#SASGF

VIRTUAL

SAS® GLOBAL FORUM 2021

AMERICAS | MAY 18 – 20   ASIA PACIFIC | MAY 19 – 20   EMEA | MAY 25 – 26

Auditing Algorithms -

The Need for New Tools to Combat Bias in Artificial Intelligence

Jason Brinkley, PhD – Abt Associates

Jason Brinkley is a Senior Data Scientist, Biostatistician, and lead of the Research Design and Analytics (RDNA) unit in the Data Science, Surveys, and Enabling Technology (DSET) division of Abt Associates. He works on a wide variety of problems in the areas of health and social sciences. He specializes in the application of advanced statistical methods and machine learning to health data with a specific focus on surveys, claims, electronic health records, and other public use sources. Dr. Brinkley is a research affiliate at the North Carolina Agromedicine Institute, serves as an officer in the Health Policy Statistics Section of the American Statistical Association, and is currently the president of the Southeast SAS Users Group.

Introduction

(Why are we gathered today?)

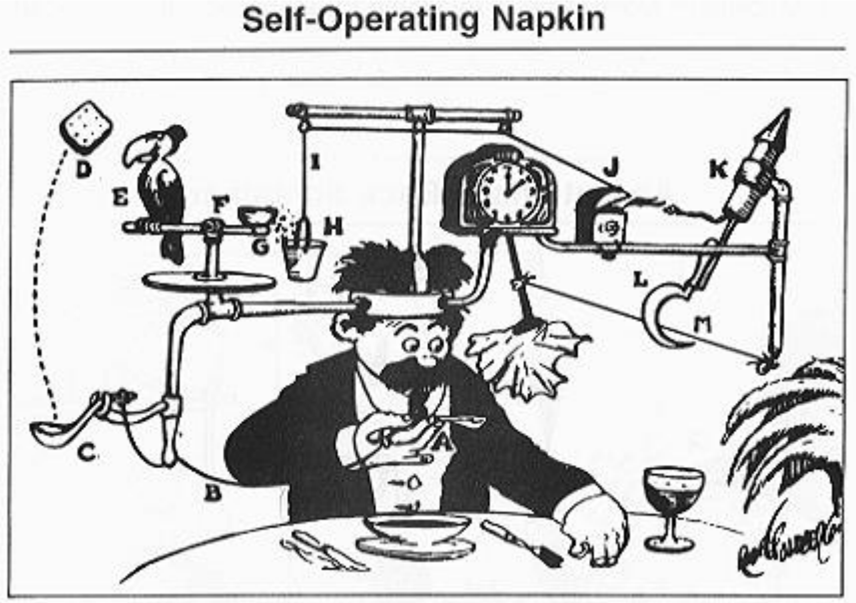
- Artificial Intelligence (AI)/Machine Learning (ML) has seen rapid growth in recent years. Businesses have turned to ML to create products faster and cheaper with higher ‘accuracy’ than human beings.
- A hyper focus on speed has created scenarios where ML has produced results that demonstrate racial, gender, or other similar biases.
- How do we deal with these biases?

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Introduction

(Did we build Rube Goldberg machines?)

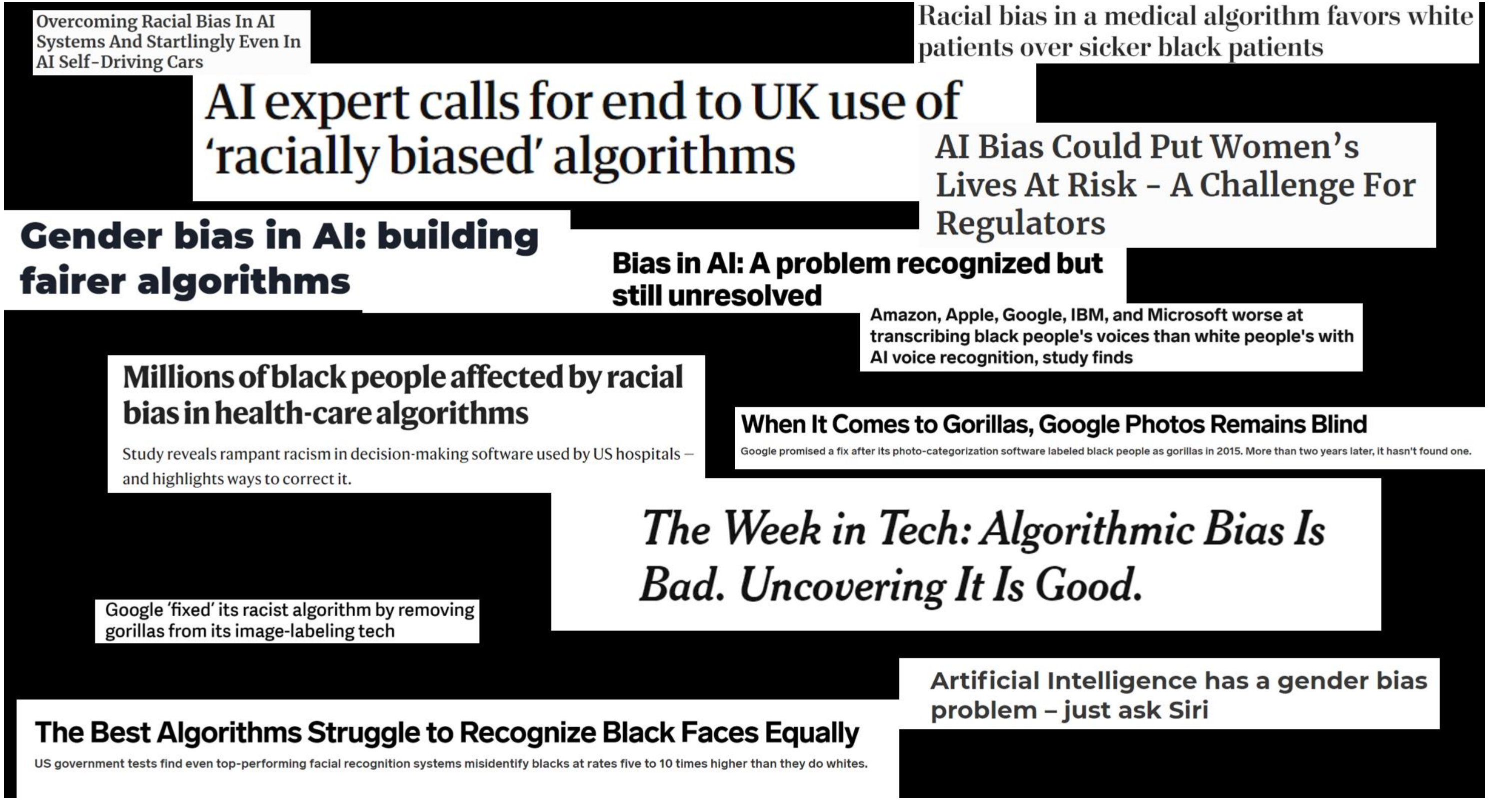
- Many ML algorithms were designed to prioritize speed and accuracy and guarantee output.
- This led to convoluted algorithms with unexpected results.
- The new world of AI generating predictions through an equity lens necessitates new evaluation measures beyond our original goals.



https://en.wikipedia.org/wiki/Rube_Goldberg_machine

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Examples



Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of ‘racially biased’ algorithms

AI Bias Could Put Women’s Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

Image blatantly stolen from https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

How did we get here?

(Decision Makers versus Programmers)

You Never Said It Was Racist!
You Said to Make it Fast and Accurate
imgflip.com

So what do we do?

(No easy fixes)

- We have to recognize that ‘fixes’ are needed at the front as well as the back end.
- Currently, the research community is focused on front end fixes, or developing algorithms that do not have these biases.
- We also have a need for back-end fixes: identify and adjust existing algorithms to remove biases.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

What we can do

- Front end fix - the *new* AI
  - Interpretable Machine Learning
  - Humble Artificial Intelligence
  - Fair Algorithms
- Back-end Fixes – we need new math
  - Metrics
  - Techniques

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Front end fix - the *new* AI

(Looks like the old AI)

- Many existing algorithms look like ‘black boxes’ in deployment: many of the individuals involved in their creation don’t understand how the algorithms make individual-level predictions.
- How would we even know if a new algorithm was better?
  - The **key** is knowing what aspects these new methods are trying to fix.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Interpretable Machine Learning

(Algorithms that humans can explain)

Expanding on the ideas of classic decision lists and decision trees to create a set of rules or an algorithm with discrete steps in implementation.

- Great for evaluating the use of race or gender (e.g.) and other closely-related features.
- Can incorporate expert guidance.
- More than traditional MART or BART trees.
- Starting point for reading – see the work of Dr. Cynthia Rudin at Duke University.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Humble Artificial Intelligence

(Robots that know when to stop)

- Setting boundaries on when AI can be used in decision making. Methodology forces algorithms to not make predictions, estimations, or decisions outside of some ‘comfort zone’.
- When potential decisions transcend the realms of the training data, algorithms are ‘forced’ into stops so that humans can take over.
- Reduces the potential for bias in unusual or extreme circumstances.
- Starting point for reading – see the work of GE Digital led by Colin Parris

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Fair Algorithms

(Ordering machines to respect diversity)

- Algorithms are given additional parameters to ensure equal probability across classes of consideration.
- Can extend to evaluations of performance by forcing ‘accuracy’ metrics to be the same across all groups. Penalizes predictions that tend to favor a specific group overall.
- Starting point for reading – see the work of Dr. Sherri Rose at Stanford

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Back End Fixes

(Don’t throw out the baby with the bathwater)

- Deploying better AI on the front end is great, but those solutions would suggest that it is *better* to ‘start over’ with new algorithms if the current solution is found to be biased.
- That isn’t always a tenable solution. But an even more important question immediately presents itself:

How do we assess whether the current algorithm is biased?

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Metrics – Disparate Impact Score

One of the few existing metrics – **We need more**!

- Suppose Y is a binary (success/fail) outcome and X is a binary class with potential disparity (white/non-white, male/female, etc.). The Disparate Impact Score for any model is defined as

$$\frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}$$

- It measures the ratio of chances between outcomes for being inside or outside the class of interest.
- Super interesting, but only useful in limited scenarios.
- See Feldman et al 'Certifying and removing disparate impact' (2015) on arXiv.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Need to Compare Algorithms

More math

Suppose I deploy an algorithm that performs this way in training:

	True Failure	True Success
Predicted Failure	998000	500
Predicted Success	0	1500

Traditional evaluation of this algorithm would have
– Precision (0.75), Recall (1.00), Accuracy (0.9995).
Overall, this would be deemed a ‘great’ algorithm : it finds 75% of the roughly 2000 successes out of 1 million opportunities.

But suppose you have to replace it with another algorithm? Suppose we go so far as to say that the other algorithm performs with exactly the same metrics. We should be good, right?

An Extreme Scenario

We *need* new math!

Suppose I look at the agreement between algorithms and I get this:

	Old Algorithm Predicted Fail	Old Algorithm Predicted Success
New Algorithm Predicted Fail	998000	500
New Algorithm Predicted Success	500	1000

The two algorithms have some extreme non-overlap. The new algorithm accurately predicts the 500 hundred successes the old algorithm misses. But in doing so, it misses 500 others. The most common ‘agreement’ statistic for such a table is a metric called **Kappa**, which for this case could be thought of as ‘correlation’. Here, we get a value of 0.67. That would suggest high agreement, but it is driven by the ability of both algorithms to prevent failures. If assessing failures is the ‘easier’ task, then we want to assess algorithms on the ‘harder’ task of predicting successes.

Techniques – Crossover Assessment

Not a real thing yet, something my colleagues and I are working on

- Imagine you have two competitor algorithms. If you want to measure ‘agreement’ between algorithms, you could run the results of one algorithm through the other algorithm and see where differences exist.
- I could deploy the results of algorithm A through algorithm B and vice versa and then make comparisons. This could be a useful idea in assessing where biased algorithms went wrong.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Working in practical settings

Lack of perfect is the enemy of good?

- None of the algorithms or assessment mechanisms discussed here are perfect or applicable in every setting.
- Diverse teams are the starting point on any of this, need different viewpoints to find potential biases.
- What can teams do now in lieu of assessing and correcting against bias in algorithms?
  - First, teams can implement ideas or versions of interpretability, humility, and fairness now.
  - Second, teams can get creative in assessment.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Example - Criminal Sentencing

Or instances where systemic issues are deep

- Countless examples in research and media of wide disparities in criminal sentencing.
- The biggest issue is the disparity in the ‘pool’ of candidates for consideration. Much farther upstream in the criminal justice system, the imbalance in white versus non-white offenders creates additional hurdles.
- Even if we apply interpretability, humility, and fairness in sentencing, the issues that led to a person breaking the law, being arrested, having being evaluated in a fair trial, all contribute to bias creation.
- We have a long way to go here.
  - Not limited to criminal justice (see disparities in dialysis and organ donation).

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Example – Bot Detection

A case against open source

- There are several ‘bot detection’ algorithms on the internet for evaluating whether social media accounts should be flagged for bot-like behavior.
- We recently did a study of vaccine-related Twitter content and wanted to ensure our language models weren’t biased by bots. We downloaded two ‘public’ packages for bot testing. Complete non-agreement (see Kappa slide).
- Conclusion – bot algorithms are ever evolving, and open-source evaluation of bot-like behavior means that both the evaluators and the creators have access to the means of detection.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Example – Public Record Redaction

‘Lou Gehrig is a great ballplayer’ versus *‘He died of Lou Gehrig’s disease’*

- There are great Natural Language Processing and AI tools for reading open text. There is a great need to scan text and remove personal identifiers, especially in health-related material.
- Literature already points to challenges in redacting non-white names. Hard to compare algorithms since they redact different words in different places of open text.
- Here is where we have been experimenting with Crossover Assessment.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Thank you!

Contact Information
Jason_Brinkley@abtassoc.com

Twitter - @DrJasonBrinkley

Abt ASSOCIATES

BOLD THINKERS DRIVING
REAL-WORLD IMPACT