Fairness in Machine Learning

A Graphical Explanation

Frontiers of Computational Journalism
Pankhuri Kumar (pk2569)

Columbia University
Columbia Journalism School

Introduction

Technology has become ubiquitous in our everyday world. Computer algorithms now decide multiple decisions that are taken for us - from school admissions and mortgage rates to crime risk assessment to self-driving cars.

When algorithms can have such far-reaching consequences for us, it is important for us to examine the motivations behind the decision making process that shapes these algorithms.

As explained in [3], machine learning has its advantages. Research has found that humans make decisions without realizing the underlying factors that subconsciously play a role in taking that decision. Research, however, has shown that when compared to human intuition, data driven decisions are more accurate.

Machine learning enables us to make decisions with relevant factors that humans might overlook in favor of intuition or expertise. Rather than starting with a pre-conceived intuition or notion of how factors are related, machine learning gives us the opportunity to delay the decision of 'relevant' factors by looking at the data, and inferring factors that have a statistical relationship with the result. In fact, machine learning is able to help us uncover factors that humans cannot specify even when they can make a decision effortlessly.

When combined with historical data, machine learning becomes powerful, but it has its own caveats. Machine learning does not just commit data to memory, rather it tries to generalize patterns from the data, creating rules that fit past cases, but can be extended to future cases. This means that data on decisions that are historically made by humans, will often have prejudices against certain social groups, or demographics. Machine learning could easily learn, and replicate, these biases.

Machine learning can potentially learn biases that exist in historical data and create a rule from it, instead of correcting for human biases that have now becoming systemic. Thus, machine learning requires *good* examples, large enough in number to find patterns but also diverse enough to show different ways a factor can appear and well-annotated ones to give machine learning reliable information [3]. Data-based decision making is only as good as the data, and does not necessarily mean all decisions will be accurate, or fair.

Over the past few years, the notion of fair machine learning has come up, a concept which develops an "understanding of when disparities are harmful, unjustified, or other unacceptable, and to develop interventions to mitigate such disparities" [3]. Though this goal has many challenges, data-driven decision making is more transparent than human decision-making, which gives us greater control over our objectives and the tradeoffs we make.

More importantly, algorithms make it necessary for us to be explicit and honest about what needs to be achieved, and using which factors. This often leads to a discussion about the fairness, relevance and impact of policies and decisions.

Methodology

This project started out from a conversation with Prof. Stray. I wanted to do a project that used explanatory joournalism, and dealt with a topic that would use some of my technical knowledge.

This was further bolstered by one of the assignments for the class, which dealt with this exact topic, by asking us to recreate and re-evaluate ProPublica's methodology for evaluating the COMPAS algorithm. A major portion of the code for my project comes from this assignment.

My research started with re-reading ProPublica's article and methodology on *Machine Bias* [4]. This was followed by reading a large number of research papers – from which it quickly emerged that though the conversation on algorithms and fairness is popular, research on the topic has been confined to the academic industry.

Very notably, three professors dominate this research space, and most reputable papers name one of them as their authors ([2], [3], [7], [9], among others). They often collaborate with one another, and the online textbook, Fairness and Machine Learning [3], which I frequently quote from, is their product. Solon Barocas at Cornell, Moritz Hardt at Berkeley, and Arvind Narayan at Princeton have all conducted fairness and ethics in machine learning courses in their universities, and the book "emerged from the notes we created for these three courses, and is the result of an ongoing dialog between us."

Along with their work, my project also relies heavily on *Fairness in Criminal Justice Risk Assessments: The State of the Art* by Berk et al. [4], which meticulously defines the different types of fairness and outlines their relationship to the mathematical concept of confusion matrices. I have used the same model, and definitions, in my project.

My goal was to take the research that is confined to academia, and create a project that is accessible to anyone with minimal background of the research. I do, however, assume basic understanding of mathematics and the topic of fairness and machine learning. I wanted this project to be able to demystify the math that goes behind evaluating the fairness of algorithms.

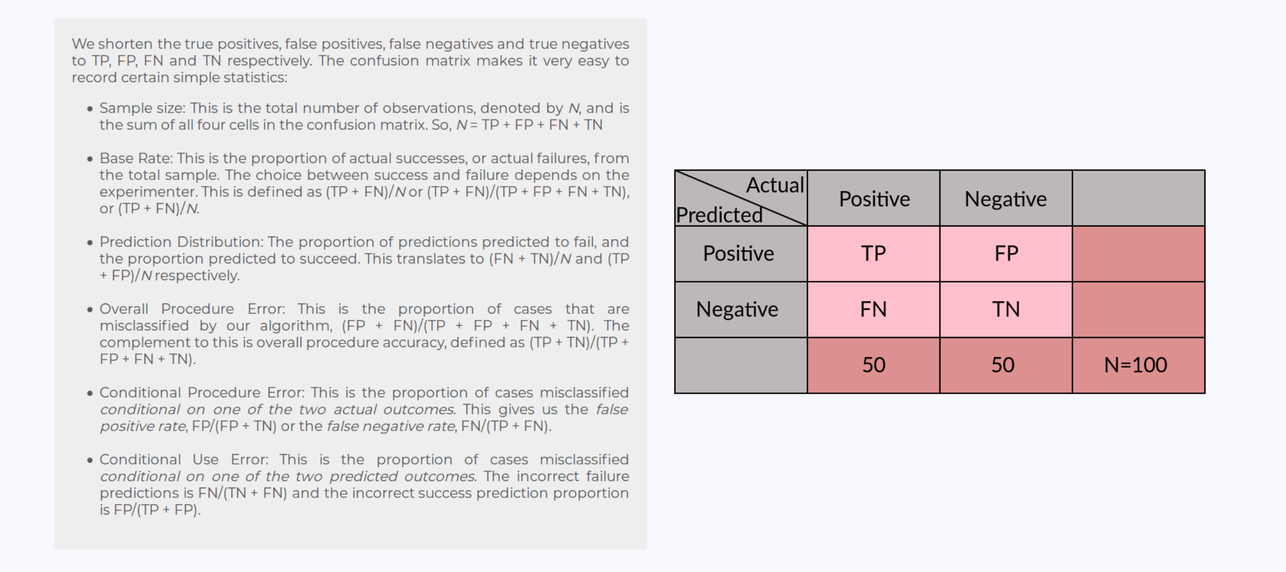
Research over the years has made it clear that fairness does not have a single definition, and most of the conversation on discrimination and impact is highly contextual. Fairness is not a general concept, and varies according to the domain and history of the issue under consideration. As [4] explains, it is now painfully obvious that achieving all kinds of fairness simultaneously is not possible. Some definitions of fairness are naturally at odds with one-another, and the trade-off between them is an important conversation with no definite answer.

My project aims to explain these complex relationships, bringing out their inherent clash through mathematical formulae, while explaining the real-world implications of choosing one fairness over the another.

Project Outline

My project is designed as a infinite-scroll (scrollytelling), which begins with an introduction to the topic of fairness and machine learning (not dissimilar from the introduction of this report), and goes on to explain its implications.

Drawing from [4], I next introduce the concept of a confusion matrix, and the information one can glean from it. This is followed by defining the different kinds of fairness, and how approximations of each can be calculated from the confusion matrix.



We shorten the true positives, false positives, false negatives and true negatives to TP, FP, FN and TN respectively. The confusion matrix makes it very easy to record certain simple statistics:

- Sample size: This is the total number of observations, denoted by *N*, and is the sum of all four cells in the confusion matrix. So, *N* = TP + FP + FN + TN
- Base Rate: This is the proportion of actual successes, or actual failures, from the total sample. The choice between success and failure depends on the experimenter. This is defined as (TP + FN)/*N* or (TP + FN)/(TP + FP + FN + TN), or (TP + FN)/*N*.
- Prediction Distribution: The proportion of predictions predicted to fail, and the proportion predicted to succeed. This translates to (FN + TN)/*N* and (TP + FP)/*N* respectively.
- Overall Procedure Error: This is the proportion of cases that are misclassified by our algorithm, (FP + FN)/(TP + FP + FN + TN). The complement to this is overall procedure accuracy, defined as (TP + TN)/(TP + FP + FN + TN).
- Conditional Procedure Error: This is the proportion of cases misclassified *conditional on one of the two actual outcomes*. This gives us the *false positive rate*, FP/(FP + TN) or the *false negative rate*, FN/(TP + FN).
- Conditional Use Error: This is the proportion of cases misclassified *conditional on one of the two predicted outcomes*. The incorrect failure predictions is FN/(TN + FN) and the incorrect success prediction proportion is FP/(TP + FP).

Actual Predicted	Positive	Negative	
Positive	TP	FP	
Negative	FN	TN	
	50	50	N=100

Fig 1. Screenshot of scrolly-telling explanation of confusion matrices

At this point, I introduce an example to display the trade-off between fairness: predicting whether customers at a cafe will take coffee with sugar, or without. The goal is to familiarize the audience with the confusion matrix, while having a conversation about fairness and how different fairness can imply different things.

I finally introduce ProPublica's article, and their dataset, with a quick brief on the definitions of fairness they thought were more important that the ones Northpointe's COMPAS algorithm optimized for.

This is followed by two different interactives, both of which use ProPublica's [8] dataset and methodology to create the datapoints. The code for this interactive is borrowed (and modified slightly) from Assignment 4 conducted as part of this class.

My algorithm does not filter out all other races except Caucasians and African Americans – this was done to get accurate numbers of men and women, which would be filtered out if the data was filtered. The confusion matrices are this made over 5278 datapoints for race, and 6172 datapoints over gender.

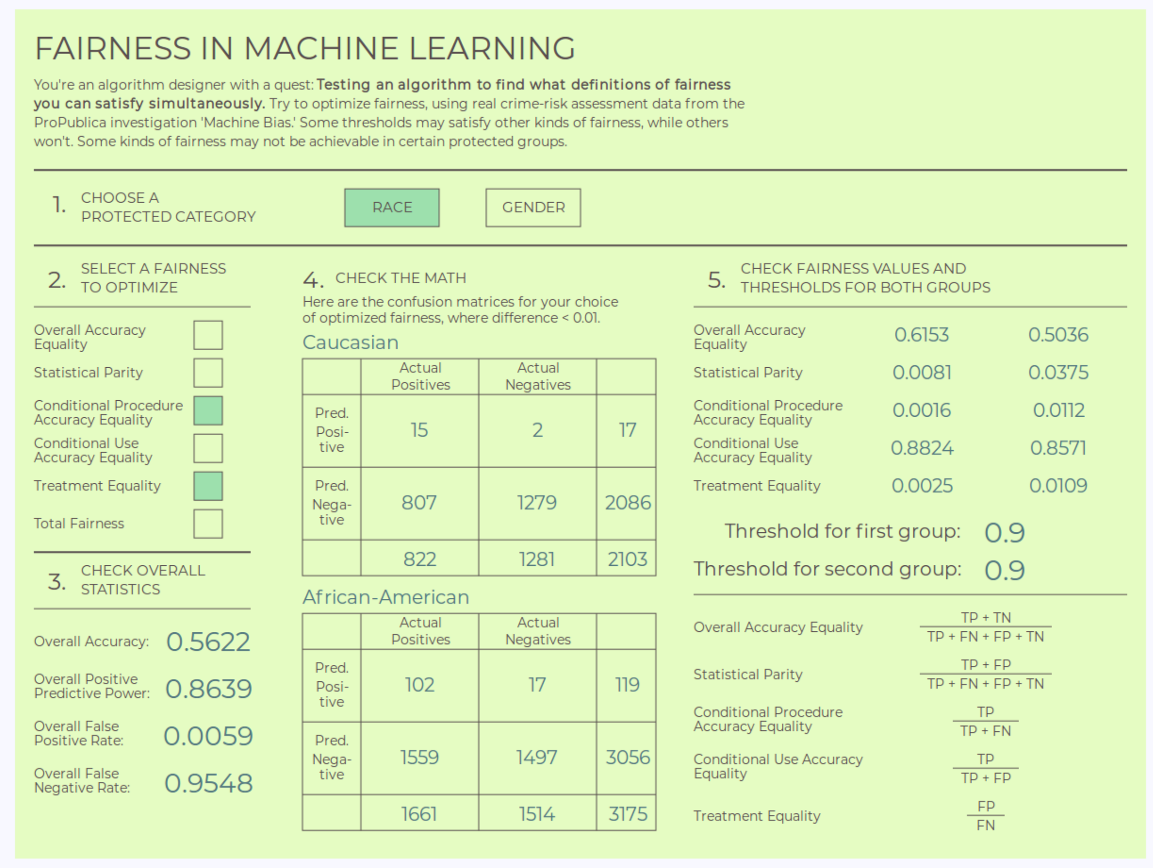
Both interactives allow the user to choose which protected group (race or gender) they want to explore and the fairness they want to optimize for, and show how the threshold for recidivism will wary depending on their choices.

The first visualization focuses on the mathematics behind the calculation, showing the confusion matrices, while providing the formulae for each fairness, and showing the calculated values for other definitions of fairness at that threshold.

The second visualization tries to visualize the implications of the threshold. It does not go into the math of the calculations, merely showing a confusion matrix with the right values. Instead, it displays the distribution of risk scores for the chosen protected group, and a threshold line, to the left of which everyone is considered 'medium' or 'high' risk.

The contrast of thresholds across the two groups (in terms of the number of people considered to be a risk) makes it easier to see the real-world implications of algorithmic (or data-based) decisions.

The second visualization, thus, superimposes the decile scores calculated by the COMPAS algorithm with a line for the threshold which signifies the number of people who might have been denied bail under the algorithm which optimized for a certain fairness.



FAIRNESS IN MACHINE LEARNING

You're an algorithm designer with a quest: **Testing an algorithm to find what definitions of fairness you can satisfy simultaneously.** Try to optimize fairness, using real crime-risk assessment data from the ProPublica investigation 'Machine Bias.' Some thresholds may satisfy other kinds of fairness, while others won't. Some kinds of fairness may not be achievable in certain protected groups.

1. CHOOSE A PROTECTED CATEGORY

RACE | GENDER

2. SELECT A FAIRNESS TO OPTIMIZE

Overall Accuracy Equality	□
Statistical Parity	□
Conditional Procedure Accuracy Equality	🟩
Conditional Use Accuracy Equality	□
Treatment Equality	🟩
Total Fairness	□

3. CHECK OVERALL STATISTICS

Overall Accuracy:	0.5622
Overall Positive Predictive Power:	0.8639
Overall False Positive Rate:	0.0059
Overall False Negative Rate:	0.9548

4. CHECK THE MATH

Here are the confusion matrices for your choice of optimized fairness, where difference < 0.01.

Caucasian

	Actual Positives	Actual Negatives	
Pred. Positive	15	2	17
Pred. Negative	807	1279	2086
	822	1281	2103

African-American

	Actual Positives	Actual Negatives	
Pred. Positive	102	17	119
Pred. Negative	1559	1497	3056
	1661	1514	3175

5. CHECK FAIRNESS VALUES AND THRESHOLDS FOR BOTH GROUPS

Overall Accuracy Equality	0.6153	0.5036
Statistical Parity	0.0081	0.0375
Conditional Procedure Accuracy Equality	0.0016	0.0112
Conditional Use Accuracy Equality	0.8824	0.8571
Treatment Equality	0.0025	0.0109

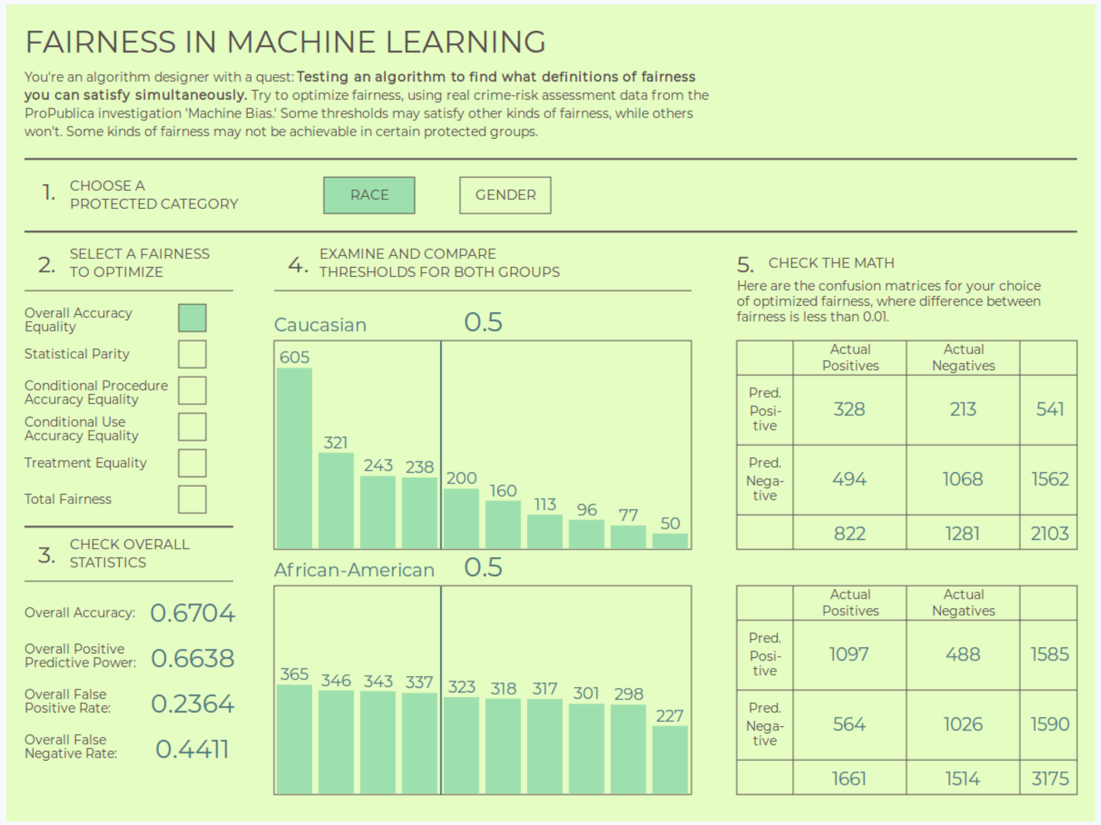
Threshold for first group: 0.9

Threshold for second group: 0.9

Overall Accuracy Equality	$\frac{TP + TN}{TP + FN + FP + TN}$
Statistical Parity	$\frac{TP + FP}{TP + FN + FP + TN}$
Conditional Procedure Accuracy Equality	$\frac{TP}{TP + FN}$
Conditional Use Accuracy Equality	$\frac{TP}{TP + FP}$
Treatment Equality	$\frac{FP}{FN}$

Fig 2. Screenshot of Interactive 1

The design for both the visualization borrows heavily from the design and layout of the interactive in *Science Isn't Broken* [1] by FiveThirtyEight, which talks about (in)significance of the p-value in research and academia.



FAIRNESS IN MACHINE LEARNING

You're an algorithm designer with a quest: **Testing an algorithm to find what definitions of fairness you can satisfy simultaneously.** Try to optimize fairness, using real crime-risk assessment data from the ProPublica investigation 'Machine Bias.' Some thresholds may satisfy other kinds of fairness, while others won't. Some kinds of fairness may not be achievable in certain protected groups.

1. CHOOSE A PROTECTED CATEGORY — RACE | GENDER

2. SELECT A FAIRNESS TO OPTIMIZE

- Overall Accuracy Equality ■
- Statistical Parity □
- Conditional Procedure Accuracy Equality □
- Conditional Use Accuracy Equality □
- Treatment Equality □
- Total Fairness □

3. CHECK OVERALL STATISTICS

Overall Accuracy:	0.6704
Overall Positive Predictive Power:	0.6638
Overall False Positive Rate:	0.2364
Overall False Negative Rate:	0.4411

4. EXAMINE AND COMPARE THRESHOLDS FOR BOTH GROUPS

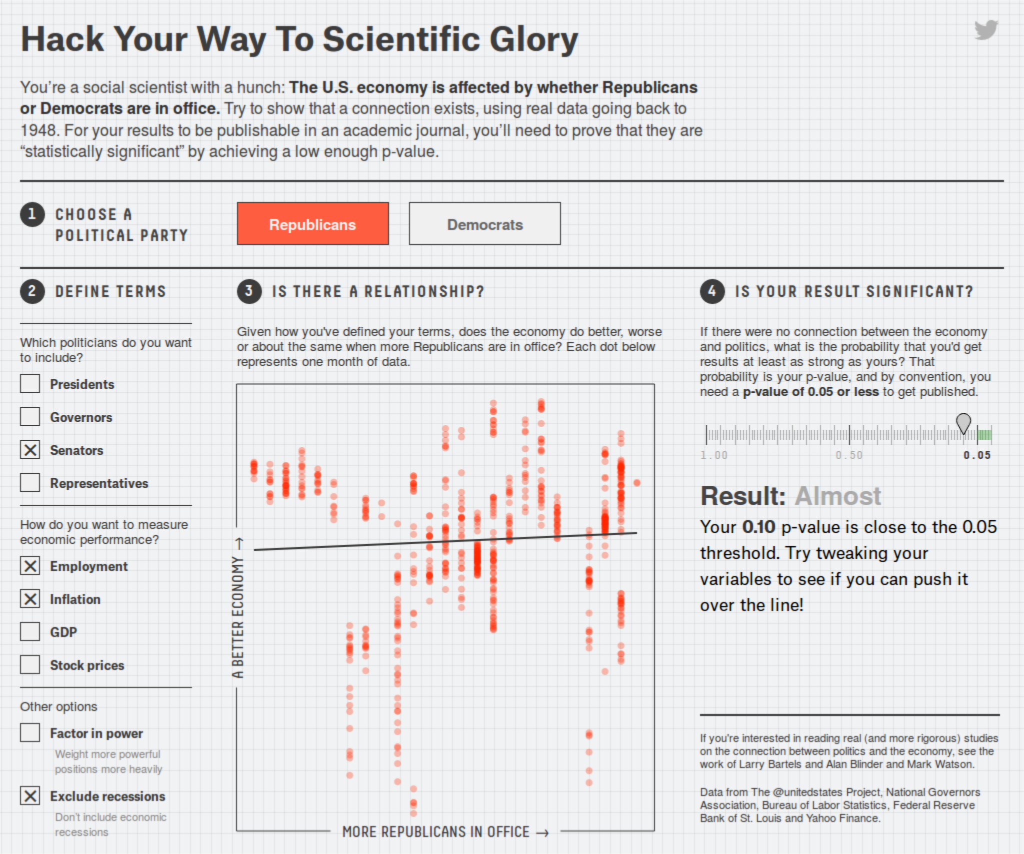
5. CHECK THE MATH

Here are the confusion matrices for your choice of optimized fairness, where difference between fairness is less than 0.01.

	Actual Positives	Actual Negatives	
Pred. Positive	328	213	541
Pred. Negative	494	1068	1562
	822	1281	2103

	Actual Positives	Actual Negatives	
Pred. Positive	1097	488	1585
Pred. Negative	564	1026	1590
	1661	1514	3175

Fig 3. Screenshot of Interactive 2



Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY — Republicans | Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- □ Presidents
- □ Governors
- ☒ Senators
- □ Representatives

How do you want to measure economic performance?

- ☒ Employment
- ☒ Inflation
- □ GDP
- □ Stock prices

Other options

- □ Factor in power — Weight more powerful positions more heavily
- ☒ Exclude recessions — Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in office? Each dot below represents one month of data.

4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Almost

Your **0.10** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Fig 4. Screenshot of *Science Isn't Broken*'s interactive visualization

The project can be found here: https://pankhurikumar23.github.io/frontiers/index

References

[1] Aschwanden, Christie, and Ritchie King. “Science Isn't Broken.” *FiveThirtyEight*, FiveThirtyEight, 19 Aug. 2015, fivethirtyeight.com/features/science-isnt-broken/.

[2] Barocas, Solon, and Moritz Hardt. “Fairness in Machine Learning NIPS 2017 Tutorial.” *nips17tutorial*, 2017, fairml.how/.

[3] Barocas, Solon, et al. *Fairness and Machine Learning*. Fairmlbook.org, 2018, fairmlbook.org/index.html.

[4] Berk, Richard, et al. “Fairness in Criminal Justice Risk Assessments.” *Sociological Methods & Research*, 2018, doi:10.1177/0049124118782533.

[5] Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R.. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR* 81:134-148

[6] Corbett-Davies, Sam, and Sharad Goel. “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.” *CoRR*, abs/1808.00023, 2018, doi:http://arxiv.org/abs/1808.00023.

[7] Hardt, Moritz. “Fairness, Accountability, and Transparency in Machine Learning.” *FAT ML 2016*, www.fatml.org/.

[8] Larson, Jeff, et al. “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*, ProPublica, 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[9] Narayanan, Arvind. “Tutorial: 21 Fairness Definitions and Their Politics.” *YouTube*, YouTube, 1 Mar. 2018, www.youtube.com/watch?v=jIXIuYdnyyk.

[10] Zhong, Ziyuan. “A Tutorial on Fairness in Machine Learning – Towards Data Science.” *Towards Data Science*, Medium, 22 Oct. 2018, towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.