Auditing Algorithms - The Need for New Tools to Combat Bias in Artificial Intelligence

BOLD
THINKERS
DRIVING
REAL-WORLD
IMPACT

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?

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

How did we get here?

Example - Medicare

- Total Medicare Enrollment in 2020 was over 61.5 million Americans.
- 38.6 million of those were enrolled in ‘Traditional Medicare’
- Approximately 74% of Medicare beneficiaries are white.

Changing US Demographics*

- In 1990, there were 32 million white Americans aged 35 to 44.
- In 2020, there were 27 million white Americans aged 65 to 74.
- Much of this change (~5 million) can be attributed to deaths.

*Metrics based on Bridged-Race Population Estimates from US Census published to CDC WONDER

What about for blacks?

- In 1990 there were 4.3 million Americans coded as black/African American aged 35 to 44.
- In 2020 that number was 3.4 million aged 65 to 74.
- So we have a decline of about 860,000 individuals. Versus 5 million decline in the white population.

Why is disparity research hard

- Combining whites and blacks in the same age group we see that that whites out number blacks 8 to 1 both in 1990 and 2020.
- However, the population <u>change</u> among whites was 15% compared to a 20% <u>change</u> in blacks.
- That is to say that if the black population had changed at the same rate as whites, then only 650,000 individuals would be gone. That leaves about 200,000 individuals that can be potentially attributed to the disparity in white versus black population change.

What does this have to do with AI?

- We can think of the cost of disparity here as 200,000 individuals among a combined 38 million pop of whites and blacks.

- Our AI infrastructure is not equipped to find or service those 200,000 individuals. We build algorithms for the 38 million.
  - Algorithm development frameworks that are built on speed and accuracy will always lean into predicting the majority.
  - One can build an entirely racist algorithm that caters to the 74% of white Americans who are on Medicare and have a highly predictive and accurate model.

So what do we do?

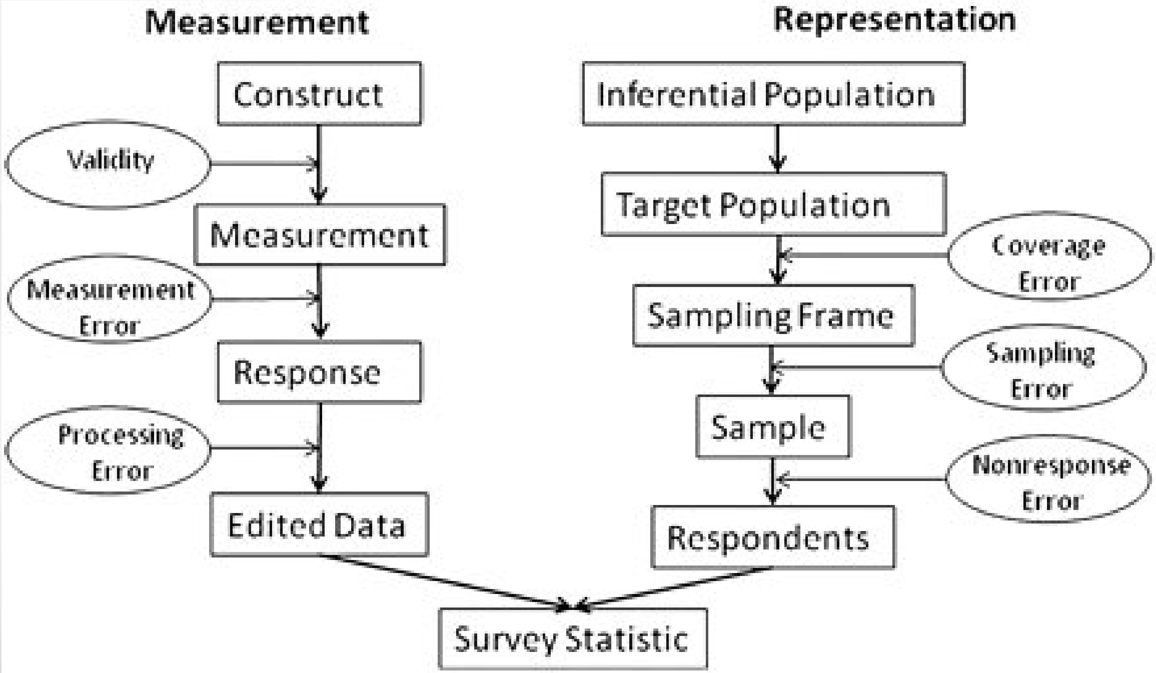
- First, we need more philosophical models for creating prediction algorithms. We need frameworks that are tested and validated.

- Currently, algorithms are developed individually or in silos with an observed data focus.

- Mostly we say ‘does this work well on the data we have? Then deploy and tweak along the way.’

An Example Framework – Total Survey Error

- This is a well developed and studied framework for assessing multi-source bias in survey data.
- Researchers use this as a focal point for discussion, evaluation, and assessment.
- We need more of these for the world of AI and for people to USE them.



The rest of this talk will focus on things related to the measurement side of algorithm development.

What else can we do?

- Second, 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.

Fixes – An Overview

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

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.

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.

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

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

None of these are perfect

- Interpretable algorithms sometimes trade accuracy for interpretability. But can be key in a world full of disparities across underrepresented groups.

- Humble algorithms may be hard to deploy since much of the decision making may have to still be left to humans.

- Fair algorithms can overcorrect the problem or make heavy handed penalties in the face of messy or unavailable data.

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?

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.

Algorithm Assessment: Scenario

- 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.

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

- 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.

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.

- **Visualize Visualize Visualize**!

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).

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.

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.

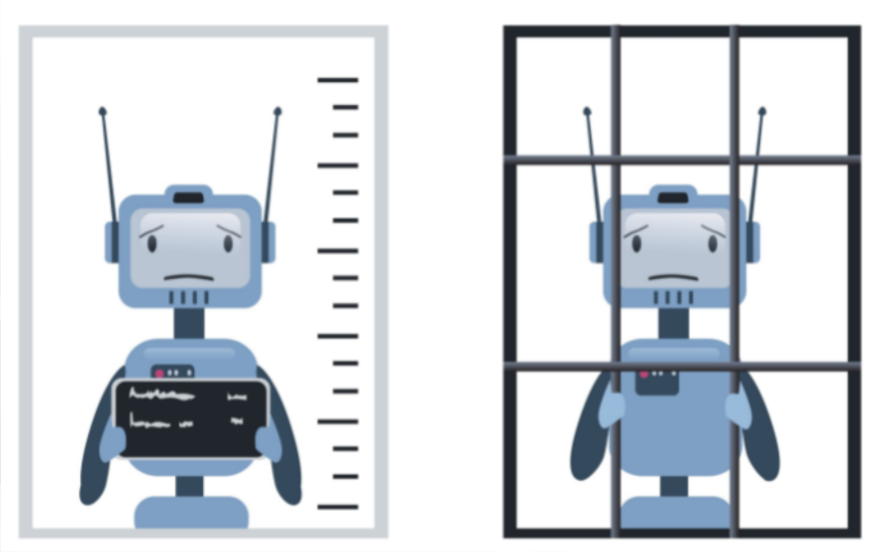
Conclusions

- The current global AI development framework favors the majority. As such there will always be representation issues until we change the framework.

- We need more work in this area on all fronts:
  - Philosophical frameworks
  - New Algorithms that are Equity-Specific
  - Metrics, techniques, and visuals that are AI-Specific

Final Remark: AI needs to be policed

- We need to evoke strong imagery on the dangers of unregulated AI.
- Algorithms should not be deployed and unmonitored.
- We need a certification or license process for background algorithms that are charged with decision-making.



See my recent blog: https://www.abtassociates.com/insights/perspectives-blog/check-the-math-if-we-want-equitable-society-we-need-to-police-ai

Questions or Comments:

Jason Brinkley | Principal Data Scientist
Abt Associates
Central Park West | Suite 210
5001 South Miami Boulevard | Durham, NC 27703
O: 919.294.7745 | F: 919.294.7701 |
Jason_Brinkley@abtassoc.com

Twitter - @DrJasonBrinkley