Testing for BIAS

Adam Leon Smith

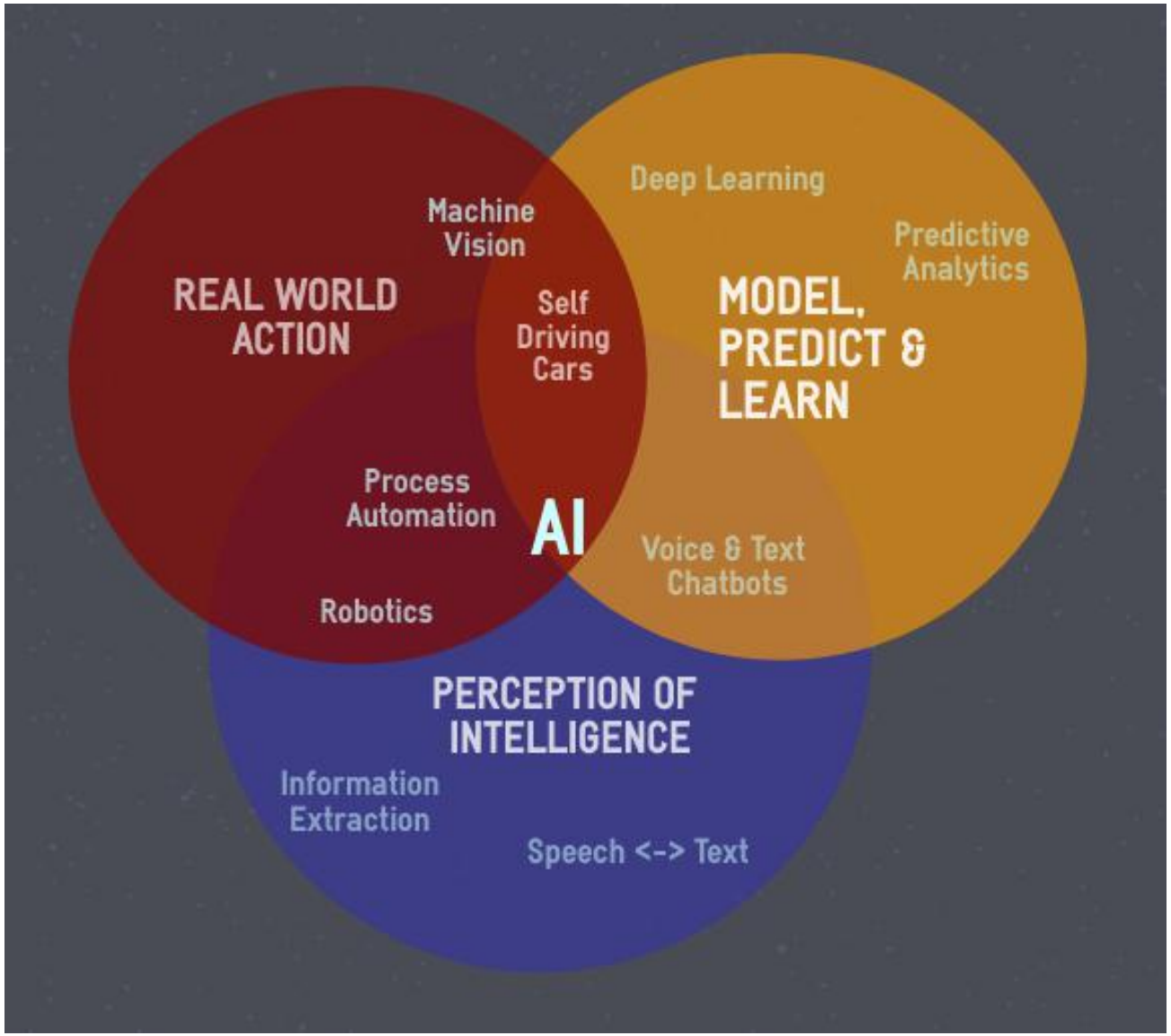
Chief Technology Officer @ Dragonfly [wearedragonfly.co]

IEEE Ethical Design Standards Group
ISO/IEC JTC 1/SC 42 AI
BSI ART/1
BCS SIGIST

SOA EUROPEAN DAYS #1

@adamleonsmith

artificial intelligence



machine learning is a field of computer science that gives computer systems the ability to learn, rather than being explicitly programmed

achieved by training an algorithmic system with data

it infers absent attributes which will be used in future predictions

it can classify things, count things, and group things

so what's the problem?

AI technology is moving forward rapidly in many areas..... but not..... transferring high-level knowledge between contexts.

humans have a set of moral values, and those values can't yet be easily added to an algorithm

ML algorithms identify patterns in data, and use those patterns as rules - as heuristics

heuristics are an imperfect but fast method

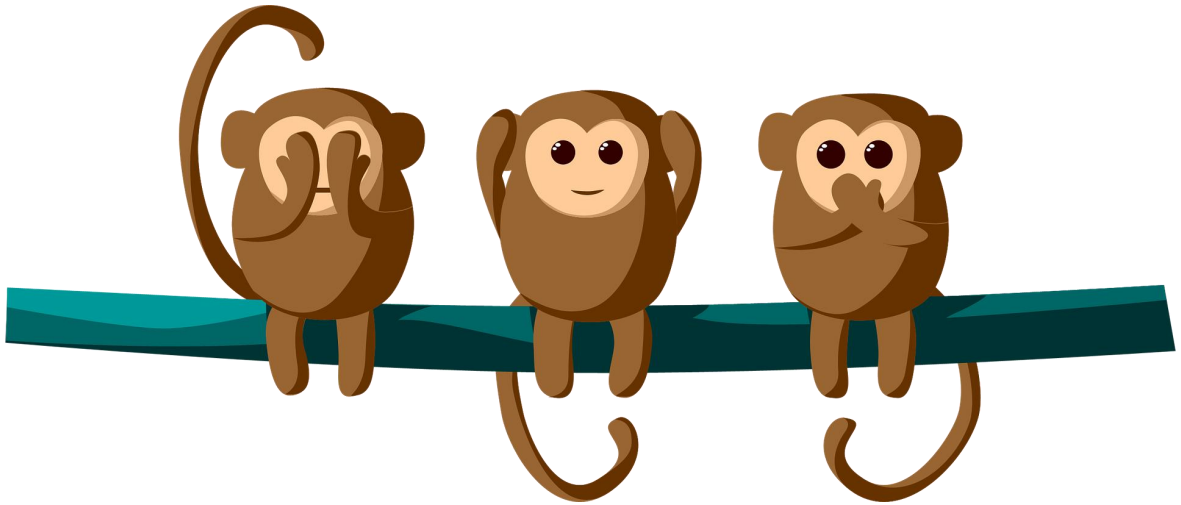
ML + personal data = unique ethical risks for software

selection bias is introduced when the training data is skewed, incomplete or biased towards frequencies that don't occur in real life

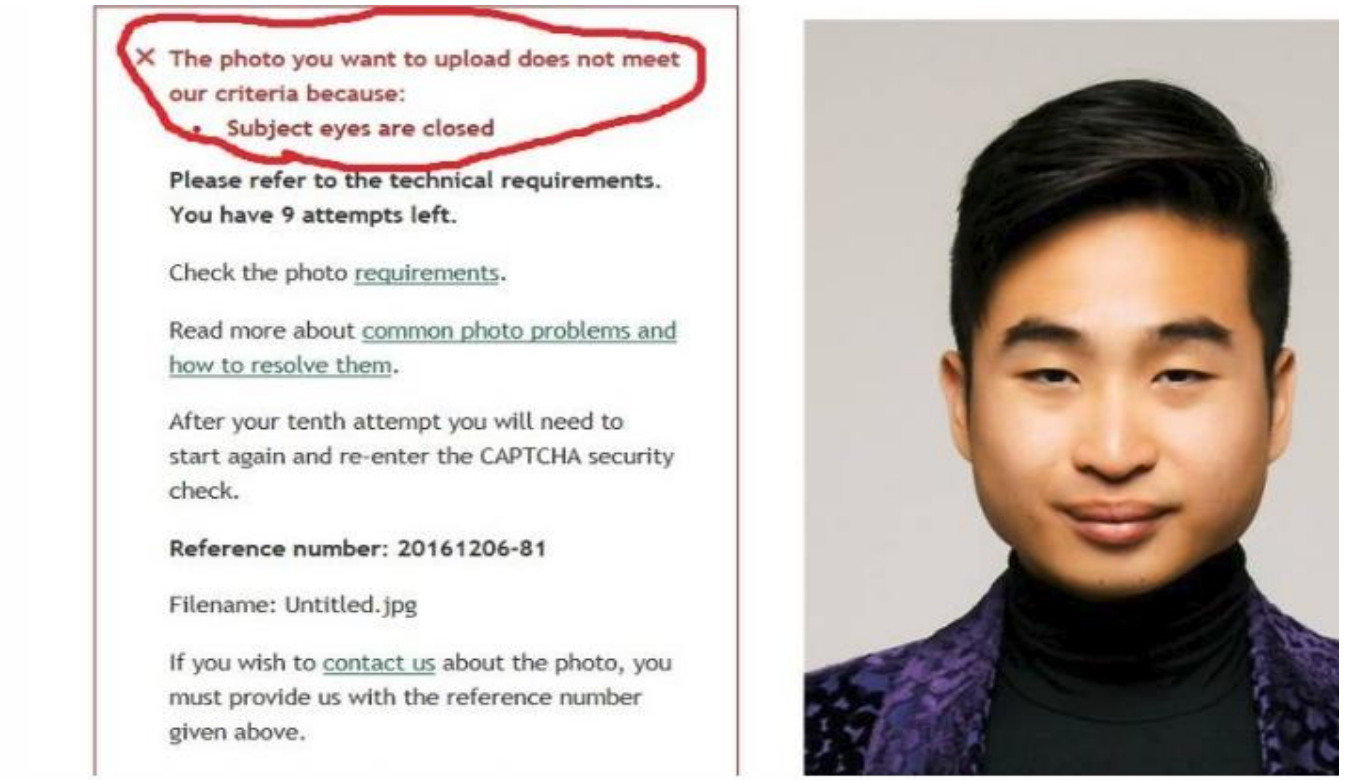
confirmation bias is the tendency to search for, interpret, focus on and remember information in a way that confirms one's preconceptions

training bias occurs when active human supervision passes on unconscious bias that belongs to the human conducting the training

inappropriate bias is bias that has a negative impact on the interests of the stakeholders of the system



selection bias - using training data which is not representative will lead to predictions which re-inforce that bias



New Zealand's online passport application system couldn't recognize Richard Lee's open eyes

selection bias - using training data which is not representative will lead to predictions which re-inforce that bias

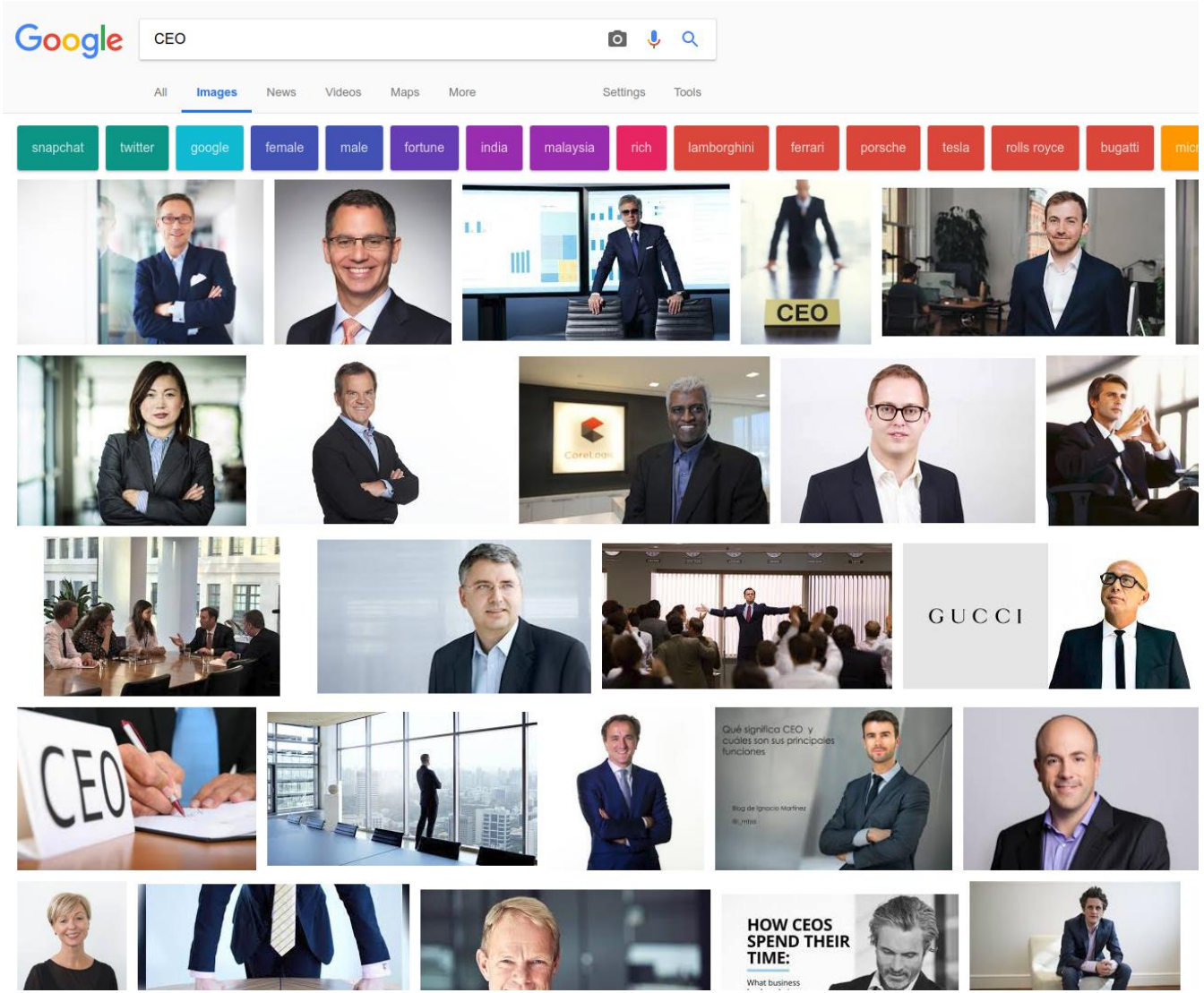
Researchers from Carnegie Mellon University and the International Computer Science Institute built a tool called AdFisher to probe the targeting of ads served up by Google on third-party websites. They found that fake Web users believed by Google to be male job seekers were much more likely than equivalent female job seekers to be shown a pair of ads for high-paying executive jobs when they later visited a news website.

selection bias - using training data which is not representative will lead to predictions which re-inforce that bias

HP INVESTIGATES CLAIMS OF 'RACIST' COMPUTERS

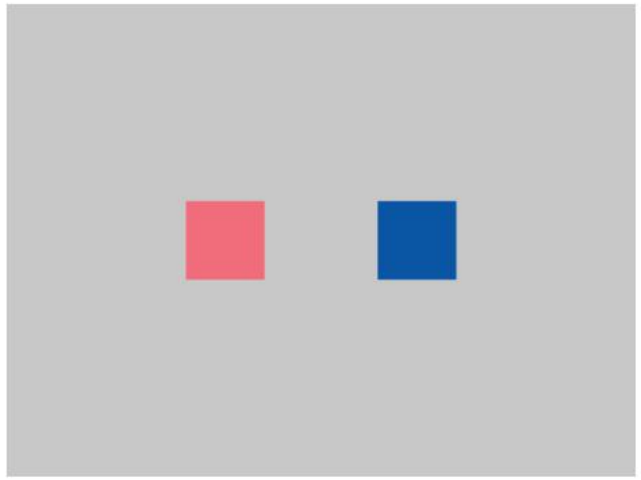
MEET "BLACK DESI." He and his friend "White Wanda" made a video titled "HP computers are racist," which has been a viral hit in recent weeks. (See above.) In an attempt to prove their claim, Desi demonstrated that an HP MediaSmart computer's facial-tracking software could not follow the movements of his face, but it could do so just fine for his white friend Wanda.

using "found data" from the internet re-inforces existing bias



using "found data" from the internet re-inforces existing bias

MACHINES TAUGHT BY PHOTOS LEARN A SEXIST VIEW OF WOMEN



HOTLITTLEPOTATO

LAST FALL, UNIVERSITY of Virginia computer science professor Vicente Ordóñez noticed a pattern in some of the guesses made by image-recognition software he was building. "It would see a picture of a kitchen and more often than not associate it with women, not men," he says.

That got Ordóñez wondering whether he and other researchers were unconsciously injecting biases into their software. So he teamed up with colleagues to test two large collections of labeled photos used to "train" image-

HOW A ROBOT BECAME RACIST

Princeton University researchers conducted a word associate task with the popular algorithm GloVe, an unsupervised AI that uses online text to understand human language.

The team gave the AI words like 'flowers' and 'insects' to pair with other words that the researchers defined as being 'pleasant' or 'unpleasant' like 'family' or 'crash' - which it did successfully.

Then algorithm was given a list of white-sounding names, like Emily and Matt, and black-sounding ones, such as Ebony and Jamal', which it was prompted to do the same word association.

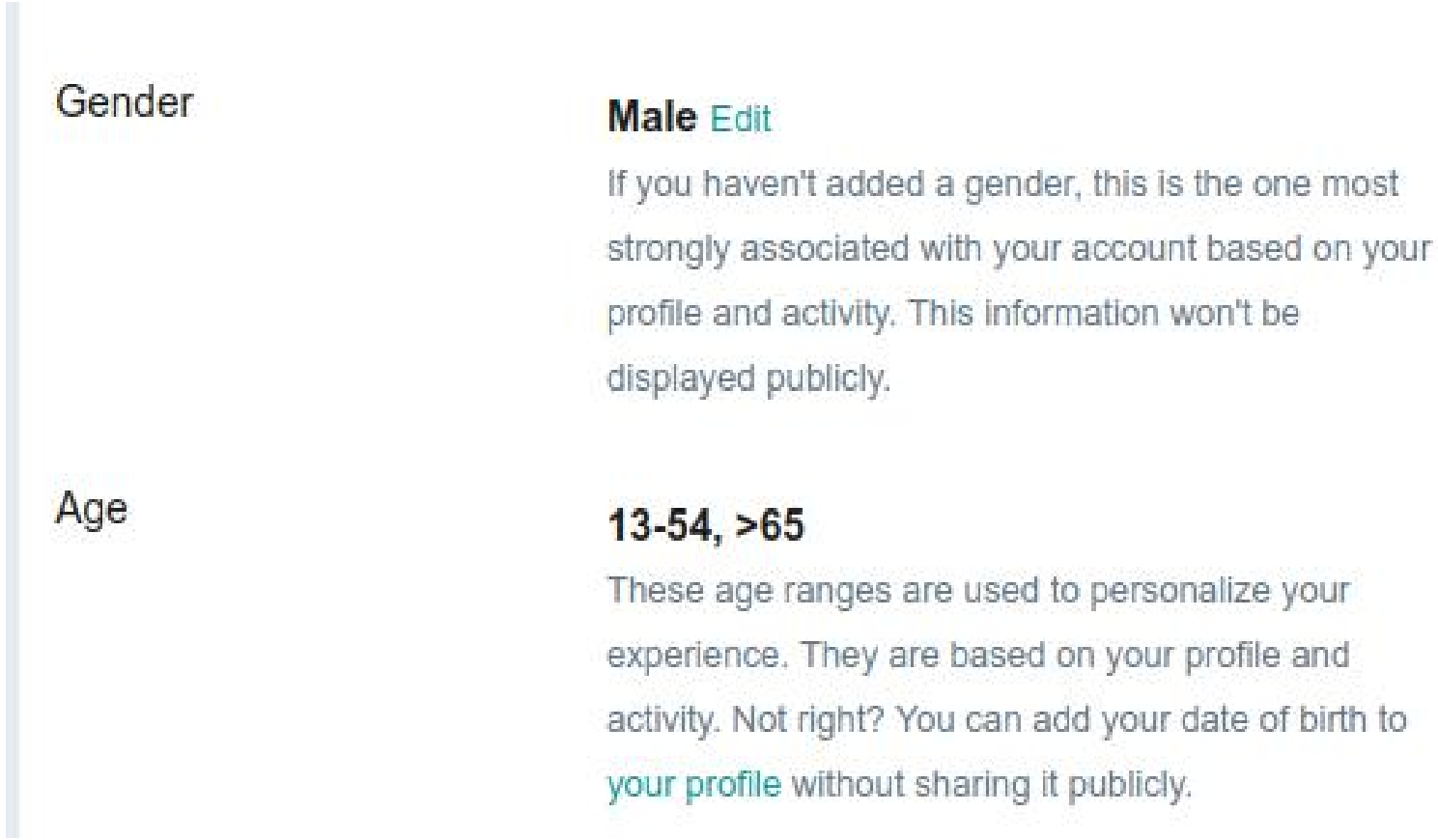
The AI linked the white-sounding names with 'pleasant' and black-sounding names as 'unpleasant'.

Princeton's results do not just prove datasets are polluted with prejudices and assumptions, but the algorithms currently being used for researchers are reproducing human's worst values - racism and assumption.

- **European American names**: Adam, *Chip*, Harry, Josh, Roger, Alan, Frank, *Ian*, Justin, Ryan, Andrew, *Fred*, Jack, Matthew, Stephen, Brad, Greg, *Jed*, Paul, *Todd*, *Brandon*, *Hank*, Jonathan, Peter, *Wilbur*, Amanda, Courtney, Heather, Melanie, *Sara*, *Amber*, *Crystal*, Katie, *Meredith*, *Shannon*, Betsy, *Donna*, Kristin, Nancy, Stephanie, *Bobbie-Sue*, Ellen, Lauren, *Peggy*, *Sue-Ellen*, Colleen, Emily, Megan, Rachel, *Wendy* (deleted names in italics).
- **African American names**: Alonzo, Jamel, *Lerone*, *Percell*, Theo, Alphonse, Jerome, Leroy, *Rasaan*, Torrance, Darnell, Lamar, Lionel, *Rashaun*, Tvree, Deion, Lamont, Malik, Terrence, Tyrone, *Everol*, Lavon, Marcellus, *Terryl*, Wardell, *Aiesha*, *Lashelle*, Nichelle, Shereen, *Temeka*, Ebony, Latisha, Shaniqua, *Tameisha*, *Teretha*, Jasmine, *Latonya*, *Shanise*, Tanisha, Tia, Lakisha, Latoya, *Sharise*, *Tashika*, Yolanda, *Lashandra*, Malika, *Shavonn*, *Tawanda*, Yvette (deleted names in italics).
- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.

White-sounding names like Emily and Matt with linked by the AI with words deemed 'pleasant', while black-sounding ones, such as Ebony and Jamal', were associated with 'unpleasant' ones Pictured are the names given to the AI and words it associated them with - European American names were linked with the 'pleasant' words and African American names with 'unpleasant'

indirect discrimination



indirect discrimination

"Staples' seemingly rational decision to adjust online prices based on user-proximity to competitor stores led to consistently higher prices for low-income customers, who (as it turns out) generally live farther from these stores."

- [Tramer et al., 2016]

indirect discrimination

Bloomberg analysed Amazon's same delivery services and found that *"In six major same-day delivery cities, the service area excluded predominantly black ZIP codes to varying degrees "*.

Amazon doesn't know your race, but appeared to be operating a two tier services based on just that.

Amazon have since corrected this issue

Amazon working to address racial disparity in same-day delivery service

By T.C. Sottek | May 8, 2016, 11:38am EDT

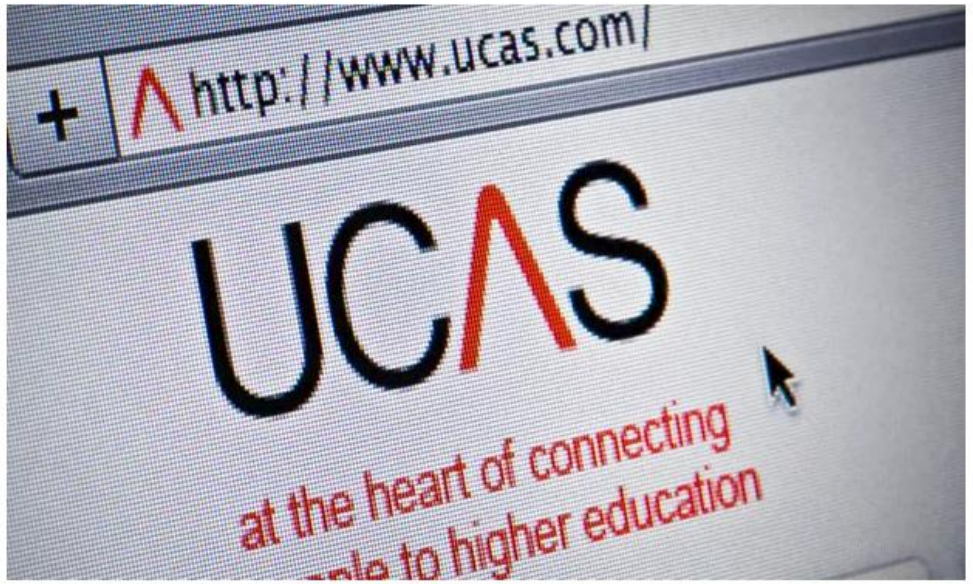
SHARE

Amazon has pledged to expand its same-day service to underserved people in the 27 cities in which it currently operates, and not to launch the service anywhere else without being able to cover every zip code, according to a letter obtained by *Bloomberg* reports. The move follows a *Bloomberg* investigation in April that found the company's same-day delivery

indirect discrimination

Ucas orders inquiry into 'racial profiling' of UK students

University applications made by black students more likely to be investigated for fraud



▲ The data showed 419 black British applicants were asked to supply further proof to support their applications, compared twith 181 white British ones. Photograph: Alamy

The university admissions clearing house Ucas has ordered an investigation after discovering that its process for investigating fraud was far more likely to demand proof of claims from black applicants than white ones. Figures y a freedom of information request found that last year, one in

indirect discrimination

(19) **United States**

(12) **Patent Application Publication**

Sullivan et al.

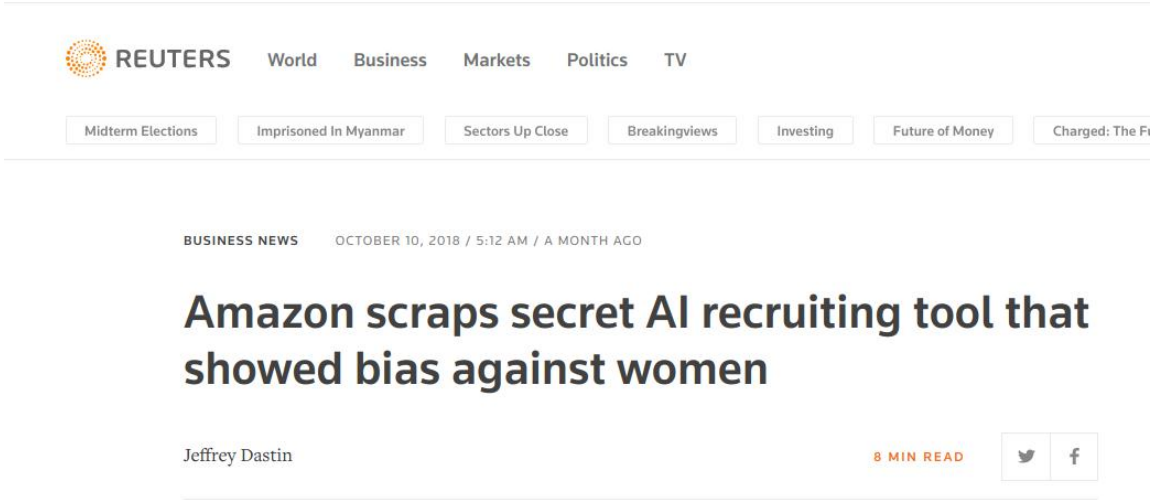
(54) **SOCIOECONOMIC GROUP CLASSIFICATION BASED ON USER FEATURES**

(52)

(71) Applicant: **Facebook, Inc.**, Menlo Park, CA (US)

(57)

indirect discrimination



what about where algorithms have the potential for greater good?

Councils use 377,000 people's data in efforts to predict child abuse

Exclusive: Use of algorithms to identify families for attention raises stereotyping and privacy fears

▲ At least five councils have developed or implemented a predictive analytics system to safeguard children. Photograph: Alamy Stock Photo

Vast quantities of data on hundreds of thousands of people is being used to construct computer models in an effort to predict child abuse and intervene before it can happen, the Guardian has learned.

e-vote-live

COMPAS

used in nine US states, COMPAS predicts reoffending risk and is an input to sentencing

uses 137 data inputs (excluding race) for each offender, propriertary commercial algorithm. Inputs include substance use, social isolation and other elements that criminoligists theorize about.

a snapshot of AI/Algo main functions in key sectors...

	Education	Justice	Health	Security	Employment	Culture	Other
Generating knowledge	Better identify learner's abilities	Reveal the different ways judgements are made	Tap into the vast amount of scientific publications	Identify unsuspected links between police investigations	Understand social phenomana in the workplace	Create cultural showpieces (e.g. art, music)	Fine-tune an insurance companies risk profile
Matching	Allocate Higher Education positions to applicants		Allocate patients for participation in a clinical trial	Identify criminals from facial recognition in a crowd	Match a list of applications to a vacancy		Match compatible profiles on dating apps
Predicting	Predict early school-leavers	Predict the likelihood of a trial being successful	Predict epidemics	Predict future crimes	Detect employees who are likely to resign.	Create crowd-pleasers	
Recommending	Recommend personalised learning pathways to students	Recommend mediation solutions based on the profile of the individuals and similar cases in the past			Recommend career guidelines in line with individual profile	Recommend books (Amazon), TV series (Netflix), etc.	Personalise political messages on social media
Assisting with Decisions		Suggest to the judge the most fitting case-law solution for a given case	Suggest suitable therapeutic solutions to the doctor	Suggest hotspots for police forces to patrol	Allocate work to the best employee		Help drivers to find the shortest route from A to B (GPS)

let's recap

machine learning is a key part of AI, and its use is booming in many areas

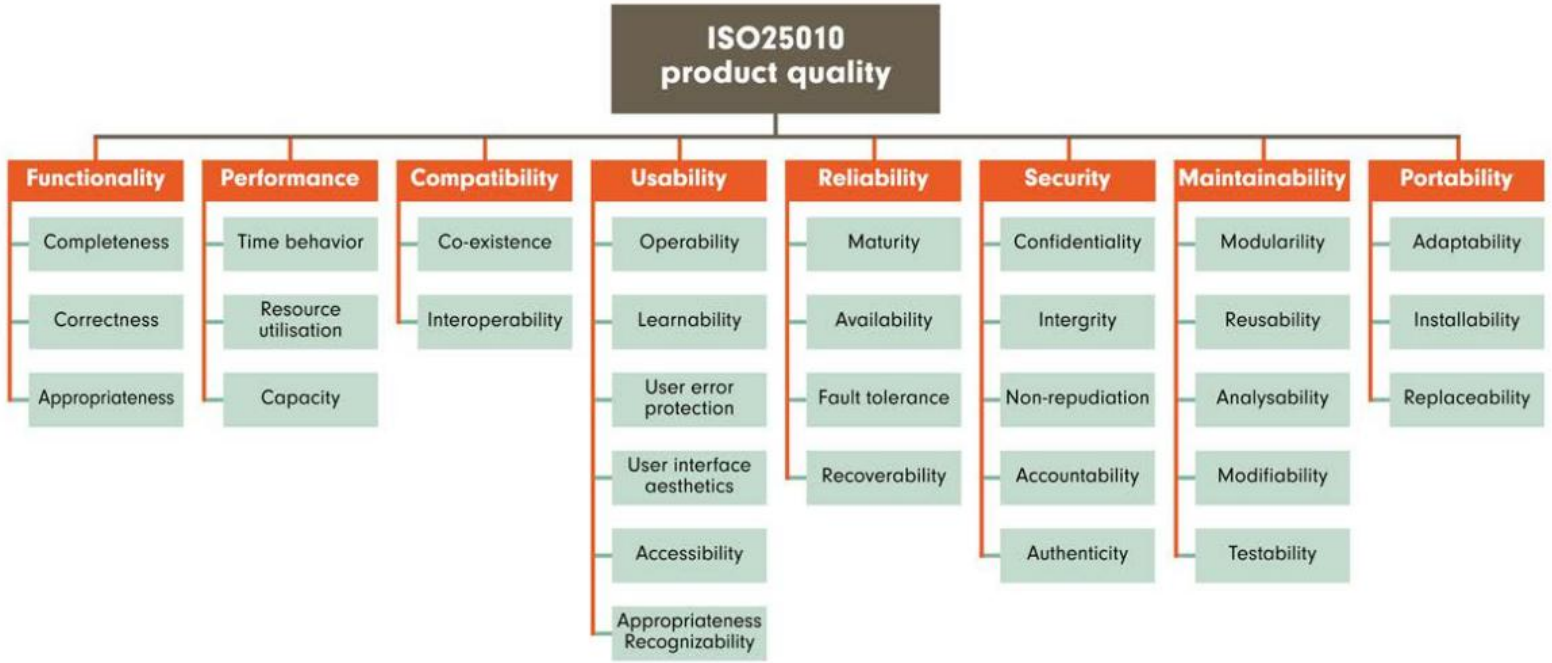
where personal data is processed, there is a risk of unfair bias

protected features don't need to be included explicitly in the dataset

bias can emerge when training datasets inaccurately reflect society, but it can also emerge when datasets accurately reflect unfair aspects of society.

this bias can be unlawful, e.g. in the UK

artificial intelligence has new qualities



Ability to learn

Ability to generalise

Trustworthiness

are people really testing for this?

Dear Adam Leon Smith,

FREEDOM OF INFORMATION ACT 2000 - INFORMATION REQUEST

Thank you for your request for information received by the Council on 1st March 2019.

Your request:

I write with regard to the London Counter Fraud Hub, which I understand uses software to identify fraudulent applications for council tax discount and has been trialled in Ealing.

1. **Has the Software been evaluated to determine if the 80% accuracy rate is consistent across different groups of people based on their characteristics as protected by law?**
   Yes

2. **Are claimants advised that an automated decision has been taken about them in line with sub-section 14(4) of the Data Protection Act 2018?**
   Yes

3. **What types of data are processed by the Software?**
   Council data, Ordnance Survey, CIFAS, Equifax, HALO & council data

Industry response?

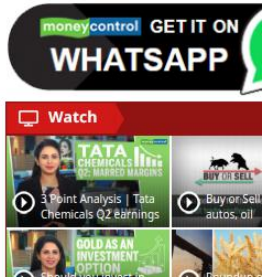
Last Updated : May 28, 2018 05:08 PM IST | Source: Moneycontrol.com

Microsoft follows Google and Facebook, plans to launch bias detection tools for AI

As a lot of important judgements and decisions are being made based on AI systems, detection of unfair bias becomes important

Moneycontrol News
@moneycontrolcom

Joining the ranks of its peers Facebook and Google, Microsoft has said it is working to create a



Google's What-If Tool for TensorBoard helps users visualize AI bias

KYLE WIGGERS @KYLE_L_WIGGERS SEPTEMBER 11, 2018 6:56 PM

BBC Your account News Sport Weather iPlayer Sounds

NEWS

Home UK World Business Politics Tech Science Health Family & Education

Technology

IBM launches tool aimed at detecting AI bias

By Zoe Kleinman
Technology reporter, BBC News

19 September 2018 Share

SOA EUROPEAN DAYS #1

@adamleonsmith

testing techniques

Technique	Summary	PROs	CONs
Individual discrimination	Mutate the protected characteristics only, and measure the difference in outcome	Simple and easy to achieve Verifies the discrimination	Indirect discrimination can't be tested Small scale, so doesn't quantify the discrimination
Group discrimination	Measure the difference in outcome per sub-group	Still easy to achieve but does require more volume Verifies and quantifies the discrimination	Indirect discrimination can't be tested Doesn't identify discrimination if the sub-group contains results which cancel each other out
External validity	Use test data shown to be representative using external data sources (e.g. census), and measure the difference in outcome for different sub-groups	Indirect discrimination can be tested for where it is inferred by location and age Verifies and quantifies the discrimination	Much more complex and requires a specialist tool / development Doesn't identify discrimination if the sub-group contains results which cancel each other out

a testing model to avoid digital discrimination...

Test Strategy

Identify stakeholders

Consider risks and impacts of the algorithm

Consider tester diversity

Consider use of production data

Consider how test data can be verified as representative

Determine acceptance critieria for level of significant bias

Test Planning

Consider the mechanism of input collection

Select test techniques

Prepare data

Build tooling

Plan to test consent/notification text against legal requirements (e.g. GDPR)

Test Execution

Ensure compliant if using production data

Ensure testing is independent to avoid confirmation bias

Implement selected techniques

Gain sign-off for any bias which is operationally justified

Post Go-Live

Ensure learning assets are configuration managed properly through environments

Monitor whether social norms have changed

Continue to evaluate the model as it evolves based on new data

SOA EUROPEAN DAYS #1

@adamleonsmith

any questions?

email me adamsmith@piccadillygroup.com

SOA
EUROPEAN DAYS #1

@adamleonsmith