Algorithmic Bias and Responsible AI

Anjana Susarla

Omura-Saxena Professor of Responsible AI
Eli Broad College of Business
Michigan State University
(asusarla@msu.edu)

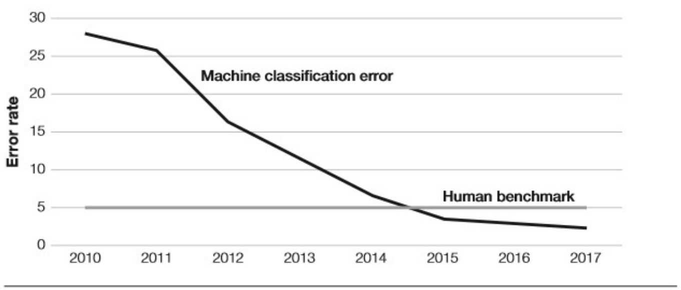
Twitter: @asusarla

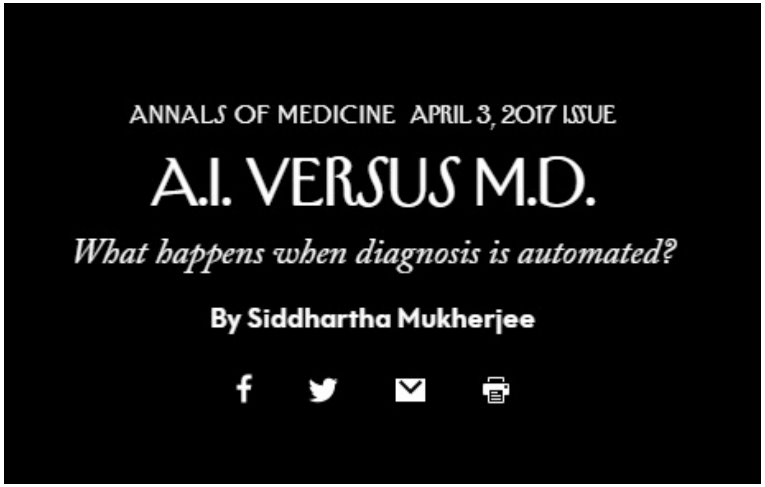
Material with grateful acknowledgements from Ashley Casovan, Virginia Dignum and Kay Firth-Butterfield

Advances in AI = Advances in Prediction!

FIGURE 3-1

Image classification error over time



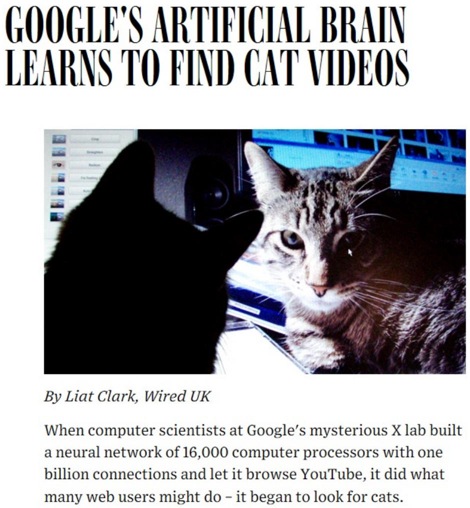


ANNALS OF MEDICINE APRIL 3, 2017 ISSUE

A.I. VERSUS M.D.

What happens when diagnosis is automated?

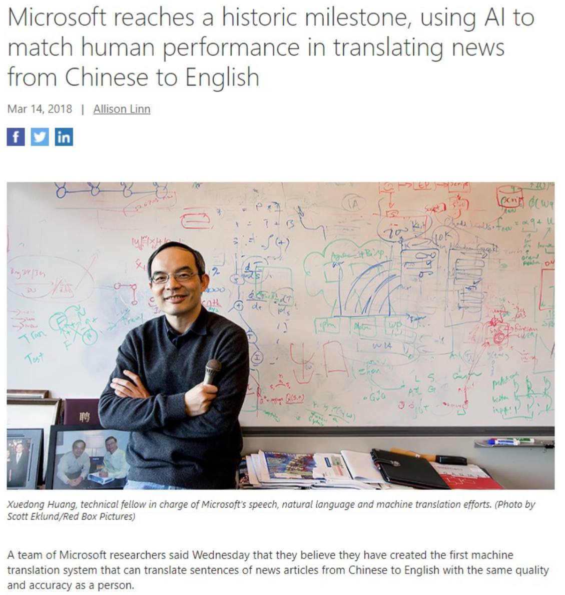
By Siddhartha Mukherjee



GOOGLE'S ARTIFICIAL BRAIN LEARNS TO FIND CAT VIDEOS

By Liat Clark, Wired UK

When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do – it began to look for cats.



Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

Mar 14, 2018 | Allison Linn

Xuedong Huang, technical fellow in charge of Microsoft's speech, natural language and machine translation efforts. (Photo by Scott Eklund/Red Box Pictures)

A team of Microsoft researchers said Wednesday that they believe they have created the first machine translation system that can translate sentences of news articles from Chinese to English with the same quality and accuracy as a person.

From Agarwal et al. (2018)

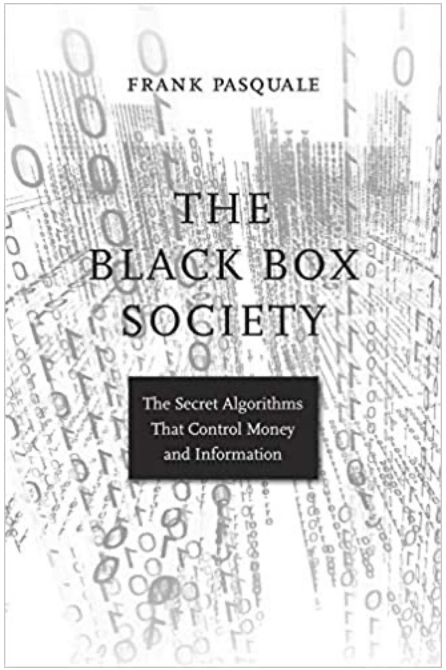
Automated Decisions

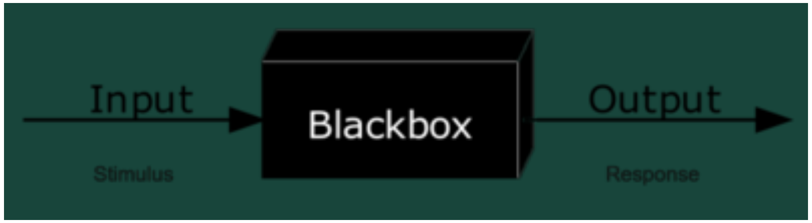
I'm letting the computer decide whether I should cut the rope!

Source: iClipart

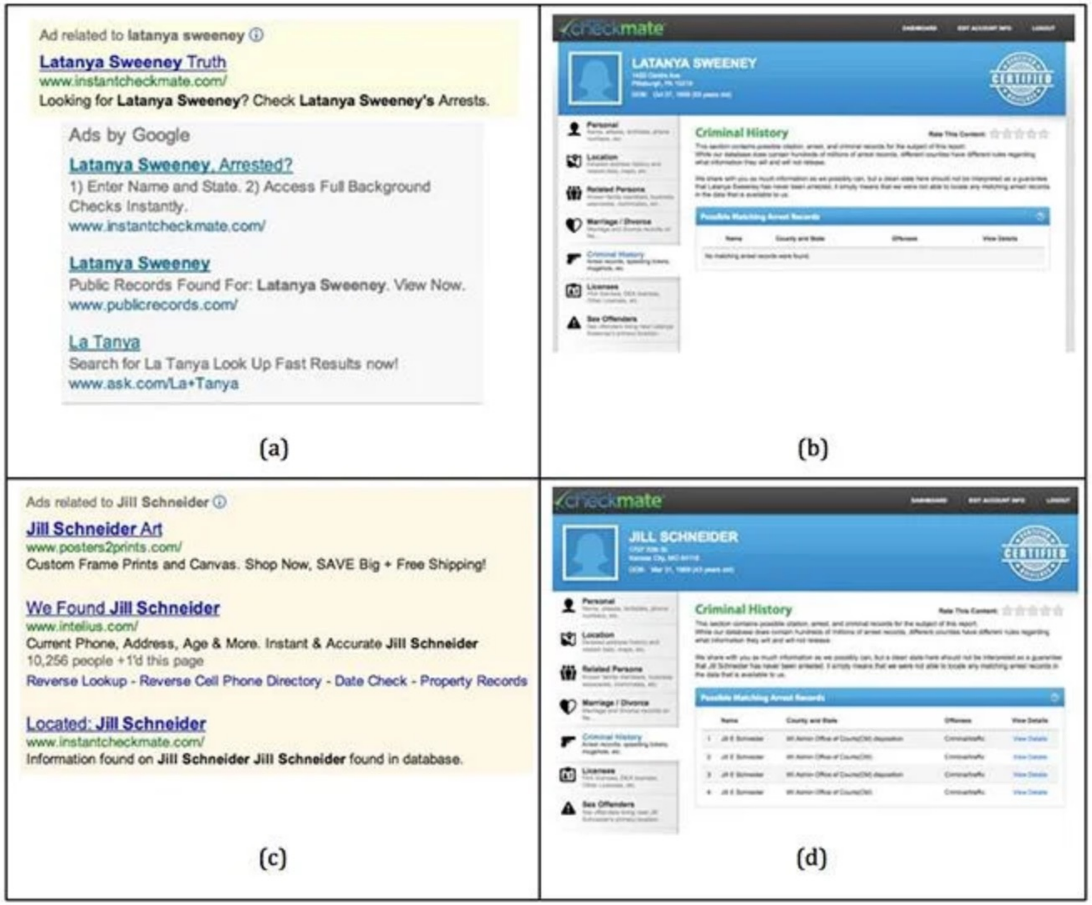
We had better be quite sure that the purpose put into the machine is the purpose which we really desire.

Norbert Wiener (1960)





The Black Box Society (Frank Pasquale, 2015)



Ad related to latanya sweeney ⓘ

Latanya Sweeney Truth
www.instantcheckmate.com/
Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.

Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya Sweeney**. View Now.
www.publicrecords.com/

La Tanya
Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

(a)

(b)

Ads related to Jill Schneider ⓘ

Jill Schneider Art
www.posters2prints.com/
Custom Frame Prints and Canvas. Shop Now, SAVE Big + Free Shipping!

We Found Jill Schneider
www.intelius.com/
Current Phone, Address, Age & More. Instant & Accurate **Jill Schneider**
10,256 people +1'd this page
Reverse Lookup - Reverse Cell Phone Directory - Date Check - Property Records

Located: Jill Schneider
www.instantcheckmate.com/
Information found on **Jill Schneider Jill Schneider** found in database.

(c)

(d)

In one study, black-identified names generated different ads than white-identified ones.

Chart courtesy Latanya Sweeney/Harvard University (http://arxiv.org/ftp/arxiv/papers/1301/1301.6822.pdf)

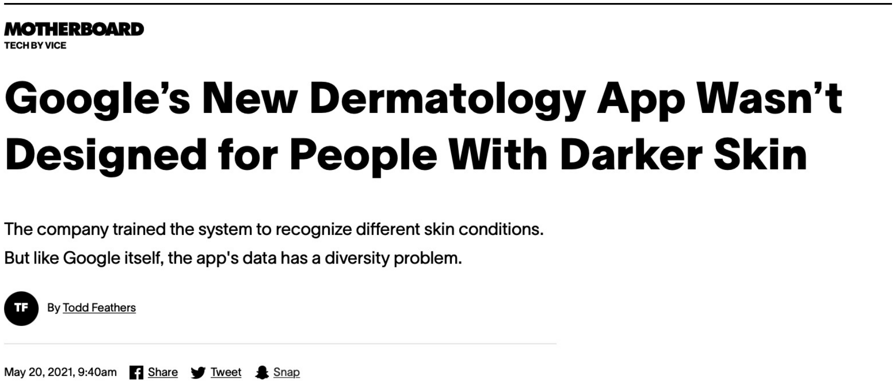
AI Biases

"All my decisions are well thought out."

I'M WAITING TO SEE HOW MANY "LIKES" MY DIAGNOSIS GETS BEFORE I SHARE IT WITH YOU.

Biases in AI

- – Unconscious biases
- – Black box decisions
- – Proxy discrimination
- – Training data



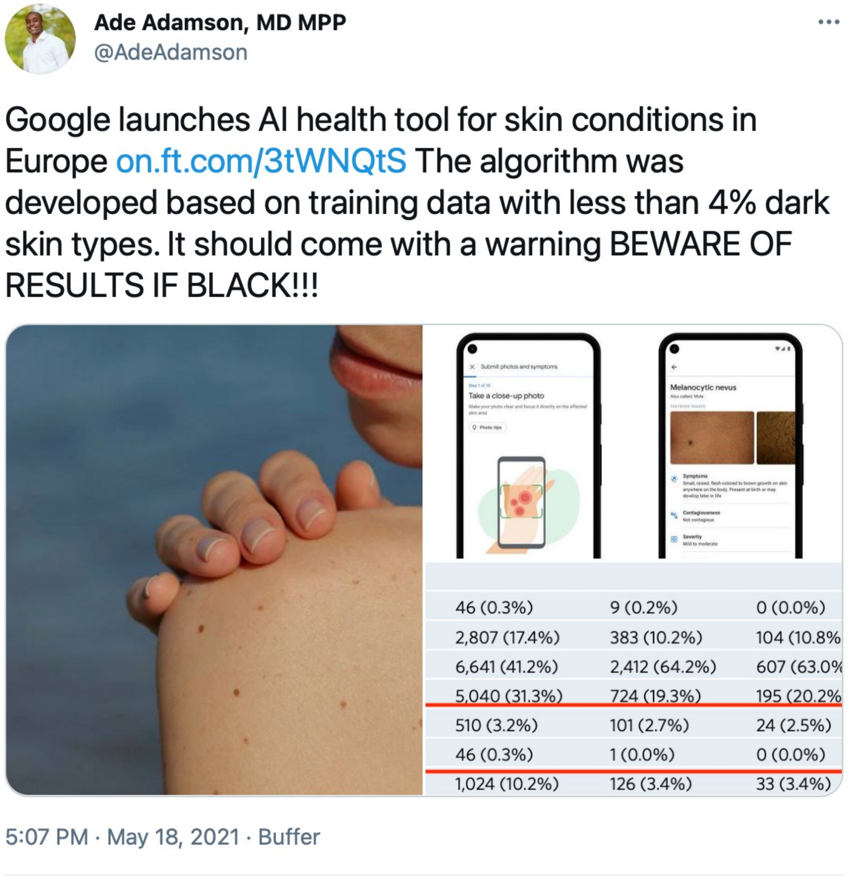
MOTHERBOARD
TECH BY VICE

Google's New Dermatology App Wasn't Designed for People With Darker Skin

The company trained the system to recognize different skin conditions. But like Google itself, the app's data has a diversity problem.

TF By Todd Feathers

May 20, 2021, 9:40am Share Tweet Snap



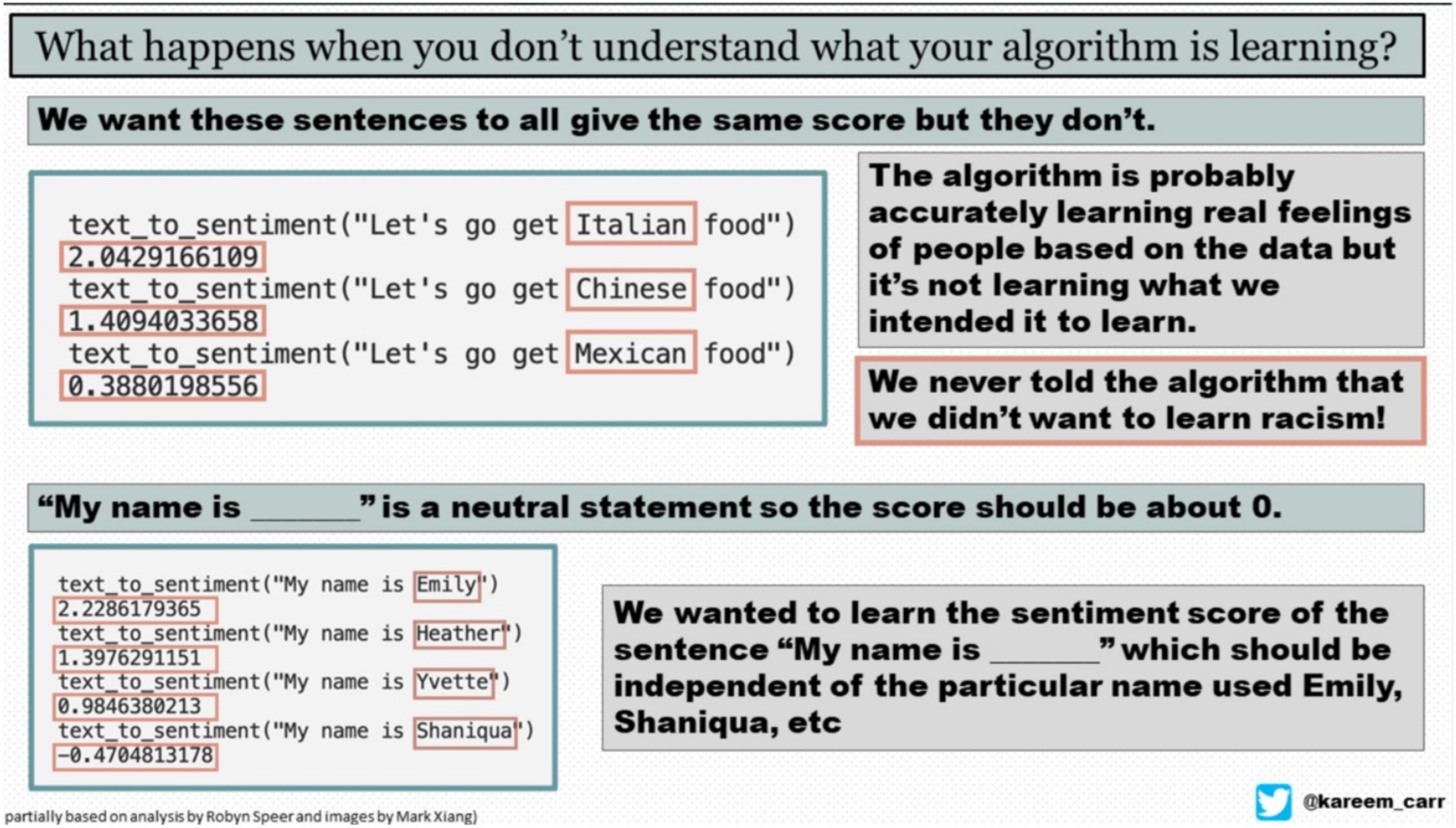
Ade Adamson, MD MPP
@AdeAdamson

Google launches AI health tool for skin conditions in Europe on.ft.com/3tWNQtS The algorithm was developed based on training data with less than 4% dark skin types. It should come with a warning BEWARE OF RESULTS IF BLACK!!!

46 (0.3%)	9 (0.2%)	0 (0.0%)
2,807 (17.4%)	383 (10.2%)	104 (10.8%
6,641 (41.2%)	2,412 (64.2%)	607 (63.0%
5,040 (31.3%)	724 (19.3%)	195 (20.2%
510 (3.2%)	101 (2.7%)	24 (2.5%)
46 (0.3%)	1 (0.0%)	0 (0.0%)
1,024 (10.2%)	126 (3.4%)	33 (3.4%)

5:07 PM · May 18, 2021 · Buffer

How to create a racist chatbot without trying



What happens when you don't understand what your algorithm is learning?

We want these sentences to all give the same score but they don't.

```
text_to_sentiment("Let's go get Italian food")
2.0429166109
text_to_sentiment("Let's go get Chinese food")
1.4094033658
text_to_sentiment("Let's go get Mexican food")
0.3880198556
```

The algorithm is probably accurately learning real feelings of people based on the data but it's not learning what we intended it to learn.

We never told the algorithm that we didn't want to learn racism!

"My name is ______" is a neutral statement so the score should be about 0.

```
text_to_sentiment("My name is Emily")
2.2286179365
text_to_sentiment("My name is Heather")
1.3976291151
text_to_sentiment("My name is Yvette")
0.9846380213
text_to_sentiment("My name is Shaniqua")
-0.4704813178
```

We wanted to learn the sentiment score of the sentence "My name is ______" which should be independent of the particular name used Emily, Shaniqua, etc

@kareem_carr

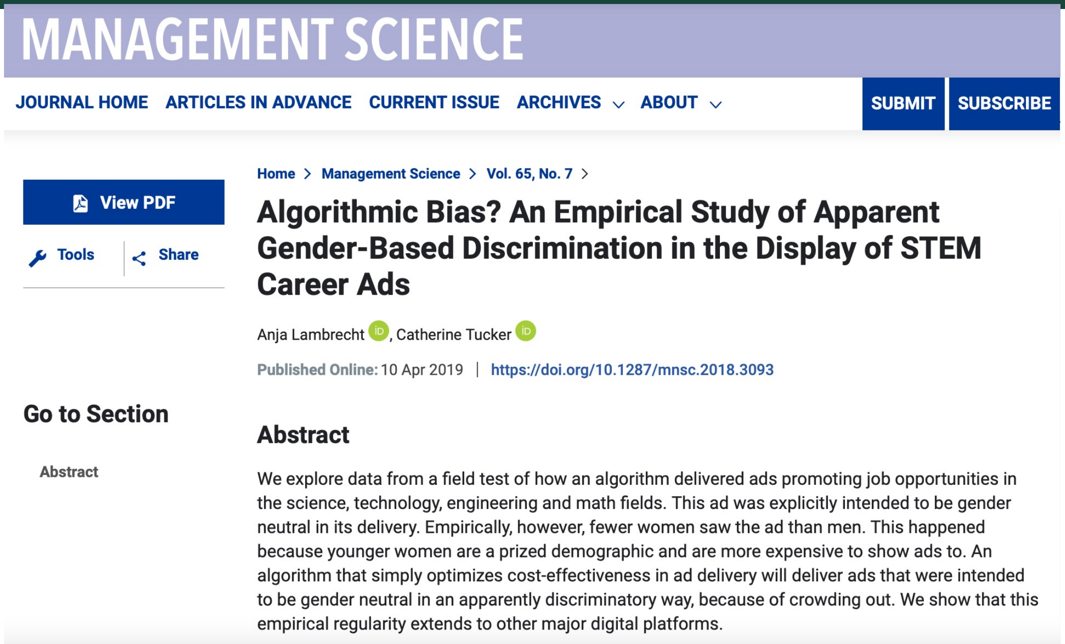
partially based on analysis by Robyn Speer and images by Mark Xiang)

Example 1: Hiring

Amazon realized its hiring system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way

That is because Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

Amazon’s system penalized resumes that included the word “women’s,” as in “women’s chess club captain” and downgraded graduates of two all-women’s colleges. And it privileged resumes with the kinds of verbs that men tend to use, like “executed” and “captured.”



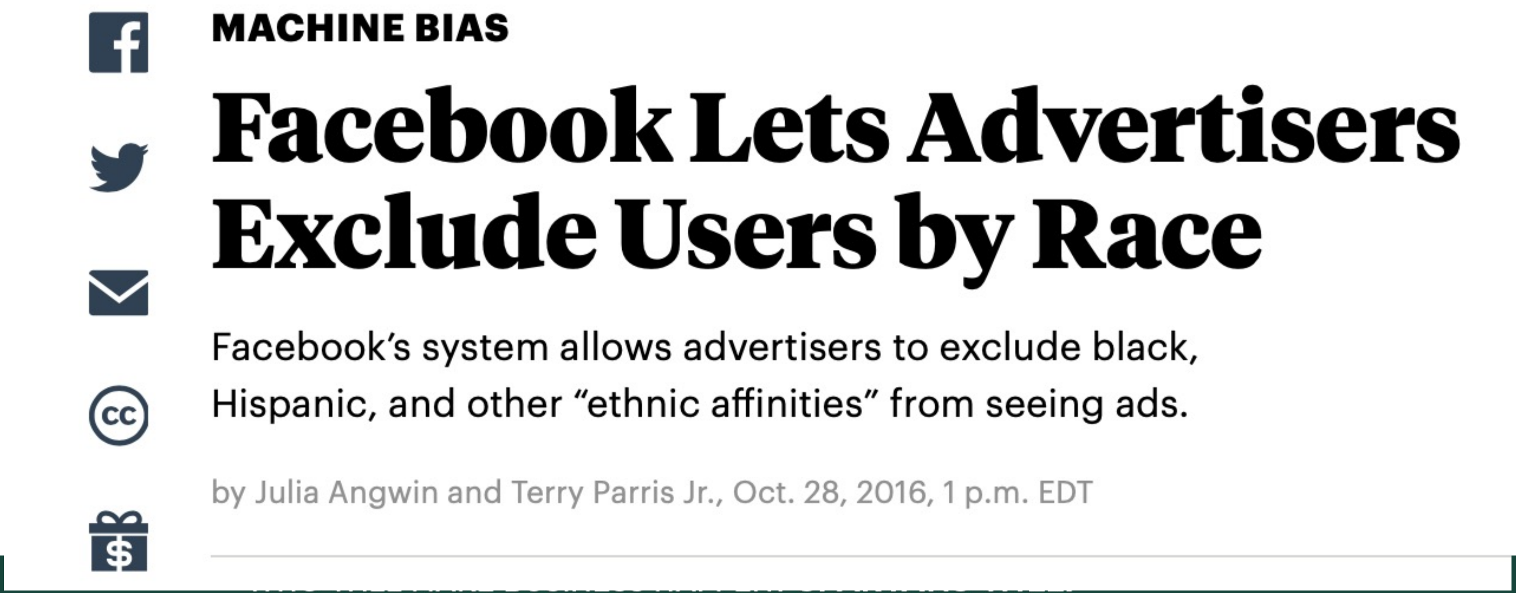
Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Anja Lambrecht, Catherine Tucker

Published Online: 10 Apr 2019 | https://doi.org/10.1287/mnsc.2018.3093

Abstract

We explore data from a field test of how an algorithm delivered ads promoting job opportunities in the science, technology, engineering and math fields. This ad was explicitly intended to be gender neutral in its delivery. Empirically, however, fewer women saw the ad than men. This happened because younger women are a prized demographic and are more expensive to show ads to. An algorithm that simply optimizes cost-effectiveness in ad delivery will deliver ads that were intended to be gender neutral in an apparently discriminatory way, because of crowding out. We show that this empirical regularity extends to other major digital platforms.



MACHINE BIAS

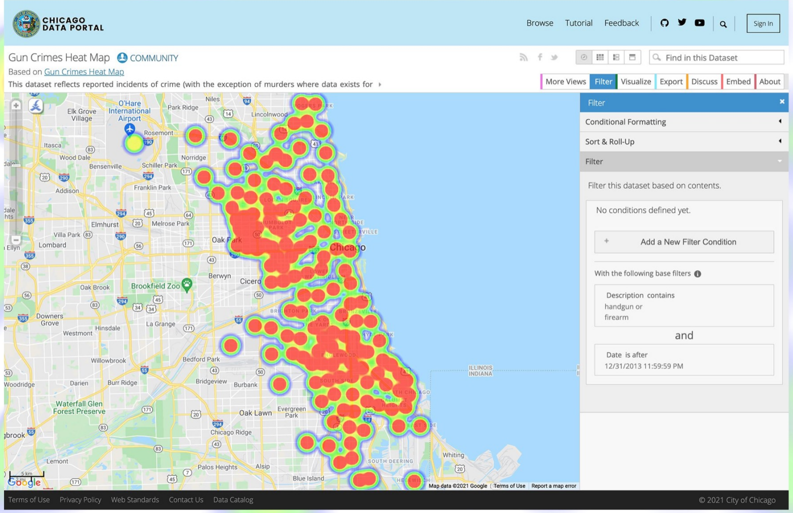
Facebook Lets Advertisers Exclude Users by Race

Facebook's system allows advertisers to exclude black, Hispanic, and other "ethnic affinities" from seeing ads.

by Julia Angwin and Terry Parris Jr., Oct. 28, 2016, 1 p.m. EDT

Example 2: Predictive Policing

- Chicago PD told a man he would be involved in a shooting
  - But could not identify on which side he would be
  - Resulted in continuous surveillance of certain individuals



Based on Gun Crimes Heat Map

This dataset reflects reported incidents of crime (with the exception of murders where data exists for

Filter this dataset based on contents.

No conditions defined yet.

With the following base filters

Description contains handgun or firearm

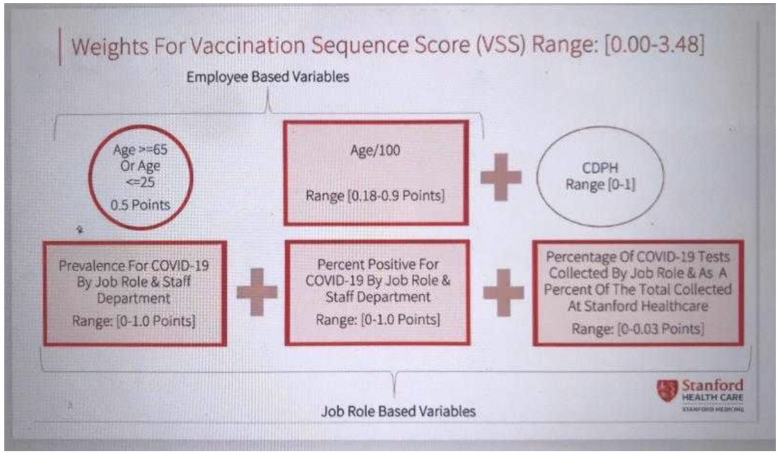
and

Date is after 12/31/2013 11:59:59 PM

Screenshot of the City of Chicago's Gun Crimes Heat Map.

Example 3: Healthcare

How the algorithm works



Weights For Vaccination Sequence Score (VSS) Range: [0.00-3.48]

Employee Based Variables

Age >=65 Or Age <=25
0.5 Points

Age/100
Range [0.18-0.9 Points]

+

CDPH
Range [0-1]

Prevalence For COVID-19 By Job Role & Staff Department
Range: [0-1.0 Points]

+

Percent Positive For COVID-19 By Job Role & Staff Department
Range: [0-1.0 Points]

+

Percentage Of COVID-19 Tests Collected By Job Role & As A Percent Of The Total Collected At Stanford Healthcare
Range: [0-0.03 Points]

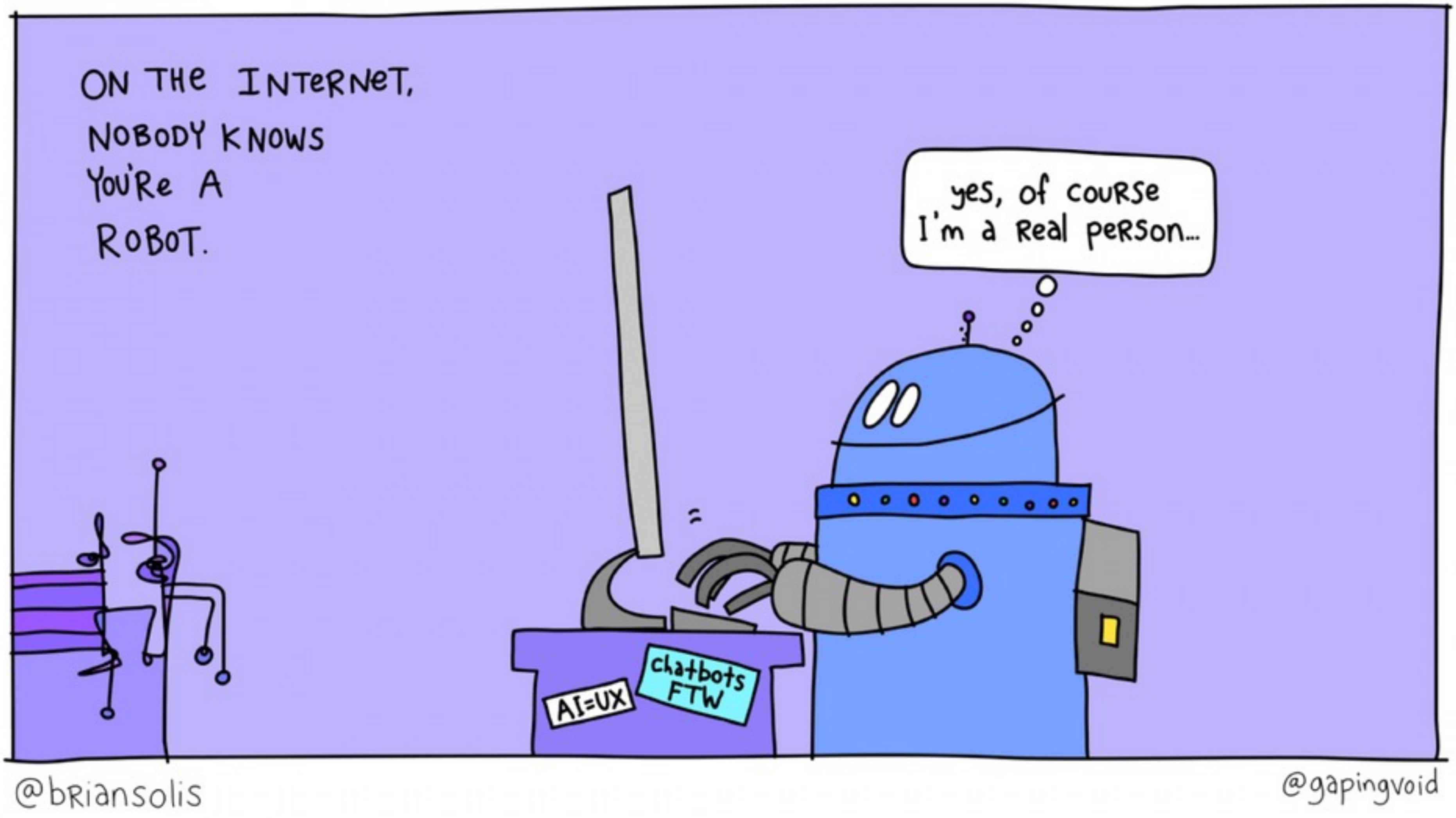
Job Role Based Variables

Stanford HEALTH CARE

Stanford used a so-called “algorithm”, which was really a rules-based formula designed to determine the order in which the thousands of medical workers at Stanford should be vaccinated. The tool took into account employee-based variables like age, job-based variables, and public health guidance, according to MIT Technology Review.

But flaws in that calculation meant hospital administrators and other employees working from home were toward the front of the line, while only seven of Stanford’s 1,300 medical residents made the list.

Fairness, accountability & transparency


ON THE INTERNET, NOBODY KNOWS YOU'RE A ROBOT.
yes, of course I'm a real person...
AI=UX
chatbots FTW
@briansolis
@gapingvoid

AI AND ETHICS

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Self-driving cars

Who is responsible for the accident by self-driving car?

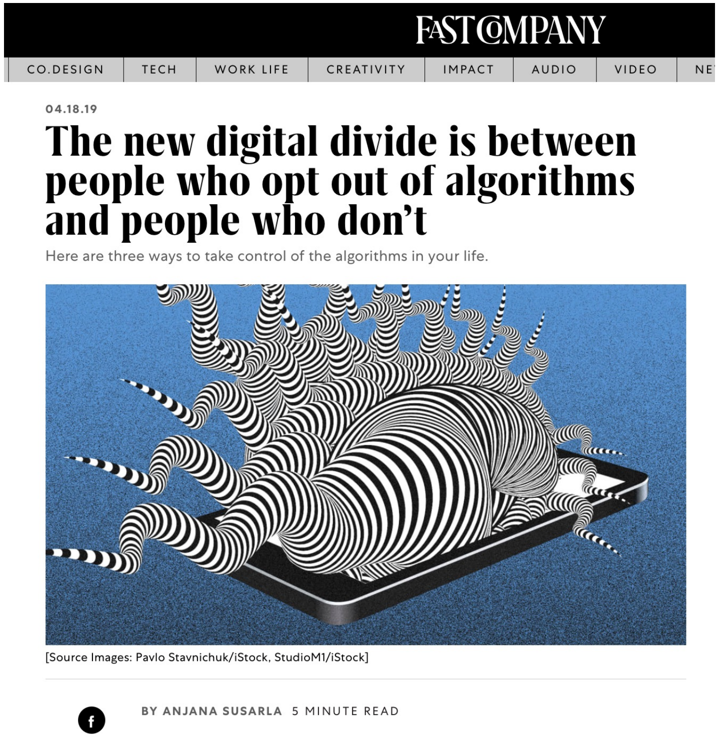
Automated manufacturing

How can workers practice new sophisticated skills so as not to lose their jobs?

Chatbots

Manipulation of emotions / nudging / behaviour change support

My journey into Responsible AI



FAST COMPANY

CO.DESIGN | TECH | WORK LIFE | CREATIVITY | IMPACT | AUDIO | VIDEO | NE

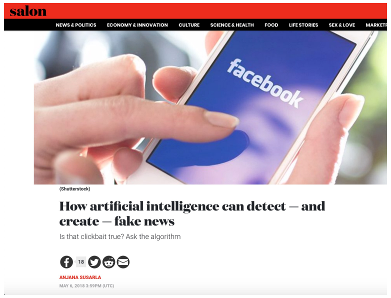
04.18.19

The new digital divide is between people who opt out of algorithms and people who don't

Here are three ways to take control of the algorithms in your life.

[Source Images: Pavlo Stavnichuk/iStock, StudioM1/iStock]

BY ANJANA SUSARLA 5 MINUTE READ



salon

NEWS & POLITICS | ECONOMY & INNOVATION | CULTURE | SCIENCE & HEALTH | FOOD | LIFE STORIES | SEX & LOVE | MARKETP

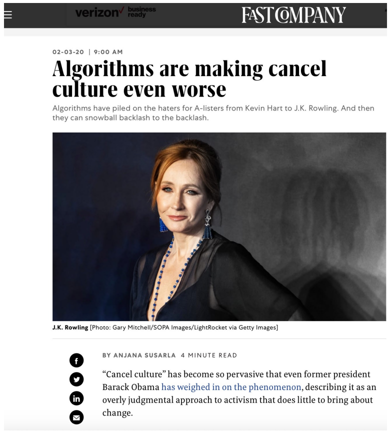
(Shutterstock)

How artificial intelligence can detect — and create — fake news

Is that clickbait true? Ask the algorithm

18

ANJANA SUSARLA
MAY 6, 2018 3:59PM (UTC)



verizon business ready | FAST COMPANY

02-03-20 | 9:00 AM

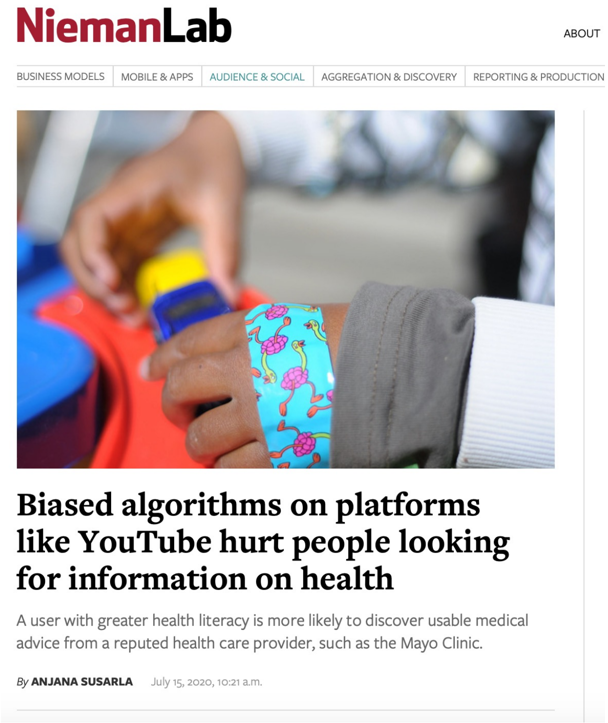
Algorithms are making cancel culture even worse

Algorithms have piled on the haters for A-listers from Kevin Hart to J.K. Rowling. And then they can snowball backlash to the backlash.

J.K. Rowling [Photo: Gary Mitchell/SOPA Images/LightRocket via Getty Images]

BY ANJANA SUSARLA 4 MINUTE READ

"Cancel culture" has become so pervasive that even former president Barack Obama has weighed in on the phenomenon, describing it as an overly judgmental approach to activism that does little to bring about change.



NiemanLab | ABOUT

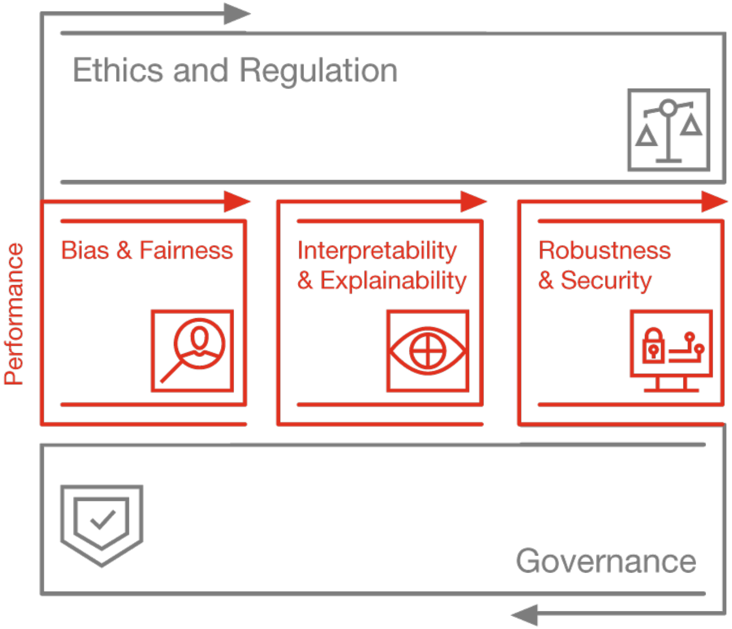
BUSINESS MODELS | MOBILE & APPS | AUDIENCE & SOCIAL | AGGREGATION & DISCOVERY | REPORTING & PRODUCTION

Biased algorithms on platforms like YouTube hurt people looking for information on health

A user with greater health literacy is more likely to discover usable medical advice from a reputed health care provider, such as the Mayo Clinic.

By **ANJANA SUSARLA** July 15, 2020, 10:21 a.m.

Agenda for Responsible AI



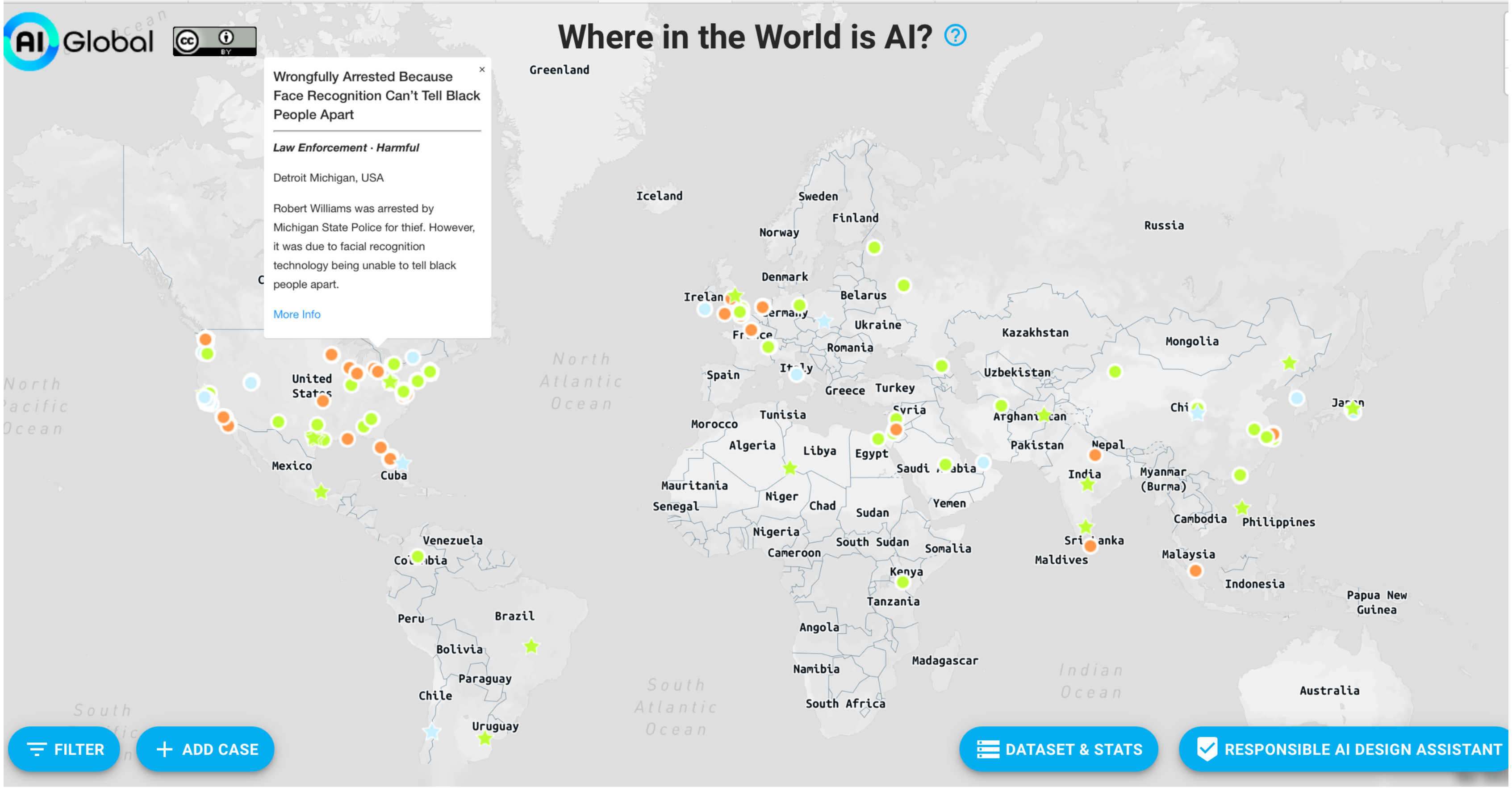
How do we understand algorithmic bias?

A feasible path..

- Develop frameworks for identifying risks
- Understanding (sources of) bias introduced in training data or machine learning models
- Algorithmic auditing and other guardrails

How do we mitigate these biases and other unintended consequences?

From the Responsible AI Institute



Wrongfully Arrested Because Face Recognition Can't Tell Black People Apart

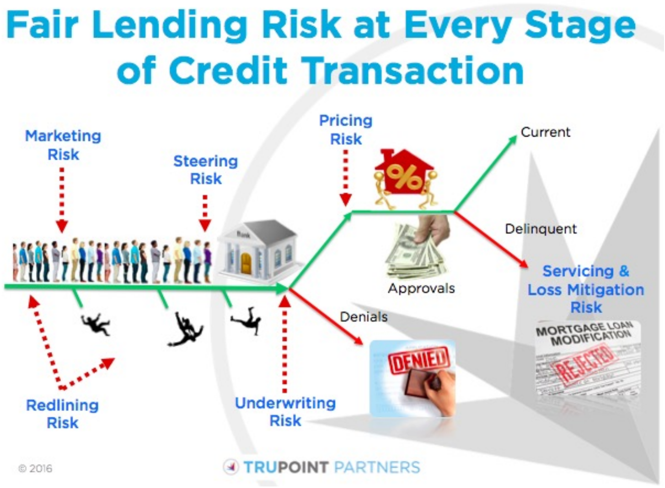
Law Enforcement · Harmful

Detroit Michigan, USA

Robert Williams was arrested by Michigan State Police for thief. However, it was due to facial recognition technology being unable to tell black people apart.

More Info

Checklist for Fair Lending



What are the key challenges in the use of AI systems in current and future lending practices?

What are the key demographics that are at a disadvantage with this work?

How can a certification program ensure that automated lending assisted or produced by an AI system is fair?

What are other measures such as accurate advertising, improved explainability and oversight of these systems that should be required as part of this certification program?

What are the barriers we need to address to accelerate sector-wide responsible AI adoption by businesses?

What type of AI systems are currently used to assist or augment AI practices?

What techniques can be used to automatically assess these systems?

What data is typically collected for lending? Is there other data that should or should not be collected to assess the responsible use of AI systems?

Evaluating fairness-aware algorithms

Which algorithm is the best?

... on which dataset?

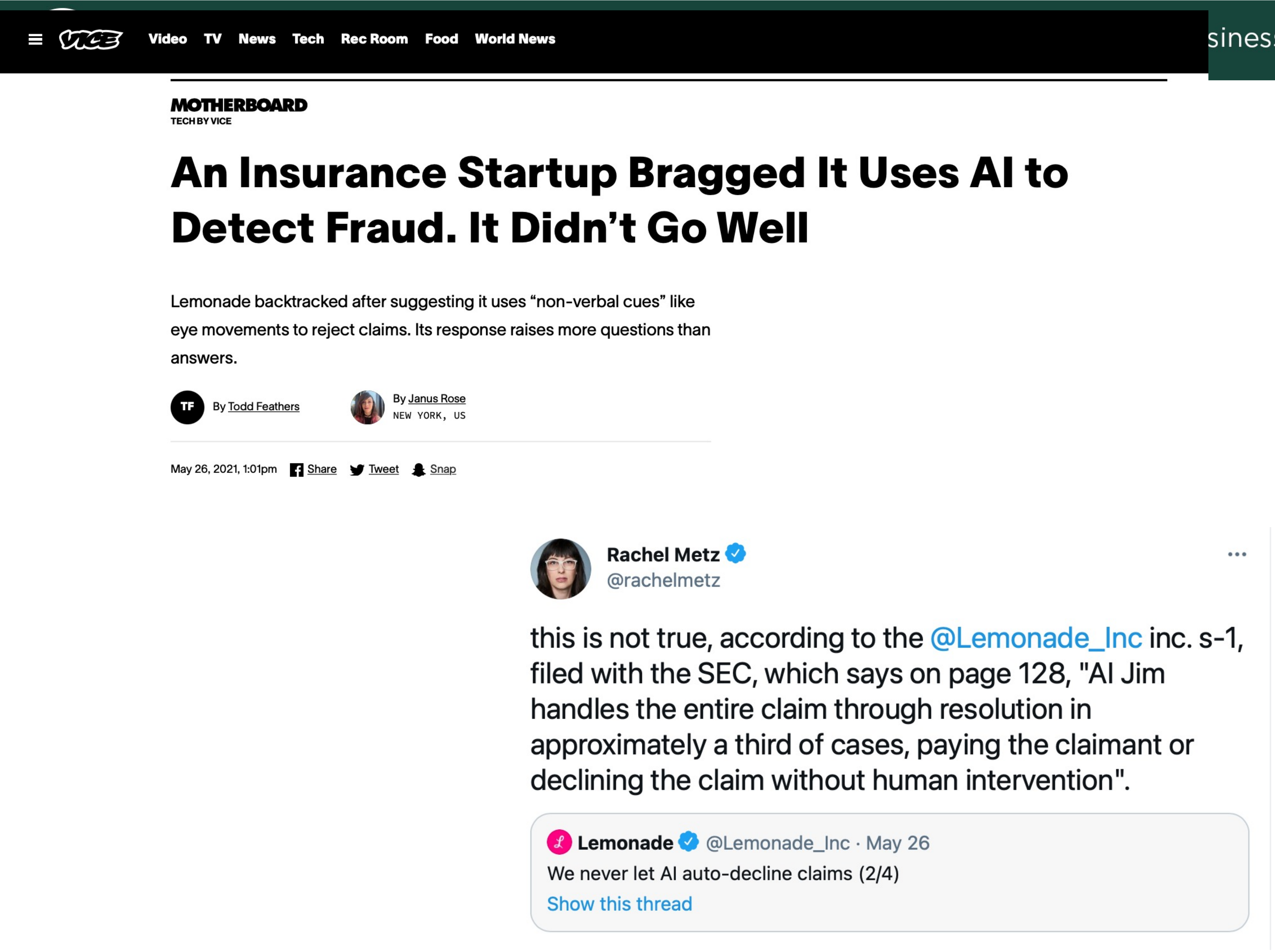
… how was it preprocessed?

… under which measure?

… with which training / test split?

… what are the right hyperparameter settings?

… what if there are multiple sensitive attributes?



An Insurance Startup Bragged It Uses AI to Detect Fraud. It Didn't Go Well

Lemonade backtracked after suggesting it uses "non-verbal cues" like eye movements to reject claims. Its response raises more questions than answers.

By Todd Feathers

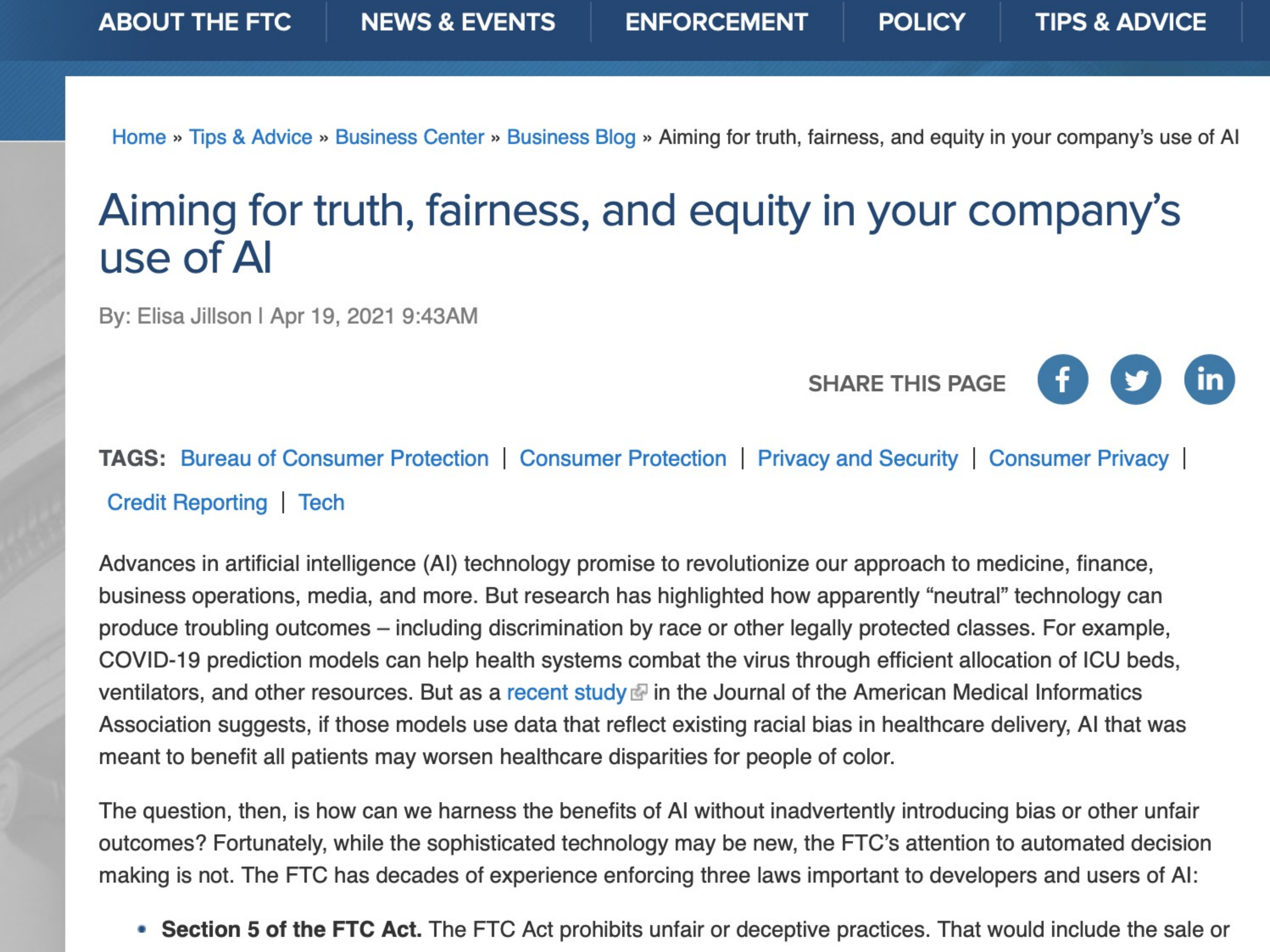
By Janus Rose
NEW YORK, US

May 26, 2021, 1:01pm

Rachel Metz
@rachelmetz

this is not true, according to the @Lemonade_Inc inc. s-1, filed with the SEC, which says on page 128, "AI Jim handles the entire claim through resolution in approximately a third of cases, paying the claimant or declining the claim without human intervention".

> Lemonade @Lemonade_Inc · May 26
> We never let AI auto-decline claims (2/4)
> Show this thread



Home » Tips & Advice » Business Center » Business Blog » Aiming for truth, fairness, and equity in your company's use of AI

Aiming for truth, fairness, and equity in your company's use of AI

By: Elisa Jillson | Apr 19, 2021 9:43AM

TAGS: Bureau of Consumer Protection | Consumer Protection | Privacy and Security | Consumer Privacy | Credit Reporting | Tech

Advances in artificial intelligence (AI) technology promise to revolutionize our approach to medicine, finance, business operations, media, and more. But research has highlighted how apparently "neutral" technology can produce troubling outcomes – including discrimination by race or other legally protected classes. For example, COVID-19 prediction models can help health systems combat the virus through efficient allocation of ICU beds, ventilators, and other resources. But as a recent study in the Journal of the American Medical Informatics Association suggests, if those models use data that reflect existing racial bias in healthcare delivery, AI that was meant to benefit all patients may worsen healthcare disparities for people of color.

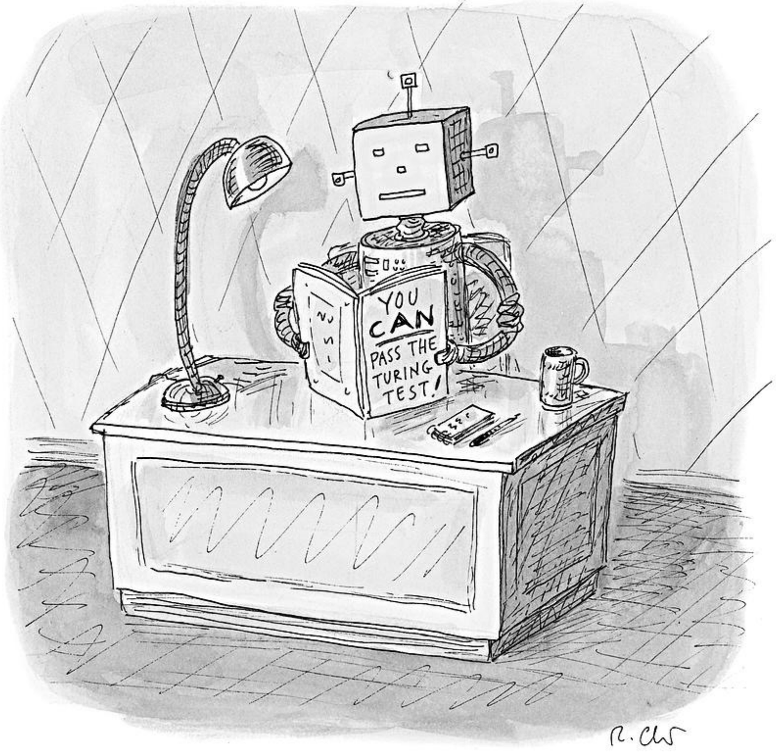
The question, then, is how can we harness the benefits of AI without inadvertently introducing bias or other unfair outcomes? Fortunately, while the sophisticated technology may be new, the FTC's attention to automated decision making is not. The FTC has decades of experience enforcing three laws important to developers and users of AI:

- **Section 5 of the FTC Act.** The FTC Act prohibits unfair or deceptive practices. That would include the sale or

Explainability?

"Does your car have any idea why my car pulled it over?"

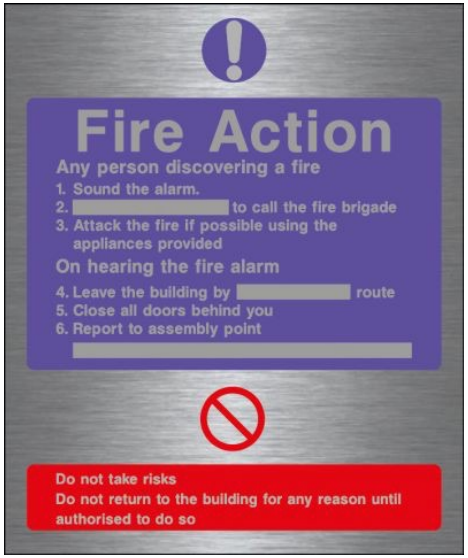
POLICE
PAUL NOTH



WHAT IS AN EXPLANATION?

Terms and Conditions

Correct
Compreensible
Timely
Complete
Parsimonous



Fire Action

Any person discovering a fire

1. Sound the alarm.
2. to call the fire brigade
3. Attack the fire if possible using the appliances provided

On hearing the fire alarm

4. Leave the building by route
5. Close all doors behind you
6. Report to assembly point

Do not take risks
Do not return to the building for any reason until authorised to do so

Email: virginia@cs.umu.se, twitter: @vdignum

RAI Internships/ Experiential Projects

Idea is the converse of typical experiential projects where students get a problem & dataset and build models

RAI Project –students get a data and model, which needs auditing on the appropriateness of the methods and validating data sources/ model training

Different approaches

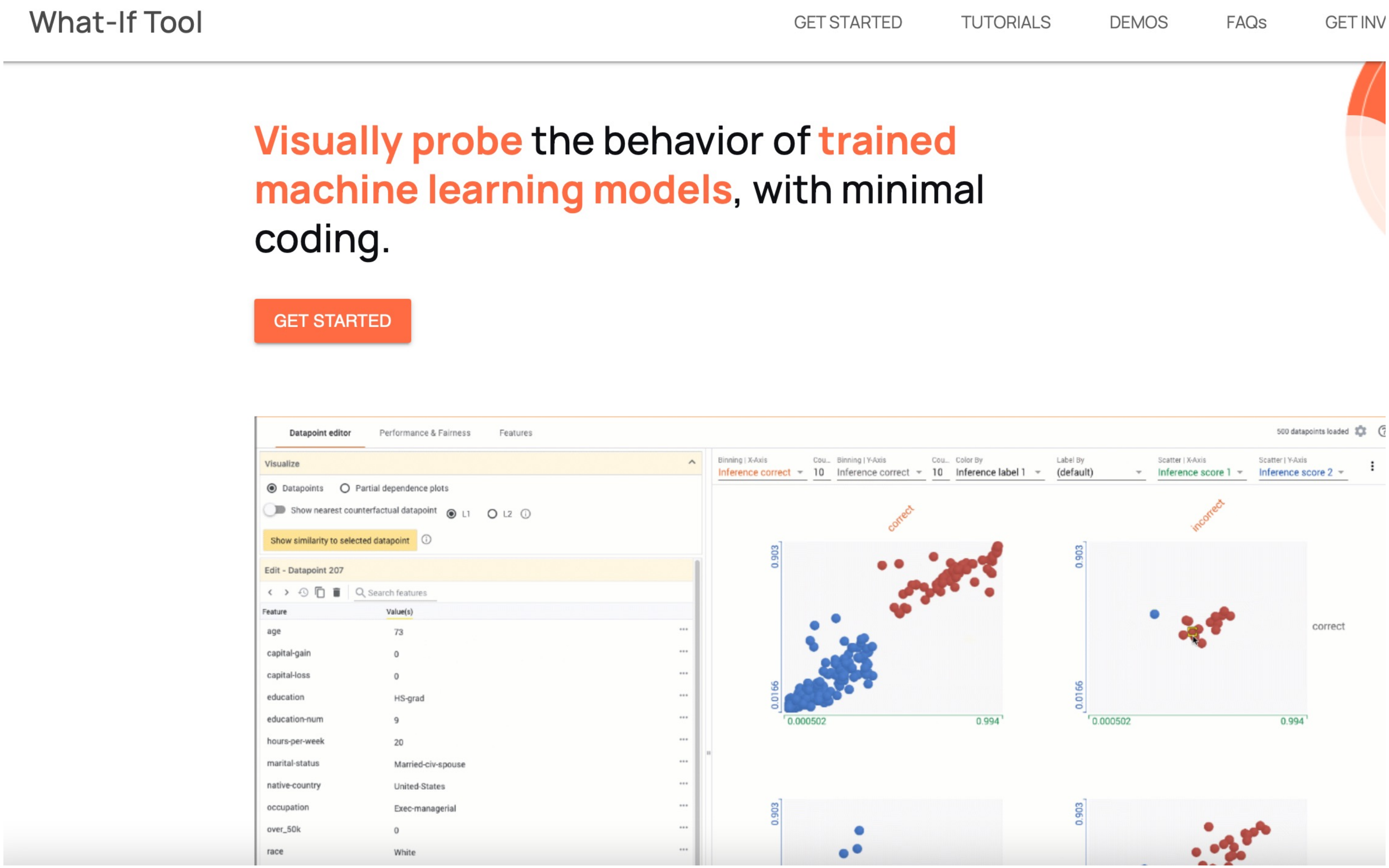
Sandbox

Counterfactual Analysis

Fairness Audits

RAI Collab

What-If from Google Research



What-If Tool

GET STARTED TUTORIALS DEMOS FAQs GET INV

Visually probe the behavior of trained machine learning models, with minimal coding.

GET STARTED

Datapoint editor | Performance & Fairness | Features

500 datapoints loaded

Visualize

● Datapoints ○ Partial dependence plots

Show nearest counterfactual datapoint ● L1 ○ L2

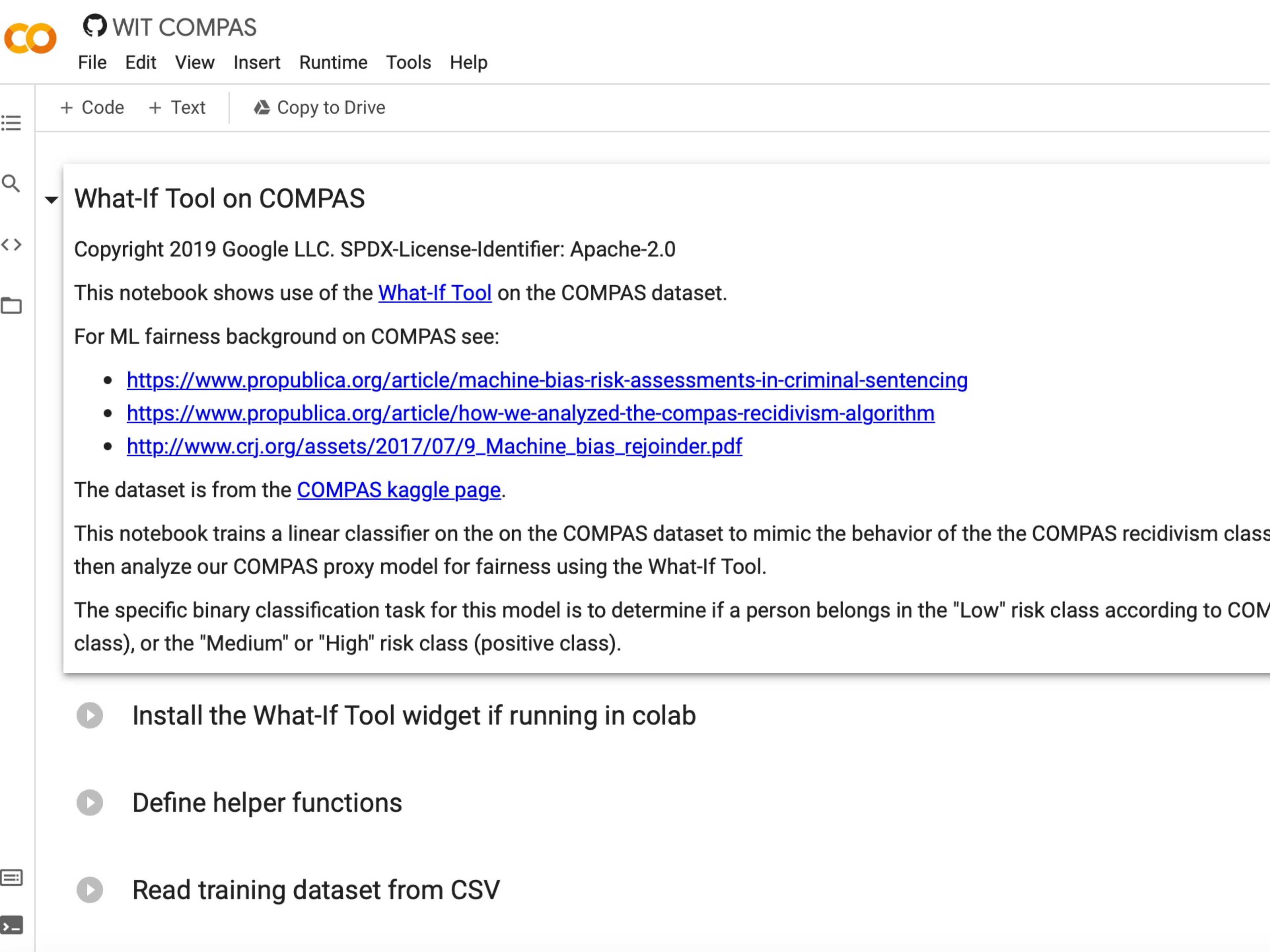
Show similarity to selected datapoint

Edit - Datapoint 207

Search features

Feature	Value(s)
age	73
capital-gain	0
capital-loss	0
education	HS-grad
education-num	9
hours-per-week	20
marital-status	Married-civ-spouse
native-country	United-States
occupation	Exec-managerial
over_50k	0
race	White

Binning | X-Axis Inference correct
Cou... 10
Binning | Y-Axis Inference correct
Cou... 10
Color By Inference label 1
Label By (default)
Scatter | X-Axis Inference score 1
Scatter | Y-Axis Inference score 2



What-If Tool on COMPAS

Copyright 2019 Google LLC. SPDX-License-Identifier: Apache-2.0

This notebook shows use of the What-If Tool on the COMPAS dataset.

For ML fairness background on COMPAS see:

- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf

The dataset is from the COMPAS kaggle page.

This notebook trains a linear classifier on the on the COMPAS dataset to mimic the behavior of the the COMPAS recidivism class then analyze our COMPAS proxy model for fairness using the What-If Tool.

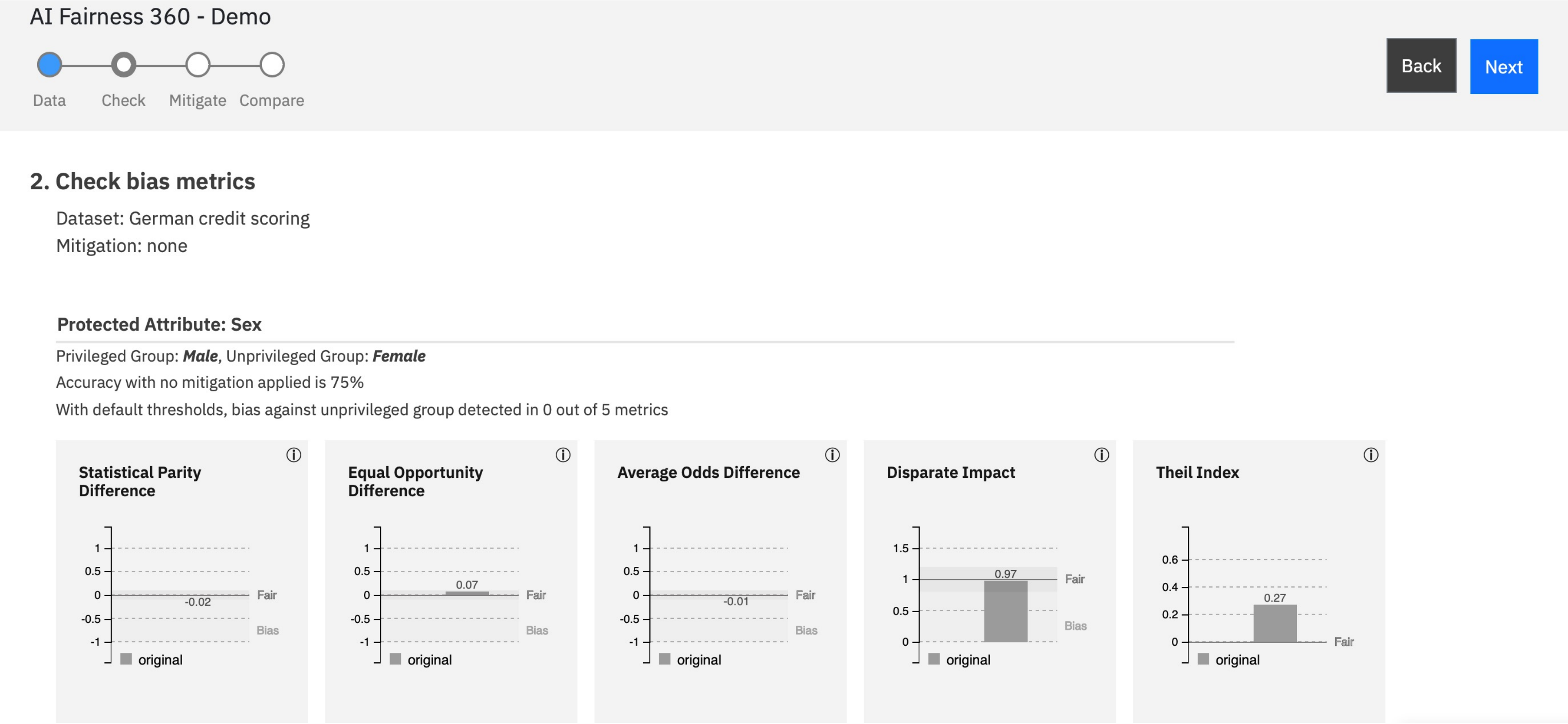
The specific binary classification task for this model is to determine if a person belongs in the "Low" risk class according to COM class), or the "Medium" or "High" risk class (positive class).

Install the What-If Tool widget if running in colab

Define helper functions

Read training dataset from CSV

IBM AI Fairness 360



AI Fairness 360 - Demo

2. Check bias metrics

Dataset: German credit scoring
Mitigation: none

Protected Attribute: Sex

Privileged Group: ***Male***, Unprivileged Group: ***Female***

Accuracy with no mitigation applied is 75%

With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

Building Bias Mitigation

1. Use case

2. Data used

3. Training models without debiasing

4. Reweighing (pre-processing bias mitigation)

5. Prejudice Remover (in-processing bias mitigation)

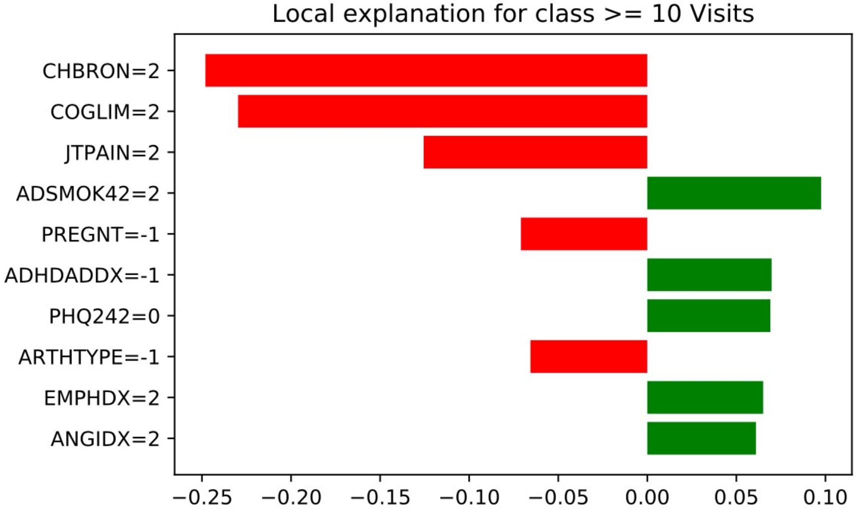
6. Summary of results

7. Deploying model

8. Generating explanations for model predictions using LIME

9. Re-deploying Model

10. Overall Summary

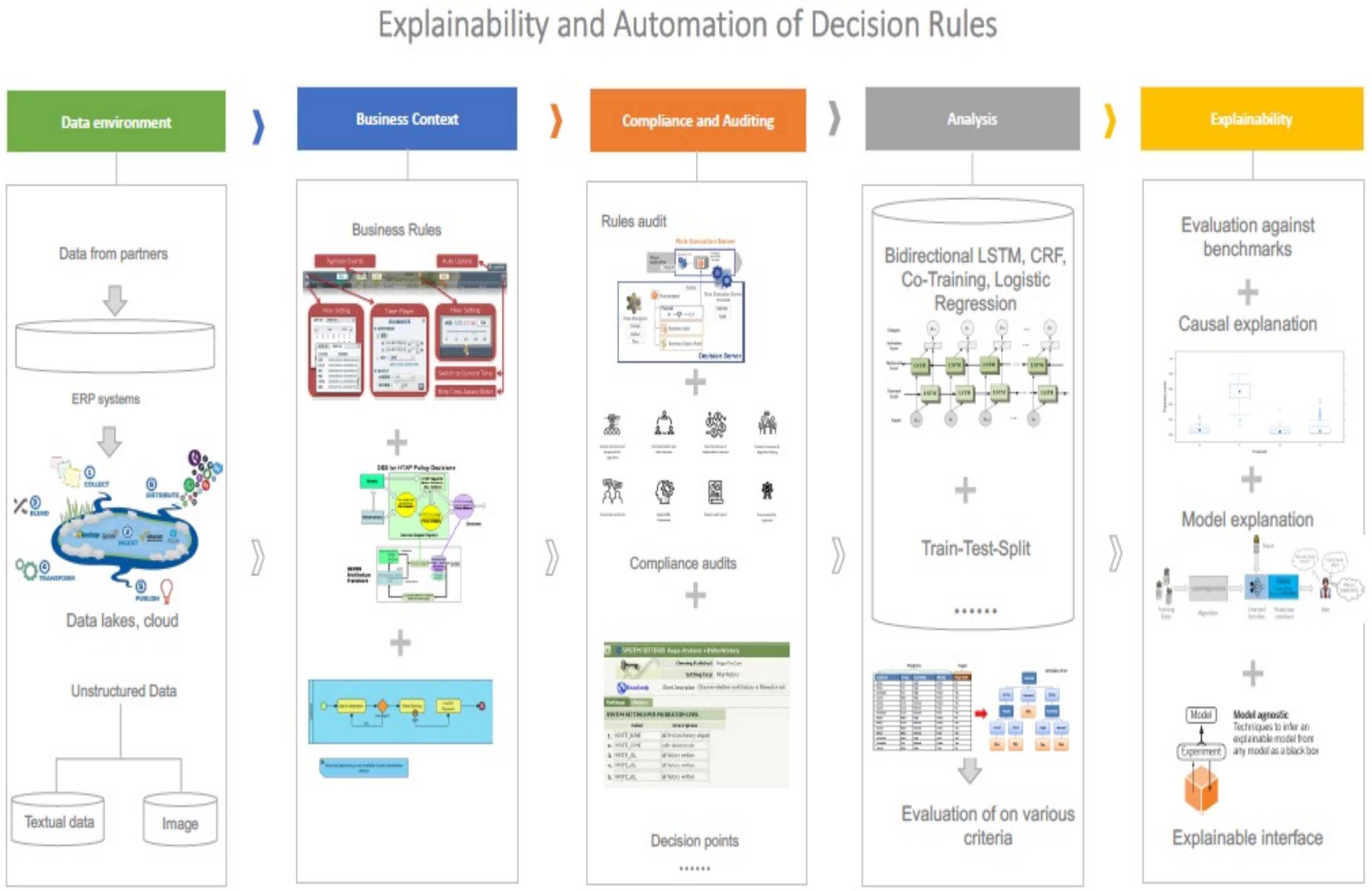


https://nbviewer.jupyter.org/github/IBM/AIF360/blob/master/examples/tutorial_medical_expenditure.ipynb

Use Case of Automated Lending

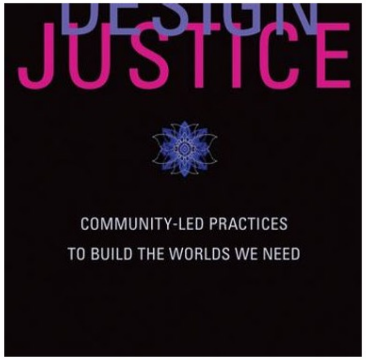
- What are the key demographics that are at a disadvantage with this work?
- How can we ensure that automated lending assisted or produced by an AI system is fair?
- What are other measures such as improved explainability and oversight of these systems that should be required?
- What data is typically collected for lending? Is there other data that should or should not be collected to assess the responsible use of AI systems?

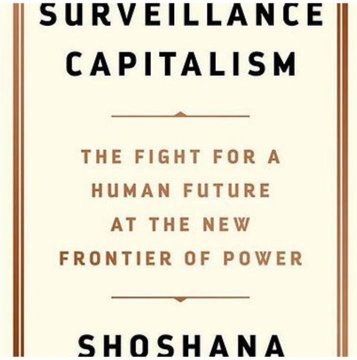
Responsible AI https://anjanasusarla.substack.com

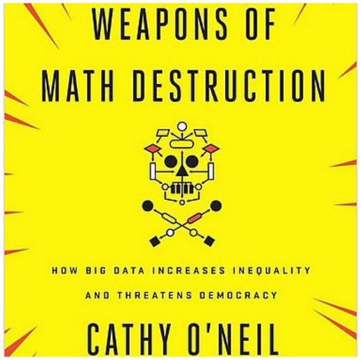


(© Susarla, 2021)

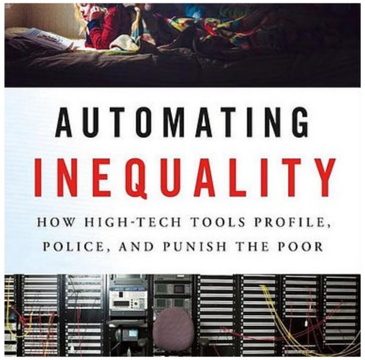
RAI Reading List

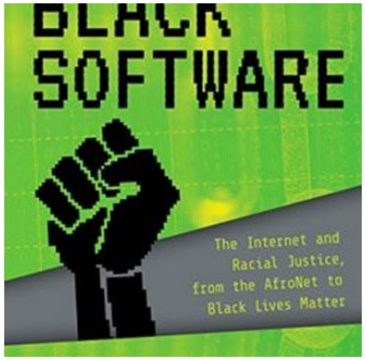




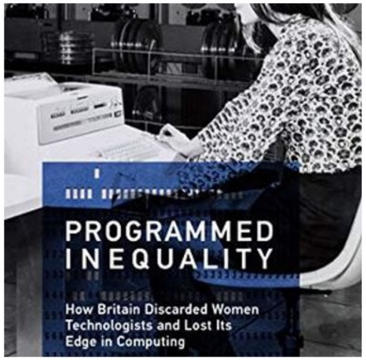


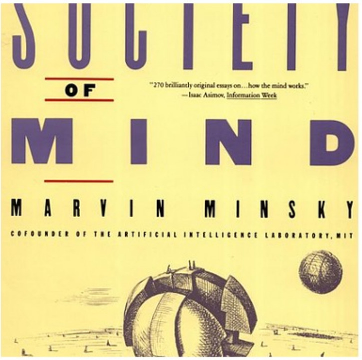
RUHA
BENJAMIN
RACE
AFTER



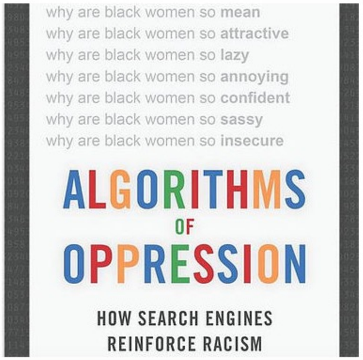


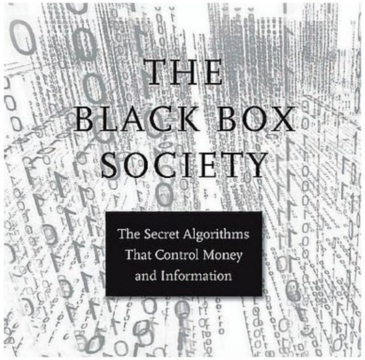
Data
Feminism
Catherine D'Ignazio and
Lauren F. Klein



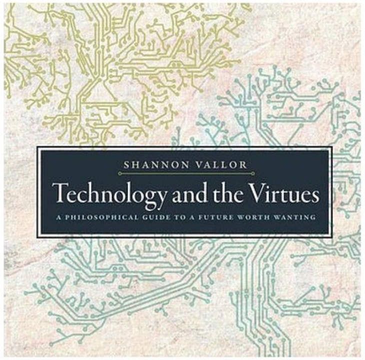

SOCIETY
OF
MIND
"270 brilliantly original essays on... how the mind works."
—Isaac Asimov, Information Week
MARVIN MINSKY
COFOUNDER OF THE ARTIFICIAL INTELLIGENCE LABORATORY, MIT

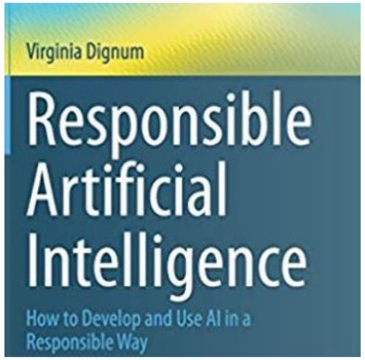
Intelligence
A Guide for
Thinking Humans


why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS
OF
OPPRESSION
HOW SEARCH ENGINES
REINFORCE RACISM



the
ethical
algorithm


SHANNON VALLOR
Technology and the Virtues
A PHILOSOPHICAL GUIDE TO A FUTURE WORTH WANTING



Questions?