Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.029451

REVIEW

AI Fairness–From Machine Learning to Federated Learning

Lalit Mohan Patnaik[1,5] **and Wenfeng Wang**[2,3,4,5,6,*]

[1]Consciousness Studies Program, School of Humanities, National Institute of Advanced Studies, Bangalore, 560012, India

[2]Research Institute of Intelligent Engineering and Data Applications, Shanghai Institute of Technology, Shanghai, 201418, China

[3]Research Center of Ecology and Environment of Central Asia, Chinese Academy of Sciences, Urumqi, 830011, China

[4]Applied Nonlinear Science Lab, Anand International College of Engineering, Jaipur, 391320, India

[5]London Institute of Technology, The ASE-London CTI of SCO, London, CR26EQ, UK

[6]Sino-Indian Joint Research Center of AI and Robotics, The IMT Institute, Bhubaneswar, 752054, India

*Corresponding Author: Wenfeng Wang. Email: wangwenfeng@sit.edu.cn

Received: 20 February 2023 Accepted: 15 November 2023 Published: 29 January 2024

ABSTRACT

This article reviews the theory of fairness in AI–from machine learning to federated learning, where the constraints on precision AI fairness and perspective solutions are also discussed. For a reliable and quantitative evaluation of AI fairness, many associated concepts have been proposed, formulated and classified. However, the inexplicability of machine learning systems makes it almost impossible to include all necessary details in the modelling stage to ensure fairness. The privacy worries induce the data unfairness and hence, the biases in the datasets for evaluating AI fairness are unavoidable. The imbalance between algorithms' utility and humanization has further reinforced such worries. Even for federated learning systems, these constraints on precision AI fairness still exist. A perspective solution is to reconcile the federated learning processes and reduce biases and imbalances accordingly.

KEYWORDS

Formulation; evaluation; classification; constraints; imbalance; biases

1 Introduction

With the development of AI, it has entered almost all spheres of our lives everywhere and the fairness of such algorithms and applications has drawn attention from many applications [1–5]. Methodologies for such studies and applications in design and other areas have been well-studied [6–10]. AI technologies not only help us to obtain more accurate predictions in advertising, credit approval, employment, education, and criminal justice but also help us to make decisions in these areas with important influence. Conceptual studies related to fairness have been reported in several papers [11–13]. Some tools and a few more applications have also been reported [14–17]. AI has been applied to many fields, such as deciding who can get scholarships, mortgages, and economic capital. Risk factor studies capturing fairness concepts and the underlying challenges have also been reported [18–21]. Ethical and societal issues with applications have also drawn attention from researchers [22–25]. Since AI has become increasingly influential in all aspects of our daily lives, it is very important to consider whether AI will adversely impact vulnerable groups when making intelligent decisions, especially those with big influences [26–28]. Alternatively, as an important tool to assist people in decision-making, AI must undoubtedly ensure fairness and inclusiveness [29–32].

But until now, the theory of precision AI fairness is still poorly understood [33–37]. It is very difficult to define such fairness. Metrics to compute fairness, some relevant tools, and applications have been well covered in the literature [38–41]. Application to learning algorithms has also been well-studied [42–45]. Approaching precision AI fairness is much more complex than simply finding technical solutions because we have to ensure that the output is independent of sensitive parameters (such as gender, race, religious belief, disability, etc.) for specific tasks that might be affected by social discrimination [46–50]. We must acknowledge that no perfect machine learning algorithms exist, even for federated learning systems. Challenges faced in quantifying fairness have been discussed in several studies [51–54]. The concept of conditional fairness and methods to mitigate unwanted bias have been well-addressed [55–58]. Some constraints on precision AI fairness still exist [59,60]. It is quite necessary to analyze the major constraints of AI fairness further, illustrate why it is a critical issue, and discuss potential solutions.

Objectives of this article are 1) to present a systematic review and discussion on the theory of fairness in AI, 2) to find the major constraints on precision fairness, and 3) to analyze the potential solutions for this problem. The organization of the whole paper is as follows. In Section 2, the theory of fairness in AI is presented, including the formulation, classification and evaluation of AI fairness. The major constraints on precision fairness are discussed in Section 3, and the potential solutions for this problem are analyzed in Section 4.

2 Theory of Fairness in AI

2.1 Formulation of AI Fairness

Various norms of fairness have been introduced to assess the extent to which AI algorithms are unfair [56]. Such an evaluation has become a topic of academic and broader interest for decades [5,18]. It was concluded that maximizing accuracy and fairness simultaneously is impossible [38]. AI Fairness can be formulated in terms of probability or statistics. Let X be the other observable attributes of any individual and U be the set of relevant latent attributes which are not observed. We symbolize the outcome to be predicted as Y, which itself might be contaminated with historical biases. Let y be predicted decisions with a category $\in \{0, 1\}$ and a predictor $\hat{Y}$. For the convenience of formulation, let L be associated with legitimate factors and A be the set of protected attributes of an individual, variables that must not be discriminated. When an AI system gives similar predictions to similar individuals, it is called "Fairness Through awareness".

That is, a fair AI task should have a similar outcome between two individuals in terms of a similarity metric with inverse distance [10]. Let i and j be two individuals represented by vectors of attribute values v_i and v_j. The similarity distance between individuals of i and j is represented by d ($v_{i,}$ v_j). Let M (v_i) represent the probability distribution over the outcomes of the prediction. For example, if the output is binary (0 or 1), M (v_i) could be [0.3; 0.7], implying that P ($\hat{Y} = 0$) = 0.3 and P ($\hat{Y} = 1$) = 0.7 for individual i. Assume that DS is a distance metric between probability distributions. For any pair of individuals i and j, if DS(M(v_i), M(v_j)) $\leq$ d($v_{i,}$ v_j), the system is fair as long as any protected attributes A are not explicitly used in the decision-making process ($X_{(y=0)} = X_{(y=1)} \wedge A_{(y=0)} \neq A_{(y=1)} \Rightarrow \hat{y}_{(y=0)} = \hat{y}_{(y=1)}$). In other words, an algorithm is fair if protected attributes are not expressly considered when making decisions [23].

Another formulation is known as "Counterfactual Fairness". The predictor $\hat{Y}$ is said to be counterfactually fair if under any context X = x and A = a, P ($\hat{Y}_{A\leftarrow-a}$ (U) = y|X = x, A = a) = P ($\hat{Y}_{A\leftarrow-a'}$ (U) = y|X = x, A = a), (for all y and for any value a′ attainable by A). According to the definition of counterfactual fairness, a decision is fair if it is the same, whether made in the real world or in a counterfactual world in which the individual belongs to a different demographic group [29]. For a set of legitimate factors L, predictor $\hat{Y}$ satisfies conditional statistical parity if P ($\hat{Y}$ |L = 1, A = 0) = P ($\hat{Y}$ |L = 1, A = 1). This means that people, either protected or unprotected (female or male) groups, should be equally likely to have a positive outcome, given a set of legitimate factors [47]. A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. P ($\hat{Y}$ = 1|A = 0, Y = y) = P($\hat{Y}$ = 1|A = 1,Y = y), y ∈ {0,1}. In other words, for both protected and unprotected (male and female) groups, the probability of a positive outcome for a person in the positive class should be equal to that for a person in the negative class. A predictor $\hat{Y}$ satisfies demographic parity if P ($\hat{Y}$|A = 0) = P ($\hat{Y}$ |A = 1). Whether a person is in the protected group or not, the chances of a positive outcome should be the same. A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if P ($\hat{Y}$ = 1|A = 0, Y = 1) = P ($\hat{Y}$ = 1|A = 1, Y = 1). This means that both unprotected (female and male) and protected (female and male) group members should have the same probability of being assigned to a positive outcome [33].

2.2 Classification and Evaluation

Based on the above formulation of AI Fairness, such fairness can be classified into different types: 1) Individual Fairness. An algorithm would be optimised for an individual, and fairness criteria could be met by ensuring that the algorithm treats people with similar characteristics in the same way. If an individual can be described mathematically by a set of parameters in a multidimensional space, then all individuals in the same parametric space will be treated similarly and will receive similar predictions from a machine learning algorithm. This is known as individual fairness, and it is also a measure of consistency [43]. In other words, they make similar predictions for similar people [29]. 2) Group Fairness. Different groups should be treated equally, but sensitive attributes and outcomes are used as measuring features [10]. For example, the authors [20] analysed Berkeley's alleged sex bias in graduate admission and found that data showed a higher rate of admission for male applicants overall, but the result differed when department choice was taken into account. Traditional notions of group fairness fail to judge fairness because they do not account for department choice. Fairness notions based on causality emerge as a result of this [20,56]. 3) Subgroup Fairness. The goal of subgroup fairness is to combine the best characteristics of group and individual fairness. It is distinguishable from these concepts, but it makes use of them to achieve better results. It chooses a group fairness constraint, such as equalizing false positives, and tests whether it holds true across a large number of subgroups [30]. Many real-world examples of the associated unfairness can be found in the previous studies [4,36,39].

Let P be the total number of positives in the dataset and N be the total number of negatives in the dataset. These types of AI fairness can be evaluated with some general metrics. For the convenience of subsequent statements, we abbreviate True Positive as TP (predicted positive and it is true) and True Negative as TN (predicted negative and it is true). For the Type 1 Error-predicted positive and it is false, we abbreviate False Positive as FP. For the Type 1 Error-predicted negative, and it is false, we abbreviate False negative as FN. Then, the accuracy of AI algorithms is calculated based on how many from the dataset have been predicted correctly, and it should be as high as possible.

Mathematically, the total number of two correct predictions (TP + TN) divided by the total number of datasets (P + N) represents accuracy [22], while the precision is estimated based on

TP/(TP+FP). A Statistical Parity Difference (SPD) is the difference between the unprivileged and privileged groups in terms of the likelihood of favourable outcomes. This can be computed from both the input dataset and the dataset output by a classifier (predicted dataset). A value of 0 means that both groups benefit equally; a value less than 0 means that the privileged group benefits more, and a value greater than 0 means that the unprivileged group benefits more. Hence, SPD = Pr(Y = 1|D = unprivileged) − Pr(Y = 1|D = privileged), as shown in Fig. 1.

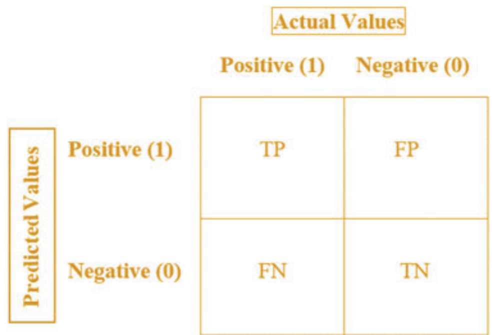

Figure 1: Confusion matrix in the classification of AI fairness

Based on the above definitions and theoretical analyses, the precision AI fairness can be evaluated by the following indexes.

1) Disparate Impact (DI):

$$DI = \frac{\Pr\left(\mathrm{Y} = 1|\mathrm{D} = \text{unprivileged}\right)}{\Pr\left(\mathrm{Y} = 1|\mathrm{D} = \text{privileged}\right)}$$

A value of one means that both groups benefit. Equally, a value less than one means that the privileged group benefits more, and a value greater than one means that the unprivileged group benefits more [39].

2) Average Odds Difference (ADE):

$$\mathrm{ADE} = \frac{1}{2}[(\mathrm{FPR}_{\mathrm{D=unprivileged}} - \mathrm{FPR}_{\mathrm{D=privileged}}) + (\mathrm{TPR}_{\mathrm{D=unprivileged}} - \mathrm{TPR}_{\mathrm{D=privileged}}))]$$

where FPR-False Positive Rate $= \frac{FP}{N}$ and TPR-True Positive Rate $= \frac{TP}{N}$.

This is the average of the differences in false positive and true positive rates between underprivileged and privileged groups. Because this is a classification metric method, it must be computed using a classifier's input and output datasets. A zero value means that both groups benefit equally; a value less than zero means that the privileged group benefits more, and a value greater than zero means that the unprivileged group benefits more [39].

3) Equal Opportunity Difference (EOD):

$$\mathrm{EOD} = \mathrm{TPR}_{\mathrm{D=unprivileged}} - \mathrm{TPR}_{\mathrm{D=privileged}}$$

This is the difference between unprivileged and privileged groups in terms of true positive rates. Because this is a method in the classification metric class, it must be computed using the input and output datasets to a classifier. A zero value means that both groups benefit equally; a value less than zero means that the privileged group benefits more, and a value greater than zero means that the unprivileged group benefits more [39].

4) Generalized Entropy Index (GEI):

$$\in(\alpha) = \begin{cases} \dfrac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left[\left(\dfrac{bi}{\mu} \right)^{\alpha} - 1 \right], & \alpha \neq 0, 1, \\ \dfrac{1}{n} \sum_{i=1}^{n} \dfrac{bi}{\mu} ln \dfrac{bi}{\mu}, & \alpha = 1, \\ -\dfrac{1}{n} \sum_{i=1}^{n} ln \dfrac{bi}{\mu}, & \alpha = 0, \end{cases}$$

where **b**–parameter over which to calculate the entropy index; $\boldsymbol{\alpha}$–the weight given to distances between values at different parts of the distribution is adjusted by this parameter. A value of 0 is equivalent to the mean log deviation, 1 is the Theil index, and 2 is half the squared coefficient of variation.

The generalised entropy index assesses inequality across a population. It is a consistent measure of individual and group fairness [54].

5) Consistency Score (CS):

$$CS = 1 - \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - \frac{1}{n\_neighbors} \sum_{j \in N_{n\_neighbors}(x_i)} \hat{y}_i|$$

where **x**–sample features; **y**–sample targets and **n_neighbors**–number of neighbors for the KNN (K-Nearest Neighbors) computation. This is an individual fairness metric that measures how similar labels are in similar cases [43].

There are many other metrics available for individual, group, and subgroup fairness measures, such as the Theil index (the generalized entropy index with $\alpha = 1$), the coefficient of variation (two times the square root of the GEI with $\alpha = 2$), etc.

3 Constraints on Precision Fairness

3.1 Inexplicability of AI Algorithms

To some extent, it is very difficult to construct responsible and ethical AI systems. Unnavigable biases exist everywhere due to the inexplicability of AI algorithms. They can affect AI systems at almost every stage. Even if the deviation caused by training data can be eliminated, it cannot be ensured that the model user's data can accurately represent the real world, which may lead to poor performance in the real world. In order to speed up or improve the explicability of training processes, we even need to introduce human bias [14,43,44].

Scientists attempted to explain how AI models predict, how AI models are queried, how the real data are collected for a particular prediction or series of predictions, and how AI mechanisms can be presented to humans in an understandable way. However, the AI model is usually a black box. We have no right to access or understand information about the underlying model. We can only use the input and output of the model to generate an explanation. Thus, we cannot see the details of the work inside the model. We try to design a simpler white-box model so that we can access the underlying model, so it is easier to provide information about the exact reason for making a specific prediction. But the price is that we cannot capture the complexity of the relationship in the data. We must face the tradeoff between interpretability and model performance.

AI explicability depends on the type of data you are exploring. In the process of exploration, we can learn which features are more important to the model than other features and which factors play a role in specific predictions. But this is not the whole story of AI fairness. In order to ensure such fairness, we need to delete all sensitive attributes in the design of the AI model. However, sensitive features may be critical to the explicability of the AI model. In addition, sensitive features for AI may be hidden in other attributes, and the combination of non-sensitive features needs to be used to determine the value of sensitive features [7,40], which imposes new biases in applications [2,36]. This offers a baseline for computing AI fairness [20,39].

3.2 Unavoidable Biases in the Datasets

Since AI algorithms learn from the data input to the machine, fairness cannot be ensured if the data is biased. On the one hand, every user wants the algorithm to be fair to himself and his group, but on the other hand, for the sake of privacy, many users are unwilling to share their personal data and sensitive information with the algorithm designer. It is very difficult to establish trust between users and AI researchers. On the premise that privacy/security is not involved, the whole society should give AI researchers as much support as possible to achieve AI fairness [23].

The real-world datasets for fairness are still very limited, including the COMPAS Dataset [53,57], UCI Adult Dataset [27,32], German Credit Dataset [8,46], Recidivism in Juvenile Justice Dataset [53], Communities and Crime Dataset [46,51], Diabetes dataset [3], Student Performance Dataset [12,26], OULAD Dataset [41], Diversity in Faces Dataset [31,34], Pilot Parliaments Benchmark Dataset [19], Dutch census dataset [1], Bank Marketing Dataset [13,27]. The algorithm designer must be clear about his own values and consider the information needs and expectations of all individual or group users. Only such a virtuous circle can minimize the deviation in data collection and effectively promote the realization of AI fairness.

It should be our obligation and responsibility to promote the development of AI towards fairness. AI fairness has been linked to everyone's daily life. Human beings use AI to decide how to interact with others, thus making all parties better. This may be a problem in a broader sense. But to achieve the fairness of the algorithm, we must go beyond the technical level and rely on social forces. This may represent a higher level of civilization. When all mankind cooperates for a higher level of civilization, AI can better serve mankind and help us in all aspects of life. The fairness of AI systems should not be determined only by data engineers and scientists. Governments, companies and institutions also need to consider the views of stakeholders and end users, including whether users feel that they are being treated by AI fairness and what is the public's expectation of AI fairness [52], as shown in Fig. 2.

Particularly, because AI systems learn from historical data, which encodes historical biases [21]. Evaluation models of product aggregation can lead to product aggregation bias [22,23]. Social bias could be caused by statistically biased algorithms or by other objective factors [23–25]. More details of measurement bias, evaluation bias, content production bias, emergent bias, ranking bias and population bias, behavioral bias, temporal bias, linking bias, and presentation bias can be found in associated publications [2]. There are several other categories of biases. Biases in engineering AI applications will bring high costs. We can partly reduce some biases in image and signal processing algorithms by analyzing the processing algorithm and improving data analysis and feature extraction [11]. Some biases can be recognized though the use of custom normalization restrictions or cost functions that determine the relative cost of making an incorrect decision, where the bias can be minimized by adversarial learning algorithms or resolved at the output stage by adjusting the labeling of specific outputs [15]. Open-source tools are also helpful for mitigating biases, such as What-if, AI Fairness 360, Local Interpretable Model-Agnostic Explanations, FairML, Aequitas, Fair learn [17,37,39,48], etc.

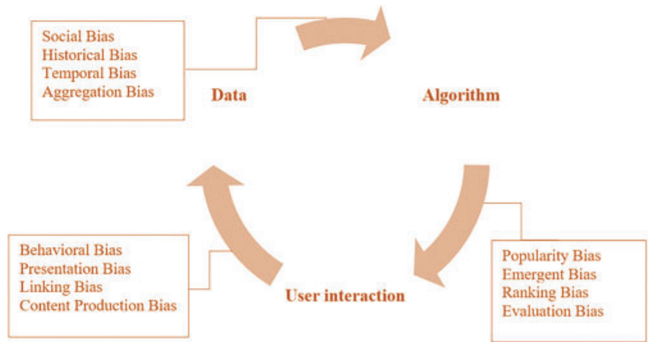

Figure 2: Block diagram of bias definitions in the data, algorithm, and user interaction processes (Feedback loops are placed on the arrows that are most appropriate for them) [35]

3.3 Imbalance of AI Utility with Humanization

We are in an era of rapid development, and we need to use AI to achieve optimal resource allocation and work management process decision-making, but from the perspective of humanity, we need some buffer because whether the AI system is fair depends on the end user. Whether the system has advanced technology is not so important. People will judge its fairness mainly according to their own views on the algorithm used to solve problems and whether the results conform to their own values. Algorithm designers must consider the balance of AI utility with humanization.

We need to consider and pay attention to the inclusiveness of the algorithm for minorities. In other words, minority data should be taken into consideration rather than regarded as abnormal data. In this sense, what people ultimately judge is the fairness of algorithm designers. Because the fairness of AI is more complex than simply finding technical solutions, the development of algorithms also needs more humanized methods. We must seek a balance between the pursuit of utility and human care so that its fairness can be recognized and accepted. We should not simply look for fairness through optimized and reasonable AI algorithms.

Only in this way can we expect that AI can perceive itself and ensure fairness in the future. Without considering the above balance relationship, the algorithm cannot be responsible for calculating results. At present, the development direction of AI is still dominated by human beings, and algorithm designers are considered to have the ability to be responsible for AI decision-making and AI use. Therefore, to establish a fair AI system, algorithm designers will be more scrutinized, and whether the AI system they develop can be trusted will also be affected by this.

4 Summary and Discussion

With the wide application of AI algorithm in various sectors of society, the fairness of the algorithm is receiving more and more attention. Research on algorithm fairness will promote AI applications to be inclusive and unbiased. Due to data bias, algorithm defects, and even human bias, existing AI algorithms generally have “discriminatory phenomena” that have unfair effects on certain specific populations. In the past few years, the industry has been gradually exploring some targeted solutions, including building more fair data sets, introducing fairness constraint losses in algorithm training, and improving the interpretability of machine learning algorithms. However, to ensure a fair and ethical result, we not only need to face the challenges from data science but also need the people who set up AI learning procedures to have great responsibility and tenacious faith to set up the fairest procedures.

Fairness is the subjective practice of using AI without bias or discrimination, especially related to human beings. But the reality is that AI fairness is a very difficult field. It requires algorithm designers to define a fair appearance for each use case. However, it is difficult to define an international label to balance group equity and individual equity. When trying to explain all or part of the machine learning model, we will find that the model contains bias. The existence of such bias may mean that the AI model is unfair. We can explain how or why the model makes such unfair predictions. However, this bias is from social cognition against specific groups, individuals, or characteristics, which has many forms and cannot be accurately reflected by simple white box models, which further limits the explicability of AI systems.

Hence, we must try our best to reduce biases in the datasets for evaluating AI fairness and do well in keeping the balance of AI utility with humanization [59,60]. In most cases, AI technology developers have no subjective will to cause bias. However, there are also some prejudices due to the tendency of decision-making because the algorithm designer may choose shortcuts based on efficiency considerations and may also be affected by the values of his circle. This will lead to deviations in the development and design of AI systems [1–4]. We hope to design an efficient AI system with fairness but inadvertently introduce bias into the system. It is necessary to build an AI fairness ecosystem. People pay too much attention to the privacy information involved in biometric identification but ignore that AI system fairness is also a component of social security and has been applied to intelligent decision-making in many security fields [10,11]. The social welfare brought by AI fairness will benefit the whole society [9].

In order to ensure that vulnerable and ordinary groups can enjoy the benefits of AI fairly, AI data sets should be diverse and require the active participation of the whole society to ensure that the algorithm design and data sets are unbiased. Federated learning may present some potential solutions [59]. However, the current federated learning processes cannot overcome the above constraints in AI fairness. Due to the self-interest of participating clients, there are potential differences in computing communication resources, data, and other factors between them [59,60]. It is crucial to maximize clients' motivation, allocate rewards reasonably, and promote the enthusiasm of federated participants by reconciling the current federated learning processes [60]. Research on AI interpretability also plays an important role in eliminating bias, and fortunately, the interpretability of AI is not a completely unsolvable black box; it has been developing with the practical application of AI [61]. One constructive example of the inexplicability of AI is to explore the working principle of the classification model. Human intervention can be added to the model when designing the portrait classification application [62]. Then, the classifier is prohibited from using skin color as the classification basis, thus avoiding the problem of racial discrimination [61,63]. This may bring new insights into the precision of AI fairness and, in turn, can be considered as a new approach to reconciling with the current federated learning processes.

5 Concluding Remarks and Outstanding Questions

AI fairness is beyond the "Theory of Everything", but ethics is endowed by human beings. We can only evaluate whether the solutions recommended by AI are fair in a specific social context. Therefore, algorithm designers also need those who understand social norms and have high moral values. Based on the consideration of the principle of AI fairness in this paper, it is necessary to divide the risk of AI fairness into multiple levels. The top level belongs to the AI system with the most prejudice, which needs to be completely or partially prohibited. The bottom layer is the AI system with no or only minimal bias, which has no special regulatory requirements. In order to ensure that the AI system serves mankind more efficiently, the public needs to participate and make more efforts to actively participate in the construction of the AI fairness evaluation data set. Finally, the theoretical research on AI fairness also needs further breakthroughs. Only by improving the interpretability of the AI system can we gain AI fairness at the minimum cost.

The outstanding questions for the subsequent studies include:

1) How can we find a more effective way to promote the theoretical research of AI fairness and help improve its interpretability and credibility?

2) How do we ensure the availability of AI fairness data and the availability of public data from the computing power layer to the algorithm layer?

3) How does integrating the data layer, application layer and solutions constitute a virtuous circle to help the AI system get better developed and applied?

4) Since the fairness of AI and AI applications will be affected by space-time factors, how to form global fairness and ensure its long-term effectiveness?

5) How do our governments and laws serve the special requirements for algorithm design, computing power, architecture and data support?

6) AI fairness is a profound philosophical problem. We hope to find a fair design pattern that integrates the data layer, application layer and user interaction. Fortunately, there have been some studies to solve the problem of unfairness in terms of data and algorithm, respectively. But how to integrate these existing methods into a virtuous circle is still an open problem. We expect subsequent studies by other researchers to supplement relevant analyses.

Acknowledgement: The first author acknowledges the support provided by the National Academy of Sciences India (NASI) to carry out the work and is thankful to the Director of the National Institute of Advanced Studies (NIAS) Bangalore for the infrastructure support provided during the work.

Funding Statement: The first author would like to thank the National Academy of Sciences India (NASI), Allahabad, India for the support and to the Director, National Institute of Advanced Studies (NIAS), Bengaluru, India for providing the infrastructure facilities to carry out this work. This research was also supported by the Shanghai High-Level Base-Building Project for Industrial Technology Innovation.

Author Contributions: Lalit Mohan Patnaik: Conducting literature survey on all aspects of fairness and bias, review of the tools and preparation of the draft version of the paper. Wenfeng Wang: Providing details on federated learning and the relavent analysis, integrating these concepts with the rest of the paper.

Availability of Data and Materials: All the data utilized to support the theory and models of the present study are available from the corresponding authors upon request. The source code and data of our project can be accessed in https://drive.google.com/drive/folders/1oXGV7l3msNk4KbNQkd4h3yf2mX5AGwlL?usp=sharing.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J. et al. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering, 7(6),* 719–742.
2. Narayanan, D., Nagpal, M., McGuire, J., Schweitzer, S., Cremer, D. D. (2023). Fairness perceptions of artificial intelligence: A review and path forward. *International Journal of Human-Computer Interaction, 40(1),* 4–23.
3. Landers, R. N., Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist, 78(1),* 36–49.
4. Li, J., Chu, Y., Xu, J. (2023). Impression transference from AI to human: The impact of AI's fairness on interpersonal perception in AI-mediated communication. *International Journal of Human-Computer Studies, 179(1),* 1–11.
5. Takan, S., Ergün, D., Katipoğlu, G. (2023). Gamified text testing for sustainable fairness. *Sustainability, 15(3),* 2292.
6. Nakao, Y., Strappelli, L., Stumpf, S., Nasser, A., Regoli, D. et al. (2023). Towards responsible AI: A design space exploration of human-centered artificial intelligence user interfaces to investigate fairness. *International Journal of Human-Computer Interaction, 39(9),* 1762–1788.
7. Zhang, J., Shu, Y., Yu, H. (2023). Fairness in design: A framework for facilitating ethical artificial intelligence designs. *International Journal of Crowd Science, 7(1),* 32–39.
8. Giovanola, B., Tiribelli, S. (2023). Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & Society, 38(2),* 549–563.
9. Choi, Y., Farnadi, G., Babaki, B., van den Broeck, G. (2020). Learning fair naive bayes classifier by discovering and eliminating discrimination patterns. *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 34, no. 6, pp. 10077–10084.
10. van, B. N., Sarsenbayeva, Z., Goncalves, J. (2023). The methodology of studying fairness perceptions in artificial intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies, 170,* 102954.
11. Shulner-Tal, A., Kuflik, T., Kliger, D. (2023). Enhancing fairness perception-towards human-centred AI and personalized explanations understanding the factors influencing laypeople's fairness perceptions of algorithmic decisions. *International Journal of Human-Computer Interaction, 39(7),* 1455–1482.
12. Deepak, P., Abraham, S. S. (2020). Fair outlier detection. *21th International Conference on Web Information Systems Engineering*, pp. 447–462. Amsterdam, The Netherlands.
13. Galhotra, S., Saisubramanian, S., Zilberstein, S. (2021). Learning to generate fair clusters from demonstrations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 491–501.
14. Harani, S., John, G. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. arXiv preprint arXiv:1901.10002.
15. Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. In: *Advances in neural information processing systems*, vol. 29, pp. 3315–3323. Barcelona: Curran Associates Inc., NY, USA.

16. Hee, J. R., Hartwig, A., Margaret, M. (2018). InclusiveFaceNet: Improving face attribute detection with race and gender diversity. *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, Stockholm, Sweden.
17. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F. et al. (2019). The What-If Tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 26(1),* 56–65.
18. Jon, K., Sendhil, M., Manish, R. (2017). Inherent tradeoffs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, Berkeley, CA, USA.
19. Buolamwini, J., Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR, New York, USA.
20. Judea, P. (2009). *Causality*. Cambridge: Cambridge University Press.
21. Townson, S. (2023). Manage AI bias instead of trying to eliminate it. *MIT Sloan Management Review, 64(2),* 1–3.
22. Garin, S. P., Parekh, V. S., Sulam, J., Yi, P. H. (2023). Medical imaging data science competitions should report dataset demographics and evaluate for bias. *Nature Medicine, 29(5),* 1038–1039.
23. Stine, A. A. K., Kavak, H. (2023). Bias, fairness, and assurance in AI: Overview and synthesis. In: *AI Assurance*, pp. 125–151. Academic Press.
24. Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E. et al. (2023). AI and the transformation of social science research. *Science, 380(6650),* 1108–1109.
25. Chan, A. (2023). GPT-3 and InstructGPT: Technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI and Ethics, 3(1),* 53–64.
26. Le Quy, T., Roy, A., Friege, G., Ntoutsi, E. (2021). Fair-capacitated clustering. *Proceedings of the 14th International Conference on Educational Data Mining (EDM21)*, pp. 407–414.
27. Mahabadi, S., Vakilian, A. (2020). Individual fairness for k-clustering. *International Conference on Machine Learning*, pp. 6586–6596.
28. Ribeiro, M. T., Singh, S., 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*, pp. 1135–1144. San Francisco, California, USA.
29. Matt, K., Joshua, L., Chris, R., Ricardo, S. (2017). Counterfactual fairness. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.
30. Kearns, M., Neel, S., Roth, A., Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 100–109. Atlanta, GA, USA.
31. Michele, M., Nalini, R., Rogerio, S. F., John, R. S. (2019). Diversity in faces. arXiv preprint arXiv:1901.10436.
32. Abbasi, M., Bhaskara, A., Venkatasubramanian, S. (2021). Fair clustering via equitable group representations. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 504–514.
33. Moritz, H., Eric, P., Nathan, S. (2016). Equality of opportunity in supervised learning. In: *Advances in neural information processing systems*, vol. 29, pp. 3315–3323. La Jolla, California.
34. Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.
35. Ninareh, M., Fred, M., Nripsuta, S., Kristina, L., Aram, G. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.
36. Osonde, A. O., William, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Santa Monica, California, USA: Rand Corporation.

37. Pedro, S., Benedict, K., Abby, S., Ari, A., Loren, H. et al. (2018). Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577.
38. Pratyush, G., John, V., Virginia, F. (2020). Fairness metrics: A comparative analysis. arXiv preprint arXiv:2001.07864.
39. Rachel, K. E. B., Kuntal, D., Michael, H., Samuel, C. H., Stephanie, H. et al. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
40. Regan, J. (2016). New Zealand passport robot tells applicant of Asian descent to open eyes. *Reuters News*. https://www.entrepreneur.com/business-news/new-zealand-passport-robot-tells-applicant-of-asian-descent/286198 (accessed on 07/12/2016).
41. Riazy, S., Simbeck, K., Schreck, V. (2020). Fairness in learning analytics: Student at-risk prediction in virtual learning environments. *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020)*, vol. 1, pp. 15–25.
42. Baeza-Yates, R. (2018). Bias on the web. *Communication of the ACM,* 54–61. https://doi.org/10.1145/3209581
43. Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M. et al. (2017). A convex framework for fair regression. arXiv preprint arXiv:1706.02409.
44. Fletcher, R. R., Nakeshimana, A., Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence, 3,* 561802.
45. Richard, Z., Yu, W., Kevin, S., Toniann, P., Cynthia, D. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333. Atlanta, Georgia, USA.
46. Ruoss, A., Balunovic, M., Fischer, M., Vechev, M. (2020). Learning certified individually fair representations. *Advances in Neural Information Processing Systems, 33,* 7584–7596.
47. Corbett-Davies, S., Emma, P., Avi, F., Sharad, G., Aziz, H. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 23, pp. 797–806. Halifax, NS, Canada.
48. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R. et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft Technical Report*, MSR-TR-2020-32.
49. Shai, D., Jonathan, L., Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, 108(17),* 6889–6892.
50. Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J. et al. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536.
51. Slack, D., Friedler, S. A., Givental, E. (2020). Fairness warnings and Fair-MAML: Learning fairly with minimal data. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, vol. 20, pp. 200–209. London, UK.
52. Sorelle, A. F., Carlos, S., Suresh, V. (2016). On the (im)possibility of fairness. arXiv preprint arXiv:1609.07236.
53. Tai, L. Q., Arjun, R., Vasileios, I. (2021). A survey on datasets for fairness-aware machine learning. arXiv preprint arXiv:2110.00530.
54. Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A. et al. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2239–2248. London, UK.
55. Ting, W., Dashun, W. (2014). Why Amazon's ratings might mislead you: The story of herding effects. *Big Data, 4,* 196–204.

56. Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L. et al. (2020). Algorithmic decision making with conditional fairness. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, vol. 26, pp. 2125–2135. California, USA.
57. Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web*, vol. 26, pp. 1171–1180. Perth, Australia.
58. Zhang, B. H., Lemoine, B., Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, vol. 18, pp. 335–340. New Orleans, USA.
59. Zhu, J., Cao, J., Saxena, D., Jiang, S., Ferradi, H. (2023). Blockchain-empowered federated learning: Challenges, solutions, and future directions. *ACM Computing Surveys, 55(11),* 1–31.
60. Pandya, S., Srivastava, G., Jhaveri, R., Babu, M. R., Bhattacharya, S. et al. (2023). Federated learning for smart cities: A comprehensive survey. *Sustainable Energy Technologies and Assessments, 55,* 102987.
61. Guo, J., Wu, J., Liu, A., Xiong, N. N. (2022). LightFed: An efficient and secure federated edge learning system on model splitting. *IEEE Transactions on Parallel and Distributed Systems, 33(11),* 2701–2713.
62. Li, Z., Zhou, Y., Wu, D., Tang, T., Wang, R. (2022). Fairness-aware federated learning with unreliable links in resource-constrained Internet of Things. *IEEE Internet of Things, 9(18),* 17359–17371.
63. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54(6),* 1–35.