Machine Bias: A Survey of Issues

Ana Farič *, Ivan Bratko
Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
E-mail: af27987@student.uni-lj.si, bratko@fri.uni-lj.si
*Corresponding author

Keywords: machine learning, artificial intelligence, bias, fairness, discrimination, COMPAS

Received: April 1, 2024

Some recent applications of Artificial Intelligence, particularly machine learning, have been strongly criticised in general media and professional literature. Applications in domains of justice, employment and banking are often mentioned in this respect. The main critic is that these applications are biased with respect to so called protected attributes, such as race, gender and age. The most notorious example is the system COMPAS which is still in use in the American justice system despite severe criticism. The aim of our paper is to analyse the trends of discussion about bias in machine learning algorithms using the COMPAS as an example. The main problem we observed is that even in the field of AI, there is no generally agreed upon definition of bias which would enable operational use in preventing bias. Our conclusions are that (1) improved general education concerning AI is needed to enable better understanding of AI methods in everyday applications, and (2) better technical methods must be developed for reliably implementing generally accepted societal values such as equality and fairness in AI systems.

Povzetek: Analizirali smo trende v diskusijah o pristranskosti odločitev strojnega učenja, kjer smo za primer vzeli sistem COMPAS.

1 Introduction

With the widespread use of machine learning, there have been cases in the last 5 to 10 years where applications received significantly negative feedback for being biased, primarily from the general media and also within professional literature. Typical applications come from domains such as judiciary, employment, and banking. Critics warn that "machine learning algorithms and systems are unfair and biased" with respect to so-called protected attributes, such as race, gender, and age of an individual. They argue that artificial intelligence recommendations depend on these attributes rather than on the objective evaluation of facts [15].

Some noteworthy article headlines describing discriminatory practices allegedly promoted by machine learning algorithms include: "There's software used across the country to predict future criminals. And it's biased against blacks [2]," "New Zealand passport robot tells applicant of Asian descent to open eyes [24]," "A beauty contest was judged by AI and the robots didn't like dark skin [19]," and "Amazon scraps secret AI recruiting tool that showed bias against women [10]." Such examples contribute to escalating concerns (and sometimes panic) about the potentially harmful impacts of artificial intelligence (AI) on our lives [20]. Experts from various fields address the issue of machine learning bias, attempting to define what bias means, where it originates, and, most importantly, what should be done about it.

In the evolving field of ethics in AI (e.g., UNESCO 2021 [26]), the topic of machine learning bias prominently appears. Policymakers often mention it in relation to regulatory principles aimed at ensuring the ethical use of AI (e.g., European AI Act, 2023, 2024 [3]). However, in these discussions, it is often not clear what exactly machine learning bias and AI bias mean. Therefore, regulatory measures in this direction are not clearly defined, except in a very abstract form. The term bias in relation to machine learning means different things to different authors. Even in the AI literature, there is no complete consensus and no universally accepted technical definitions of bias that could be operationally used to prevent bias [15]. For various meaningful definitions of bias, it has even been mathematically proven that, except in special cases, they cannot be satisfied simultaneously [17].

In this paper, we review various definitions of bias and different opinions on how to address the problem most effectively in practice. The conclusions converge towards the idea that addressing bias appropriately requires considering societal values and operationalizing them through interdisciplinary collaboration with a democratically accepted social agreement in the form of appropriate legislation. Better general education on AI and its methods would contribute to a better general understanding of bias in AI in practice. The lack of uniformity in dealing with bias in AI will be in this paper demonstrated using the example of the COMPAS system [2, 11, 12, 16].

2 COMPAS

The COMPAS system is considered in a series of publications (Figure 1) as arguably the most controversial case illustrating the bias of AI. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a decision-making system used by many American courts where judges assess the risk of recidivism, specifically estimating the likelihood that an offender will reoffend within two years if released. COMPAS was developed by an American company, then known as Northpointe (Equivant today). COMPAS takes into account 137 attributes for each offender, obtained either from the individual or their criminal record. This data is analyzed by a specific algorithm, which, as a trade secret of the company, is not publicly known. Based on this analysis, the algorithm provides a score ranging from 1 to 10, where a higher score indicates a higher risk of recidivism.

Figure 1 illustrates 10 highly cited articles on this system and the mutual citation between articles. An arrow from paper A to paper B indicates that paper B is cited in paper A.

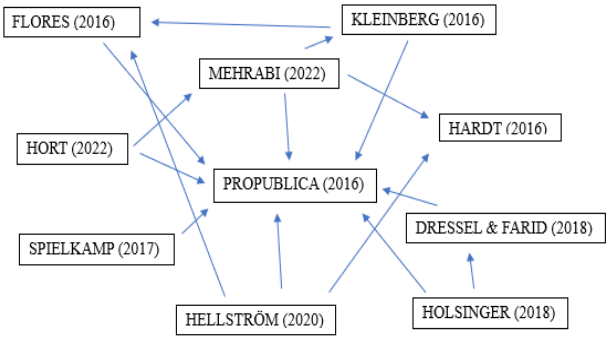

Figure 1: Interconnectedness of publications on the COMPAS system.

At the center of the graph is an article from the ProPublica newsroom [2], which, according to [1], sparked interest in studying bias in AI. In [2], a group of investigative journalists described their analysis of the COMPAS system and experiments with real data on over 7,000 defendants from Broward County, Florida in the years 2012 and 2013. They found that only 61% of people assessed as likely to reoffend actually did so. In further analysis, they focused mainly on the racial aspect, concluding that the program is biased against African American defendants. They monitored how many of them were re-convicted in the next two years and compared predictions with actual outcomes. 44.9% of African Americans marked as high risk did not reoffend. In contrast, 47.7% of whites marked as low risk reoffended within two years. These two metrics for system mispredictions are commonly referred to as (1) FPR (false positive rate), which is the proportion of the negative class incorrectly predicted as positive, and (2) FNR (false negative rate), which is the proportion of the positive class incorrectly predicted as negative. Complete results regarding the FPR and FNR rates in the COMPAS system predictions were:

	White	Black
FPR	23.5%	44.9%
FNR	47.7%	28.0%

These results were interpreted as evidently biased against African American defendants, leading to an assessment of the use of the COMPAS system as inappropriate and discriminatory [2]. This conclusion seems quite justified.

An additional issue highlighted in [2] is the fact that the decision criterion used by COMPAS is not transparent, as the algorithm is protected as a trade secret. COMPAS itself does not provide an explanation for its predictions. This article is highly cited, and consequently, COMPAS became the most well-known example of bias in machine learning, both within the professional circles of machine learning and among the public without expertise in AI. Despite this, COMPAS is still in use.

In response to the ProPublica article, a group of experts from the American justice system published a rejoinder with the telling title "False positives, false negatives and false analysis: a rejoinder to Machine Bias …" [12] They pointed out several controversial decisions in ProPublica's analysis, conducted their own experimental research, and concluded that ProPublica's assertions were incorrect. While this criticism seems justified, it would be more convincing if they clearly showed where the crucial mistake in ProPublica occurred. Instead, they presented their own experimental results, claiming that offenders are treated fairly regardless of race. They obtained these results by considering risk assessments of offenders on a scale from 1 to 10, as assessed by COMPAS. From these assessments, they calculated the AUC (Area under ROC curve), a standard performance measure in machine learning systems. AUC is interesting because it is equal to the probability that the predictive system correctly distinguishes between positive and negative cases. This means that, if we take two random cases (two defendants), one positive (did reoffended) and the other negative (did not reoffend), the system will correctly determine which is positive and which is negative with the probability AUC. [12] reported that the AUC value for whites was 0.69, and 0.70 for African Americans. The difference is not statistically significant. From this, they concluded that COMPAS is not racially discriminatory, and ProPublica's results indicating discrimination cannot be correct. However, this indirect argument allows for doubt since AUC, FPR, and FNR are not uniquely related to each other.

Dressel and Farid [11] reported on a relevant experiment where they were interested in the accuracy of predictions about the risk of recidivism achieved by randomly selected people without domain knowledge. They also compared the accuracy of COMPAS with that of a simple

linear classifier. They conducted a human prediction experiment (performed with crowd sourcing) on a subset of Broward County data (about 1000 out of a total of around 7000 defendants) from the studies [2] and [12]. Since using all 137 attributes for prediction by people would be impractical, they only used 7 selected attributes from the original set. The predictive accuracy of non-experts in these experiments was surprisingly almost the same as that of the COMPAS system. Interestingly, human predictions in this experiment were similarly biased as COMPAS, measured by FPR and FNR for whites and African Americans. These results hardly changed when additional information about the race was given to the human evaluator. They also found that a simple linear classifier achieved a similar predictive accuracy using only two attributes, and more sophisticated classifiers did not improve predictive accuracy (or fairness).

Holsinger et al. [16] criticized the study by Dressel and Farid [11]. The criticism is based on the following arguments. Participants were recruited via Amazon's Mechanical Turk, and they received payment for participation. The participants were only shown the values of seven selected attributes: age, gender, offense type, offense severity, adult convictions, juvenile felony charges, and juvenile misdemeanor charges. All these attributes are known as important risk factors for recidivism. According to Holsinger et al. [16], this reduction of the original set of 137 attributes made the prediction task easier than the original task with 137 attributes. This view is indeed justified as it is known that appropriate selection of useful attributes in machine learning may be difficult. After providing an individual rating, participants received feedback regarding the correctness of the answer and their average accuracy. Participants who achieved high accuracy were rewarded with slightly higher payment. All this significantly differs from the real context where expert decision-makers face a plethora of (often irrelevant and biased) information, which makes their task more difficult [16]. A counter argument could be used here that these decision-makers could easily inform themselves about the most important recidivism factors by a simple web search. But the authors in [16] conclude that, in the context of all circumstances, Dressel and Farid did not reveal anything new and just "rediscovered what has been well-established in a large body of risk assessment literature: Compared to unstructured human judgement, structured human judgment and actuarial approaches are more accurate. Structuring decisions limits consideration and unnecessary emphasis on factors that are unrelated to risk of recidivism (i.e. bias)."

Regardless of these results, Rudin [22] used machine learning to synthesize a very simple and completely understandable predictive model from the mentioned Florida data. This model comprises three simple if-then rules (shown below) and uses three attributes only (gender, age and number of past crimes). Unlike the COMPAS model, these rules are trivially understandable:

```
IF age between 18-20 and sex is male
  THEN predict arrest (within 2 years)
  ELSE IF age between 21-23 and 2-3 prior offenses
    THEN predict arrest
    ELSE IF more than three priors
      THEN predict arrest
      ELSE predict no arrest
```

This predictor is equally accurate for recidivism prediction as COMPAS, and it has similar FPR and FNR on data from Broward County, Florida.

From the described results, the recidivism predictive problem seems challenging despite extensive available information about the defendant, and better accuracy apparently cannot be achieved. At the same time, almost everything seems to be achieved with just three most useful attributes, and the additional 130+ attributes do not bring anything substantially new. From this, the authors in [2] and [11] jump to the conclusion that the use of machine learning in the justice system is not promising in general. This is, of course, a hasty and overly simplistic conclusion, as Spielkamp [23] points out. In many other applications, machine learning has surpassed the predictive accuracy of experts, as for example confirmed already by many early experiments with machine learning in medical diagnosis [6].

All these different opinions regarding bias and usefulness of the COMPAS system highlight the lack of universally accepted operational definitions of bias and fairness in machine learning. This situation is nicely illustrated by the highly cited article by Mehrabi et al. [20], which discusses more than a dozen relevant definitions but does not provide a synthesis that would limit this conceptual complexity and offer a practically useful approach. Additionally, this article causes further uncertainty by quickly dismissing the COMPAS system and categorizing it as evidently biased, flawed, and useless, without addressing counter arguments in [12].

3 Issues in definitions of bias and fairness

In the general media, machine learning is often accused of bias based more on intuition, without precisely defining mathematically verifiable criteria by which bias can be detected. Statements such as "the system has shown bias towards people of color in the judiciary," [2] or "the system is biased against women in employment evaluations," [5] use general phrases like "bias of algorithms," "algorithmic bias", "machine learning bias" or "artificial intelligence bias." Sometimes these statements are accompanied by a simple explanation like "machine learning systems are developed almost exclusively by white men, so ..."

Today, it is clear that the matter is not so trivial. Overly simplistic explanations are becoming rarer. It is also becoming clearer that the phrase "algorithm bias" is not suitable and gives the wrong impression that algorithms can have malicious intentions and do not operate according to mathematical and statistical principles [20]. The goal of these methods is always to discover the laws that apply in the real world from real-world data. However, a problem arises if biased practices already exist in the real world. Data collected in such a world reflects this bias, and the learning algorithm detects and reproduces this bias. If the results obtained from biased data in the real world are then used again in the real world, we will reproduce the existing bias [14].

However, it is still not precisely defined what bias actually is. Often, it is an impression of bias, where biases for or against an individual or a group manifest in a way that is perceived as unjust [21].

Let us look at the problems with defining bias. In the field of machine learning, various explanations are found, all meaningful in their own ways. The term bias in machine learning refers to several phenomena [15]:

1. Inductive Bias: This is the principle by which an algorithm chooses one of the typically large number of possible hypotheses, all in some way justified by the training data. It is a set of assumptions made by a learning algorithm to generalize a finite set of observations (training data) into a general model of the domain [17]. This type of bias is a necessary component of machine learning, without it machine learning is not possible. An example of such bias is Occam's razor, which says: if we have two explanations that explain the data equally well, we should choose the simpler one [13], [15], [21]. Although the term bias has a negative connotation, inductive bias is a positive and even an inevitable component of machine learning, as explained by [15], and is a basic concept in AI textbooks.

2. Bias in Training Data: This bias reflects actual biases in established decision-making in a given application area (e.g., bias in the judgments of experts in actual judicial practice in the environment from which the training data are drawn) [5], [10]. [1] emphasizes that biases in training data can be attributed to cognitive biases of human thinking. It is a natural phenomenon where the human brain filters infinite types of information in a way that retains what is relevant to us. Because algorithms are trained on data representing human behavior, they reflect these cognitive biases. This bias is referred to as negative legacy [5], or as historical bias [15].

3. Bias from Improper Data Collection or Sampling: For example, if there are significantly fewer examples available for a particular group of people than for other groups, according to mathematically grounded statistical and probability principles, some groups, typically minorities, appear to be discriminated against simply because probability estimation methods correctly assess probabilities differently when there is little data available. This bias is referred to as underestimation [5]. [25] emphasizes that we must also question where the training data come from. If algorithms traditionally relied on reliable labels determined by experts, today algorithms may learn from data originating from the broader society, where labels and patterns are often biased.

The above sources of bias are relatively widely accepted. However, the problem remains of how to precisely define criteria that objectively indicate whether a system is biased or to quantitatively assess that bias. There are numerous measures that seem relevant, but they may turn out to be contradictory, and for now, there is no simple, universally accepted measure.

The situation is well illustrated by the comprehensive review of different definitions of fairness by Mehrabi et al. [20]. In “fairness through awareness”, an algorithm is considered fair if it gives similar predictions to similar individuals; in “treatment equality,” ; the ratio of FNR and FPR in both groups (based on a protected attribute) is the same; in “fairness in relational domains,” ; in addition to attributes, social, organizational, and other connections between individuals are considered; in “fairness through unawareness,” ; an algorithm is considered fair as long as any protected attributes are not explicitly used in the decision-making process.

Berk et al. [4] make a similar point. They examine different ways that fairness can be formally defined, how these different kinds of fairness can be incompatible, how risk assessment accuracy can be affected, and various algorithmic remedies that have been proposed. They conclude that, except in most trivial cases, it is impossible to maximize accuracy and fairness at the same time and impossible to simultaneously satisfy all kinds of fairness. Kleinberg et al. [18] explore this problem more thoroughly. They define three natural, seemingly obvious conditions that a system must meet to be unbiased (fair). However, it turns out that these three conditions cannot be satisfied simultaneously, except in (trivial) special cases that are uninteresting for practical purposes. So, these three basic requirements together are unattainable. These three requirements are:

(1) Calibration of Probability Estimates: if the algorithm identifies a set of people who are supposed to belong to the positive class with a given probability P, then approximately proportion P of that set must actually belong to the positive class. The same condition must apply to all groups of individuals who differ in the "protected attribute," such as race or gender. In other words, estimates must mean what they are supposed to mean and must be independent of the group (based on protected attributes) to which the individual belongs.

(2) Balance of the Positive Class: the average risk score of the individuals in the positive class must be the same for all groups. In the case of COMPAS, for example, white and black convicts belonging to the positive class should have comparable risk scores.

(3) Balance of the Negative Class: analogous to the average in positive class.

Kleinberg et al. [18] mathematically prove that these three requirements, although essentially aiming for the same goal of reducing bias, are incompatible with each other, except in special cases.
When bias occurs, the question is how to eliminate it. There are various ideas for this, of which the two most obvious are (a) "protected attributes" and (b) reverse discrimination. Typical protected attributes are race and gender.

The principle of protected attributes means that we forbid the learning algorithm from using these attributes when deciding on the classification of an instance. This idea usually does not work well, as the learning algorithm effectively reconstructs their values from other, unprotected attributes that correlate with the protected ones. For example, from data on education or residence location, the algorithm may probabilistically infer a person's race [14].

The principle of reverse discrimination is to deliberately give certain advantages to underprivileged groups in treatment to counteract the effects of discrimination. This measure is obviously well-intentioned, but it actually introduces additional injustice, which is sometimes questionable [1]. Such injustice (reverse discrimination) may be justified, but not from the perspective of fairness, but from the perspective of "higher" values, such as rectifying historical injustices and achieving long-term justice through temporary injustice. Therefore, it is a strategic implementation of socially accepted values that are not easily achievable in practice due to historical reasons and persistence. The difficult question remains to what extent reverse discrimination makes sense. This should be determined by democratically accepted social consensus, formalized with appropriate laws for each case. In practice, addressing bias is approached within three phases of machine learning: (1) pre-processing phase, where we augment the minority sample, (2) mid-processing phase, where we introduce constraints to compensate for an uneven sample, and (3) post-processing phase, where we adjust thresholds for minorities [5][20][21].

When developing methods and tools, we must be aware of potential pitfalls. Alelyani [1] and Holsinger et al. [16] emphasize that certain solutions can lead to new injustices. Chakraborty et al. [7] state that the common side effect of mutating training data is the loss of significant connections between variables or degradation of learner's performance (as measured by accuracy and F1 score). They identify prior decisions that generated the training data as the root causes of bias. They propose a Fair-SMOTE tactic that makes it possible to mitigate bias while maintaining (or even improving) performance at the same time. The key point is to mutate the data in a way that extrapolates all the variables by the same amount. This way, we don't lose connections between variables. Specifically, it finds data imbalance and improperly labeled data points (by situation testing) and then use oversampling to balance the data and remove improperly labeled data points. As an outcome, it generated fair results. They reject worries by [4] that the cost of fairness is a reduction in learner performance. They conclude that it is always better to reflect on the domain and use those insights to guide improvements than blindly applying some optimization methods.

4 Conclusions

Bias has become a popular and controversial topic in some significant machine learning applications. The discussion is dominated by confusion, stemming from the fact that people have different intuitive understandings of the concepts of fairness and bias. Fairness is experienced in various ways, and there is no perfect consensus on the details. Similarly, there is no agreement on what a clear, mathematically formulated criterion should be to unequivocally quantify the bias of a specific system. There is a multitude of disagreements, controversies, and open issues where consensus is lacking. There is no consensus on the origin of bias, nor on which tool or method is most suitable for addressing bias.

The fairness of machine learning should mean producing decisions that society would be satisfied with. However, we are not unified in this regard. The case of COMPAS illustrates how crucial this unity is. COMPAS was tested by multiple experts, and their opinions are entirely contradictory. Some argue that COMPAS is biased, while others say it is not. Spielkamp [23] believes that everyone is correct because they understand fairness in various ways. The study by Kleinberg et al. [18] is particularly illuminating, where the authors mathematically prove the mutual exclusion of certain definitions of fairness that seem equivalent and necessary at first glance. We must clearly define important societal values, consider historical data and the real state of the world, educate ourselves about at least the basic workings of algorithms, and then articulate our expectations accordingly. Only after this can we decide which concept of fairness is suitable in a particular case. Corbet-Davies et al. [8] suggest that, in certain contexts, without a proper understanding of the domain and desired outcomes, acting according to popular formal conceptions of fairness can even have opposite effects than desired. The authors propose that, in algorithm development, instead of adhering to axiomatic notions of fairness, we should focus on their consequences, which strongly depend on the context.

In general literature, there doesn't seem to be anyone anticipating how challenging it will be in practice to

address the issue of bias. Expectations regarding values will need to be precisely formulated with appropriate laws. For example, should reverse discrimination be implemented in a specific application due to historical injustices, and to what extent? This formulation will have to be more technical than usual in regulations and laws, as it will be the basis for the concrete implementation in artificial intelligence algorithms. It is clear that the problem of bias is not solely of a technological nature, and therefore effective approaches to solving the problem must also include broader solutions. We must strive for interdisciplinary research, where AI engineers collaborate with disciplines dealing with ethics, legislation and decision-making [27].

For a proper general understanding and action in this field, there is a need for quality general education. The lack of it is evident in the way information is reported, in people's responses, and even in the confusion of experts. Various algorithms are becoming an inevitable part of our lives. It is unacceptable that we not only know too little about them but also have incorrect perceptions. On the other hand, we have governments and democratic institutions that do not understand the workings of artificial intelligence systems, yet they are the ones who commission and then implement such systems into their decision-making processes. Institutions often lack the knowledge and resources to know how to ask for appropriate algorithmic tools. It is imperative to educate people so that they can articulate what the algorithms should actually measure, what the output should be, and what criteria need to be met for the algorithm to be fair [9].

Acknowledgement

This work was supported by ARIS (Agency of Research and Innovation of Slovenia) and the Ministry of Digital Transition of Slovenia, as part of the research project V2-2272.

References

[1] Alelyani, S. (2021). Detection and Evaluation of Machine Learning Bias. *Applied Sciences, 11*(14). https://doi.org/10.3390/app11146271.

[2] Angwin, A., Larson, J., Mattu, J. & Kirchner, L. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. *ProPublica*.

[3] *Artificial Intelligence Act*, European Parliament, 14 June 2023; current updated unofficial version January 2024.

[4] Berk, R., Heidari, H., Jabbari, S., Kearns, M & Roth, A. (2021) Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research, 50*(1), 3-44. https://doi.org/10.1177/0049124118782533.

[5] Blanzeisky, W. & Cunningham, P. (2021). Algorithmic Factors Influencing Bias in Machine Learning. In: Kamp, M., et al. Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2021. Communications in Computer and Information Science, 1524, Springer, Cham. https://doi.org/10.1007/978-3-030-93736-2_41.

[6] Cestnik, B., Kononenko, I. & Bratko, I. (1987). ASSISTANT 86: A knowledge-elicitation tool for sophisticated users: In: Bratko, I., Lavrač, N. (eds) *Progress in Machine Learning: Proc. of European Working Session on Learning EWSL 87*. Sigma Press, 1987, 31-45.

[7] Chakraborty, J., Majumder, J. & Menzies, T. (2021). Bias in Machine Learning Software: Why? How? What to do? arXiv: 2105.12195. https://doi.org/10.48550/arXiv.2105.12195.

[8] Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R. & Goel, S. (2018). The Measure and Mismeasure of Fairness. arXiv: 1808.00023. https://doi.org/10.48550/arXiv.1808.00023.

[9] Courtland, R. (2018). Bias detectives: the researchers striving to make algorithms fair. *Nature, 558*, 357-360. https://doi.org/10.1038/d41586-018-05469-3.

[10] Dastin, J. (11.10.2018). *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[11] Dressel, J. & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1). https://doi.org/10.1126/sciadv.aao5580.

[12] Flores, A. W., Bechtel, K. & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals and It's Biased Against Blacks." *Federal Probation Journal, 80*(2).

[13] Gordon, D. F. & Desjardins, M. (1995). Evaluation and Selection of Biases in Machine Learning. *Machine Learning*, 20, 5-22. https://doi.org/10.1023/A:1022630017346.

[14] Hardt, M., Price, E. & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. arXiv: 1610.02413. https://doi.org/10.48550/arXiv.1610.02413.

[15] Hellström, T., Dignum, V. & Bensch, S. (2020). Bias in Machine Learning – What is it Good for? arXiv: 2004.00686. https://doi.org/10.48550/arXiv.2004.00686.

[16] Holsinger, A. M., Lowenkamp, C. T., Latessa, E. J., Serin, R., Cohen, T. H., Robinson, C. R., Flores, A. W. & vanBenschoten, S. W. (2018). A Rejoinder to Dressel and Farid: New Study Finds Computer Algorithm Is More Accurate Than Humans at Predicting Arrest and as Good as a Group of 20 Lay Experts. *Federal Probation Journal*, 82(2), 51-56. https://doi.org/10.2139/ssrn.3271682.

[17] Hüllermeier, E., Fober, T. & Mernberger, M. (2013). Inductive Bias. *Encyclopedia of Systems Biology*. https://doi.org/10.1007/978-1-4419-9863-7_927.

[18] Kleinberg, J., Mullainathsn, S. & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv: 1609.05807. https://doi.org/10.48550/arXiv.1609.05807.

[19] Levin, S. (8.9.2016). *A beauty contest was judged by AI and the robots didn't like dark skin*. The Guardian. https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

[20] Mehrabi, A., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR), 54*(6), 1-35. https://doi.org/10.48550/arXiv.1908.09635.

[21] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E. … Staab, S. (2020). Bias in data-driven artificial intelligence systems – An introductory survey. *WIREs Data Mining and Knowledge Discovery, 10*(3). https://doi.org/10.1002/widm.1356.

[22] Rudin, C. (2019). Stop explaining black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x.

[23] Spielkamp, M. (2017). Inspecting Algorithms for Bias. *MIT Technology Review*, July 2017.

[24] Staff, R. (7.12.2016). *New Zealand passport robot tells applicant of Asian descent to open eyes*. Reuters. https://www.reuters.com/article/us-newzealand-passport-error-idUSKBN13W0RL.

[25] Sun, O., Nasroui, O. & Shafto, P. (2020). Evolution and impact of bias in human and machine learning algorithms interaction. *PLoS ONE*, 15(18). https://doi.org/10.1371/journal.pone.0235502.

[26] UNESCO Recommendation on the Ethics of Artificial Intelligence, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[27] Yu, H., Shen, Z., Miao, C., Lesser, V. R. & Yang, Q. (2018). Building Ethics into Artificial Intelligence. arXiv: 1812.02953. https://doi.org/10.48550/arXiv.1812.02953.