Algorithmic Fairness in a Technology-Based World

Rafal Promowicz

CIS 498 Thesis

May 3, 2021

Abstract

Algorithmic fairness is an umbrella term encompassing the way in which machine-learning models classify information and ultimately make decisions sans human intervention. This literature review seeks to break the field into its sub-components and delve into the research being done in each. Outside of explaining the big-picture state of research, particular attention will be given to the current methods for defining, and then realizing fairness constraints through key algorithms at each stage of the machine-learning workflow. The associated trade-offs involved in increasing the perceived "fairness" of an algorithm from a social and mathematical perspective, as well as the different contexts with which fairness manifests itself will be addressed. Application areas and existing case studies will be discussed to convey the real-world impacts decision-making algorithms are increasingly involved in across a range of domains.

Introduction

Algorithmic fairness is a concept revolving around the usage of computer programs to make equitable, bias-free decisions across a host of application areas. The definitions of "fairness" which are sought are extremely variable and can take on different meanings depending on the context they are being considered in. For example, an algorithm may reasonably be expected to minimize any sort of discrimination stemming from the race of an individual being evaluated when it comes to a decision of whether or not to issue a credit card. However, the same situation raises a different question of fairness along a different "axis". The question of which outcome is "worse" between a qualified applicant being denied a credit card (false negative) or an unqualified applicant being given a credit card (false positive) is not one which can be readily answered despite both of these scenarios undoubtedly violating some aspect of fairness either for the applicant or bank involved.

There exists a fundamental relationship among the satisfaction of an imposed fairness condition and the accuracy of the decision-making model which is central to the concept of "maximizing fairness" while giving up as little prediction accuracy as possible. The mathematical relationship between these fairness and accuracy can often be modelled along a Pareto curve – reducing the task to an optimization problem of finding the points along such a curve which provide the highest overall accuracy and fairness along both axes.

The review will also address some of the underlying reasons discrepancies in fairness arise – looking into the inherent bias of training-set data in particular. Various methods have gained popularity in addressing this concerns including the notion of "bolt-on fairness" ie. applying filters to the algorithm after it has produced an initial result, and simply restricting the type of information an algorithm can use to produce a decision by redacting information such as gender, age, etc. The trade-offs among these and other approaches need to be well-understood and contrasted to effectively apply them to the impactful decision-making scenarios.

Algorithmic Fairness

Varying Definitions

Pinning down a particular definition of fairness is an extremely difficult task which ultimately comes down to context. From a purely legal standpoint, issues of direct and indirect discrimination may arise in the context of algorithmic decision-making. Direct discrimination deals with the a "substantially significant" difference in treatment of members of a particular protected class (Barocas, 2016). This type of discrimination is less likely to take place in practice as algorithmic designers rarely have an expressed intent to harm a particular demographic. The second, more common, form of discrimination deals with the indirect impact the decisions of such algorithms may have as an unintended consequence. Due to the nature of how underlying data impact results of algorithms which are agnostic to common discrimination factors like gender or race, this type of discrimination tends to be what becomes evident in research studies and real-world examples (such as those of the "Examples" section below). It is clear that the concept of "fairness" is specific to the situation, and the next section will explore current approaches to making it measurable.

Quantifying Fairness

Fairness, as the name suggests, is generally determined by the "amount" of injustice inflicted onto a group of interest. In statistical terms, a popular and relatively simple-to-understand method for this is using false positive and false negative calculations for evaluating machine learning models. A confusion matrix is a tool which represents the 4 main outcomes when comparing a machine learning classifier's output to that of the correct result when running it on a labelled testing set. Under the assumption that the space of output results for a model only consists of binary "yes" or "no" decisions, the confusion matrix is comprised of the following outcomes:

- True Positive [TP]: Prediction "yes" matches the correct "yes" label
- True Negative [TN]: Prediction "no" matches the correct "no" label
- False Positive [FP]: Prediction "yes" does not match the correct "no" label
- False Negative [FN]: Prediction "no" does not match the correct "yes" label

The adjacency matrix displaying these figures is has become a popular way to quantify the judgement of machine learning models more broadly through its interpretability and easy-of-application to all kinds of domains and algorithms. The relation to the notion of fairness here comes from the matrix being able to convey what each of these rates are when considering only the data of a particular group of interest as opposed to the entire population. For instance, if a study were interested in a model's handling of input data regarding Asian-American subjects, the confusion matrix for this subset of the population could be computed and compared to the matrix for the general population as an indicator of whether the percentages for each of the 4 occurrences are similar in a statistically significant sense.
Of course, one of the primary drawbacks here is the need for a structured, labeled data set containing information about what the "correct" decision should have been. This poses a varying

degree of difficulty depending on the context of topic addressed. An example that will be explored in the "Examples" section deals with recidivism rates of inmates who may be considered for parole with a machine learning model providing input on what risk category they will likely be in. Acquiring data for whether granting parole was the correct decision is a process which may take years to evaluate – especially since if they do not recidivate upon release (a true positive) some reasonable time frame must be established for which to label that person's correct label as "yes".

Proposed Enforcement Strategies

Researchers have considered various strategies to mitigate unfairness ranging from imposing certain constraints after calculating results to rethinking the way data is fed into machine learning models altogether. Some of the popular approaches to the problem and their associated trade-offs are outlined below. It is important to keep in mind that for any given scenario, a decision must first be made in regard to which of the 4 entries in the confusion matrix is are the source of the greatest "unfairness" and harm to the disproportionately affected population in the context of the scenario – i.e. whether False Positives or False Negatives are the more damaging statistic to try to reduce. The following 3 approaches consider ways to ensure group fairness notations. To clarify their intent, a running example will be used with the total population divided into a majority and minority group where each individual represents a candidate for a position seeking to be hired.

Equality of Odds Approach

This approach was described in 2016 by Moritz Hardt, Eric Price, Nathan Srebro and centers around the goal of ensuring that if a model has the same error rate for each subgroup being studied in the result as they would have individually, then it can be said that it treats each such subgroup the same way (Hardt et. al, 2016). This approach is formalized with the following equality:

$$\Pr(\acute{Y}=1|A=0, Y=y) = \Pr(\acute{Y}=1|A=1, Y=y),\ y\in\{0,1\}$$

Ý is the model's output
A is 1 for the demographic of interest and 0 otherwise
Y is the true outcome

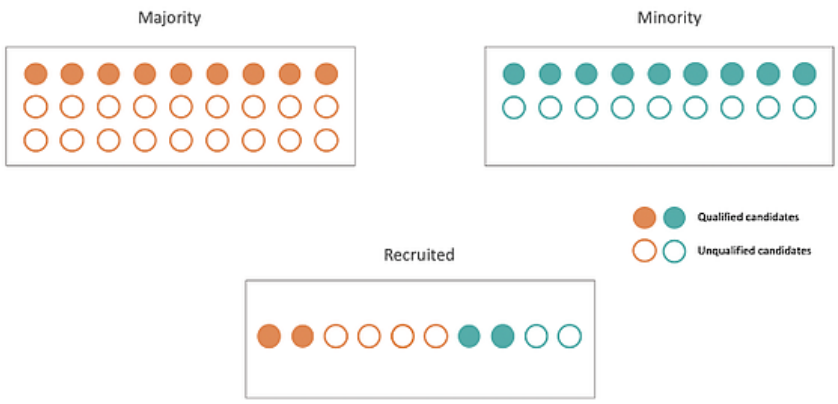

Figure 1: Equal Odds Illustration (Landeau, 2020)

This equality implies that the error rate given either a desirable or undesirable outcome (y=1 vs y=0), is the same for both the population of interest and any other population chosen from the same (A=1 vs A=0). The Equal Odds constraint is, however, a strong enough guarantee that satisfying it can introduce significant decreases in accuracy of the model as a whole.

Equality of Opportunity Approach

Given the limitations described in the Equal Odds method, the authors of the paper created a weaker constraint which is more realistic to apply in practice. The Equal Opportunity method has a similar goal of ensuring that if a model has the same true positive rate for each subgroup being studied in the result as they would have individually, then it can be said that it treats each such subgroup the same way (Hardt et. al, 2016). This is formally defined as:

$$Pr(\acute{Y}=1|A=0, Y=1) = Pr(\acute{Y}=1|A=1, Y=1)$$

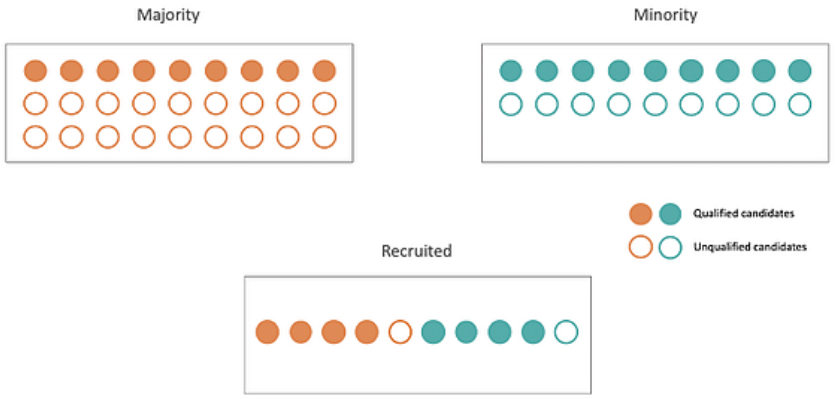

Figure 2: Equal Opportunity Illustration (Landeau, 2020)

The key difference translates to the probability of the model producing an incorrect output being the same between the group of interest and all other groups – but only if the outcome studied is positive or desirable (represented by the Y=1 term).
However, in both the Equal Odds and Equal Opportunity approaches, some limiting factors inhibit the goal of truly non-biased results. One limitation is that the data which make up training sets will often themselves be disproportionately composed of information about one group over another (such as gender). If, for example, females have far less access to X than males and they are consequently only represented in 25% of the data then only this percentage will be chosen when in an algorithm dealing with X which adheres to the equality constraints given (Landeau, 2020). Essentially, the equality constraints do not account for bias which is already represented in the data given at the onset. In addition, both of the constraint can introduce significant decreases in accuracy of the model as a whole – a property which will be explored further in the "Trade-Offs" section.

Predictive Rate Parity

In 2016, Dieterich, Mendoza, and Brennan published a paper describing a new constraint for enforcing a notion of fairness called Predictive Rate Parity. The more popular application of this constraint refers to *Positive* Predictive Rate Parity in particular, which is the idea that among each of the groups in the result, the proportion of people with an affirmative label for the topic studied (the Y=1 term) is the same (Dietrich et. al, 2016). This is defined as:

$$P(Y=1|A=0, \acute{Y}=1) = P(Y=1|A=1, \acute{Y}=1)$$

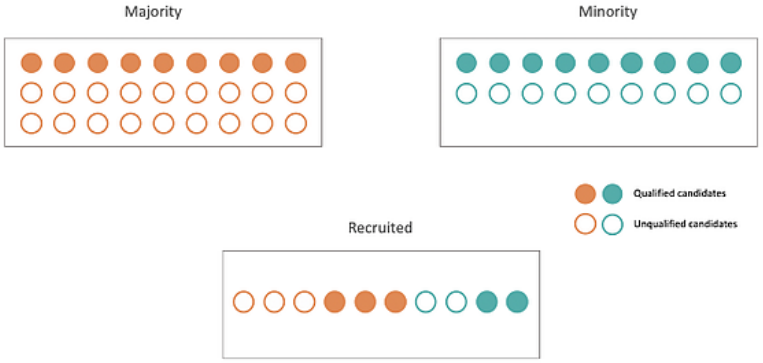

Figure 3: Positive Predictive Rate Parity Illustration (Landeau, 2020)

Individual Fairness Approach

Deviating from the group fairness constraints above, a paper by Dwork et. al in 2012 outlined a notion of fairness on an individual level. The paper defined this type of fairness as the situation where if two individuals being considered are substantially similar by some appropriate metric, then the outputs of the classifier for these two individuals is also substantially similar (Dwork et.

al., 2012). While this formulation has become popular amongst the literature from an abstract perspective, the question about determining "substantial similarity" quickly arises as a non-trivial problem for implementing such a constraint (Chouldechova, 2018).

A visual representation of the concept of individual fairness is as follows:

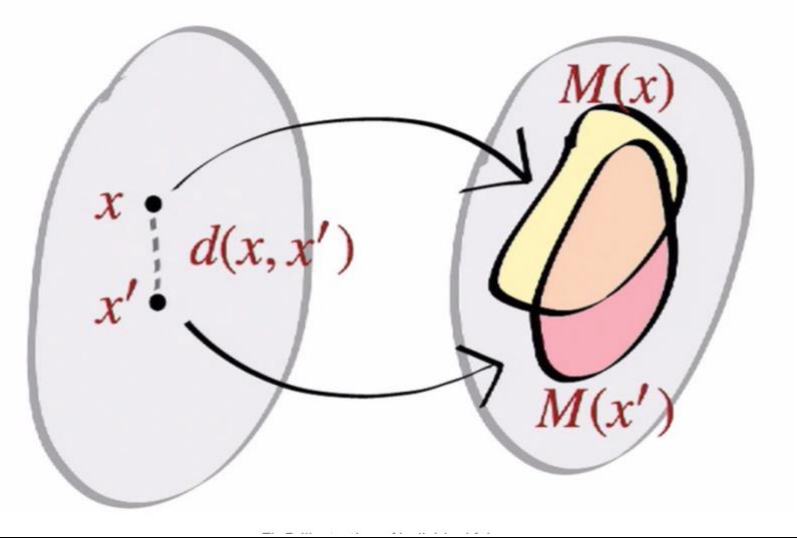

Figure 4: Mapping of Individuals to Illustrate Fairness (Zhong, 2020)

Preservation of fairness at the individual level means that the inequality D(M(X), M(X')) ≤ d(X,X') is satisfied, where D and d are both metric [distance] functions on the input and output spaces respectively, and M is the mapping from an individual data point (like x or x') to the outcome. Figure 4 essentially illustrates the earlier statement of "substantial similarity" and the equation involving metric functions quantifies this notion.
The primary drawback with the above individualized fairness constraint is the difficulty of applying it in practice. While the concept of "substantial similarity" is useful in principle, it becomes arduous to find suitable metric functions that will correctly provide a sense of how far apart two similar inputs ended up being in terms of their output labels after categorization by the model (Kim, 2018). A relevant illustrative example is a scenario in which multiple candidates are being considered for a job posting: one has an undergraduate degree and two years of experience, another has a graduate degree with one year of experience, and yet another has a graduate degree with no work experience. A model can take these inputs and return a suggestion on who to hire, though delving into this decision, it becomes very context-dependent to determine how far apart the candidates are using a metric function as described above – directly making it difficult to evaluate if the definition of individual fairness given is preserved.

Trade-Offs

There exists a fundamental relationship among the satisfaction of an imposed fairness condition and the accuracy of the decision-making model. The Impossibility Theorem of Fairness states that when considering the various definitions of fairness mentioned above, it generally (outside of trivial situations) must be the case that only one of these definitions is satisfied (Saravanakumar, 2021). This theorem poses the natural question of which definition to apply each time, at which point the context of the problem the classifier is dealing with becomes important to establishing which definition can truly lead to an equitable outcome.

Part of the reason for which multiple definitions of fairness cannot be simultaneously satisfied is the trade-off which exists between accuracy and fairness metrics. A paper by Susan Wei and Marc Niethammer detailed the tools required for estimating a Pareto Front in the context of algorithmic bias – essentially a curve formed of points on a fairness-accuracy coordinate grid (Wei, 2020). Plotting such a curve shows results in a concave shape shown below in Figure 4.

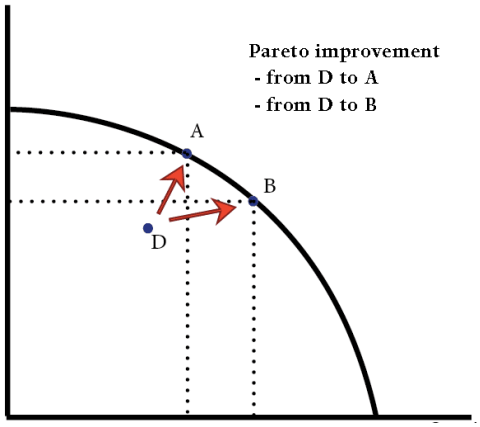

Figure 5: Sample Pareto Curve w/ Improvements Demonstrated

The key insight from the Pareto Frontier is its depiction of the most favorable points a classifier outputs on a spectrum. In Figure 5, taking the x axis to be some quantified scale of fairness and the y axis to be the accuracy of the classifier, only points found along the Pareto Frontier are worth considering when concerned with fining the "best" possible output. It is the case that for any point on the inside of the Pareto Frontier, moving to a point on the Frontier itself will always result in a better outcome. In the case of point D, moving directly upward would result in the same fairness – yet with greater accuracy, while moving directly to the right till the Frontier would result in the same accuracy – yet with greater fairness. As opposed to the strictly 90-degree movements, it is clear that movement from point D to any other point on the Pareto Frontier would constitute an overall improvement, it is a matter of choice for the classifier designers which dimension is more important to maximize in the current context. The shape of the Frontier is itself indicative of the generally inverse-relationship that accuracy and fairness have under the Impossibility Theorem mentioned above – with the highest rates of change in the slope of the curve occurring at the extremes along both axes.

Noteworthy Mitigation Algorithms

The literature organizes attempts at implementing fairness into three categories depending on what stage of the ml-model workflow they operate in: pre-processing data, training-optimizations, and post-processing results.

Pre-Processing and the Reweighing Algorithm

Pre-processing algorithms generally tend to focus on the attributes of the underlying dataset being used and address the inherent bias often present in this data before it is fed into any machine-learning models. These algorithms methodically remove the information they decide may be detrimental to fairness, and the extent of this removal is often controlled by a threshold parameter the user can vary (Feldman et. al, 2015). The reason certain bias-inducing data cannot simply be removed entirely is two-fold. The first, as described in the "Trade-Offs" section, is the inherent loss in accuracy associated with completely leaving out data which is available yet chosen to be disregarded. In addition, researchers have repeatedly found that other attributes in a dataset being studied tend to act as a proxy for the attributes explicitly removed simply based on societal norms or the history of certain marginalized groups. Models often exhibit similarly unfair results even after the removal of the attributes which explicitly identify a minority group.

In 2012 Faisal and Calders published a paper detailing the Reweighing Algorithm aimed at converting input data into a relatively unbiased dataset through adding weights to "correct" the bias in a pre-processing step. The algorithm is shown below.

Input: $(D, S, Class)$
Output: Classifier learned on reweighed D
1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: $D_W := \{\}$
7: **for** X in D **do**
8: Add $(X, W(X(S), X(Class)))$ to D_W
9: **end for**
10: Train a classifier C on training set D_W, taking onto account the weights
11: **return** Classifier C

Figure 6: Reweighing Algorithm Pseudocode (Faisal, 2012)
S represents the relevant sensitive variable studied
X is an entry of the dataset D
Class is the target variable (either – or + here)

The essence of the Reweighing algorithm boils down to computing an appropriate weight for each entry in the input dataset in order to "equalize" whether or not each entry was favored at first. If there was no bias, it is expected that S and Class be independent variables, and so the expected joint probability distribution for the instance were the data point was in the b group with a + class would be:

$$P_{exp}(S = b \wedge Class = +) := \frac{|\{X \in D \mid X(S) = b\}|}{|D|} \times \frac{|\{X \in D \mid X(Class) = +\}|}{|D|}$$

However, the reality is that there is a statistical dependence between the two variables leading to a true joint probability distribution for this data point to be:

$$P_{obs}(S = b \wedge Class = +) := \frac{|\{X \in D \mid X(S) = b \wedge X(Class) = +\}|}{|D|}$$

The Reweighing algorithm uses these two probability figures to compute an appropriate weight for each entry based on the ratio of the two:

$$W(X) := \frac{P_{exp}(S = X(S) \wedge Class = X(Class))}{P_{obs}(S = X(S) \wedge Class = X(Class))}$$

This ratio generally encompasses every permutation of the protected variable S and Class and is mathematically equivalent to the calculation on Line 3 of Figure 6. It is important to note that the weights being added themselves follow a distribution on which elements exhibiting higher bias will have different weights than their counterparts depending on the magnitude of this difference between the expected and observed probabilities. Faisal and Calders demonstrated that quantifying discrimination by subtracting the product of the weight and frequency of a particular permutation part of the protected group (for example, minority) from the product of the weight and frequency of a permutation part of the unprotected group (for example, majority) yielded 0 with this approach, showing the weighed dataset did not exhibit discrimination by such a metric. Reweighing has proven to be a useful tool in its ability to maintain the structural integrity of the dataset and simply add an additional field to help address discrimination as opposed to reformulating the existing labels or throwing away data.

Training Optimization

In-Process, or Training Opinations are explicit adjustments to the way a machine-learning model learns meant to immediately address concerns about discrimination ‘on-the-go’ instead of performing specialized pre-processing of training data or post-processing of results. The have been many results in the literature which tend towards this goal, though some of the most prevalent simply include additional constraints which must be satisfied in addition to the traditional minimization of a loss function which takes place. For example, consider the following constraints from a paper written in demonstrating this approach (Zafar et al., 2017):

$$\begin{aligned} \text{minimize:} \quad & L(\theta) \\ \text{subject to:} \quad & P(\hat{y} = 1 | z = 0, y = -1) - P(\hat{y} = 1 | z = 1, y = -1) \leq \epsilon \\ & P(\hat{y} = 1 | z = 0, y = -1) - P(\hat{y} = 1 | z = 1, y = -1) \geq -\epsilon \end{aligned}$$

In this convex optimization problem, L(θ) is a loss function being minimized, while the additional constraints represent fairness guarantees. These guarantees are:

$$\text{Overall misclassification rate:}$$
$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1)$$
$$\text{false positive rate:}$$
$$P(\hat{y} \neq y | z = 0, y = -1) = P(\hat{y} \neq y | z = 1, y = -1)$$
$$\text{false negative rate:}$$
$$P(\hat{y} \neq y | z = 0, y = 1) = P(\hat{y} \neq y | z = 1, y = 1)$$

The constraints thus enforce that the false positive rate equality the model can allow during training. The two formulations are given of the same constraint (which disallow exceeding the scalar ε or –ε) since this is a standard form for the convex optimization process. In the case that equal false negativity rate made more sense the corresponding constraints could easily be adopted. Making the optimization problem above computationally feasible requires the usage of additional heuristics and re-formulations but the fundamental goal remains unchanged.

Post-Processing and Reject-Option Classification

Post-processing algorithms allow a machine-learning model to run as it normally would including any sort of bias it has while computing scores for each input between 0 and 1. Scores closer to the bounds of that range suggest a fairly strong conviction in the result, though scores within some pre-determined distance of 0.5 require further analysis. Post-processing algorithms generally work by establishing some threshold for each type of protected group and applying it to the score mentioned above as a means of assigning a final classification to that data point. In 2012, Faisal Kamiran, Asim Karim and Xiangliang Zhang published an algorithm known as Reject-Option based Classification, or ROC (Kamiran et al., 2012). The pseudocode for the algorithm is as follows:

Input: $\{\mathcal{F}_k\}_{k=1}^{K}$ ($K \geq 1$ probabilistic classifiers trained on $\mathcal{D}$), $\mathcal{X}$ (test set), $\mathcal{X}^d$ (deprived group), θ
Output: $\{C_i\}_{i=1}^{M}$ (labels for instances in $\mathcal{X}$)
** Critical region **
$\forall X_i \in \{Z | Z \in \mathcal{X}, \max[p(C^+|Z), 1 - p(C^+|Z)] < \theta\}$
  If $X \in \mathcal{X}^d$ **then** $C_i = C^+$
  If $X \notin \mathcal{X}^d$ **then** $C_i = C^-$
** Standard decision rule **
$\forall X_i \in \{Z | Z \in \mathcal{X}, \max[p(C^+|Z), 1 - p(C^+|Z)] \geq \theta\}$
  $C_i = \text{argmax}_{\{C^+, C^-\}}[p(C^+|X_i), p(C^-|X_i)]$

Figure 7: Reject Option based Classification Algorithm Pseudocode (Kamiran et al., 2012)

The critical insight is that the algorithm declares the data points as ambiguous if they satisfy the following inequality:

$$max(P(+|X), 1 - P(+|X)) \leq \theta$$

P(+|X) is the probability
Ø is a tolerance value between 0.5 and 1

In the case that the condition is satisfied, the instance is "rejected" and so if it is part of the deprived group X^d then it is given a positive label and if not, is given a negative label. It is important to note that the tolerance value need not be a fixed scalar and can instead take on a different value depending on which protected group the current data point is a part of – a method which maintains the necessary flexibility to correct for each group appropriately.

Significance

Examples

Instances of algorithmic fairness are becoming increasingly common, popular examples including Amazon's Rekognition A.I. recruiting tool, NorthPointe's Compass recidivism prediction system, and the Apple-Goldman Sachs credit card.

The Apple-Goldman Sachs incident deals with applications for credit card limit raises or personal loans. A process that was once based on a representative reviewing the applicant's financial history and personal credibility is now almost exclusively offloaded to a model converting these parameters into an acceptance/rejection and particular limit based on data the institution has access to. While the companies have publicly denied any sort of difference in treatment among male and female applicants, the Apple Card's credit limits were often significantly higher for men in spite of comparable income levels (Vigdor, 2019). One of the defenses of this situation is that the algorithm making the credit-limit decisions did not 'see' the gender data associated with each applicant. Despite Goldman Sach's claim of making use of factors like credit score and income level and not explicitly considering gender data in their decisions, the situation involves a flaw in approach. The explicit suppression of certain data fields and running a classifier produces less favorable results than making use of a mechanism for ensuring some degree of fairness would (Johndrow, 2019). Despite the gender data being removed from the dataset, implicit bias may still exist based on other fields present so removing a field itself isn't an adequate form of fairness enforcement.

In the infamous NorthPoint Compass algorithm, proprietary algorithms use information about a convicted offender's gender, crime and extensive personal information to generate a prediction of how likely the offender will recidivate (repeat his/her behavior) if they were to be released on parole – a data point presented to the presiding judge during the trial. Northpoint's Compass

algorithm was found to severely discriminate against African American defendants giving them higher "risk-of-repeat-offense" scores than their white counterparts for largely similar crimes and historical risk profiles (Angwin, 2019).

Amazon's Rekognition tool was another instance of algorithmic decision-making comes with a resume-filtering algorithm which would remove many of the time requirements and guesswork from the recruiting process by providing managers with only the select few resumes which pass the algorithm's screening methods. Amazon's resume-screening project yielded massively disproportionate suggestions favoring white male hires over women with similar or even more suitable qualifications (Dastin, 2018). This instance demonstrated the significant impact a training dataset with pre-existing bias can have on the classification process of a trained model on test data. Because the technical employees at Amazon and Amazon Web Services have historically been white men with college degrees in engineering fields, and this was the population which the algorithm used as a guide for what to look for in promising candidates who would fit well into the company, it unsurprisingly showed heavy discrimination against women in particular.

Costs of Ignoring Fairness

The costs of ignoring fairness in algorithmic decisions is extremely impactful to the livelihood of the afflicted when considering that more and more decisions are being offloaded to machines. While it may sound like a semi-theoretical and complex notion (which to some extent – it is), algorithmic fairness is an area of research which is rapidly increasing in relevance to the everyday livelihood of ordinary people. It is worthwhile to highlight the particular ways in which unfair decisions can impact the lives of the masses. The examples in the previous section range from depriving a qualified applicant a chance to get a job they are interested in to significantly affecting an individual's freedom. These are by no means the only examples where bias has been uncovered in the output of widely used classifiers. An article published in *Science* found that an algorithm used in medical systems disproportionately recommended white patients with similar health conditions to patients of color for getting treatment (Ledford, 2019). The algorithm categorized patients into higher or lower tiers of risk based on their expenditure on healthcare related costs from the most recent year on record in the patient's medical profile – which while well-intentioned – turned out to highlight the discrepancy between the average medical expenses for patients in each of these subgroups despite having the same chronic health problems. The unwillingness to issue referral for advanced care by the algorithm imposes very real risks to the livelihood of the protected sub-population in this instance.

It is clear that corporations, governments and universities alike are investing capital and working hours into developing machine-learning based algorithms to improve facets of their operations. Researchers and independent interested parties are increasingly providing evidence of concrete situations in which the decisions of classifiers have caused profound harm to particular subsets of the population. This isn't to say that algorithms for such tasks do not have a future – though it is clear they need to be fundamentally re-evaluated through the lens of fairness and their assumptions and goals made clear from the onset instead of acting as a black box for decision-making.

Looking Forward

Developing Approaches

Research is underway on a variety of novel approaches towards mitigating the unfairness risk which underlies current models – particularly in the "types" of learning implemented.

Causal-Related Learning

One of the main challenges with enforcing fairness deals with the ambiguity surrounding what actually causes the discrimination. 'Causal models' are those which demonstrate an ability to study how data are generated and the effects of interventions on outputs (Kusner, 2020). Current machine learning models tend to highlight correlations between some set of variables in their input data yet struggle with explaining the underlying reason behind these connections – the well-known correlation versus causation dilemma. The work regarding causal models has been broken into three primary tests which hint at causal relationships among training data and a classifier's outputs.

The first deals with counterfactuals – which studies whether changing a particular data point from the past would change the output of the model (Kusner, 2020). Researchers can tweak individual parameters and determine which ones have a measurable effect on the model's output and conclude whether these results are in line with what they expect. For example, if changing the gender of an applicant consistently resulted in different outputs in a context where this factor should not matter, counterfactual fairness would not be satisfied. Another approach deals with the sensitivity of a model's results to variables which cannot be adequately measured or controlled for. Though not every variable to a model can be studied thoroughly, a sensitivity analysis can inform an approximate proportionality between causes and effects. The third test deals with the impacts of so-called 'interventions' to study potential ripple effects in the future that are tied to the alteration of a variable being trained on – in particular impacts which arise a couple of steps away from a direct cause-effect relationship (Kusner, 2020).

Fair Adversarial Learning

Adversarial-trained neural networks are generally composed of a source of data called a generator and an adversarial model (Wadsworth, 2018). The generator fabricates data samples (independent of any real data) and the discriminator determines whether the samples are "fake" or not – feedback which the generator uses in a loop to improve its outputs (Pessach, 2021). The idea is to maximize the predictor's outcome accuracy while also minimizing the discriminator's ability to predict the protected attribute. A similar version may also include a second discriminator – the first decides whether generated samples are real like above and the second decides whether the sample is a part of the "privileged group" or not. This modification introduces the second layer to the feedback loop and has shown promising improvement in reducing biases (Xu, 2018).

Fair Sequential Learning

Sequential learning is based on the necessity for dealing with streams of real-time data as opposed to having an entire training dataset in advance. These systems pose a challenge since

decisions need to constantly be made after processing each new data point that will preserve a notion of fairness, as the state of the model will influence the way new streaming data is handled in the future. An attempt at tackling this scenario was made in 2018 by Hoda Heidari and Andreas Krause with the introduction of time-dependent individual fairness metrics which required a model's decisions to be consistent over time (Heidari, 2018). The solution was based on the principle that similar data points which were introduced to the model at a similar time should have similar outcomes given to them. The enforcement of this idea was carried out with a post-processing mechanism for maintaining the consistency among predictions and the effectiveness of the method was validated with trials on sample data sets (Heidari, 2018).

Milestones to Reach

Although perfect fairness is generally regarded to be infeasible, it is important to highlight fairness milestones which future research will be able to consistently implement. One of the major obstacles is the lack of representative datasets for most topics of study. Often the method of data collection is subject to some sort of bias or the data simply do not include adequate information about all of the possible protected groups – for example, a lack of African American women applicants to a particular job posting – meaning the model will not have a sense of how to treat such an application. Another, more logistical issue, involves creating some sort of standard for fairness enforcement. With the many notions of fairness used in different contexts, it is difficult to say with any real certainty whether or not a particular algorithm has been shown to be fair. It may satisfy one definition while completely failing a preserving a different "type" of fairness. While this is an imposing problem, some effort towards standardizing fairness guidelines would be a significant improvement over the highly segmented approach currently taken.

Other Dimensions of Interest

Algorithmic Ethics

Algorithmic ethics is an emerging field of study which attempts to address what constitutes "acceptable" data-usage from a more philosophical standpoint – a question growing in importance with the rise of big-data processing capabilities. There is much discourse in academic publications about what sort of ethical standards algorithms must meet, and whether they should be expected to be the same or stricter as those for human decision-makers. An influential paper in the Big Data and Society journal outlined six key concerns that future work in algorithmic fairness needs to be able to address in order to establish a strong ethical basis (Mittelstadt et al., 2016). They are depicted as follows:

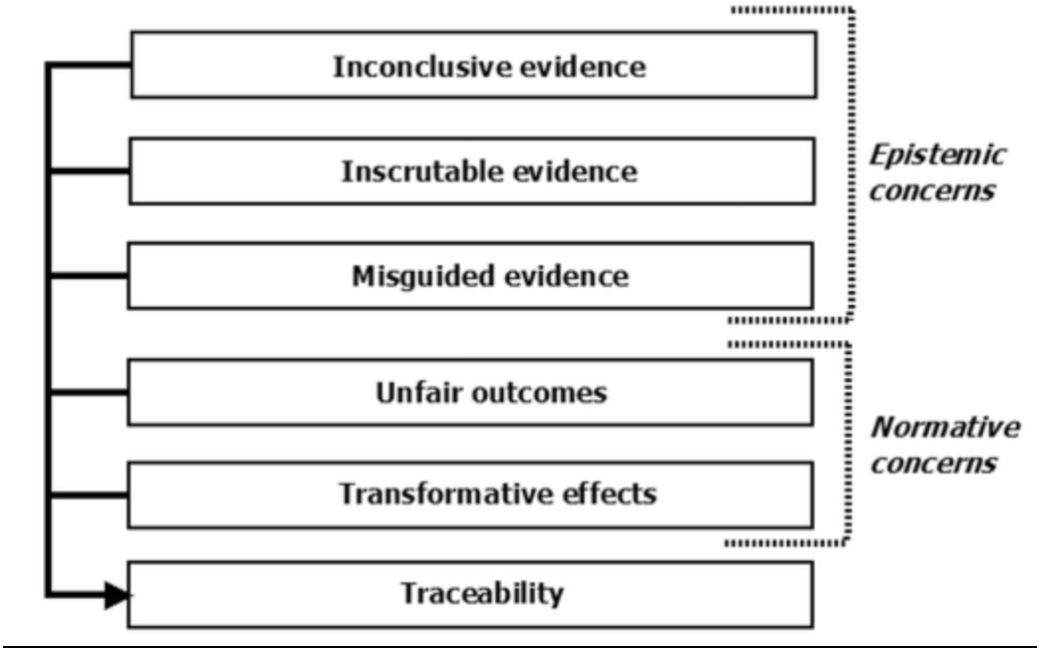

Figure 8: Six Ethical Concerns for Algorithms (Mittelstadt, 2016)

The epistemic concerns group here refer to the underlying data on which classifiers are trained. *Inconclusive evidence* deals with the notion that every machine-learning model, regardless of the statistical thresholds of error and other error-reduction mechanisms, will always be based on data that is not fully representative and produce conclusions which are uncertain.
Inscrutable evidence deals with the inevitable difficulty for validation/reproduction of a model's decisions since it is very difficult to determine what the effect of the any particular data point in the training set was on the model's label in a test set.
Misguided evidence refers to a fundamental limitation of every model in its inability to ever produce results which are "better" than the type of data they are based on. Low-quality (many factors can indicate this such as size, representability, etc.) training data will place a severe limitation on a model's ability to adhere to fairness definitions and by extension ethical expectations.
The *unfair outcomes* category that is simple to understand but difficult to address appropriately. It simply refers to observed discrimination in the results of an algorithm, even if it only occurs in one particular protected group, despite it satisfying various measures of fairness elsewhere.
Transformative effects is perhaps one of the most difficult of the principles to understand, and include activities such as profiling which may not have clearly demonstrable negative effects like discrimination against particular groups but may still be ethically questionable in their practices.
The final principle is *Traceability*, which refers to the ability of harm or bias caused by an algorithm to be traced back to its cause or responsible party.
These six concerns provide a framework for the study of algorithmic ethics, though it is clear that answering whether or not an algorithm can be thought of as ethical is a function of a multitude of factors which may be just as difficult to define as to implement in model.

Differential Privacy

Differential privacy refers to the security and anonymity of personally identifiable user data analyzed or used in machine learning research. The field has increased in popularity due to the nature of how classifiers work – if the underlying data belongs to individuals, there is growing concern that models can threaten privacy rights. A widely held formulation of privacy is that of ε -differential privacy, which is defined as follows:

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(D_2) \in S]$$

A is the decision-making algorithm
D_1 and D_2 are datasets which differ by a single element (the element is not present in D_2)
S is a subset of all possible values A may produce
ε is a positive real value

This formulation guarantees that in a differentially private dataset the personal information for any particular individual in the dataset will not be distinguishable from that of others – and to put it another way, any analysis done on the dataset will have effectively the same result whether or not that particular individual's data was using the in analysis (Dwork, et al., 2014).

The notion above refers to datasets in which D_1 and D_2 differ by only a single datapoint. However, it is easily extendable to larger groups of datapoints with the following modification:

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon c) \cdot \Pr[\mathcal{A}(D_2) \in S]$$

c is a scalar representing the number of differences among D_1 and D_2

Satisfying this definition, every group of c items is now (ε/c-differentially private), exemplifying the tradeoff between larger-scale privacy and the strength of the guarantee. There are many reasons for which differential privacy is useful in an applied context. The randomness embedded within the data-collection process that ensures a dataset is differentially private also ensures that no post-processing work can be done to reduce the "amount" of privacy the set offers. Another extremely important benefit differential privacy offers is the ability to quantify how much privacy is lost due to the various processing steps performed. This becomes very useful when considering whether multiple computations being done on the data in sequence (modification, removal of data, etc.) are directly tied to an increased risk of privacy – i.e. the question of whether a composition of actions on differentially-private datasets adequately maintain this property.
These notions make up a small part of the research into differential privacy and the variety of attacks and counter-measures researchers are working towards addressing. The field is important to develop in tandem with algorithmic fairness (and the depth of available research results points to it being well ahead) as it similarly addresses an aspect of machine learning other than the over-arching accuracy question – yet no less important for society as a whole.

Conclusion

Algorithmic fairness is a complicated yet critical field in the modern age of technology-driven decision making. A steadily increasing dependence on machine learning models has moved them away from theoretical tools and toward becoming a cornerstone of financial, legal, medical, and many other systems which impact society as a whole. This paper analyzed some of the ways in which these models must be modified to truly become suited for their emerging role in these domains. Fairness in the context of algorithms, much like in the many non-algorithmic scenarios it already appears in, is a difficult notion to define and enforce well. Progress in the field has, however, steadily brought ways to address to the issues of inequality, discrimination, bias, etc. to light, and the research continuing to be done is instrumental to avoiding instances of systematic harm inflicted on particular groups which have arisen in the past.

References

Angwin, Julia, et al. “Machine Bias.” ProPublica, 9 Mar. 2019, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas S., Selbst, A. (2016). Big data’s disparate impact. Calif. L. Rev. 104 (2016), 671.

Chouldechova, A., & Roth, A. (2018, October 23). The Frontiers of Fairness in Machine Learning. Retrieved April 14, 2021, from https://arxiv.org/pdf/1810.08810.pdf

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference. ACM, 214–226.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved February 01, 2021, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dieterich, W., Mendoza, C., & Brennan, T. (2016, July 8). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Retrieved April 13, 2021, from https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy (Vol. 9). Boston: Now Publ. doi:10.1561/0400000042

Faisal K., Calders T. 2012. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1 (2012), 1–33.

Heidari, H., Krause, A. 2018. Preventing Disparate Treatment in Sequential Decision Making. In IJCAI. 2248–2254.

Johndrow, J. (2019, October 7). Removing bias from predictive modeling. Retrieved April 22, 2021, from https://knowledge.wharton.upenn.edu/article/removing-bias-from-predictive-modeling/

Kamiran, F., Karim, A., & Zhang, X. (2012, April 1). Decision Theory for Discrimination-aware Classification. Retrieved April 26, 2021, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf

Kim, M., Reingold, O., & Rothblum, G. (2018, November 28). Fairness Through Computationally-Bounded Awareness. Retrieved April 26, 2021, from https://arxiv.org/pdf/1803.03239.pdf

Kusner, M. J., & Loftus, J. R. (2020, February 6). Retrieved April 24, 2021, from https://media.nature.com/original/magazine-assets/d41586-020-00274-3/d41586-020-00274-3.pdf

Landeau, A. (November 11, 2020). Measuring fairness in machine learning models. Retrieved April 12, 2021, from https://blog.dataiku.com/measuring-fairness-in-machine-learning-models

Ledford, H. (2019, October 24). Millions of black people affected by racial bias in Health-care algorithms. Retrieved April 23, 2021, from https://www.nature.com/articles/d41586-019-03228-6

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 259–268.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016, July 21). The ethics of algorithms: Mapping the debate. Retrieved April 27, 2021, from https://journals.sagepub.com/doi/pdf/10.1177/2053951716679679

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems. 3315–3323.

Nguyen, A. (2019, July 06). Understanding differential privacy. Retrieved May 01, 2021, from https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a

Pessach, D., & Shmueli, E. (2020, January 21). Algorithmic Fairness. Retrieved April 22, 2021, from https://arxiv.org/pdf/2001.09784.pdf

Saravanakumar, K. K. (2021, January 29). THE IMPOSSIBILITY THEOREM OF MACHINE FAIRNESS A CAUSAL PERSPECTIVE. Retrieved April 22, 2021, from https://arxiv.org/pdf/2007.06024.pdf

Vigdor, N. (2019, November 10). Apple card investigated After gender discrimination complaints. Retrieved April 22, 2021, from https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

Wadsworth, C., Vera, F., & Piech, C. (2018, June 30). Achieving Fairness through Adversarial Learning: An Application to Recidivism Prediction. Retrieved April 25, 2021, from https://arxiv.org/pdf/1807.00199.pdf

Xu, D., Shuhan, Y., Lu, Z., and Xintao, W. 2018. Fairgan: Fairness-aware generative adversarial networks. In 2018 IEEE International Conference on Big Data (Big Data). IEEE, 570–575.

Zafar, M., Valera, I., Rodriguez, M., & Gummadi, K. (2017, March 08). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. Retrieved May 02, 2021, from https://arxiv.org/abs/1610.08452

Zhong, Z. (2020, June 19). A tutorial on fairness in machine learning. Retrieved May 01, 2021, from https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb