FAIRNESS IN MACHINE LEARNING: WHY AND HOW?

Mladen Nikolić[1], **Andrija Petrović**[2]

[1] Faculty of Mathematics,
University of Belgrade, Studentski Trg 16, 11000 Belgrade, Serbia Email: nikolic@math.rs
[2] Technical Faculty,
Singidunum University, Danijelova 32, 11000 Belgrade, Serbia
Email: apetrovic@singidunum.ac.rs

Abstract

Services based on machine learning are increasingly present in our everyday lives. While such application make promises of its improvement, they also pose considerable risks if machine learning models do not perform as expected. One specific issue related to the quality of learnt models which has recently gained considerable visibility is their unfairness. Namely, it has been noted that the decisions of machine learning models sometimes reflect human biases against some historically discriminated groups of people, thus unintendedly perpetuating the discrimination. In this paper we discuss why is the fairness of machine learning models important, by revisiting some notable examples of discrimination committed by the models and discuss different notions of fairness. We discuss how to measure the fairness of such models and how to achieve it, reflecting on both algorithmic and non-technical aspects of this effort. We present several fairness ensuring methods representative of different fairness paradigms, one of them being our own.
Key words: fairness, machine learning, ethical artificial intelligence.

1 Introduction

In the previous decade, machine learning achieved great success in numerous applications, starting from computer vision [20, 9], but quickly penetrating other fields like natural language processing [24, 8], game playing [29, 30, 23], and many others, often surpassing expectations and even surpassing the performance of human experts.

Many of these achievements were quickly integrated in to services we use at everyday basis and can influence our life experience. Some examples include personal assistants on mobile phones, face recognition systems, information retrieval systems, language processing and translation services, content and product recommendation systems, etc. In recent years, steps have been taken to introduce machine learning in health, legal, financial, security, and employment related decision making [17, 4, 6, 25, 12, 3, 31]. Thus, machine learning might have an opportunity to affect the functioning of the society for better or for worse.

One specific issue related to machine learning models of critical importance in this context is their (un)fairness [22, 7]. Namely, in recent years it has been observed that machine learning models tend to learn human biases present in the data from which they learn and to perpetuate discrimination against historically discriminated groups. One might have hoped that use of artificial intelligence would lead to less biased decision making, but due to learning from human generated data, the issues persist. However,

this observation sparked a new field of machine learning, which deals with mitigation of such biases and fair decision making. Numerous advances have been achieved both in algorithm design and in societal awareness of the problem [22, 7].

In this paper, we aim at pointing out *why* fairness is an important topic in machine learning and we discuss different notions of fairness. Besides explaining *why*, we also discuss *how* is fairness achieved, emphasizing importance of both the societal and technical aspects. We discuss a number of important methods for improving fairness of machine learning models, and present our own method which merges two existing paradigms while aiming at better interpretability of the learnt model [26]. We also reflect on the existing challenges in the field and argue that societal aspects of machine learning applications are at least as important as technical aspects which currently draw most attention in machine learning related discourse.

2 Why?

Fairness is sometimes considered a fringe topic of machine learning, while the spotlight is taken by the development of new models with better generalization performance, specific inductive biases suitable for different kinds of data, greater computational efficiency, etc. While such an approach might be understandable in the infancy of a technology, in its mature phase in which it performs well enough to influence the everyday experience of millions of people, societal impact of the technology should be considered as one of its primary aspects. In order to answer why is the fairness in machine learning important, we inspect some of the most notable examples of unfairness of machine learning models. They indicate which kind of negative societal impact such models can make. While a sequence of examples may convince us of the importance of the issue, no discussion of this kind would be complete without considering what we mean by fairness. Therefore, we also briefly reflect on the notions of fairness considered in the field.

2.1 Notable Examples of Unfairness in Machine Learning

There has been a considerable amount of reports on the unfairness of machine learning models and decision making algorithms in general. Well known example is related to the system COMPAS used at US courts in order to assist decision making regarding release or detention based on the estimated risk of a person committing another crime [2]. It has been determined that the system has higher false alarm rate for African-Americans than for Caucasian individuals.

Until recently, when translating from the languages which lack grammatical gender (e.g. Turkish) to languages which have it, Google Translate preferred masculine form for some occupations (e.g. doctor) and feminine form for some others (e.g. nurse) [27]. Recently, the issue was resolved by offering translations for both grammatical genders.

In the evaluation conducted by the US National Institute for Standards and Technology it has been noted that many facial recognition models exhibit different error rates over different demographic groups, differing even in the order of magnitude [14]. On the other hand, the best performing models exhibited negligible differences. Facial recognition software in digital cameras has exhibited higher false detection rates of blinking for Asian individuals.[1] The consequences of machine learning bias in future might be even more serious. For instance, an autonomous vehicle which would not recognize individuals of a specific demographic group, might endanger them in traffic.

[1] http://content.time.com/time/business/article/0,8599,1954643,00.html

2.2 Notions of Fairness

There is no single definition of fairness. In different times or in different cultures today, the notion can be understood in different ways and has strong ideological undertones. Therefore, the question belongs to a broad spectrum of social sciences and philosophy and has been discussed many times before the advent of artificial intelligence. However, in order to deal with fairness in in the context of artificial intelligence, one needs to operationalize the notion. In the light of the previous discussion, the field, does not deal with a single definition, but instead provides a collection of tools which can be used in hope of achieving fair model-based decisions with respect to different mathematical definitions. This approach has its own pitfalls as will be discussed later.

Still, the notions of fairness come in two major groups. The first one is so called *individual fairness* [11, 21, 5]. It postulates that similar individuals should be treated similarly. Of course, the question of adequate similarity measure can be a subject of debate. The principle may sound sensible, but it is not without its shortcomings. While it can guard against random whimsical injustice against an individual, it does not guard against structural and historical discrimination towards some vulnerable groups defined by some specific feature like gender or race. Two persons may be similar in some respects for instance due to their origins in a segregated minority community and for the same reasons be dissimilar in those respects from people outside of the community. Using a similarity metric sensitive to such distinctions will allow discrimination against the members of the community, although it would treat similar individuals in a similar way. Another major understanding of fairness is *group fairness* [11, 15, 26] which stresses an importance of treating different groups of people equally in order to avoid historical discrimination based on membership of individuals in specific groups. A well intentioned effort might aim at achieving both kinds of fairness. However, there exist impossibility results which prove that different kinds of fairness cannot be achieved simultaneously [10, 19]. Therefore, defining fairness, making trade-offs, and taking responsibility which such decisions involve will not be resolved by technology alone. This conclusion puts the question of fairness in artificial intelligence at hands of the larger society.

3 How?

In this section we discuss the means of achieving fairness in machine learning models. Such discussions usually focus on technical aspects like models and algorithms. However, recently it has been convincingly argued that in order to ensure fairness a sequence of ethical and societal considerations should precede technical ones [28, 13]. Therefore, we proceed in such fashion.

3.1 Ethics Before Technicalities

A major flaw in current approach to fairness in machine learning, as recently recognized, is disproportionate focus on mathematical abstractions which do not capture the complexity of the issue and the reliance on engineering practices commonly used in software development, which can even exacerbate it or at least hinder its resolution [13, 28]. An alternative approach which is suggested is a sociotechnical approach which insists on including societal considerations when designing a system which will exist and act in a societal context. A recommended sequence of considerations when building a machine learning solution is as follows.

- Should the solution be built in the first place? If a purpose of the system is itself not legitimate and if it will hurt individuals in general or exacerbate social inequity and hurt vulnerable groups, no algorithmic improvement will make it fair [13]. For instance, mass surveillance technologies which governments can use to intrude into peoples privacy, subvert their rights, etc., cannot be made fair by any kind of improvement.

- If it is built, should it be banned? The fact that a system is in operation does not justify its use if it should not have been built in the first place [13].

- Is the proposed solution really a solution to the problem? Maybe the best solution is not based on technology at all [28].

- Does the solution affect the society in a predictable way or will its introduction cause different problems [28]?

- Is the formal mathematical definition of fairness used in the solution adequate? Fairness metrics are formulated over outcomes and it is very hard to properly take into consideration full context in which such outcomes occur, but the context affects the judgement of outcomes' fairness [28].

- Does the solution design accounts for all relevant societal factors? Most often technical approach abstracts away most of the societal context. However, the context in which it will operate will determine if it will mitigate the problem or even exacerbate it. While the fairness ensuring method might ensure that the fairness constraints on the outputs are satisfied, the final outcome of the process in which the system operates also depends on the way such outcomes will be treated by other actors in the process and failure to account for that finally does not lead to fair outcomes [28].

- Can the solution developed for one societal context be ported to another one? If the new context in which the solution should operate considerably changes from the previous one, the solution might not be adequate for the new one [28].

- Who's values are being incorporated in the system? If a natural language processing system (e.g., for question answering) is trained on large text corpora available at the internet (e.g., Wikipedia, news portals, etc.), it will incorporate the dominant views of potentially small but privileged groups which generate most of such content, with views of other groups being disregarded [13].

The previous points do not mean that the technical methods are worthless, but that more elements should be considered than they are considering today.

3.2 Metrics

Metrics serve a purpose of mathematical formalizations of different notions of fairness. Shortcomings of such efforts are already highlighted. We distinguish between group fairness metrics and individual fairness metrics.

Group fairness These metrics are meant to indicate if decisions $\hat{\mathbf{y}}$ of a classifier are disproportional between different groups as defined by the value of some sensitive feature $\mathbf{s}$ (e.g, gender or race). In this paper we assume that $\mathbf{s}$ is a single binary variable, although more general approach is possible. There is a multitude of such metrics, but we focus on three commonly used ones for illustration purposes. First such metric is *absolute statistical parity difference* [22, 7]:

$$\mathbf{ASD} = |P(\hat{\mathbf{y}} = 1|\mathbf{s} = 0) - P(\hat{\mathbf{y}} = 1|\mathbf{s} = 1)| \tag{1}$$

Low values of **ASD** mean that both groups have approximately the same probability of being labeled 1 (e.g., bank loan granted) by the model. Such notion of fairness is called *statistical parity* or *demographic parity*. Second common metric is *absolute equal opportunity difference* [22, 7]:

$$\mathbf{AEOD} = |P(\hat{\mathbf{y}} = 1|\mathbf{s} = 0, \mathbf{y} = 1) - P(\hat{\mathbf{y}} = 1|\mathbf{s} = 1, \mathbf{y} = 1)| \tag{2}$$

This measure can be interpreted as a difference of opportunities between unprivileged and privileged group. Values of AEOD close to zero are desirable. Such notion of fairness is called *equal opportunity*. The third often used metric is *average odds difference* [22, 7]. It can be formulated as:

$$\begin{aligned}\mathbf{AOD} =& \frac{1}{2}(|P(\hat{\mathbf{y}} = 1|\mathbf{s} = 0, \mathbf{y} = 0) - P(\hat{\mathbf{y}} = 1|\mathbf{s} = 1, \mathbf{y} = 0)| &(3)\\ &+ |P(\hat{\mathbf{y}} = 1|\mathbf{s} = 0, \mathbf{y} = 1) - P(\hat{\mathbf{y}} = 1|\mathbf{s} = 1, \mathbf{y} = 1)|) &(4)\end{aligned}$$

Values of **AOD** close to zero are desirable. Such notion of fairness is called *equalized odds*.

Individual fairness These metrics focus on differences between individuals. One intuitive metric is computed by setting a threshold on similarity of the individuals and averaging the absolute differences between outcomes for individuals which satisfy the given similarity threshold [21]. Of course, there is an issue of selecting appropriate similarity threshold, which has to be done based on domain knowledge. Another fairness metric is based on couterfactuals and causal models. Let $\mathbf{u}$ be unobserved features and $\mathbf{x}$ and $\mathbf{s}$ observed features of which feature $\mathbf{s}$ is considered sensitive. Each observed feature is assumed to be a function of other observed features and of unobserved features. Let F be a set of all such functions. Let $\mathbf{y}_{\mathbf{s}\leftarrow a}(\mathbf{u})$ denote the value $\mathbf{y}$ would take if the feature $\mathbf{s}$ had taken value a and the value of features $\mathbf{u}$ did not change. This value is called a counterfactual and can be explicitly computed given $\mathbf{u}$, $\mathbf{x}$, and F. Then, the difference

$$|P(\hat{\mathbf{y}}_{\mathbf{s}\leftarrow 0}(\mathbf{u}) = \mathbf{y}|\mathbf{x}, \mathbf{s}) - P(\hat{\mathbf{y}}_{\mathbf{s}\leftarrow 1}(\mathbf{u}) = \mathbf{y}|\mathbf{x}, \mathbf{s})|$$

reflects the difference in treatment which the same person would receive had it belonged to a different group [21]. However, the computation of counterfactuals is a hard problem in practice.

Again, one should be aware that these metrics cannot be optimized simultaneously [10, 19] and that they alone can hardly fully capture the intended meaning of fairness, so they should be used with caution.

3.3 Methods

There is a multitude of methods aiming at different notions of fairness [22, 7]. A naive idea of how to obtain fairness would be to refrain from using sensitive features in training. This is called *fairness through unawareness*. Such an approach is thwarted by correlations of other features which we perceive as not sensitive themselves with the sensitive feature. For instance, persons address might correlate with its race. Therefore, more involved approaches are needed, but approaches which take sensitive features into consideration and are therefore able to check the fairness of their predictions. This is called *fairness through awareness* [11]. We will discuss representative examples of different kinds of methods, and also present our own method which merges some of the existing paradigms.

3.3.1 Group Fairness Methods

There are three major groups of group fairness methods: pre-processing based, in-processing based, and post-processing based ones. We discuss their representative methods in turn.

Pre-processing The idea of pre-processing approaches is to treat the learning method as a black box and achieve the fairness of the resulting model by manipulating its inputs. For instance, the instances from the training set themselves can be altered or given different weights. One simple approach proposes reweighing the instances in order to compensate for the bias they exhibit [18]. Specifically, if the sensitive variable $\mathbf{s}$ and the class variable $\mathbf{y}$ were statistically independent, their joint probability would factorise as $P(\mathbf{s}, \mathbf{y}) = P(\mathbf{s})P(\mathbf{y})$. Therefore, in order to compensate for their dependence, each instance should be weighted by a term

$$\frac{P(\mathbf{s})P(\mathbf{y})}{P(\mathbf{s}, \mathbf{y})} \tag{5}$$

where specific values from the instance are used for the variables. All probabilities involved are easily estimated from the training data by counting, given that $\mathbf{s}$ and $\mathbf{y}$ are categorical.

In-processing In-processing approaches alter existing or propose new optimization problems and methods in order to achieve the fairness of the learnt model. Possibly a predominant approach aims at learning a mapping which will provide fair representation of the input data (e.g., with reduced or eliminated bias) and learning the classifier over such representations. One specific method proposes solving the following optimization problem [1]:

$$\min_{\phi,\theta} \left(\mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{s}\sim P(\mathbf{x},\mathbf{y},\mathbf{s})}[L(f_\phi(g_\theta(\mathbf{x})), \mathbf{y})] - \beta \max_{\psi} \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{s}\sim P(\mathbf{x},\mathbf{y},\mathbf{s})}[L(h_\psi(g_\theta(\mathbf{x})), \mathbf{s})] \right) \tag{6}$$

where $P(\mathbf{x}, \mathbf{y}, \mathbf{s})$ is the distribution of the data, L is the loss function, g_θ is a function producing fair representation of the data, f_ϕ predicts the class of an instance, h_ψ predicts the sensitive feature, and β is a hyperparameter. Obviously, the optimization problem is solved by finding the representation of the data from which the class can be well predicted, but even the best model cannot predict the sensitive feature well. The importance of these two objectives is weighed by β. The problem is solved by optimizing adversarially [1].

Post-processing Post-processing techniques treat learning algorithms as black boxes, but instead of manipulating the data to achieve fairness (like pre-processing techniques), they manipulate the outputs of the classifier. Such manipulation can be performed in multiple ways. For instance by modifying the thresholds which translate the score which model provides into its predictions (e.g., for logistic regression the threshold need not be 0.5) or by adding constraints the predictions need to satisfy and by finding the most similar predictions to the ones the model has provided, but which satisfy such constraints. One such approach aims at achieving equalized odds (or equal opportunity by relaxing constraints), mentioned before [16]. Let $\mathbf{y}$ be the true value of the target variable and $\hat{\mathbf{y}}$ the prediction given by the model. Denote

$$\gamma_s(\hat{\mathbf{y}}) = (P(\hat{\mathbf{y}} = 1|\mathbf{s} = s, \mathbf{y} = 0), P(\hat{\mathbf{y}} = 1|\mathbf{s} = s, \mathbf{y} = 1)) \tag{7}$$

The first element of the pair is a false positive rate and the second one is a true positive rate, for a group satisfying $\mathbf{s} = s$. Also let $P_s(\hat{\mathbf{y}})$ denote the convex hull of the set $\{(0,0), \gamma_s(\hat{\mathbf{y}}), \gamma_s(1-\hat{\mathbf{y}}), (1,1)\}$ (we assume that it holds $\mathbf{y}, \hat{\mathbf{y}} \in \{0,1\}$). The points of this set represent all easily achievable models given the trained model. Namely, it is easy to obtain the model achieving the performance $(0,0)$ by classifying all instances as negative and to achieve the performance $(1,1)$ by classifying all instances as positive. The performance $\gamma_s(\hat{\mathbf{y}})$ is already achieved by the original model and the performance $\gamma_s(\hat{\mathbf{y}})$ by the model predicting the opposite of the original one. Other points in the convex hull are easily achieved by interpolating in between, for instance, by adjusting the threshold on the score the model provides. Then, the proposed optimization method is:

$$\min_{\hat{\mathbf{y}}} \mathbb{E}L(\hat{\mathbf{y}}, \mathbf{y}) \tag{8}$$

$$\text{s.t. } \gamma_0(\hat{\mathbf{y}}) \in P_0(\hat{\mathbf{y}}),\ \gamma_1(\hat{\mathbf{y}}) \in P_1(\hat{\mathbf{y}}) \tag{9}$$

$$\gamma_0(\hat{\mathbf{y}}) = \gamma_1(\hat{\mathbf{y}}) \tag{10}$$

The objective function requires the difference between the original and post-processed predictions to be small. The fist constraint means that the model is achievable in a sense defined before. The second constraint obviously enforces equalized odds.

Reweighing in-processing The advantage of the pre-processing reweighing approach is that it provides interpretable weights (which can indicate which instances are the source of bias). However, it does not perform end-to-end optimization of the objective function and is therefore suboptimal. The in-processing representation learning approach is not interpretable, but performs end-to-end optimization. Recently, we proposed to merge these two paradigms and obtain the best of two worlds in one method [26]. In its simplest form, the proposed optimization problem is stated as:

$$\min_{\theta,\phi} \max_{\psi} \mathbb{E}_{\substack{\mathbf{x},\mathbf{y},\mathbf{s} \sim P(\mathbf{x},\mathbf{y},\mathbf{s}) \\ w = f_\theta(\mathbf{x})}} [w \cdot (\alpha \log P_\psi(\mathbf{s}|\mathbf{x}) - \log P_\phi(\mathbf{y}|\mathbf{x}))] \tag{11}$$

where f_θ is a function which computes weights for the given instance, ϕ and ψ parametrize learnable conditional probability distributions of the class and the sensitive feature, respectively, and α is a hyperparameter. The optimal weights w reweigh the training instances in order to obtain the best prediction of the class while even the best model cannot accurately predict the sensitive feature. The importance of these two objectives is weighed by α. The problem is, again, solved by optimizing adversarially [26].

3.3.2 Individual Fairness Methods

Individual fairness methods are based on the idea that similar individuals should obtain similar treatment. In their seminal paper [11], the authors propose the following approach. Let $P_\theta(\mathbf{x}, \mathbf{y})$ be the distribution of the data, $P_\theta(\mathbf{y}|\mathbf{x})$ the parametrized conditional distribution over a categorical variable $\mathbf{y}$, d the distance over feature vectors $\mathbf{x}$, and D the divergence between distributions over $\mathbf{y}$. Then, the fair classifier is obtained by solving the following optimization problem:

$$\min_\theta \mathbb{E}_{\substack{\mathbf{x},\mathbf{y}\sim P(\mathbf{x},\mathbf{y}) \\ \mathbf{y}'\sim P_\theta(\mathbf{y}'|\mathbf{x})}} L(y, y') \tag{12}$$

$$\text{s.t. } D(P(\cdot|\mathbf{x}_1), P(\cdot|\mathbf{x}_2)) \leq d(\mathbf{x}_1, \mathbf{x}_2) \text{ for all } \mathbf{x}_1, \mathbf{x}_2 \tag{13}$$

An approach for fair regression was formulated with a similar idea, but instead of constraint based formulation, a fairness related penalty was proposed [5]. Let f_θ be a parametrized regression model and let S_1 and S_2 be two datasets, each representing one group with respect to the values of some sensitive binary variable. Then, the penalty is formulated as:

$$\frac{1}{|S_1||S_2|} \sum_{\substack{(\mathbf{x}_1,\mathbf{y}_1)\in S_1 \\ (\mathbf{x}_2,\mathbf{y}_2)\in S_2}} d(\mathbf{y}_1, \mathbf{y}_2)(f_\theta(\mathbf{x}_1) - f_\theta(\mathbf{x}_2))^2 \tag{14}$$

If a causal model of the phenomenon of interest is known, one can rely on counterfactual fairness method [21]. However, such models are hard to establish, so we do not delve deeper into the details of the method.

4 Conclusions

In this paper we discussed the issue of fairness of machine learning algorithms. Given their current and even greater future influence on humans, we argue that their fairness should be considered one of the topics of primary interest when considering the future of these algorithms. We discussed notions of fairness, the metrics used to measure and formalize them, and methods used to optimize them in order to achieve fair outcomes, one of them being our own method which merges two existing paradigms. We also stressed the existing challenges related to the societal context of application of these algorithms which will have to be paid more attention to in future work on this topic.

References

[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[3] Alexander Babuta, Marion Oswald, and Christine Rinik. Machine learning algorithms and police decision-making: Legal, ethical and regulatory challenges. 2018.

[4] Richard Berk. Accuracy and fairness for juvenile justice risk assessments. *Journal of Empirical Legal Studies*, 16:175–194, 03 2019.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *Arxiv*, 2017.

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[7] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ArXiv*, abs/2010.04053, 2020.

[8] Junyi Chai and Anming Li. Deep learning in natural language processing: A state-of-the-art survey. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6, 2019.

[9] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

[10] Richard B. Darlington. Another look at "cultural fairness". *Journal of Educational Measurement*, 8(2):71–82, 1971.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, page 214–226. Association for Computing Machinery, 2012.

[12] John M. Fossaceca and Stuart H. Young. Artificial intelligence and machine learning for future army applications. In Michael A. Kolodny, Dietrich M. Wiegmann, and Tien Pham, editors, *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*, volume 10635, pages 8 – 25. International Society for Optics and Photonics, SPIE, 2018.

[13] Timnit Gebru. Machine learning in practice: Who is benefiting? Who is being harmed? NeurIPS keynote, 2021.

[14] P. Grother, M. Ngan, and K. Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects, 2019.

[15] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[17] Senerath Mudalige Don Alexis Chinthaka Jayatilake and Gamage Upeksha Ganegoda. Involvement of machine learning tools in healthcare decision making. *Journal of Healthcare Engineering*, 2021.

[18] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, 2012.

[19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pages 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 54(6), 2021.

[23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[24] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2021.

[25] Felipe Dias Paiva, Rodrigo Tomás Nogueira Cardoso, Gustavo Peixoto Hanaoka, and Wendel Moreira Duarte. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications*, 115:635–655, 2019.

[26] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. Fair: Fair adversarial instance re-weighting. *Neurocomputing*, 476:14–37, 2022.

[27] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9, 2021.

[28] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68. Association for Computing Machinery, 2019.

[29] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[30] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[31] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. Hiring algorithms: An ethnography of fairness in practice. In *ICIS 2019 Proceedings*, ICIS Proceedings, pages 1–9. Association for Information Systems, 2020.