Addressing Algorithmic Bias

MUNIBA TALHA

Can you read this ???

Ocne uopn a tmie, terhe was a redaer. Tehy weer socllrnig thorguh the itnrenet and cmae aorcss a parrgaaph full of julbemd wrdos. Cruosuily, they funod tehy culod uesdnatnrd the mgessae, eevn tuhgoh ervteyihng was mexid up. The redaer pusead, tnhiinkg aoubt tihs altbiiy … hwo deos our biran do tihs? Tehy thhgout lnog and hrad. In the end, tehy dcdeied it was tmie for lncuh and mdae tehlemesvs a deoluciis beagl!

Heuristis & Biases

Heuristics are mental shortcuts that help us function as efficiently as possible.

Familiarity Heuristic: the pattern of favouring that which is familiar over something novel.

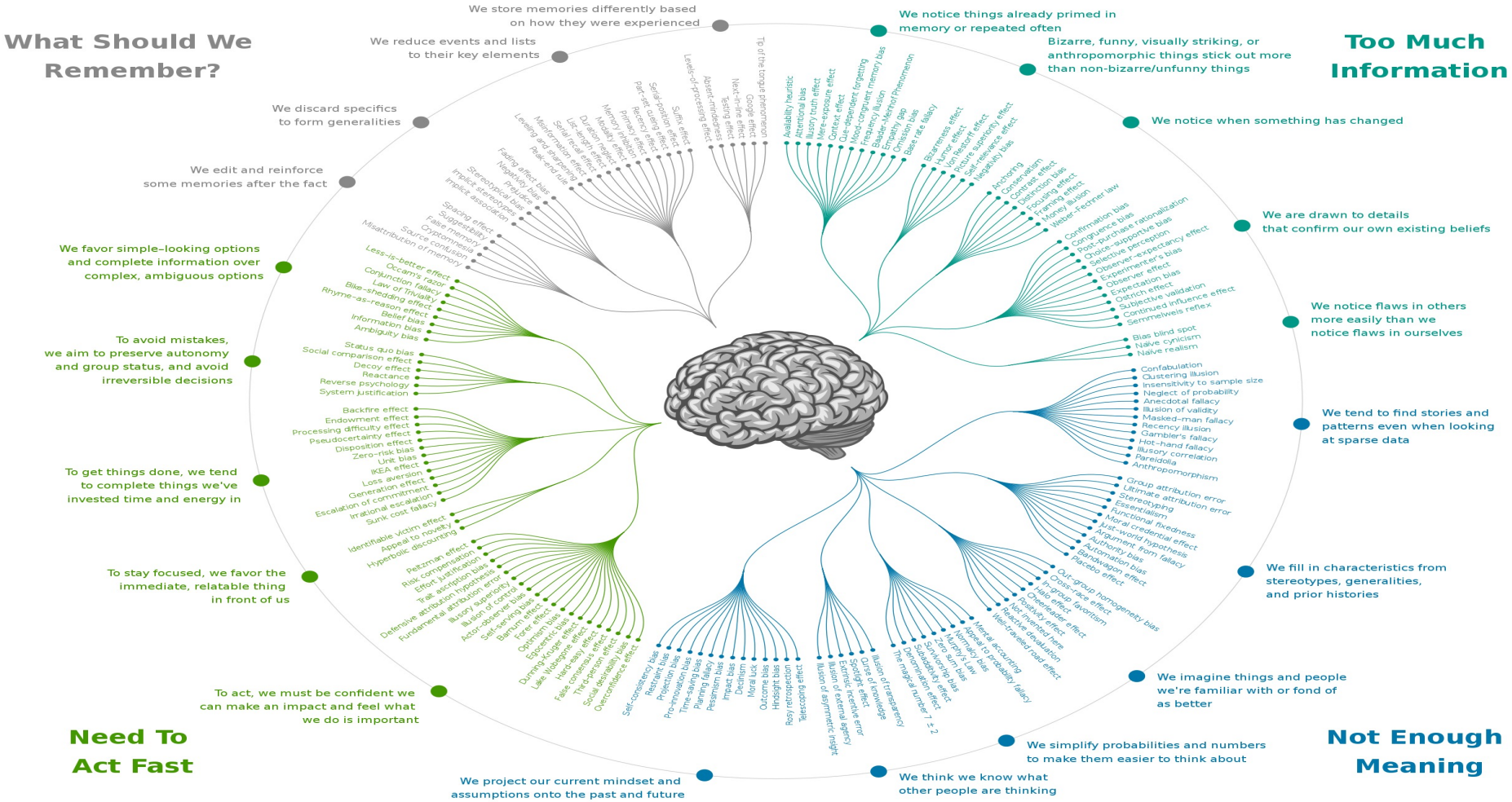
Representativeness Heuristic: the shortcut of grouping objects by similarity and organizing them based around the category prototype (e.g. like goes with like)

Bias is defined as an inclination towards or against an idea, a thing, a person or a group, usually in a way that is closed-minded, prejudicial, or unfair.

Status Quo: A preference for the current state of affairs or our comfort zone. For example: **86% of Fortune 500 CEOs are white and male, the status quo bias is what will compel board leaders and directors to continue to hire white men for leadership roles**.

Stereotyping: the unconscious attribution of particular qualities to a member of a certain social group. Example is Imagine a Nurse: Is it woman ?

THE COGNITIVE BIAS CODEX



Algorithmic Bias

Algorithmic bias can be described as a systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others. Also, occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the data or machine learning process.

Algorithmic Bias

- In 2015 Google's image-recognition system **labeled African Americans as "gorillas."**
- In 2018, Amazon's Rekognition system drew criticism for **matching 28 members of Congress to criminal mugshots**

Bias in Facial Recognition Technology

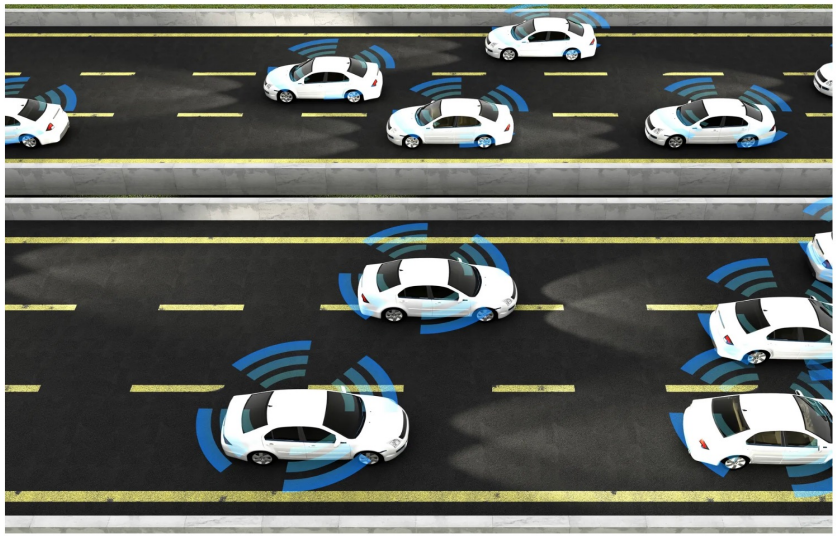
- MIT researcher Joy Buolamwini found that the algorithms powering three commercially available facial recognition software systems were failing to recognize darker-skinned complexions [7]

A new study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians

The findings speak to a bigger problem in the development of automated systems: algorithmic bias.

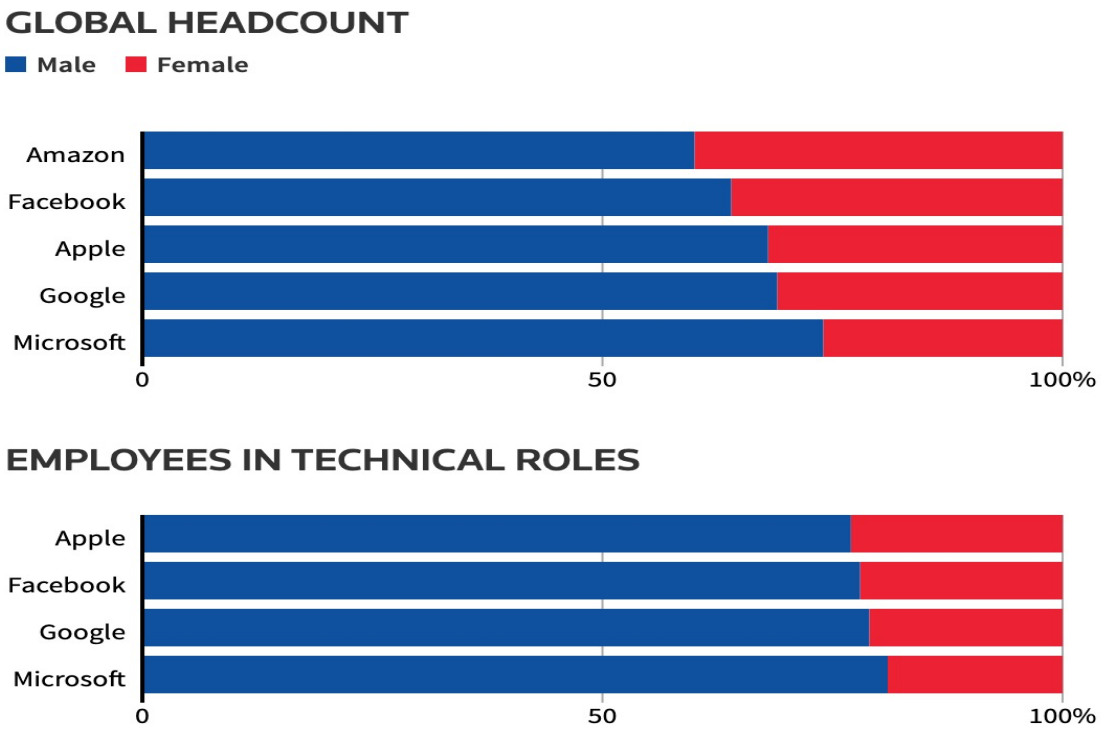
By Sigal Samuel | Updated Mar 6, 2019, 10:50am EST

SHARE



Autonomous vehicles may drive racial inequity on the highway if we're not careful. | Shutterstock

Amazon AI Hiring Tool Was Discriminating Against Women



Screenshot from Reuters
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS [3]

Algorithmic Bias in Health Care

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Health care algorithms can reinforce existing inequality –
Screen shot from Scientific American[1]

How AI Reinforced Biases In The Criminal Justice System in America

Screen Shots taken from Propublica [2]

Algorithmic Bias is everywhere

goto;

ImageNet, developed by researchers at Stanford, is a widely used database with millions of images that computer vision AI technologies learn from. Historically, images mainly included photos from the US, and various photos were classified problematically – including labels like “nerd” and “slut”. ImageNet Roulette, an art project by Kate Crawford and Trevor Paglen exposed the deep gender, racial and other biases embedded in the database [1].

Screenshot from NY Times[1]

Twitter taught Microsoft's AI chatbot to be a racist in less than a day [4]

Where does bias enter the algorithms?

goto;

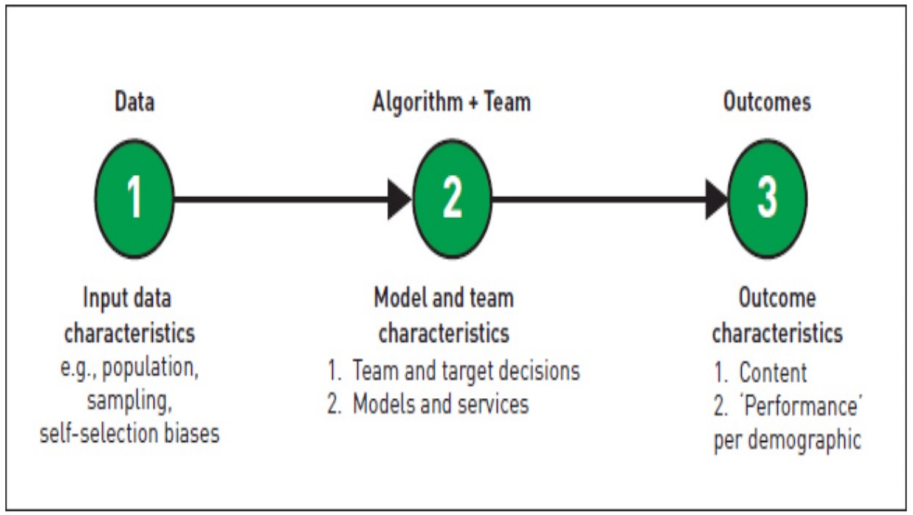
"...whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors [8]."

Where does bias enter the algorithms?

Data: characteristics of the input data

Algorithm + Team: model characteristics as well as team decisions

Desired Outcomes such as recommendation content and served populations



ASSESSING AND ADDRESSING ALGORITHMIC BIAS IN PRACTICE [6]

Fair Algorithms

- Models without discrimination are models that makes the same number of mistakes with each group, which essentially means fairness while acknowledging merit.

		Actual values	
		Positive (1)	Negative (0)
Predicted values	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Fair ALgorithms

goto;

We can think about models without discrimination
by thinking of a model that makes the same number of mistakes
with each group, which essentially means fairness while
acknowledging merit.

- Equality on inputs = Statistical parity
- Equality on outputs = Error rate parity

Analysing trade-offs when choosing who to protect from algorithmic bias

STATISTICAL PARITY: GREAT FOR RANDOM DRAW, FAILS UNDER MERIT OPTIMIZATION

Distributing season tickets

Parity Model

1. Goal is to optimize for profit while preserving parity
2. Model needs to distribute fairly
3. Capacity: 50% of ticket requests
4. Dataset: 40% of skier cohort, 10% of snowboarder cohort spends 5x more on average on concessions

Distributing 100 season tickets

700 applied, 50 accepted, 40 from high-spend cohort, 10 from avg spend cohort

300 applied, 50 accepted, 10 from high-spend cohort, 40 from avg spend cohort

Protecting Groups, Protecting Individuals

EQUALITY OF FALSE NEGATIVE: BETTER FOR PROTECTING GROUPS & MERITS, STILL IMPERFECT FOR INDIVIDUALS IN GROUPS

Distributing loans to pay for season tickets

Error rate Model

1. Goal is to optimize for profit while preserving parity
2. Model needs to distribute false rejections evenly (creditworthy applicants denied loan)
3. Dataset: 40% of snowboarders predicted to repay loans, 20% of skiers

Distributing season ticket loans

random applicant, 10% chance of false rejection

random applicant, 10% chance of false rejection

Accuracy VS Fairness

goto;

Accuracy: In the absence of fairness, reflect the training data, minimize error rate

It is ability of traditional system to make the accurate decisions.

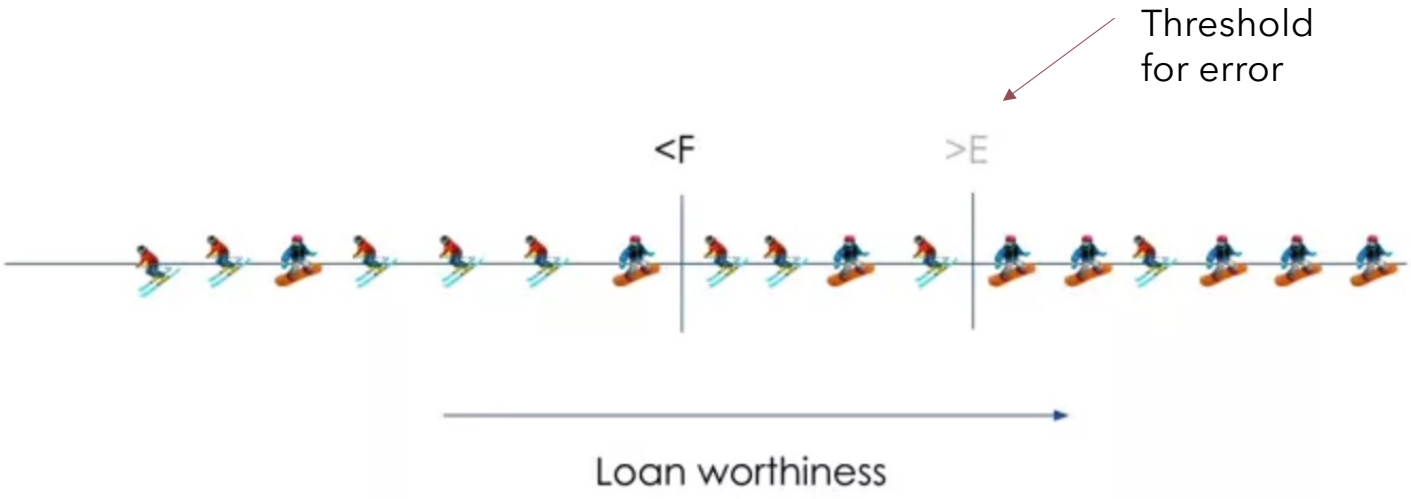
Accuracy VS Fairness

goto;

Accuracy + Fairness: Reflect the training data, minimize error rate as long as parity is obtained

Fairness always comes at the expense of accuracy, unless you have perfect training data, which is often in reality not the case.

An Accurate Model



Combating Bias in Algorithms

goto;

	Inventory algorithms	Screen Each Algorithm for Bias	Retrain Biased Algorithms	Prevention
Step-I	An inventory listing all algorithms your organization is currently using or developing	Articulate the algorithm's ideal target vs. its actual target	Try re-training the model on a label closer to the ideal target	Implement best practices for organizations working with algorithms
Step-II	Talk to relevant stakeholders about how and when algorithms are used	Analyse and interrogate bias	Consider alternative options (if necessary)	
Step-III			Consider suspending or discontinuing use of the algorithm (if necessary)	

ALGORITHMIC BIAS AUDIT PROCESS GUIDE [9]

Screening Algorithms for Bias

goto;

- Identify how to recognize fairness issues and deploy solutions in real world scenarios
- Appraise a predictive model for fairness issue
- Discover auditing model attributes

What can we control?

goto;

- **Change inputs and evaluate outputs**
- **Identify attributes, create auditing data**

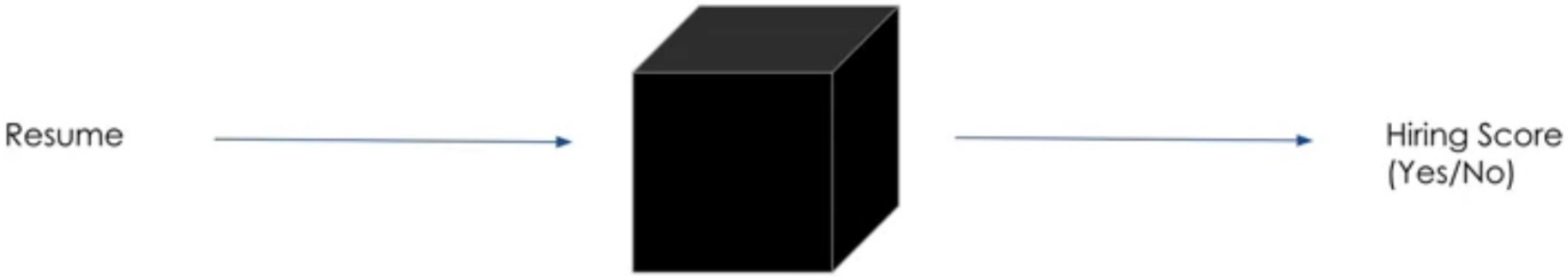
What can we control?

goto;

In a blind model, one that does not explicitly know the race, or gender, or other group categories of those applicants, how do we ensure fairness?

Algorithmic Auditing

goto;



Artificial Intelligence Data Fairness and Bias [10]

Algorithmic Auditing

goto;

Artificial Intelligence Data Fairness and Bias [10]

Auditing Algorithms

goto;

- Change one input, keep others constant
- Score a weight of input attributes on output
- Assemble a picture of the model's true blind spots
- Present audit repot and begin investigating into biased data/fairness metrics

Challenges with Addressing Algorithmic Bias

goto;

- In the absence of standards that apply universally and the diversity of AI usage, each organization should determine what kinds of bias are more likely to skew the algorithms it uses.

- Defining and evaluating bias is simply too dependent on each organization's algorithms and stakeholders

Mitigating Algorithmic Bias

Identify	• Identify your unique vulnerabilities.
Control	• Control your data
Govern	• Govern AI at AI Speed
Diversify	• Diversify your team
Validate	• Validate Independantly and Continously

goto;

GOTO
Copenhagen 2022

THANK YOU

#GOTOcph

References

goto;

1- Mitigating Bias in Artificial Intelligence – A play book for business leaders who build & use AI to unlock value responsibly & equitably by Berkely Haas

2- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

3- https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

4- https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

5- Steps Businesses Can Take To Mitigate AI Discrimination Bias By David Lorimer

6- ASSESSING AND ADDRESSING ALGORITHMIC BIAS IN PRACTICE BY Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, Sravana Reddy

References

goto;

7- Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems by Hardest Larry - MIT News, February 11, 2018. Available at http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

8- Big Data's Disparate Impact by by Solon Barocas and Andrew D. Selbst, SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2016)

9- ALGORITHMIC BIAS PLAYBOOK Center for Applied AI at Chicago Booth

10- Artificial Intelligence Data Fairness and Bias Learn Quest