935 (2021).

[23]. Moglia, A., Georgiou, K., Georgiou, E., Satava, R. M. & Cuschieri, A. A systematic review on artificial intelligence in robot-assisted surgery. Int. J. Surg. 95, 106151 (2021).

[24]. Theunissen, M. & Browning, J. Putting explainable AI in context: institutional explanations for medical AI. Ethics Inf. Technol. 24, 23 (2022).

[25]. Benjamens, S., Dhunnoo, P. & Mesko, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digital Med. 3, 118 (2020).

[26]. Decamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. J. Am. Med. Inform. Assoc. 27, 2020–2023 (2020).