#TampaCC

Tampa Community Connect

# A Lap around Algorithmic bias, and AI's Ethical Imperative

Adnan Masood, PhD.

October 2018

# Adnan Masood, PhD.

Dr. Adnan Masood is an Artificial Intelligence and Machine Learning researcher, visiting scholar at Stanford AI Lab, software architect, and Microsoft MVP (Most Valuable Professional) for Artificial Intelligence. As Chief Architect of AI and Machine Learning, at UST Global, he collaborates with Stanford Artificial Intelligence Lab, and MIT AI Lab for building enterprise solutions

Author of Amazon bestseller in programming languages, **"Functional Programming with F#",** Dr. Masood teaches Data Science at Park University, and has taught Windows Communication Foundation (WCF) courses at the University of California, San Diego. He is a regular speaker to various academic and technology conferences (WICT, DevIntersection, IEEE-HST, IASA, and DevConnections), local code camps, and user groups. He also volunteers as STEM (Science Technology, Engineering and Math) robotics coach for elementary and middle school students.

# TRIGGER
# WARNING

# A Lap around Algorithmic bias, and AI's Ethical Imperative

Algorithmic bias is shaping up to be a major societal issue as Artificial Intelligence and Machine Learning continue to rapidly transform the industries. Implicit algorithmic bias poses a threat to fairness, diversity, transparency, and neutrality associated with data driven decision making. It is easy to say that the Algorithms Aren't Biased, we (humans) Are, but is the kind of prejudice and discrimination that already prevails in society inscrutable? GDBR's right of explanation for all individuals to obtain "meaningful explanations of the logic involved" when automated (algorithmic) individual decision is involved is making leadership across industries think long and hard about upcoming regulations pertaining to black-box automated decision-making systems. In this talk, we will explore the question of why do algorithms discriminate? What is unfair bias, Who is in control of the data, How can outsiders validate algorithms and given these risks, how should we use algorithms? Fairness and Bias in an Algorithmic Age has countless examples from Norman's Rorschach inkblots to COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), flawed and misrepresentative systems used to rank teachers, gender-biased models for natural language processing, and voice interfaces, chatbots, and other systems are discriminating against certain minority dialects. Algorithms that may conceal hidden biases are already routinely used to make vital financial and legal decisions. Proprietary algorithms are used to decide, for instance, who gets a job interview, who gets granted parole, and who gets a loan. This talk focuses on questions like controlling machine-learning algorithms and their biases, the merit of approximation models as a reasonable way to get insight, right of explanation, and how to apply AI within many domains which requires transparency and responsibility such as health care, finance, surveillance, autonomous vehicles, and government. We will briefly cover concepts around algorithmic discrimination, sources of algorithmic bias, measures of discrimination and finally ACM's guidelines for detecting and preventing algorithmic bias. This is an active area of research and this talk manifests tip of the ice-berg; by exposing spectrum of hard questions around algorithmic bias we need to answer if we expect to benefit from advances in algorithmic technology.


"If I had to guess at what our biggest existential threat is, it's probably AI." ~Elon Musk
"AI is likely to be either the best or worst thing ever to happen to humanity." ~Stephen Hawking
"When a few people control a platform with extreme intelligence, it creates dangers in terms of power and control." ~Bill Gates

IS AI A THREAT TO HUMANITY?

# Why Algorithmic Bias

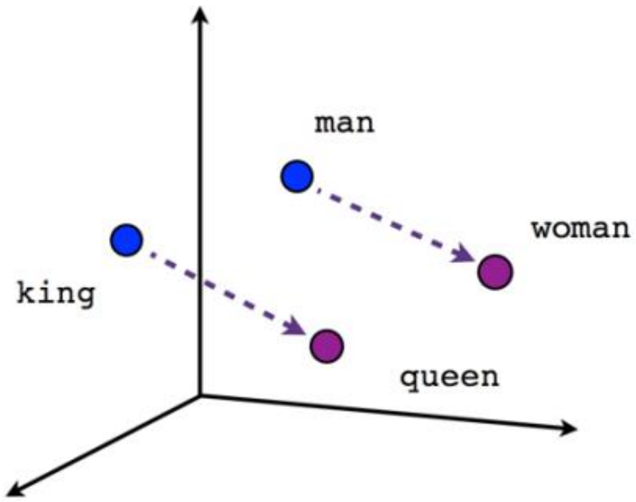
- Optimality i.e. 'Right/Good'= Maximized Utility Function which deduces complex value environments but risks stasis when optimality reached
- Efficiency i.e. All values instrumentalized relative to system goals/function and speed
- Decisional advantage over humans
- Precision calculation advantage over humans
- Reliability Stability advantage over humans
- Readability Informational advantage over humans
- Compressibility - Lossless reduction of informational complexity
- Computational advantage
- Replicability - Economic advantage (force multiplier)
- Invulnerability - Non-biological advantages over humans (physical affective in-vulnerabilities)

**Impact**

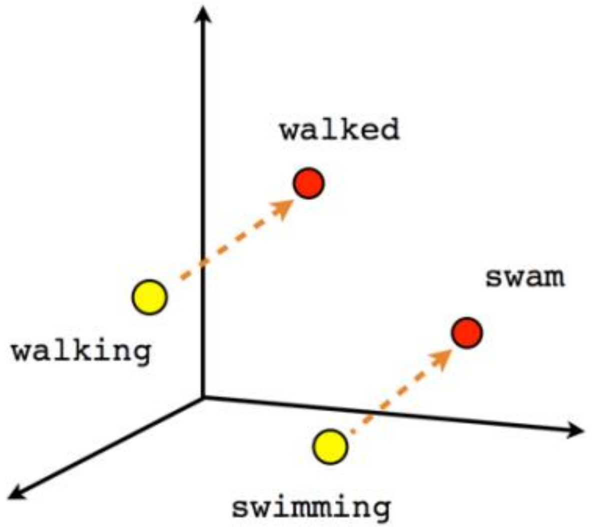
- Commercial influences
- Voting behavior
- Gender discrimination
- Racial and ethnic discrimination
  - Online hate speech
  - Surveillance
- Sexual discrimination

**Obstacles to research**

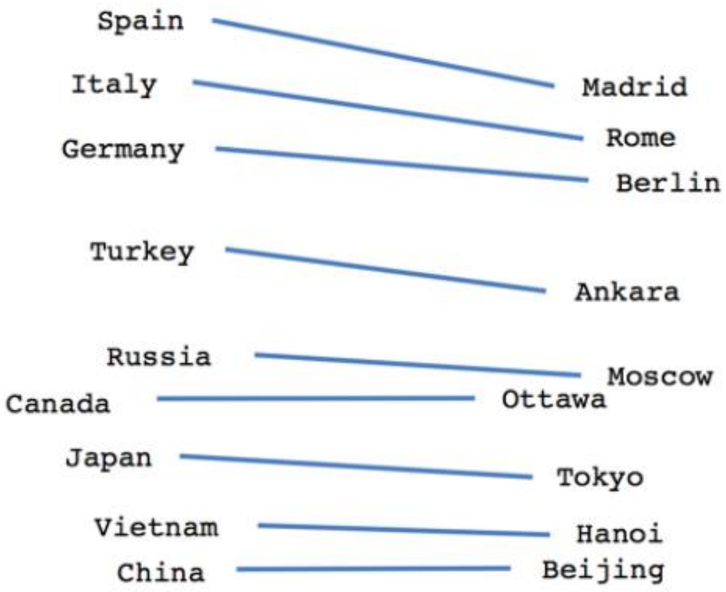
- Lack of transparency
- Complexity
- Lack of data about sensitive categories



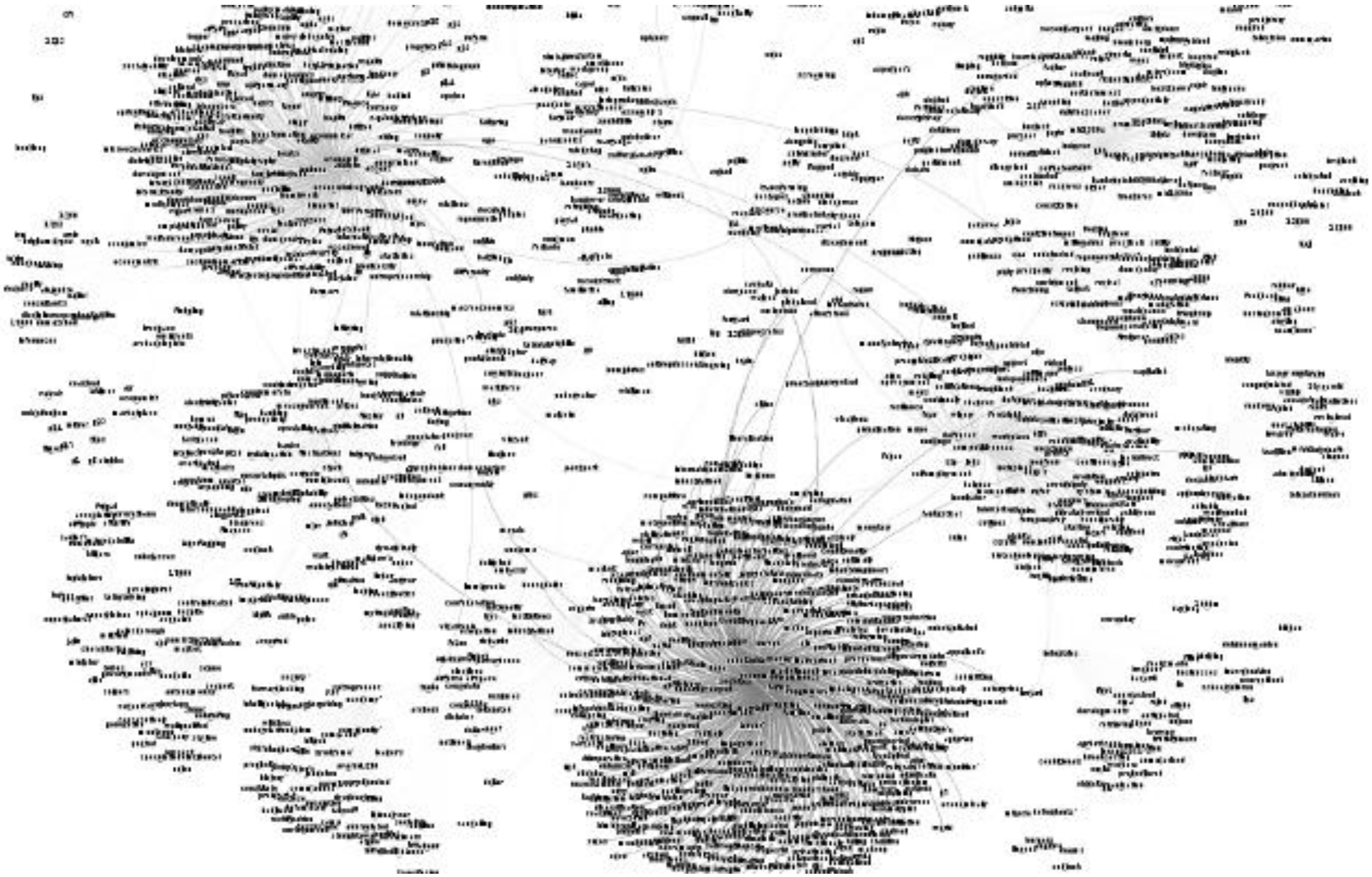

Male-Female




Verb tense




Country-Capital



# Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

*(Submitted on 21 Jul 2016)*

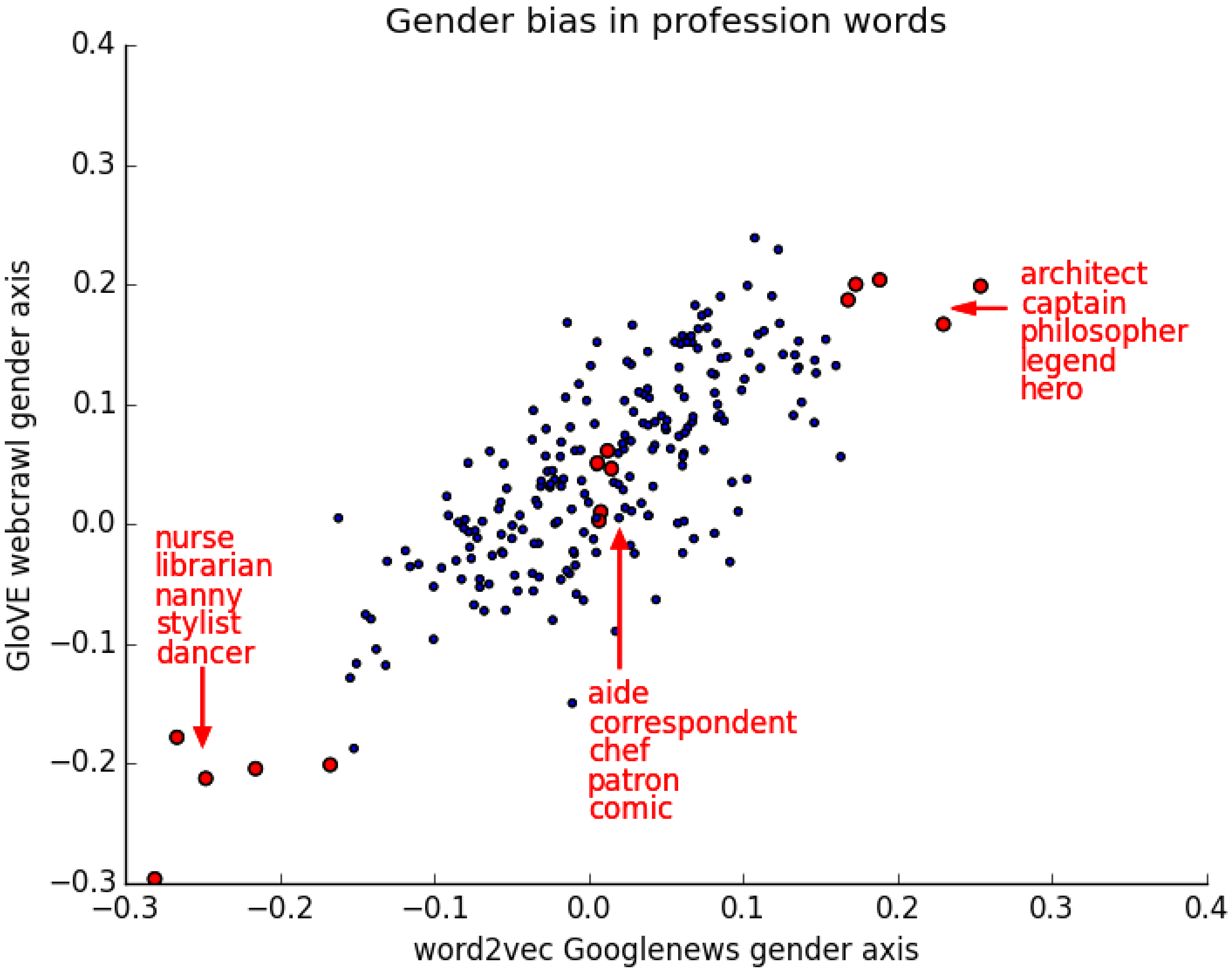
The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.


**Extreme *she* occupations**

| | | |
|---|---|---|
| 1. homemaker | 2. nurse | 3. receptionist |
| 4. librarian | 5. socialite | 6. hairdresser |
| 7. nanny | 8. bookkeeper | 9. stylist |
| 10. housekeeper | 11. interior designer | 12. guidance counselor |

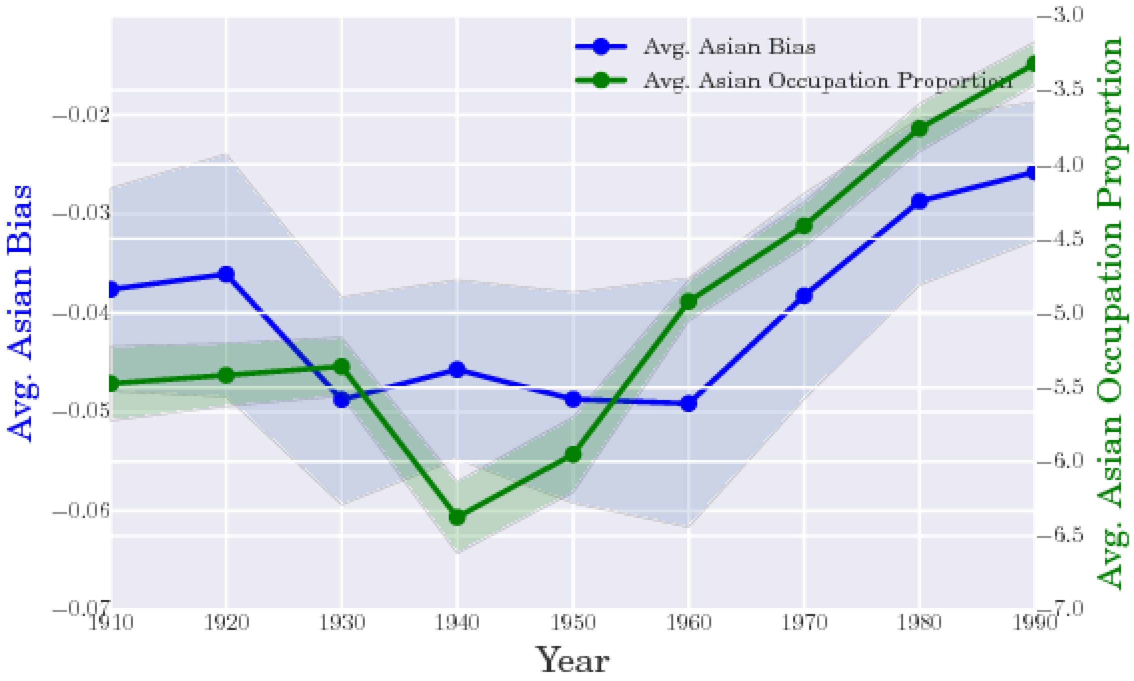
**Extreme *he* occupations**

| | | |
|---|---|---|
| 1. maestro | 2. skipper | 3. protege |
| 4. philosopher | 5. captain | 6. architect |
| 7. financier | 8. warrior | 9. broadcaster |
| 10. magician | 11. figher pilot | 12. boss |


Gender bias in profession words
GloVE webcrawl gender axis
word2vec Googlenews gender axis
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
architect
captain
philosopher
legend
hero
nurse
librarian
nanny
stylist
dancer
aide
correspondent
chef
patron
comic

| Hispanic | Asian | White |
|---|---|---|
| housekeeper | professor | smith |
| mason | official | blacksmith |
| artist | secretary | surveyor |
| janitor | conductor | sheriff |
| dancer | physicist | weaver |
| mechanic | scientist | administrator |
| photographer | chemist | mason |
| baker | tailor | statistician |
| cashier | accountant | clergy |
| driver | engineer | photographer |

) The top ten occupations most closely associated with each




(d) Average ethnic (Asian vs White) bias score over time for oc-
cupations in COHA (blue) vs the average conditional log pro-

## SOFTWARE SCANDALS

Prominent incidents that highlight the effect of algorithmic bias

| | |
|---|---|
| December 2009 | Hewlett-Packard investigates instances of so-called "racist camera software" which had trouble recognizing dark-skinned people |
| March 2015 | A Carnegie Mellon University study determines that some personalized ads from sites such as Google and Facebook are gender-biased |
| July 2015 | A Google algorithm mistakenly captions photos of black people as "Gorillas" |
| March 2016 | Microsoft shuts down AI chatbot Tay after it starts using racist language |
| May 2016 | ProPublica investigation finds that a computer program used to track future criminals in the US is racially biased |
| September 2016 | Machine-learning algorithms used to judge an international beauty contest displays bias against dark-skinned contestants |

PewResearchCenter

NUMBERS, FACTS AND TRENDS SHAPING THE WORLD

FEBRUARY 8, 2017

# Code-Dependent: Pros and Cons of the Algorithm Age

Algorithms are aimed at optimizing everything. They can save lives, make things easier, and conquer chaos. Still, experts worry they can also put too much control in the hands of corporations and governments, perpetuate bias, create filter bubbles, cut choices, creativity, and serendipity, and could result in greater unemployment.

By Lee Rainie and Janna Anderson

**FOR MEDIA OR OTHER INQUIRIES:**

Lee Rainie, Director, Pew Research Internet, Science and Technology Project
Janna Anderson, Director, Elon University's Imagining the Internet Center
Dana Page, Senior Communications Manager
202.419.4372

www.pewresearch.org

MANAGE

# Deep learning models hampered by black box functionality

**A lack of transparency into how deep learning models work is keeping some businesses from embracing them fully, but there are ways around the interpretability problem.**

Ed Burns
Site Editor

Follow:

Deep learning models have a potentially big problem -- a lack of interpretability -- that could keep some enterprises from getting much value from them.

One of the great things about deep learning is that users can essentially just feed data to a neural network, or some other type of learning model, and the model eventually delivers an answer or recommendation. The user doesn't have to understand how or why the model delivers its results; it just does.

But some enterprises are finding that the black box nature of some deep learning models -- where their functionality isn't seen or understood by the user -- isn't quite good enough when it comes to their most important business decisions.

"When you're in a black box, you don't know what's going to happen. You can't have that," said



Related Content

# Addressing the ethical issues of AI is key to effective use

**Enterprises must confront the ethical implications of AI use as they increasingly roll out technology that has the potential to reshape how humans interact with machines.**

Kathleen Walch
Cognilytica

Follow:

Many enterprises are exploring how AI can help move their business forward, save time and money, and provide more value to all their stakeholders. However, most companies are missing the conversation about the ethical issues of AI use and adoption.

Even at this early stage of AI adoption, it's important for enterprises to take ethical and responsible approaches when creating AI systems because the industry is already starting to see backlash against AI implementations that play loose with ethical concerns.

For example, Google recently saw pushback with its Google Duplex release that seems to show AI-enabled systems pretending to be humans. Microsoft saw significant issues with its Tay bot that started going off the rails. And, of course, who can ignore what Elon Musk and others are saying about the use of AI.

Yet enterprises are already starting to pay attention to the ethical issues of AI use. Microsoft, for example, has created the AI and Ethics in Engineering and Research Committee to make sure the company's core values are included in the AI systems it creates.

## How AI systems can be biased

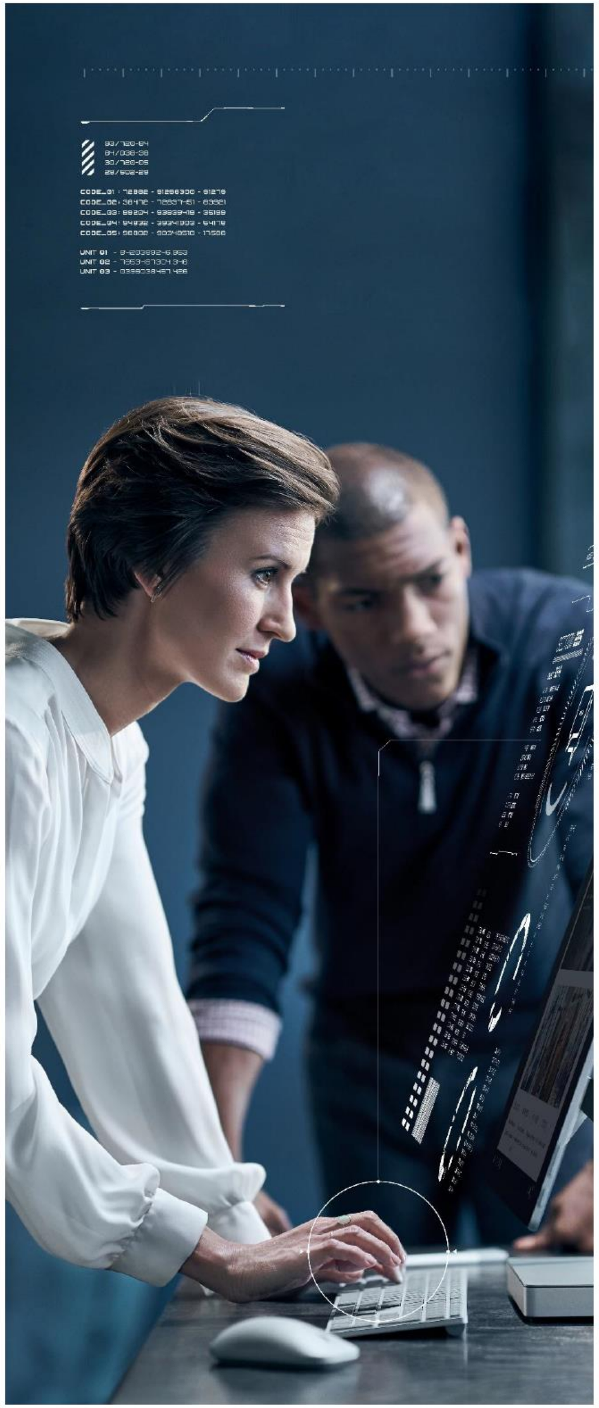


Related Content

Bot security through AI openness
– SearchEnterpriseAI

Australia's chief scientist calls for AI regulations
– ComputerWeekly.com

Expert panel warns developers to beware of AI bias
– SearchEnterpriseAI



# Artificial Intelligence

Microsoft Practice Development Playbook

Microsoft

aka.ms/practiceplaybooks

# Table of Contents


Religion
Gender
Age
Physical Ability
Upbringing
Income
Education
Parental Status
Culture
Family
Appearance
Sexual Orientation
Ethnicity
Marital Status
Race

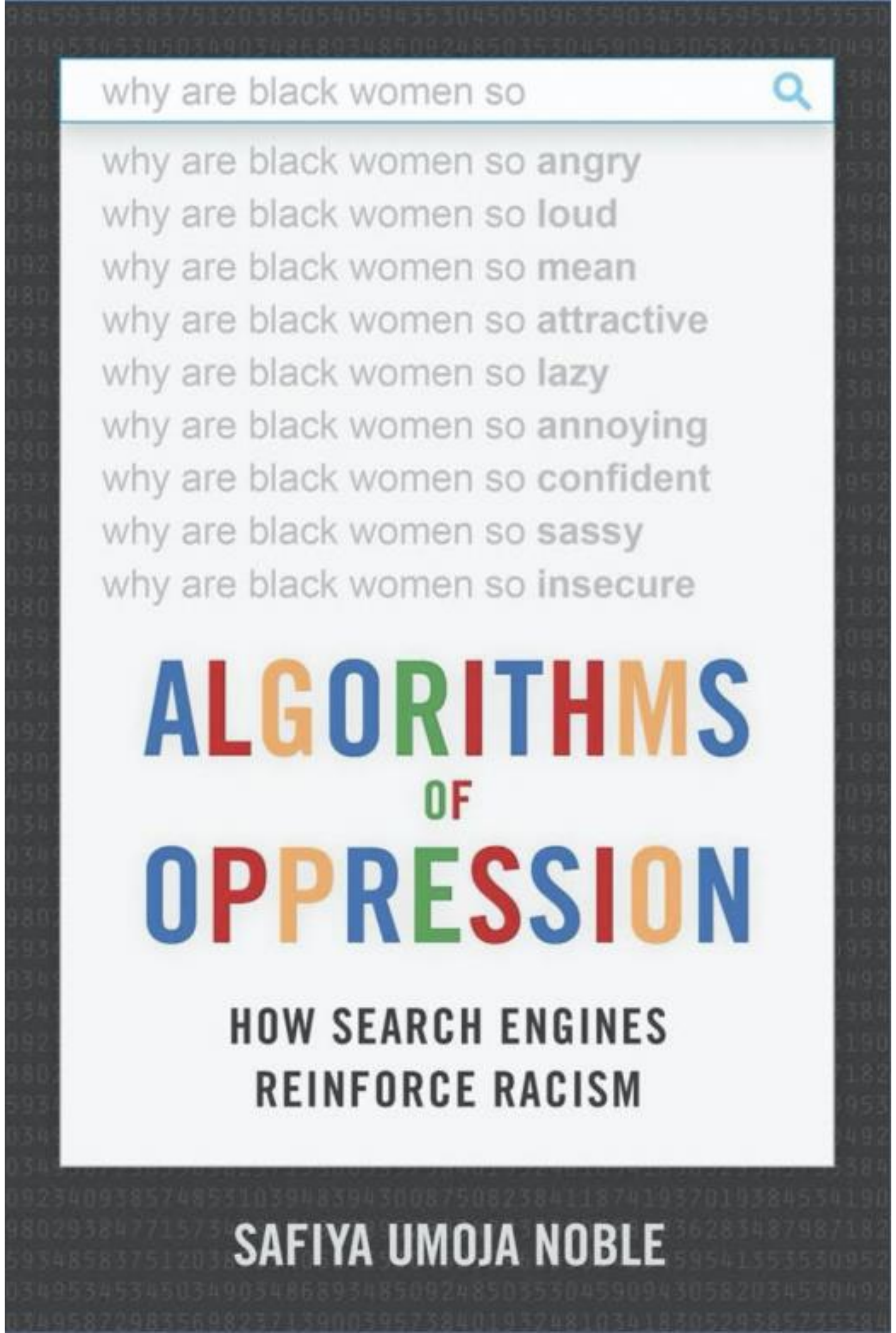
# Implicit Bias is...

**Attitudes, Stereotypes, & Beliefs**
that can affect how we treat others.

Implicit bias is not intentional, but it can still impact how we judge others based on factors, such as:

- **Race**
- **Ability**
- **Gender**
- **Culture**
- **Language**

In early childhood settings, implicit biases can affect how providers perceive and respond to children, which can lead to unfair differences in the use of exclusionary discipline practices, such as suspension and expulsion.


why are black women so
why are black women so angry
why are black women so loud
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
SAFIYA UMOJA NOBLE

# Microsoft's Tay & Twitter: A 24-hour Story


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

# Microsoft's Tay & Twitter: A 24-hour Story


https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref_src=twsrc%5Etfw

# Microsoft's Tay & Twitter: A 24-hour Story


https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref_src=twsrc%5Etfw

# Microsoft's Tay & Twitter: A 24-hour Story


https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref_src=twsrc%5Etfw

# Microsoft's Tay & Twitter: A 24-hour Story


https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref_src=twsrc%5Etfw

# AI & Ethics

Carnegie Mellon created a research center to study the ethics of AI

- http://www.nytimes.com/2016/11/02/technology/new-research-center-to-explore-ethics-of-artificial-intelligence.html

# Popular applications that use data predictive models

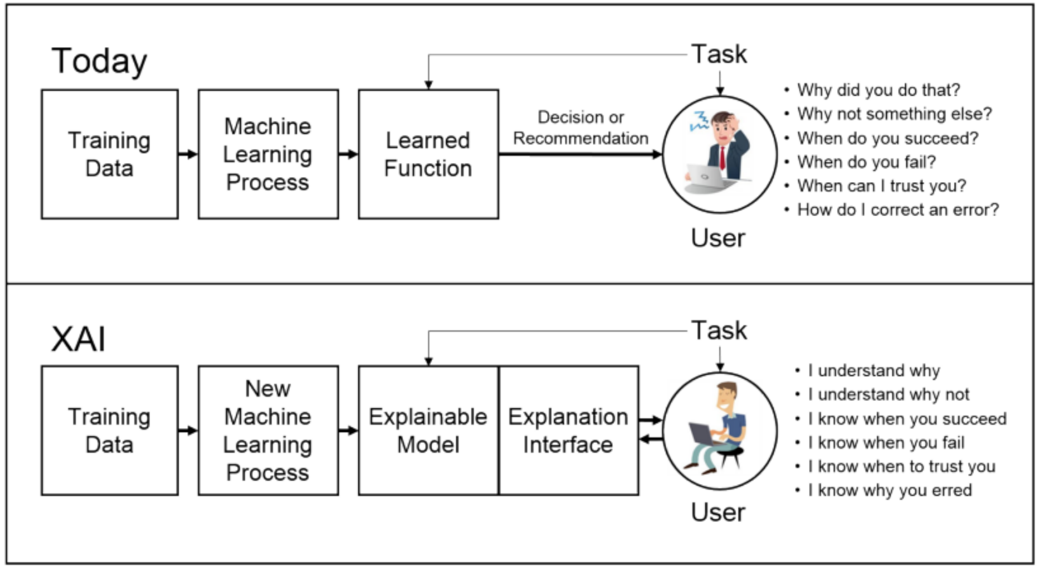
## Typical examples would include

- Product recommendation systems
  - Ex. Amazon, Netflix, etc
- Search tools
  - Ex. Google, Bing, etc
- AI personal assistants
  - Ex. Siri, Alexa, Cortana, etc
- Automobile sector
  - Ex. Autonomous vehicles, Self-parking systems, etc
- Financial Industry
  - Credit risk assessment, fraud detection, portfolio management/recommendation, etc

# Explainable AI (XAI)
# defense advanced research projects agency

**Broad Agency Announcement**
**Explainable Artificial Intelligence (XAI)**
**DARPA-BAA-16-53, August 10, 2016**

**"DARPA is soliciting innovative research proposals in the areas of machine learning and human computer interaction. The goal of Explainable Artificial Intelligence (XAI) is to create a suite of new or modified machine learning techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of Artificial Intelligence (AI) systems. Proposed research should investigate innovative approaches that enable revolutionary advances in science, or systems."**



Goodman, B. and Flaxman, S. 2017. European Union Regulations on Algorithmic Decision Making and a “Right to Explanation”. AI Magazine 38(3): 50-57, Association for the Advancement of Artificial Intelligence.

- **Goes into effect on May 25, 2018**
- **The bulk of the language deals with how data is collected and stored, the regulation contains Article 22: Automated individual decision making, including profiling, potentially prohibiting a wide swath of algorithms currently in use in recommendation systems, credit and insurance risk assessments, computational advertising, and social networks, for example.**
- **Citizens have the right to receive an explanation for algorithmic decisions.**
- **ML depends upon data that has been collected from society, and to the extent that society contains inequality, exclusion, or other traces of discrimination, so too will the data.**

# Ethical artificial intelligence

**IEEE Tech Ethics Program  launched in 2016**

- Who should be held responsible for the harm an application causes by its actions.
- Researchers placed 4 black and white stickers on stop sign. A self-driving car interpreted the sign to be a speed limit sign and sped up. How
- A Microsoft bot named Tay began learning to engage in pleasant and playful conversations on Twitter and within 24 hours was tweeting misogynist and racist comments it picked up from other Twitter users.
- A Facebook research project, in which bots where tasked to learn to negotiate with other bots, the bots developed a language to use to replace English since they deemed it too inefficient.  The project was terminated.
- A researcher trained a commonly used AI ML technique using public Facebook data to identify Homosexual individuals.  Its accuracy was better than a human's and the learning was unsupervised and therefore not understood.

AI

# IBM researchers propose 'factsheets' for AI transparency

KYLE WIGGERS @KYLE_L_WIGGERS AUGUST 22, 2018 6:00 AM

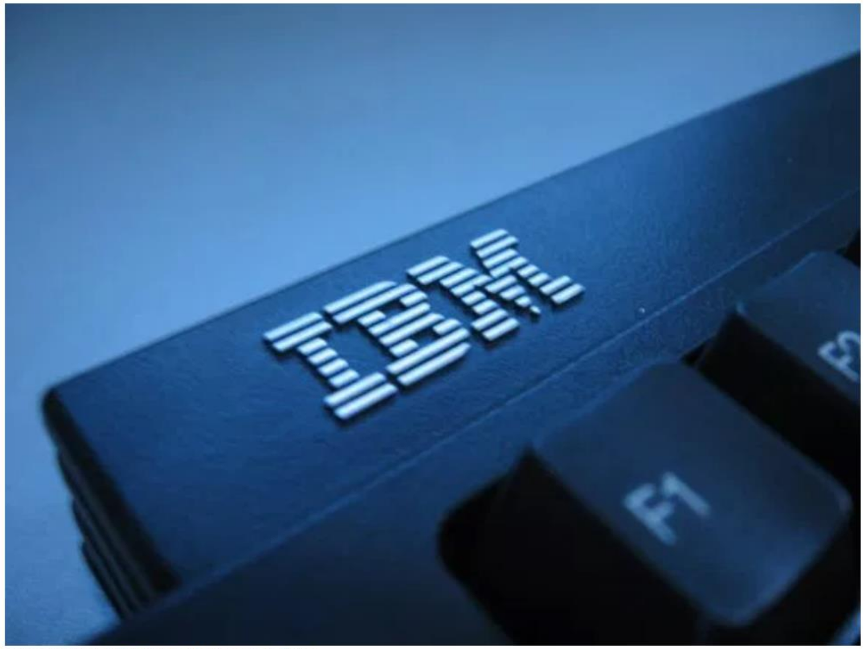



Image Credit: Esteban Maringolo/Flickr

We're at a pivotal moment in the path to mass adoption of artificial intelligence (AI). Google subsidiary DeepMind is leveraging AI to determine how to refer optometry patients. Haven Life is using AI to extend life insurance policies to people who wouldn't traditionally be eligible, such as people with chronic illnesses and non-U.S. citizens. And Google self-driving car spinoff Waymo is tapping it to provide mobility to elderly and disabled people. But despite the good AI is clearly capable of doing, doubts abound over its safety, transparency, and bias.

https://ai.google/principles/

# Artificial Intelligence at Google

## Our Principles

Google aspires to create technologies that solve important problems and help people in their daily lives. We are optimistic about the incredible potential for AI and other advanced technologies to empower people, widely benefit current and future generations, and work for the common good. We believe that these technologies will promote innovation and further our mission to organize the world's information and make it universally accessible and useful.

We recognize that these same technologies also raise important challenges that we need to address clearly, thoughtfully, and affirmatively. These principles set out our commitment to develop technology responsibly and establish specific application areas we will not pursue.

### 2. Avoid creating or reinforcing unfair bias.

AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. We recognize that distinguishing fair from unfair biases is not always simple, and differs across cultures and societies. We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.

### 3. Be built and tested for safety.

We will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm. We will design our AI systems to be appropriately cautious, and seek to develop them in accordance with best practices in AI safety research. In appropriate cases, we will test AI technologies in constrained environments and monitor their operation after deployment.

### 4. Be accountable to people.

We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control.

AJL -ALGORITHMIC JUST | WIT | Playing with AI Fair

Adnan

Secure | https://www.ajlunited.org

Home  Gender Shades  The Coded Gaze  AI, Ain't I A Woman  Fight Bias  Newsletter  GET INVOLVED

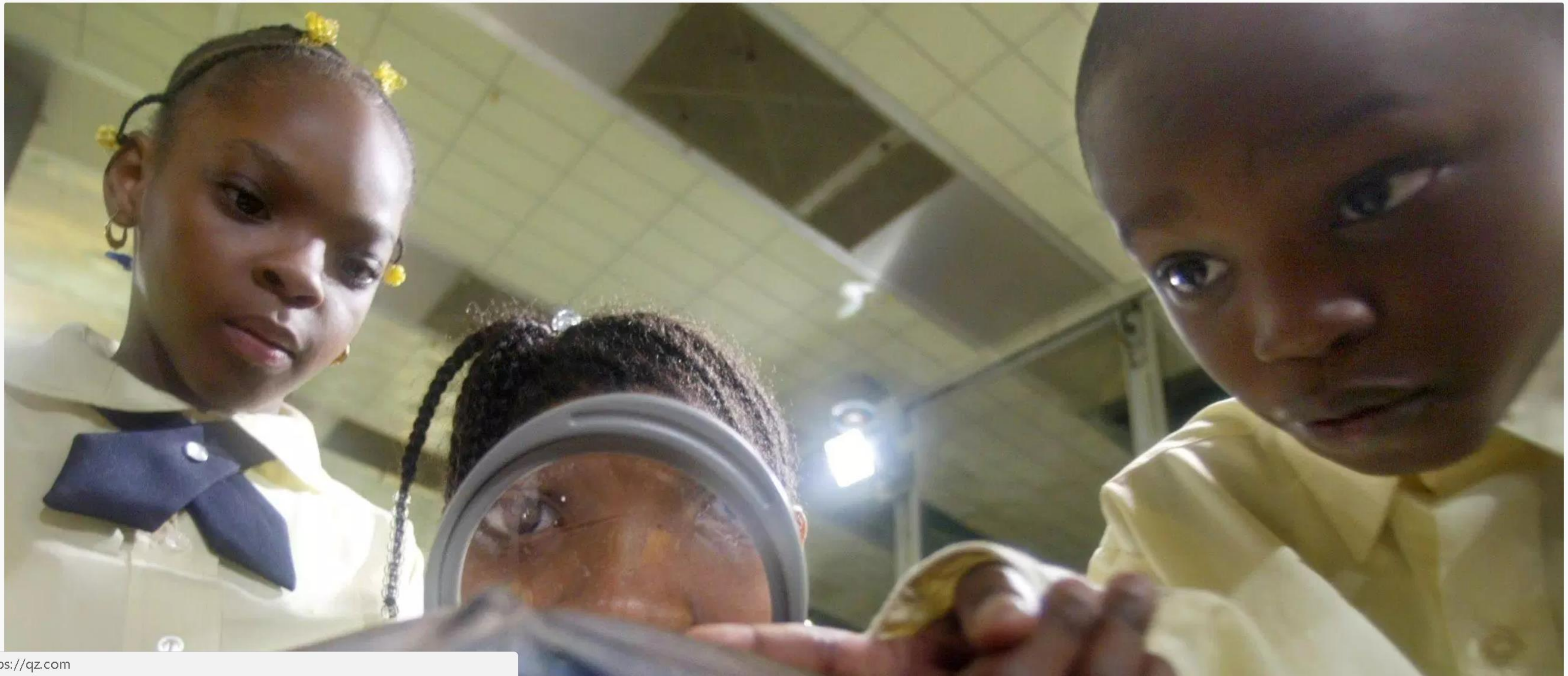
# ALGORITHMIC JUSTICE LEAGUE


LEARN MORE

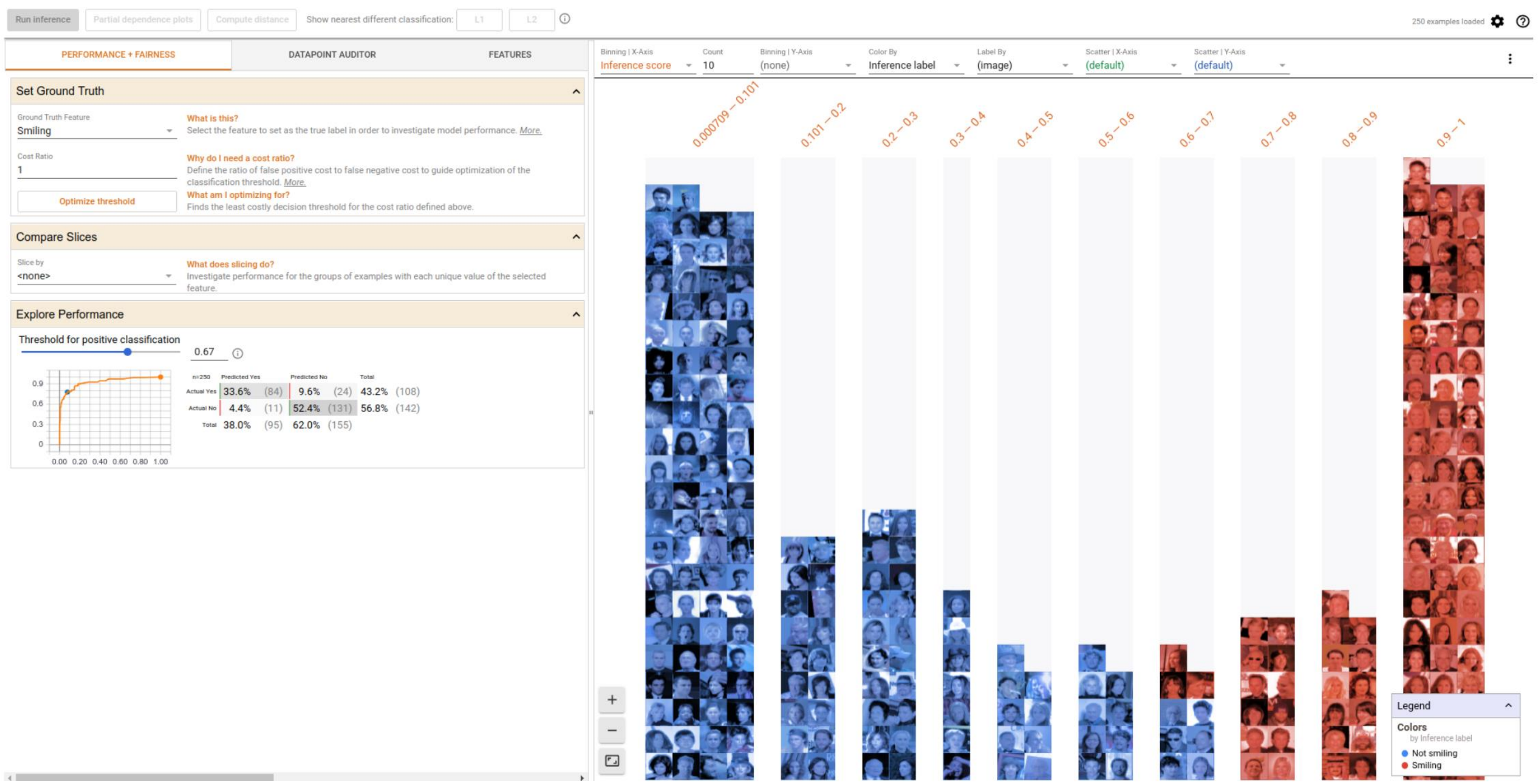
AI, KNOW THYSELF

# Google created a tool to test for biases in AI data

By Dave Gershgorn · September 13, 2018



# What-If Tool



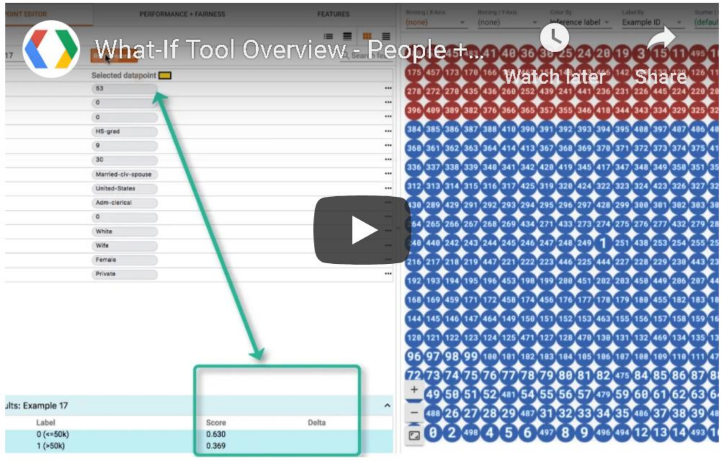

The What-If Tool (WIT) provides an easy-to-use interface for expanding understanding of a black-box ML model. With the plugin, you can perform inference on a large set of examples and immediately visualize the results in a variety of ways. Additionally, examples can be edited manually or programatically and re-run through the model in order to see the results of the changes. It contains tooling for investigating model performance and fairness over subsets of a dataset.

The purpose of the tool is that give people a simple, intuitive, and powerful way to play with a trained ML model on a set of ...ed

# What If...

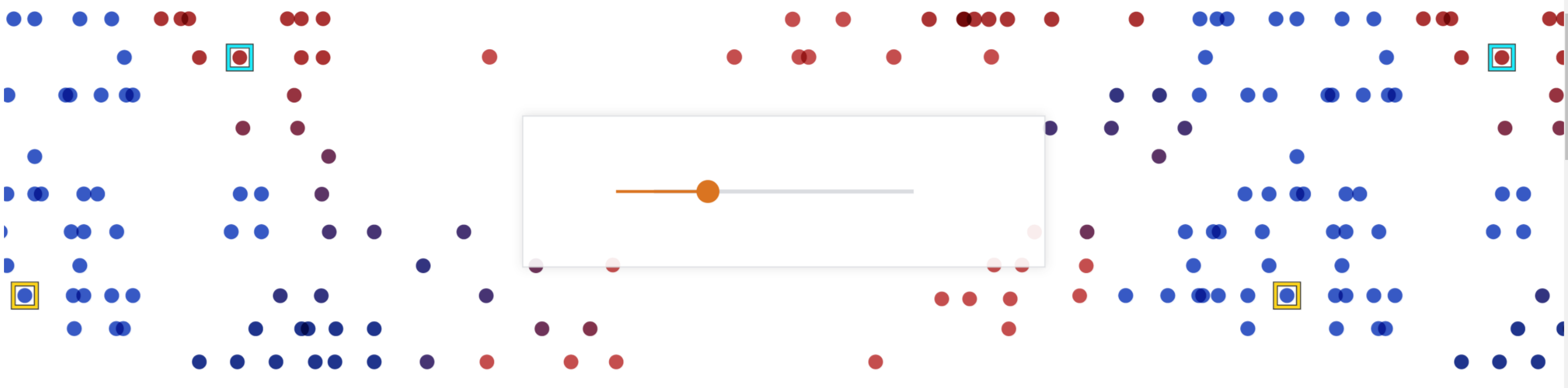
## you could inspect a machine learning model, with no coding required?

Building effective machine learning systems means asking a lot of questions. It's not enough to train a model and walk away. Instead, good practitioners act as detectives, probing to understand their model better.




But answering these kinds of questions isn't easy. Probing "what if" scenarios often means writing custom, one-off code to analyze a specific model. Not only is this process inefficient, it makes it hard for non-programmers to participate in the process of shaping and improving machine learning models. For us, making it easier for a broad set of people to examine, evaluate, and debug machine learning systems is a key concern.

That's why we built the What-If Tool. Built into the open-source TensorBoard web application - a standard part of the TensorFlow platform - the tool allows users to analyze an machine learning model without the need for writing any further code. Given pointers



# Playing with AI Fairness

## Google's new machine learning diagnostic tool lets users try on five different types of fairness

Posted by David Weinberger, writer-in-residence at PAIR

*David is an independent author and currently a writer in residence within Google's People + AI Research initiative. During his residency, he will be*

# AI Fairness 360

The AI Fairness 360 toolkit (AIF360) is an open source software toolkit that can help detect and remove bias in machine learning models.

Get the code

The AI Fairness 360 toolkit (AIF360) is an open source software toolkit that can help detect and remove bias in machine learning models. It enables developers to use state-of-the-art algorithms to regularly check for unwanted biases from entering their machine learning pipeline and to mitigate any biases that are discovered.

AIF360 enables AI developers and data scientists to easily check for biases at multiple points along their machine learning pipeline, using the appropriate bias metric for their circumstances. It also provides a range of state-of-the-art bias mitigation techniques that enable the developer or data scientist to reduce any discovered bias. These bias detection techniques can be deployed automatically to enable an AI development team to perform systematic checking for biases similar to checks for development bugs or security violations in a continuous integration pipeline.


The diagram above represents a simple machine learning pipeline. Bias might exist in the initial training data, in the algorithm that creates the classifier, or in the predictions the classifier makes. The AI Fairness 360 toolkit can measure and mitigate bias in all three stages of the machine learning pipeline.

## GitHub repository

AIF360

| | |
|---|---|
| **Language**<br>Python | **Modified**<br>Aug 22, 2018 |
| **Watchers**<br>35 | **Stars**<br>289 |
| **Contributors**<br>0 | **Issues**<br>2 |
| **Pull requests**<br>0 | **Forks**<br>45 |
| **Branches**<br>0 | **Releases**<br>0 |

# AI Fairness 360 (AIF360 v0.1.1)

build passing

The AI Fairness 360 toolkit is an open-source library to help detect and remove bias in machine learning models. The AI Fairness 360 Python package includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models.

The AI Fairness 360 interactive experience provides a gentle introduction to the concepts and capabilities. The tutorials and other notebooks offer a deeper, data scientist-oriented introduction. The complete API is also available.

Being a comprehensive set of capabilities, it may be confusing to figure out which metrics and algorithms are most appropriate for a given use case. To help, we have created some guidance material that can be consulted.

We have developed the package with extensibility in mind. This library is still in development. We encourage the contribution of your metrics, explainers, and debiasing algorithms.

Get in touch with us on Slack (invitation here)!

## Supported bias mitigation algorithms

- Optimized Preprocessing (Calmon et al., 2017)
- Disparate Impact Remover (Feldman et al., 2015)
- Equalized Odds Postprocessing (Hardt et al., 2016)
- Reweighing (Kamiran and Calders, 2012)
- Reject Option Classification (Kamiran et al., 2012)
- Prejudice Remover Regularizer (Kamishima et al., 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
- Learning Fair Representations (Zemel et al., 2013)
- Adversarial Debiasing (Zhang et al., 2018)
- Meta-Algorithm for Fair Classification (Celis et al.. 2018)

## Supported fairness metrics

- Comprehensive set of group fairness metrics derived from selection rates and error rates
- Comprehensive set of sample distortion metrics
- Generalized Entropy Index (Speicher et al., 2018)

**Intelligent Machines**

# Microsoft is creating an oracle for catching biased AI algorithms

As more people use artificial intelligence, they will need tools that detect unfairness in the underlying algorithms.

by Will Knight May 25, 2018

**Microsoft is building a tool to automatically identify bias in a** range of different AI algorithms. It is the boldest effort yet to automate the detection of unfairness that may creep into machine learning—and it could help businesses make use of AI without inadvertently discriminating against certain people.

Big tech companies are racing to sell off-the-shelf machine-learning technology that can be accessed via the cloud. As more customers make use of these algorithms to automate important judgements and decisions, the issue of bias will become crucial. And since bias can easily creep into machine-learning models, ways to automate the detection of unfairness could become a valuable part of the AI toolkit.

“Things like transparency, intelligibility, and explanation are new

Microsoft is developing a | Inspecting Algorithms fo

Secure | https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

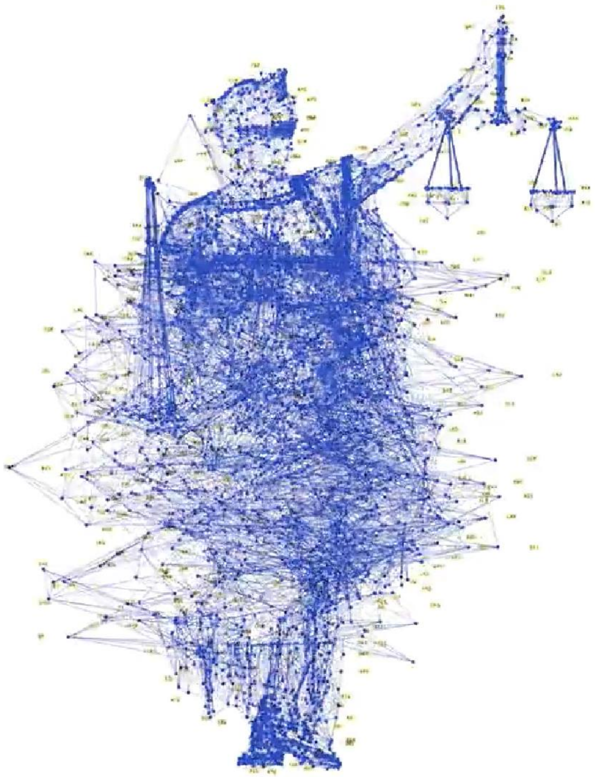
MIT Technology Review

Adnan M. Search

Topics+ The Download Magazine Events More+

**Business Impact**

# Inspecting Algorithms for Bias



PABLO DELCAN

Courts, banks, and other institutions are using automated data analysis systems to make decisions about your life. Let's not leave it up to the algorithm makers to decide whether they're doing it appropriately.

by Matthias Spielkamp June 12, 2017

FAIR SHAKE

# Facebook says it has a tool to detect bias in its artificial intelligence

By Dave Gershgorn • May 3, 2018

# Recidivism prediction Example - Context

Recidivism : “a tendency to relapse into a previous condition or mode of behavior; *especially* : relapse into criminal behavior”

Use of recidivism in judicial system

- To decide the details of probation, every case is assigned a ranking/category. The category depends on the case itself as well as historical data
- Traditional approach
  - ‘One size fits all approach’ - Using basic statistical tools , the probation period was fixed
- Using predictive models
  - The ranking/category is more specific to the case and less generic
  - Better accuracy and automated

# Why discuss about use of ML in recidivism?

- Imagine a scenario in the not too distant scenario.
- You and your friend are caught breaking the same traffic rule – lets say overspeeding
  - and the parameters are same - same speed, traffic zone, county, etc
  - Lets assume, it's the first reported offence for both
- Lets say, the judge sentences
  - Your friend to pay fine $$
  - You to pay fine 2x$$ and x points on Drivers license
- Another assumption, judge is objective and only follows recommendations of state-of-the-art predictive model
- Question now becomes, why the difference in sentencing ?

# Why discuss about use of ML in recidivism?

- Continuing with the same example
- If such instances occur frequently to people of similar backgrounds, neighborhood , community, etc.
- Then, either
  - Model built has loopholes
  - Data has inherent bias
- Most importantly, it becomes imperative to know which features in the data unduly influence outcomes
- If model is too complex or accessible only as service, then how to find out answer to the above question?
  - The authors of this paper suggest methods of auditing black-box models by obscuring features of the data set

# Instances where ML models produced biased results

Users discovered that Google's photo app, which applies automatic labels to pictures in digital photo albums, was classifying images of black people as gorillas

Nikon's camera software, which misread images of Asian people as blinking

Amazon's same-day delivery service was unavailable for ZIP codes in predominantly black neighborhoods. The areas overlooked were remarkably similar to those affected by mortgage redlining in the mid-20th century

When Microsoft released the "millennial" chatbot named Tay in March, it quickly began using racist language and promoting neo-Nazi views on Twitter.

And after Facebook eliminated human editors who had curated "trending" news stories last month, the algorithm immediately promoted fake and vulgar stories on news feeds

It's a growing concern that is affecting the products/services of the best tech companies

# References:


Philip Adler, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing Black-box Models by Obscuring Features. 2016. arXiv:1602.07043

Brandon Smith, Sorelle Friedler
Auditing Deep Neural Networks to Understand Recidivism Predictions
http://thesis.haverford.edu/dspace/handle/10066/18664

Certifying and Removing Disparate Impact talk by Suresh Venkatasubramanian in ACM KDD 15 meet. https://youtu.be/4ds9fBDtMmU

http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/

http://gizmodo.com/5256650/camera-misses-the-mark-on-racial-sensitivity

http://www.wired.com/2009/12/hp-notebooks-racist/

http://www.bloomberg.com/graphics/2016-amazon-same-day/

https://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay

https://www.theguardian.com/technology/2016/may/12/facebook-trending-news-leaked-documents-editor-guidelines

Supplementary reading:

http://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html?_r=1
https://www.ncjrs.gov/pdffiles1/nij/240696.pdf

# Q&A