Framework for developing algorithmic fairness

Dedy Prasetya Kristiadi[1], Po Abas Sunarya[2], Melvin Ismanto[3], Joshua Dylan[4], Ignasius Raffael Santoso[5], Harco Leslie Hendric Spits Warnars[6]
[1]Computer Systems Department, Raharja University, Indonesia
[2]Informatics Engineering Department, Raharja University, Indonesia
[3,4,5]Computer Science Department, School of Computer Science, Bina Nusantara University, Indonesia
[6]Computer Science Department, BINUS Graduate Program–Doctor of Computer Science,
Bina Nusantara University, Indonesia

Article Info

Article history:

Received Aug 19, 2019
Revised Nov 1, 2019
Accepted Dec 4, 2019

Keywords:

Algorithmic discrimination
Fair algorithm
Fairness in algorithm
Fairness pattern

ABSTRACT

In a world where the algorithm can control the lives of society, it is not surprising that specific complications in determining the fairness in the algorithmic decision will arise at some point. A framework was proposed for defining a fair algorithm metric by compiling information and propositions from various papers into a single summarized list of fairness requirements (guideline alike). The researcher can then adopt it as a foundation or reference to aid them in developing their interpretation of algorithmic fairness. Therefore, future work for this domain would have a more straightforward development process. We also found while structuring this framework that to develop a concept of fairness that everyone can accept, it would require collaboration with other domain expertise (e.g., social science, law, etc.) to avoid any misinformation or naivety that might occur from that particular subject. That is because this field of algorithmic fairness is far broader than one would think initially; various problems from the multiple points of view could come by unnoticed to the novice's eye. In the real world, using active discriminator attributes such as religion, race, nation, tribe, religion, and gender become the problems, but in the algorithm, it becomes the fairness reason.

Corresponding Author:

Harco Leslie Hendric Spits Warnars,
Computer Science Department, BINUS Graduate Program–Doctor of Computer Science,
Bina Nusantara University,
Anggrek Campus, Jl. Kebon Jeruk Raya No.27, Jakarta, Indonesia 11480,
Email: spits.hendric@binus.ac.id

1. INTRODUCTION

We are living in a society where the invisible algorithm can affect various aspects of our lives, whether we realize it or not. Things such as the filtering of email spams, page-ranking [1] search queries, or showing the recommended ads based on our purchase history, are pervasive in this era that most people often do not think twice at how it works. Furthermore, we also may not have realized how big the data that has flowed over the global internet, how it is stored, and how it is processed [2]. Based on the statistic conducted in 2018, internet users reached up to 3.8 billion users in 2017 alone, and the number of data flowing over the internet has passed over the Zettabyte threshold (1,000,000,000,000,000,000, 000 bytes) [3][4].

The rationale for introducing more automated processes is due to the "rational" nature of an algorithm, with the hopes of eliminating discrimination and inequality side effects of the current conventional system caused by human error and irrationality. However, those algorithms are starting to be subject to skepticism when it shows a sign of biases and discriminatory behavior. One trending example is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk tool of Broward County [5]. COMPAS is a system to assess a defendant's probability of recidivism (reoffending their crime) based on the 100 chosen attributes (e.g., age, sex, and criminal history) of an individual [6]. ProPublica's previous investigation of the risk tool found that the system is biassed against black defendants [7]. Such that it would falsely flag black defendants as future criminals and would rate African-American defendants a higher risk score than the white defendants.

In this paper, we are going first to analyze some standard definitions of algorithmic fairness, a problem that has been a favorite topic among computer scientists and ethics scholars. The first step is to acknowledge that Artificial Intelligence (AI) will always have its limitations, and will not still work as we intended, here in particular with biases. The word "bias" in AI application can have multiple intersections with each definition (e.g., statistical bias, society bias, machine learning bias variable, error bias, etc.), and that will confuse when we want to discuss it [8]. Therefore, in this paper, when we say "bias," we are referring to the statistical bias connotation that is commonly related to machine learning and statistical information. Then, we will mention some popular views and theorem of right decision-making strategy, and attempt to think objectively how it can be realistically and practically implemented in a real-world situation. We want to learn if it can work as far as it was designed for and look for some limitations and drawbacks while considering some possible complexities with it.

The objective of this paper is not to construct our interpretation of algorithmic fairness definition, but to help other scholars and various domain experts discover an alternative explanation to it. This paper is created in the hope of providing a guideline that can contribute an invaluable resource to aid other researchers in their research. Moreover, we also wanted to remind the reader that this paper is far from being comprehensive. There is essential more definition of fairness proposed than this paper will or can mention. The discussion in this paper will be generally explained in layman's term (i.e., non-mathematical-heavy and theoretical), as formally defining what "fairness" means in an algorithmic term is not the main objective of this paper. We will also avoid common debate of "What is the objective of using algorithm for analyzing data" (i.e. is it to maximize accuracy, or is it trying to understand the data and its cause-and-effect), and will instead focus on "What are the set of things that researchers should be worrying about when defining their interpretation of algorithmic fairness definition." This is because many things can go unnoticed to try to design a definition of algorithmic fairness, or even deliberately skimped over for whatever the reason may be.

2. BACKGROUND

To prepare for the discussion in this paper, we will start by setting up the environment and defining some basic terminology that can establish the foundation of this paper to avoid misconception and thus enforce correct understanding between the reader and the author. Then, we will also mention some related works that have been made before this paper to help us ground our knowledge and establish a civilized discussion with their proposed arguments and ideas. Note that starting this point forward, when we say "algorithm" as a singular noun, we are referring to the machine learning algorithm that is used in a decision-making system. In particular, the classifier section of machine learning.

2.1. Problem setup

To help the reader visualize how different definitions of fairness will affect the view of a model's (i.e., an estimator) performance, we will use a binary classification 2x2 confusion matrix, as shown in Table 1. It is simple yet very effective at delivering a message and presenting an idea, and it also yields useful insights that apply to any context. For example, it can be of loans (default vs. did not), hiring (succeed or not), or pretrial case (recidivate or not). The pretrial case (criminal justice) is the one that this paper will mostly use since we assume that the reader will have a basic understanding of this paper's problem, and be familiar with the illustration as the consequence of the COMPAS controversy.

In this scenario, we assume there are some N number of defendants, grouped A or B (i.e., white, or black race), that will be assessed by a risk-assessment tool (i.e., an estimator). This tool will rate each defendant by their features for their risk-score. This risk-score will be a value that can either be a rating of high-risk or low-risk (i.e., the probability that a defendant has a high or low chance of recidivating, respectively). The decision-maker will then decide what the outcome will be for the defendants. It can either be positive (detained, because of high-risk rating), or negative (released, because of low-risk score). Here we assume that the positive label is the desired result of a classifier. The features of each defendant can further be separated into two different attributes, protected (e.g., race, gender, religion, etc.) and unprotected (e.g., age, criminal history, nationality, etc.). How the tool will assess this information is invisible to the user, and the target outcome for the prediction is also unknown at the time of making the decision (i.e., the model will not know the "right" result for each defendant). In a real-world scenario, the prediction will not be as clear-cut as outputting 1 and 0 exclusively. This means that rather than a discrete set of output, a continuous set is expected (ranging from 0 to 1 inclusive) as the probability of an individual being high-risk.

The four metrics that will be used are the one that comes from this confusion matrix. That is False Positive (FP), True Positive (TP), False Negative (FP), and True Negative (TN). We use this because every definition of fairness can be trivially measured by algebra with one or more of these four metrics (e.g., Precision will be TP/(TP+FP)). We are thus defining more than these four bases definition is unnecessary.

Table 1. Confusion matrix

	Labeled Low-Risk	Labeled High-Risk
Do not Recidivate	True Negative (TN)	False Positive (FP)
Recidivate	False Negative (FN)	True Positive (TP)

2.2. Related work

The study of developing an appropriate decision-making algorithm to prevent disparate impact is not new by any means, though it is arguably still on its adolescent (development stage). Kleinberg et al. specify their three desired property of fairness as [9]

a. Calibration within groups: For every group, a score of x should mean an x fraction of people in category b is positive, independent of each group’s protected attribute.
b. Balancing for the positive class is when the scoring average of positive (+) members in one group is equaled with the scoring average of positive members in another group.
c. Balancing for the negative class is when the scoring average of negative (-) members in one group is equaled with the scoring average of negative members in another group.

Kleinberg’s impossibility theorem states, “In any instance of risk score assignment where all three properties can be achieved, we must have either perfect prediction or equal base rates.” Perfect prediction in this regard means that for each feature vector, either everyone is in the negative class, or everyone is in a particular class. Then, we can assign everyone a score of either 0 or 1 to everyone, nothing in between. Same base rates as in both groups have the same fraction of positive instances. However, this also means that there is a little risk score corresponding to this base rate for everyone (rating everyone the same score) [10].

Chouldechova in her article [11] explore the statistical bias in Recidivism Prediction Instruments (RPI). She also discovers a similar conclusion, where the criteria that she chose – positive predictive value, false-positive rate (FPR), and false-negative rate (FNR) – cannot all be simultaneously satisfied when recidivism prevalence (number of positive result over the total population) differs across groups. She also adds that disparate impact can occur when the RPI fails to satisfy the fairness criterion of predictive parity.

Corbett-Davies et al. continues the work [12], and try to take it from a different perspective, while still being related with the previous impossibility theorem for fairness in an algorithm, taking the view of a Utilitarian. He reformulated algorithmic fairness to maximize public safety by including race as fairness attribute as a risk score with race-specific threshold[12]. The author shows that there are (or will be) a tension between upholding the short- and long-term utility, which questions “Should we choose to maximize public safety, but at the cost of an inequitable future,” posing the burden on the policymaker. He also highlights a drawback of the previous “three-space” model by Sorelle (Construct Space, Observed Space, and Decision Space) [13], in that algorithms, are generally a subset of a more extensive system with a very specific task and requirement in mind (set by the organization), which often disregards “what data is being used” in the process.

Further studies have also been made by several authors of diverse domain expertise. Some author suggesting a stronger notion on individual’s protected attributes in order to avoid inference of proxies utilizing vulnerable unprotected characteristics [14], and some proposed the method of “adaptive” decision making (called active framework) [15] which leverage higher degree of freedom compared to randomization-based classifier which previously considered optimal [16]. In present research done by Corbett-Davies et al. [17] however, shows that there is still a limitation to that idea, and the author recommends separating the statistical (data) problem from the policy problem of designing interventions.

3. DISCUSSION

In this section, we will go through each sub-topics with a basic level of understanding and provide critical reasoning and arguments to the problem. There is also a distinction with the more popular Group Fairness, and the less popular Individual Fairness. In this part, we are going to clarify those differences and how they can influence the definition of fairness, and while we are at it, a couple of thought-provoking questions will be rhetorically asked the reader, systemically building and formulating the framework of our algorithmic fairness guidelines.

3.1. Impossibility theorem and group fairness

Group fairness is often correlated with statistical bias. On a technical standpoint, statistical bias means the value of differences or disparities between an estimator's expected value (target) and the actual value, which can be caused by a variety of reasons (note that this is different from the other bias, society bias, which implies implicitly biased behavior or opinion of humanity or society). The mathematical definition of fairness is often correlated with one definition of fairness, demographic parity, or calibration, which often defined with sensitive protected attributes[18]. This definition ensures that a risk score x on group A means the same thing if given on group B. In this sense, COMPAS risk scores are not biased. That is, the “40%” risk score produced by the estimator means that 40% of the individuals on that group is positive, independent on a protected attribute (i.e., race).

We need to reframe this problem: “It is not all about mathematical correctness, or even about political correctness. It is about finding a perfectly balanced notion that can account for every party's opinions.” Because, as Kleinberg theorem implies, what everyone collectively wanted is not realistically achievable. There is an inherent trade-off: you have to choose out of these satisfying notion of fairness, of which do you want to satisfy, and which do you not want to fulfill. One comprehensive insight into this is that different metrics matter to a different perspective. For example, the decision-maker might care for those it labeled high-risk, how many of them will recidivate (i.e., Predictive value). The innocent defendants, on the other hand, might wonder what the probability is for them to be incorrectly classified as high-risk (i.e., false-positive rate). Moreover, society might want the selected sets of people in each category/group to be demographically balanced.

Chouldechova suggests picking two metrics we "care most about" (i.e., FPR and FNR). However, even that might not be possible. Because the impossibility theorem makes very few or minimal assumptions of how it will handle the difference in prevalence (e.g., will it even treat everybody the same way, and how do we know if itis not using the protected attribute of a group in a discriminatory way). Corbett-Davies’ recent paper shows that even eliminating protected attributes (blindness or anti-classification) from the model’s data can still cause a disparate impact on the result as machine learning is very good at picking up proxies of those attributes. For example, one might look at the ZIP code of an individual to determine if the individual is living in a black or white neighborhood.

Hardt et al. [19] also come to a similar conclusion that being race-blind is proved to be ineffective at enforcing the ideal equity. The author also states that machine learning bias is "just" a side effect of maximizing accuracy. Typically, the decision-maker would not even look at the race (i.e., protected attributes) when maximizing accuracy. There might be conscious or unconscious bias on the model itself. That is, again, because machine learning is much better at picking proxies in the data. How should we be able to know if the results of classification do not infer the skin color of the people by analyzing their geographic location (e.g., neighborhood or zip code)?

Looking at that way, there is indeed no "right" definition. Thus, the question that needs to be answered for the practitioners (or the decision-maker) is "What metrics should be used as the variable in the decision-making process?”, and to an algorithmic fairness designer “To what extent is it acceptable to discriminate a certain group of people to uphold the other?” The first problem correlates with the next sub-topic about individual fairness.

Another question then leads nicely from the previous complication, which is, "How much should we trust the decision-makers intentions?" There two ways to look at this: The first view is that the decision-makers have a callous if not malicious intent; compliance with the fairness definition is the only thing preventing discriminatory behavior. The other view is that the decision-maker has fundamentally good intentions, and the fairness definition can help them avoid unintentional discrimination. This debate should be self-explanatory, and a different perspective will have a different subjective view on it.

3.2. Utilitarianism and individual fairness

Problem with group fairness, however, is that as it is defined with the basis of group averages. To put it, group fairness can discriminate against the known-"average" people, in favor of the average one. It does not promise the non-averages anything about their individual qualities. The same thresholds (assuming the scores are well-calibrated) are generally impossible to pick while still equalizing both FPR and FNR on all groups, which explained from the previous section, is caused by the statistical bias that comes into play. One idea to circumvent this problem is by using the notion of individual fairness. First introduced by Dwork et al. [20], it states “similar individuals should be treated (and grouped) similarly” (based on a metric that defines how similar two given individuals are in the context of a decision-making task).

A critical property of this definition is the formula of “constrained optimization" (i.e., satisfying a particular interpretation of fairness require a restricted set of decision rule). Those constraints are reduced into three central part:

a. Individual A who is labeled as unfavorable must have the same FPR as individual B, who is also marked as negative.
b. Individual A who is labeled as positive must have the same FNR as individual B who is also marked as positive.
c. Individual A who is labeled as positive must not have a TPR lower than the FPR of individual B who is marked as negative.

Individual fairness has a semantically robust and intuitive definition but has yet to gain much progress because of the same problem that group fairness also manifests, which is about specifying the accepted metrics. In making this decision, it is hard not to see the parallelism with a Utilitarianism point of view, focuses on the results and consequences of the actions, treating intentions as (generally) irrelevant. Utilitarian agree that a moral theory should apply equally to everyone, and thought that there is nothing more fundamental than the primal desire to seek pleasure and avoid pain. Fundamentally, the answer comes down to "because I want what I want, which will make me happy," and hopefully also "makes us happy." In this view, one should always act so to produce the highest good for the most significant number, even if it means sacrificing someone for the "greater other."

This is why this algorithmic fairness problem is synonymous with the “trolley problem," as it also has much correlation with the Utilitarian view. This ultimately means that one way or another, a sacrifice will undoubtedly take place. Look at it this way: Fairness is inherently costly. Some trade-offs are bound to happen, and a decision will ultimately need to be made, shown by the proofs produced by Kleinberg on his impossibility theorem. It is not easy to find the right balance of individual fairness (equity) and group safety (utility). This problem of maximizing capital and utility is still of an ongoing debate until now, and it is long before the tension will go away anytime soon.

3.3. Data fairness and other complexities

In many settings, such as what has been observed on the COMPAS case, it is very often the case that if we want to harmonize or balance the outcome between different groups, then we have to treat those people from different groups disparately (e.g. setting different threshold, different base rates, or even having different set of features being used). In many “traditional” scenario, however, we did not need to resolve this tension in some way, such that it demote either disparate treatment or disparate impact (known famously from the court case of Ricci v. DeStefano). We can still uphold them both equally as a principle, and find the clever "workaround" for some occasional cases where they did come into conflicts. In machine learning, we will come into these conflicts all the time. This "case-by-case" workaround merely is impractical to do. That is because the objective of machine learning being employed in these contacts is so that we can have a uniform way of making this decision making a task, while not changing any criteria specific to some instances. By not conforming to this basic idea, we will disregard the initial intention of machine learning itself.

With that regard, it is also important to remember that the data used for the learning of the model will influence the outcome of the prediction heavily. Such a problem is the definition of Data Fairness; that is, how do we determine if the data being used is fair to be used for benchmark and measurements. Another layer of complexity thus further deepens when we are dealing with unrecognized error (e.g., past biased data). Such that it is not only that the accuracy rate might differ, but also the evaluation technique we used will not even tell us that it is going to affect the accuracy in any way. Adding to that complexity, another thought-provoking problem arises: “How should we deal with a system that uses the protected attribute during training, but not during the prediction?” [21].

Moreover, it might not just be that there is bias in the way we collect the data, but the data can encode past discrimination. Barocas, on its AI Now 2017 talk, argues that the score (data) we have trained (the machine) on can misrepresenting what past people were able to do [22]. It can be hard or even impossible to tell how much of a difference in prevalence is affected by measurement bias (e.g., in crime situation, we might be biased to sample recidivism only on a "high-risk” territory), and how much is it due to historical prejudice (e.g. poverty, crime, unemployment), versus other reasons. Therefore, what we might want to consider when designing a metric of algorithmic fairness is, "How do you make an accurate prediction when the circumstances of these collected past data themselves are not fair?” [23].

It is for this reason that some group of the scholar will suggest against using demographic parity as a notion of fairness, and decided to form their alternative approach [24]. This approach separates the data problem (mitigating measurement bias) with the algorithmic problem (minimizing bias concerning input data), instead of clumping it together. One counterpoint to this is that there is possibly no way to solve the data problem. In a criminal justice context, there is a noted lack of information [25,26] with which to correct systematic measurement error in arrest data. Consequently, if we cannot possibly solve the data problem, we should assume – as a null hypothesis – that there are indeed no intrinsic differences, and we should use demographic parity.

3.4. Constructing the framework

We model everything as an optimization today. That is, an answer to the question, "Is a decision-making process fair?" can be an arbitrary number that has general meaning with no formal definition available yet. One important thought-provoking question is, "Can algorithmic fairness even be modeled as an optimization problem, or is the current vocabulary of machine learning too limiting to address it in a meaningful way?" Conversely, "Does the inscrutability of machine learning limit the usefulness of human intuition as a guide to algorithmic fairness?" We do not have a good intuition of knowing if a particular feature, in conjunction with other elements can lead to an unfair outcome or not.

Before defining any metric of fairness, we must first be able to answer that very fundamental question. After it is resolved, assuming the answer is that it is, in fact, possible, we should move on to the next set of problems. Reiterating our previous points, we must decide which notion of fairness will be satisfied, and which one will not. Specifying this is not easy, as we also need to realize that different perspectives will have a different set of requirements. We cannot satisfy everyone's needs; some trade-offs need to be made. This is why it is analogous to the "trolley problem." We should also consider the results of the decision-making process. Do we want to maximize public safety, but at the cost of future equity?

On the other hand, is it the contrary, where we want to maximize social equity, in hopes of creating an equitable community, costing invulnerability in the process? Who should have the authority to make this crucial decision? A question of determining the individual jurisdiction who have those power is a real problem, and should not be ignored when designing a definition which can account for every group. None of these questions has a comfortable, clear-cut, answer to it. That is why we still cling to the hope of creating a perfectly balanced notion of fairness, where everyone can be treated equally, with minimal trade-offs and approved by the whole society. To achieve that, we listed some guidelines that can be of help for researchers in seeking that answer. All of these questions will come from the three major categories from the previous discussion, that is Group Fairness, Individual Fairness, and Data Fairness. We also added a class for Community and Authority to ensure that the community and authority here, as both a group and individuals, also take responsibility or be accountable for any action taken during this crucial decision-making process [27]. It is categorized in such a way so that other researchers reading this paper can tackle the problem of algorithmic fairness incrementally (segmenting it to different focus groups), rather than trying to address it all at once. Also, note that some of these problem questions can intersect or converge with multiple categories, but we group them only to a maximum of two groups for simplicity.

It is still a work in progress, and some questions are not explained thoroughly from this paper, but every single development made by the community will only bring improvement to this framework. The simplified framework structure is described in figure 1. Nonetheless, here is the list of questions that should be answered by algorithmic fairness designer to avoid any pothole or blind spot during the development process, such as one is: to what extent is it acceptable to discriminate a particular group of people to uphold the other? Two: Should proxy discrimination be a concern even if it is not possible to know if machine learning is inferring any biased information from the proxies? Three: Should we "just trust" in the prediction of an estimator, as it has a better average success rate compared to if it had a human intervention? Four: How to navigate these tradeoffs between different measures of group fairness, between-group fairness and individual fairness, and between fairness and utility? Five: How should we deal with a system that uses the protected attribute during training, but not during the prediction? Six: How do you make an accurate prediction when the circumstances of these collected past data themselves are not fair? How can we even determine if it is appropriate data or not? Seven: To what extent should the machine-learning model reflect society stereotypes and historical data? Eight: Should dataset curators be required to do a biased assessment, and should researchers who release pre-trained models have the same obligations? If so, should such models be "de-biased"? Nine: Who should have the authority to be the decision-maker and intervention?

Question 4 is initially thought of a problem of both individual fairness and group fairness but is moved to the group fairness category, because of the nature of individual fairness, which can be expanded and simplified into a problem of grouping those individuals into a single homogenous group (i.e. group fairness). It also belongs to the community and authority category, because these problem needs to be addressed not only by the one with the authority but also must have a mutual agreement with all parties. Proxy discrimination (question 2) can belong to both the individual and group fairness category, as both are interchangeable in this context. We put this to the individual fairness to clarify that proxy discrimination could also occur on an individual level of fairness. Also, keep in mind that proxy is also *better* at generalizing data for a group of people than individual metrics.

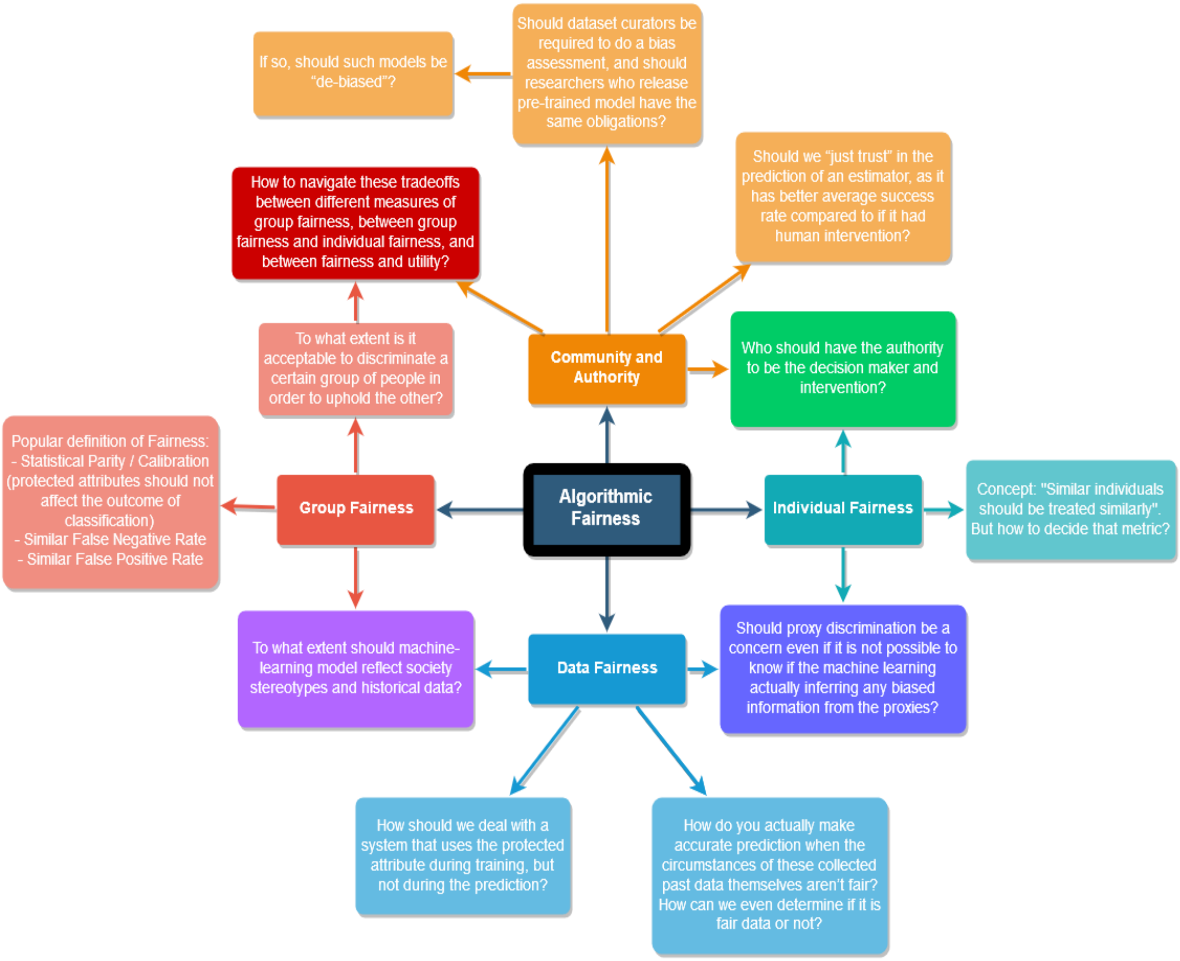
If so, should such models be "de-biased"?
Should dataset curators be required to do a bias assessment, and should researchers who release pre-trained model have the same obligations?
How to navigate these tradeoffs between different measures of group fairness, between group fairness and individual fairness, and between fairness and utility?
Should we "just trust" in the prediction of an estimator, as it has better average success rate compared to if it had human intervention?
To what extent is it acceptable to discriminate a certain group of people in order to uphold the other?
Community and Authority
Who should have the authority to be the decision maker and intervention?
Popular definition of Fairness:
- Statistical Parity / Calibration (protected attributes should not affect the outcome of classification)
- Similar False Negative Rate
- Similar False Positive Rate
Group Fairness
Algorithmic Fairness
Individual Fairness
Concept: "Similar individuals should be treated similarly". But how to decide that metric?
To what extent should machine-learning model reflect society stereotypes and historical data?
Data Fairness
Should proxy discrimination be a concern even if it is not possible to know if the machine learning actually inferring any biased information from the proxies?
How should we deal with a system that uses the protected attribute during training, but not during the prediction?
How do you actually make accurate prediction when the circumstances of these collected past data themselves aren't fair? How can we even determine if it is fair data or not?

Figure 1. Our Framework for the building of Algorithmic Fairness, structured in four primary categories such as Group Fairness, Individual Fairness, Data Fairness, and Community and Authority, within a single simplified diagram. Connecting to that four categories will be some of the critical questions (objectives) that will be correlated with the parent category. Some of the targets that are derived from more than one type will be colored differently for clarity.

4. CONCLUSION

The massive advancement in machine learning has provided us with plenty of benefits throughout various aspects of our lives. This is the reason why we should care about a problem, such as the one being discussed in this paper, fairness in an algorithm. This is a genuine problem that we are currently facing, in that we wanted to maximize prediction, yet we also wanted to ask whether these tools and techniques we are utilizing are not a mirror of human biases (i.e., is it doing the things it is doing somewhat).

There has been an ongoing debate about the "right" metrics used for defining fairness in an algorithm. Many papers have suggested the impossibility of satisfying every desired attribute for fairness, suggesting that trade-offs are inevitable. Nonetheless, it has not stopped other researchers in their attempt of finding a fair algorithm or defining their version of fairness metric.

In this paper, we have demonstrated various conditions and cases of fairness problem and introduced a list of open questions or guidelines that can help other researchers and domain experts in executing their investigation. As shown in this paper, been many challenging situations require deep understanding not only on one domain but also on many others. Conflict after conflict will keep occurring if we do not come up with a solution that can satisfy everyone's interest. This domain of ethical machine learning is still currently at its youth and has a long way in advancing to its maturity. We do not know how close we are in finding the perfect solution for this problem, and we do not even know if there is one. However, we believe that one day, a universally-accepted notion of fairness will be discovered, and that future will without a doubt reach the desired milestone of social equity that many of us still hold on into. Our work in this regard is only a small part of the large community that is working on this topic. With this paper, we encourage other passionate researchers in this community to continue our work and attempt to fulfill the same goal we all have a safe, fair, and equitable future to create diverse communities with different backgrounds. The implementation of algorithmic fairness will give something different for society, where the fairness attribute will come from the current population in society. For example. When dealing with a Scholarship recommender/ recommendation system for university will apply differently in any study program or university. Let say, in the engineering study program; it will have distinct lack of women students in any university, so gender attribute can be used as fairness attribute to find the right scholarship's recipient will be applied differently based on the current number of women students in the engineering study program.

REFERENCES

[1] A. Langville, Google's PageRank and beyond. Princeton [u.a.]: Princeton Univ. Press, 2012.

[2] V. N. Inukollu, S. Arsi, and S. Rao Ravuri, "Security Issues Associated with Big Data in Cloud Computing," in *Cloud Computing. International Journal of Network Security & Its Applications*, vol. 6, pp. 45-56, 2014.

[3] "Data Never Sleeps 6.0", Domo.com, 2018. [Online]. Available: www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf. [Accessed: 08-December- 2019].

[4] J. Winter and R. Ono, "The Future of Internet - Alternative Visions." *Springer International Publishing*, Switzerland, 2015.

[5] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias — ProPublica," ProPublica, 2018. [Online]. Available:www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.[Accessed:1-l1-2019].

[6] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, pp. 1-5, 2018.

[7] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm — ProPublica," *ProPublica*, pp. 1-16, 2016.

[8] B. Dickson, "What is algorithmic bias?" TechTalks, 2018. [Online]. Available: www.bdtechtalks.com/2018/03/26/racist-sexist-ai-deep-learning-algorithms/. [Accessed: 11- June- 2019].

[9] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS),* Berkeley, CA, USA, pp 43:1–43:23, 2017.

[10] J. Kleinberg, "Inherent Trade-Offs in Algorithmic Fairness," in *SIGMETRICS '18 Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, *Irvine*, CA, USA, no. 46, pp. 40-40, 2018.

[11] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," Big Data, vol. 5, no. 2, pp. 153-163, 2017.

[12] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proc. of the 23rd Conf. on Knowledge Discovery and Data Mining (KDD)*, Halifax, NS, Canada, 2017.

[13] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im)possibility of fairness," *Arxiv.org*, 2016.

[14] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti, "Exposing the probabilistic causal structure of discrimination," *International Journal of Data Science and Analytics*, vol.3. no. 1, pp.1-21, 2017.

[15] A. Noriega-Campero, M. Bakker, B. Garcia-Bulle, and A. Pentland, "Active Fairness in Algorithmic Decision Making," *Research Gate*, pp. 77-83, 2019.

[16] "Ricci v. DeStefano," Oyez, 2018. [Online]. Available: www.oyez.org/cases/2008/07-1428.[Accessed 1Dec 2019]

[17] S. Corbett-Davies and S. Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," *Arxiv.org*, 2018. [Online]. Available: www.arxiv.org/abs/1808.00023. [Accessed: 1- August- 2019].

[18] G. Pleiss, M. Raghayan, F.Wu, J. Kleinberg, and K.Q. Weinberger, "On fairness and calibration," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[19] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in 30th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 2016.

[20] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness Through Awareness," *ACM DL Digital Library*, pp. 214-226, 2018.

[21] Z. Lipton, A. Chouldechova and J. McAuley, "Does mitigate ML's impact disparity require treatment disparity?," Arxiv.org, 2018. [Online]. Available: www.arxiv.org/abs/1711.07076v2 [Accessed: 23- Sept- 2019].

[22] S.Barocas, "What is the Problem to Which Fair Machine Learning is the Solution?," *AI Now Experts Workshop on Bias and Inclusion*, 2017.

[23] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic Fairness," in AEA Papers and Proceedings, vol. 108, pp. 22–27, 2018.

[24] J. Johndrow and K. Lum, "An algorithm for removing sensitive information: application to race-independent recidivism prediction," *Arxiv.org*, 2017.[Online].Available:www.arxiv.org/abs/1703.04957.[Accessed: 2-11-2019].

[25] T. Ruggero, J. Dougherty, and J. Klofas, "Measuring Recidivism: Definitions, Errors, and Data Sources," Rit.edu, 2015. [Online]. Available: www.www.rit.edu/cla/criminaljustice/sites/rit.edu.cla.criminaljustice/files/images/2015-03%20-%20Measuring%20Recidivism%20-%20Definitions,%20Errors,%20and%20Data%20Sources.pdf. [Accessed: 29- Nov- 2019].'

[26] K. Crawford, "The Trouble with Bias," in *31th Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[27] B. W. Goodman, "A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the Europian Union General Data Protection," *29th Conf. on Neural Information Processing Systems (NIPS)*, Spain, 2016.