Artificial Intelligence

Steve James

Ethics and Societal Impact

Disclaimer

- This is a huge subject, spanning many disciplines and addressing many real different problems
- This is HARD!
- I am not an expert

Facial feature discovery for ethnicity recognition

January 2019 · Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(11)
DOI:10.1002/widm.1278

Authors:

Cunrui Wang
Dalian Nationalities University

Qingling Zhang
Northeastern Univers

Yu Liu

ELSEVIER

Image and Vision Computing
Volume 121, May 2022, 104404

Intelligent deep learning based ethnicity recognition and classification using facial images

Geetha [b], S. Neelakandan [c], Aditya Kumar Singh Pundir [d],
mar [f]

Backgrounder

China's Repression of Uyghurs in Xinjiang

More than a million Muslims have been arbitrarily detained in China's Xinjiang region. The reeducation camps are just one part of the government's crackdown on Uyghurs.

A Uyghur man works at his
Frayer/Getty Images

WRITTEN BY
Lindsay Maizland

UPDATED
Last updated September 22, 2022 11:30 am (EST)

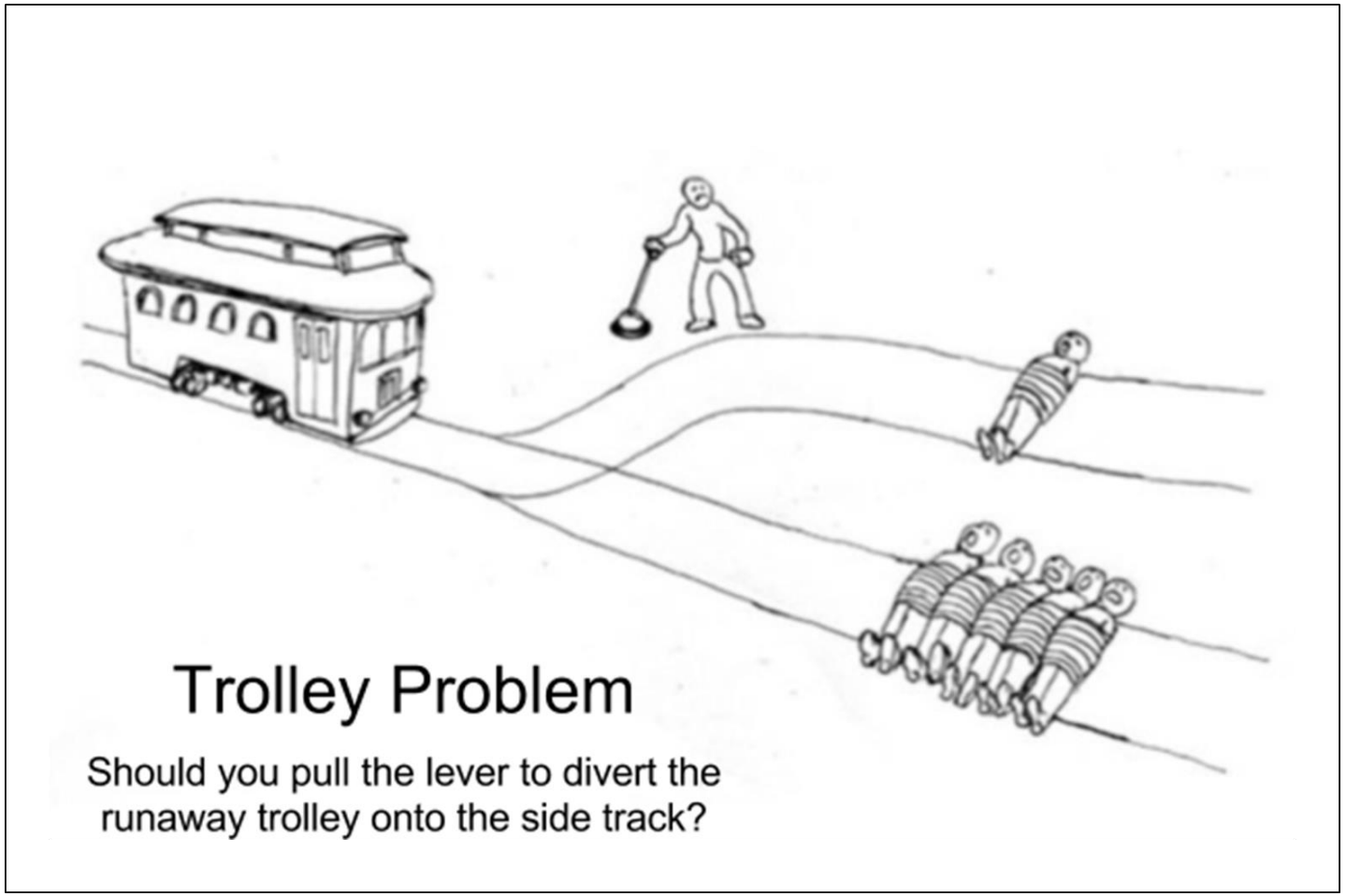
Ethical theories

- Virtue ethics:
  - Moral behaviours uphold the person's virtues
  - Criticism: increasing evidence that character traits are illusory
- Deontology (Duty)
  - Moral behaviours are those that satisfy the categorical imperative
  - Criticism: unacceptable inflexibility
- Utilitarianism (consequentialism)
  - Moral behaviours are those that bring the most good to the most people
  - Criticism: How to measure utility?

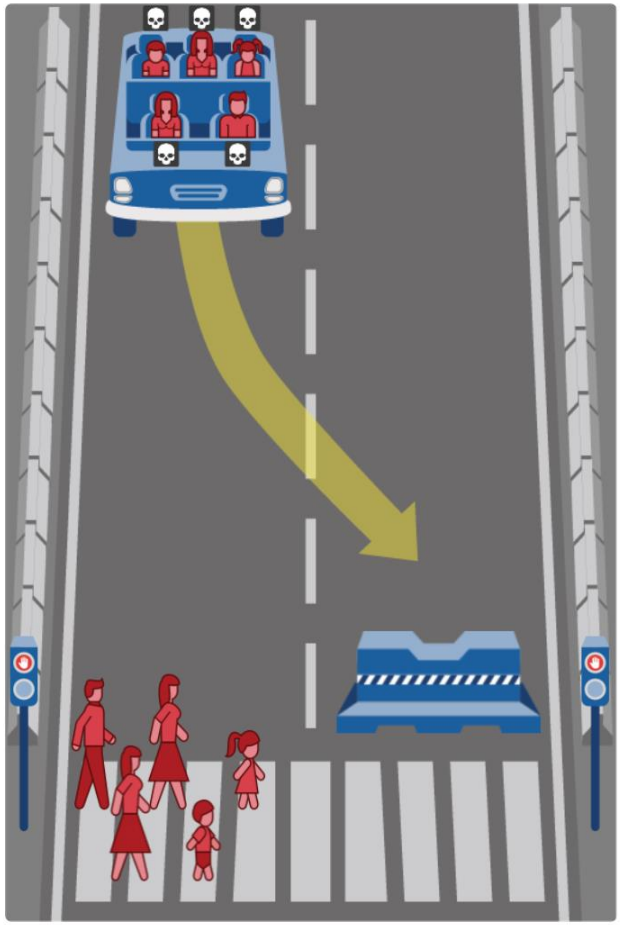
Ethics of technology

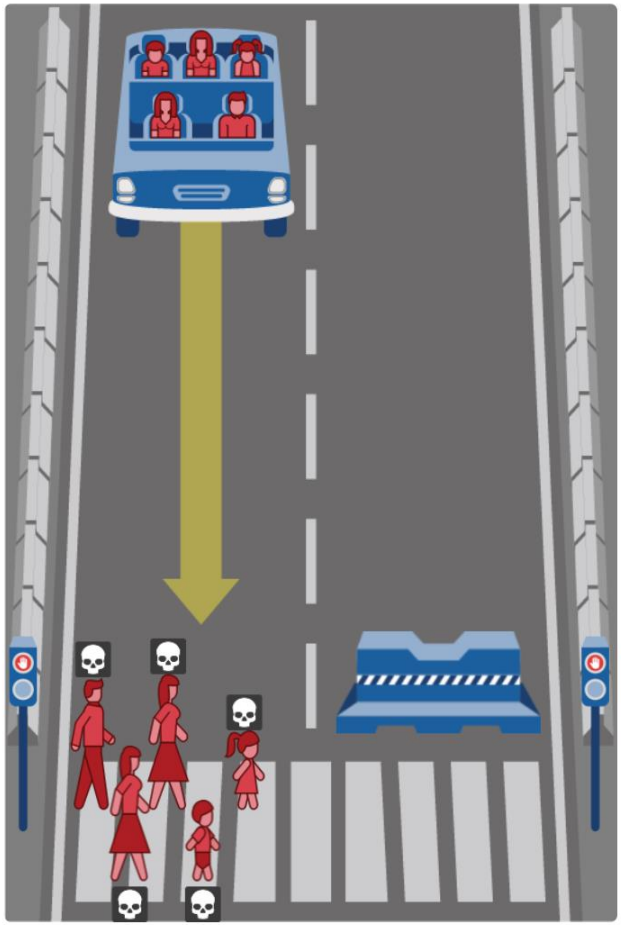
- Ethics change with technological progress
  - Right to internet access
  - Birth control
  - Surrogate pregnancy
  - Embryo selection
  - Artificial womb
  - Lab-grown meat

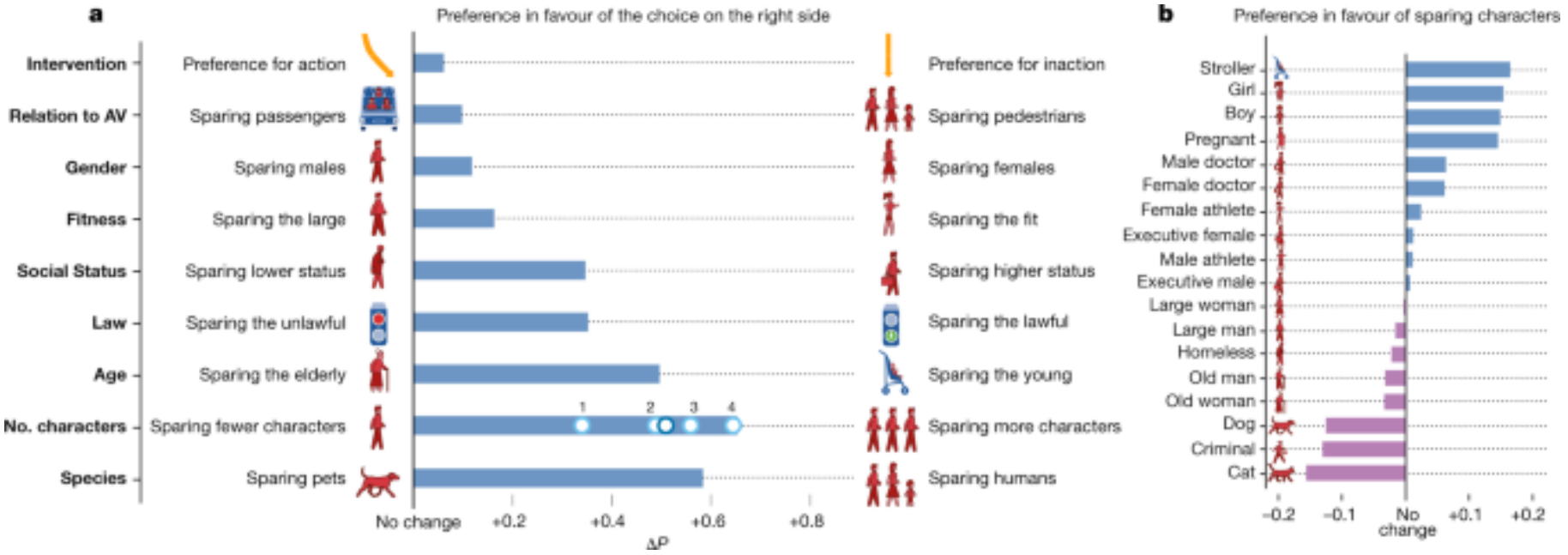
Trolley problems

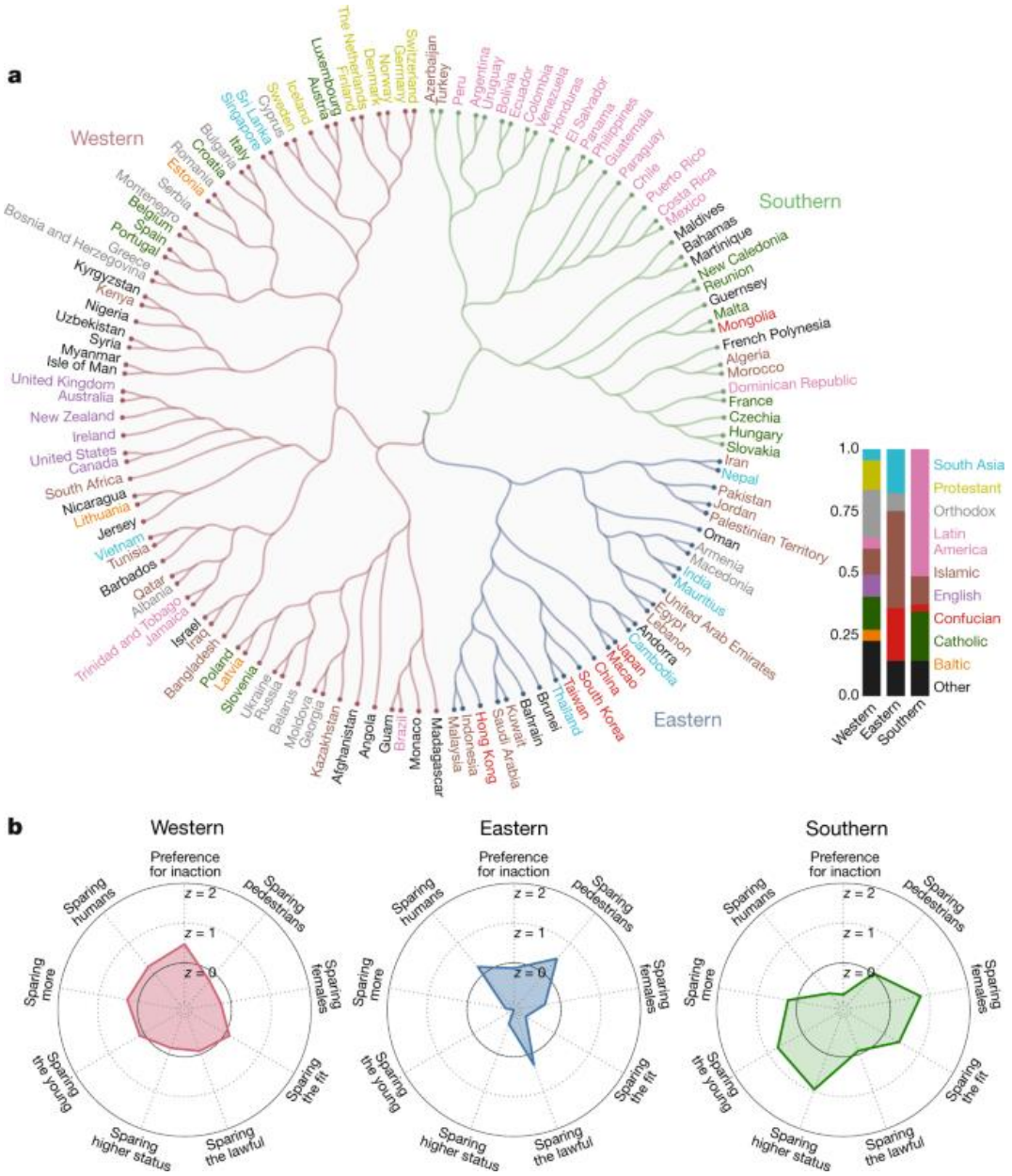


The Moral Machine






a
Preference in favour of the choice on the right side
Intervention
Preference for action
Preference for inaction
Relation to AV
Sparing passengers
Sparing pedestrians
Gender
Sparing males
Sparing females
Fitness
Sparing the large
Sparing the fit
Social Status
Sparing lower status
Sparing higher status
Law
Sparing the unlawful
Sparing the lawful
Age
Sparing the elderly
Sparing the young
No. characters
Sparing fewer characters
1
2
3
4
Sparing more characters
Species
Sparing pets
Sparing humans
No change
+0.2
+0.4
+0.6
+0.8
ΔP
b
Preference in favour of sparing characters
Stroller
Girl
Boy
Pregnant
Male doctor
Female doctor
Female athlete
Executive female
Male athlete
Executive male
Large woman
Large man
Homeless
Old man
Old woman
Dog
Criminal
Cat
−0.2
−0.1
No change
+0.1
+0.2


a
Western
Southern
Eastern
Luxembourg
Austria
The Netherlands
Finland
Denmark
Norway
Germany
Switzerland
Azerbaijan
Turkey
Peru
Argentina
Uruguay
Bolivia
Ecuador
Colombia
Venezuela
Honduras
El Salvador
Panama
Philippines
Guatemala
Paraguay
Chile
Puerto Rico
Costa Rica
Mexico
Maldives
Bahamas
Martinique
New Caledonia
Reunion
Guernsey
Malta
Mongolia
French Polynesia
Algeria
Morocco
Dominican Republic
France
Czechia
Hungary
Slovakia
Iran
Nepal
Pakistan
Jordan
Palestinian Territory
Oman
Armenia
Macedonia
India
Mauritius
United Arab Emirates
Egypt
Lebanon
Andorra
Cambodia
Japan
Macao
China
South Korea
Taiwan
Thailand
Brunei
Bahrain
Kuwait
Saudi Arabia
Hong Kong
Malaysia
Indonesia
Madagascar
Monaco
Guam
Brazil
Angola
Afghanistan
Kazakhstan
Georgia
Moldova
Belarus
Russia
Ukraine
Slovenia
Latvia
Poland
Bangladesh
Iraq
Israel
Jamaica
Trinidad and Tobago
Albania
Qatar
Barbados
Tunisia
Vietnam
Jersey
Lithuania
Nicaragua
South Africa
Canada
United States
Ireland
New Zealand
Australia
United Kingdom
Isle of Man
Myanmar
Syria
Uzbekistan
Nigeria
Kenya
Kyrgyzstan
Greece
Bosnia and Herzegovina
Portugal
Spain
Belgium
Montenegro
Serbia
Estonia
Romania
Croatia
Bulgaria
Italy
Singapore
Sri Lanka
Cyprus
Sweden
Iceland
1.0
0.75
0.5
0.25
0.0
South Asia
Protestant
Orthodox
Latin America
Islamic
English
Confucian
Catholic
Baltic
Other
Western
Eastern
Southern
b
Western
Eastern
Southern
Preference for inaction
Sparing pedestrians
Sparing females
Sparing the fit
Sparing the lawful
Sparing higher status
Sparing the young
Sparing more
Sparing humans
z = 2
z = 1
z = 0

LONG TERM PROBLEMS

Autonomous weapons

- Tempting to dismiss as far-fetched at this time
- But "the future is already here, just not evenly distributed"

Experts Shocked by Military Robodog With Sniper Rifle Attachment

"This crosses a moral, legal and technical line, taking us to a dark and dangerous world."

/ Robots & Machines / Boston Dynamics / Ghost Robotics / Killer Robots

Image by Ghost Robotics/Futurism

Replacing human labour

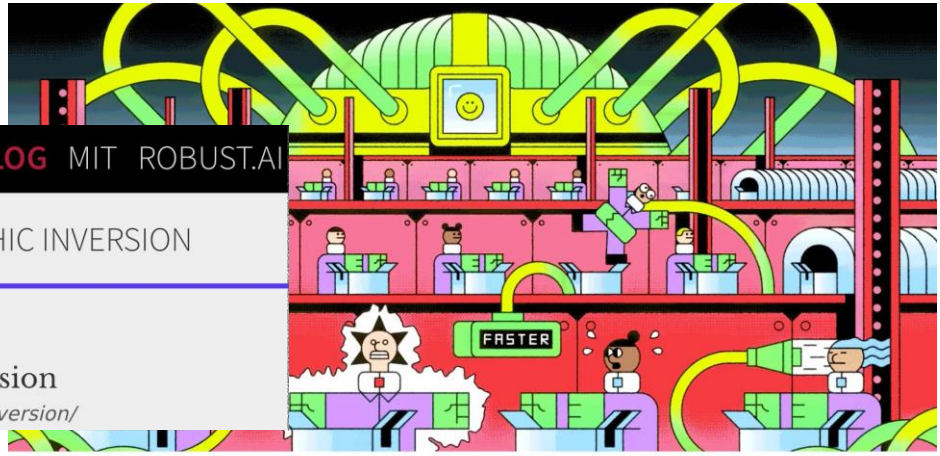
RODNEY BROOKS *Robots, AI, and other stuff* BLOG MIT ROBUST.AI

POST: MEGATREND: THE DEMOGRAPHIC INVERSION

APRIL 9, 2017 — ESSAYS

Megatrend: The Demographic Inversion

rodneybrooks.com/megatrend-the-demographic-inversion/



FEATURES

HOW HARD WILL THE ROBOTS MAKE US WORK?

In warehouses, call centers, and other sectors, intelligent machines are managing humans, and they're making work more stressful, grueling, and dangerous

By Josh Dzieza | @joshdzieza | Feb 27, 2020, 8:00am EST

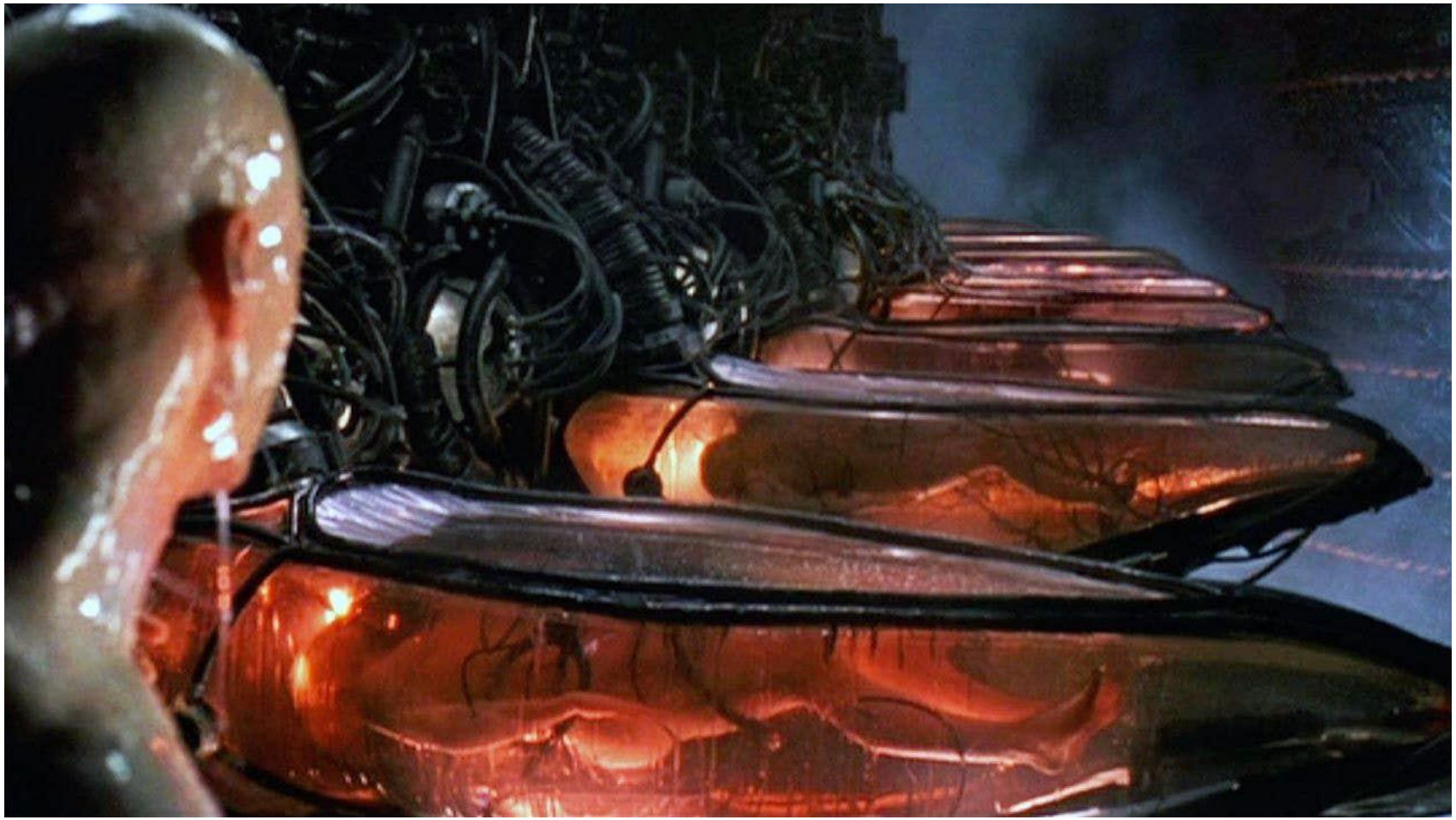
Illustrations by Joel Plosz

TIME

BUSINESS • COVID-19

Millions of Americans Have Lost Jobs in the Pandemic—And Robots and AI Are Replacing Them Faster Than Ever

Replacing humans entirely



Alignment

- "Paperclip maximiser"
  - AGI given the goal of producing paperclips eventually turns every atom in space into paperclips
- Old lesson: establishing and communicating our goals and values is hard, and technology amplifies the difficulty

A photograph of an Aladdin figure meeting the genie after rubbing the magic lamp (photo by JD Hancock/Flickr)

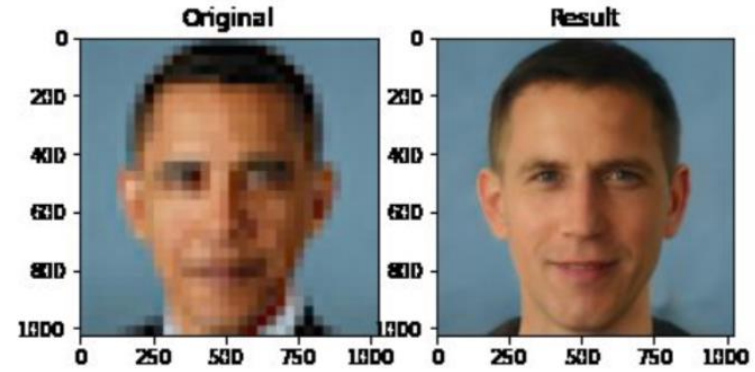
Where does it manifest?

BIAS

PULSE

@Chicken3gg

Replying to @tg_bomze

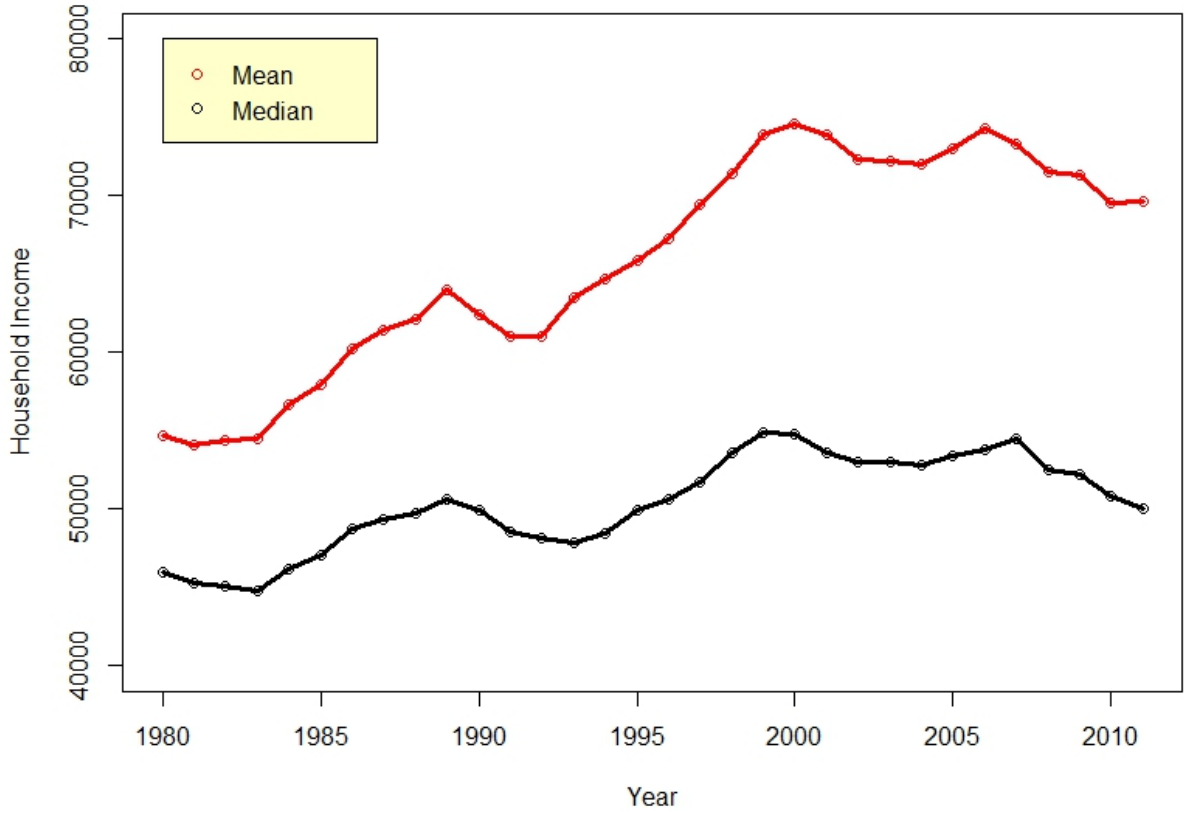


♡ 24.3K 8:14 AM - Jun 20, 2020

Yann LeCun @ylecun

ML systems are biased when data is biased.
This face upsampling system makes everyone look white because the network was pretrained on FlickFaceHQ, which mainly contains white people pics.
Train the *exact* same system on a dataset from Senegal, and everyone will look African.

An aside: Average household income



An aside: algorithmic choices

$$L1 = \frac{1}{k}\sum_{i=1}^{k}|x_i - \beta|$$

$$\frac{\partial L1}{\partial \beta} = -\frac{1}{k}\sum_{i=1}^{k} sign(x_i - \beta)$$

- $\frac{\partial L1}{\partial \beta} = 0$ when β is median

$$L2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \beta)^2$$

$$\frac{\partial L2}{\partial \beta} = -\frac{2}{k}\sum_{i=1}^{k}(x_i - \beta)$$

- $\frac{\partial L2}{\partial \beta} = 0$ when β is mean

What do you see?

What do you see?

Prototype Theory

- Categorisation:
  - reduce the infinite differences to behaviourally and cognitively usable proportions
- There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975)
  - May also store exemplars (Wu & Barsalou, 2009)

A riddle

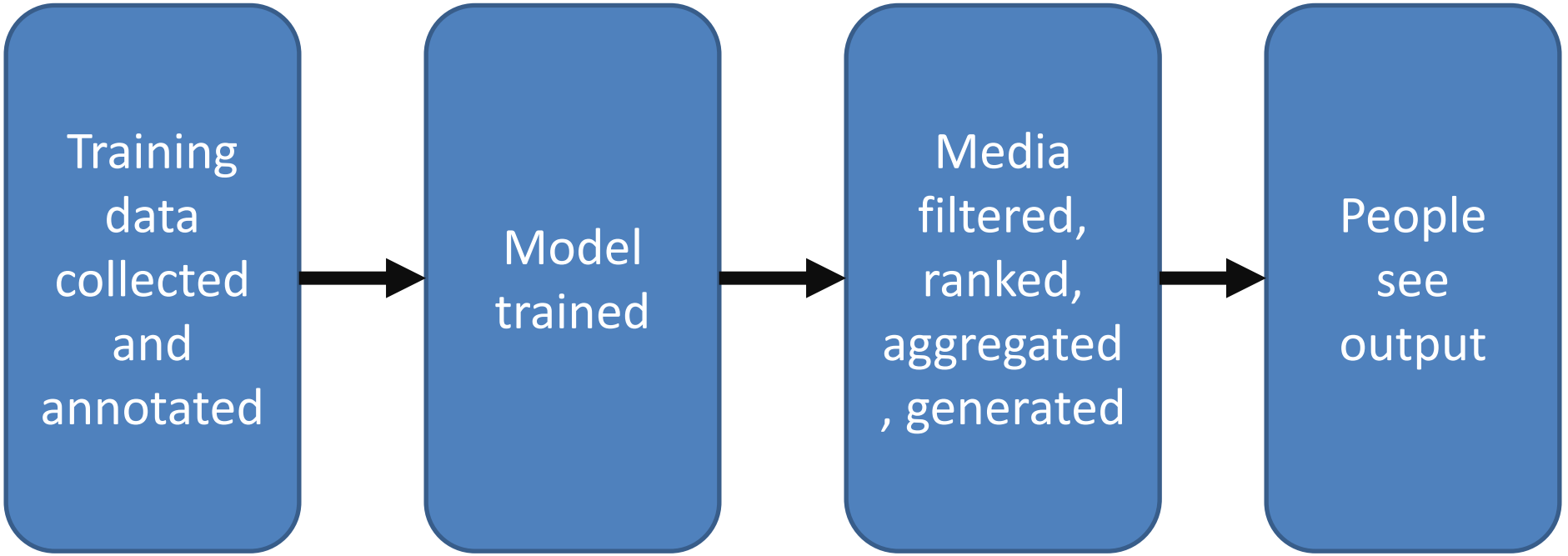
- *A man and his son are in a terrible accident and are rushed to the hospital in critical care.*
- *The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"*
- *How could this be?*

- The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists

Learning from text

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

Gordon and Van Durme, 2013


Training data collected and annotated
Model trained
Media filtered, ranked, aggregated, generated
People see output

Training data collected and annotated

- Human biases in data
  - Reporting bias
  - Selection bias
  - Overgeneralisation
  - Outgroup homogeneity bias
  - Stereotypical bias
  - Implicit associations
  - etc

Training data collected and annotated

- Human biases in collection/annotation
  - Sampling error
  - Non-sampling error
  - Correspondence bias
  - In-group bias
  - Confirmation bias
  - Experimenter's bias
  - Anecdotal fallacy
  - etc

Biases

- Reporting bias: What people share is not a reflection of real-world frequencies
- Selection Bias: Selection does not reflect a random sample
- Out-group homogeneity bias: People tend to see outgroup members as more alike
- Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation
- Confirmation bias: The tendency to search for, interpret, favour, and recall information in a way that confirms one's preexisting beliefs or hypotheses

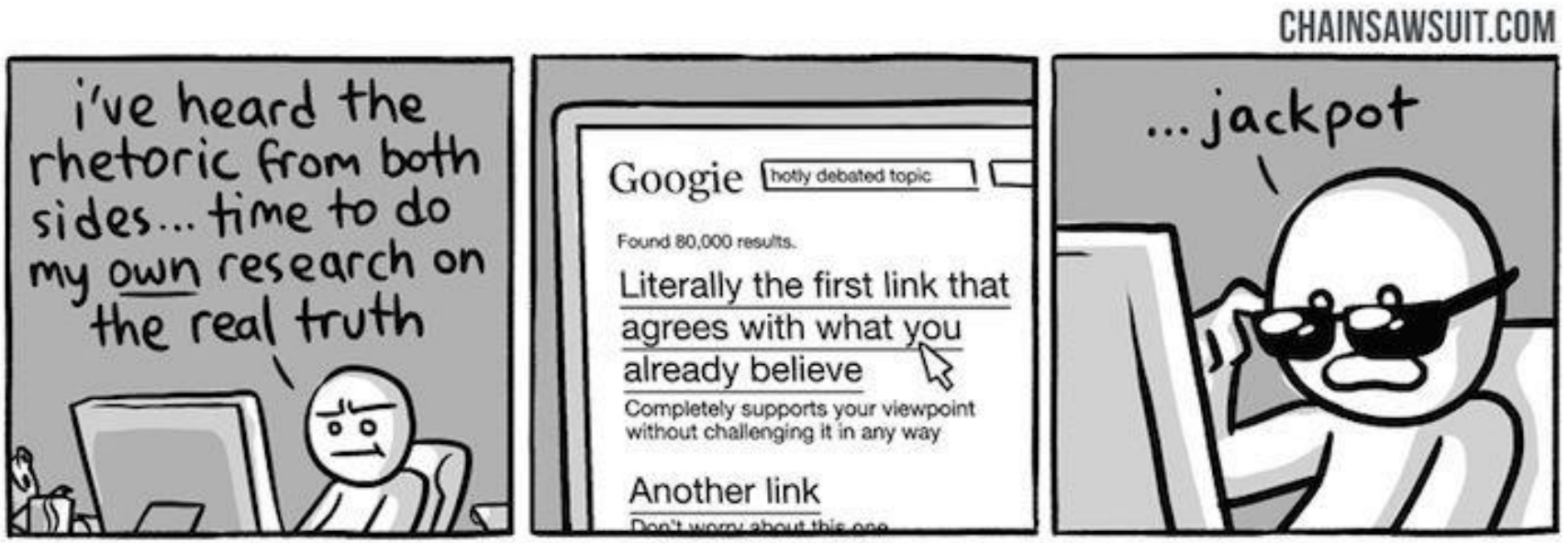
Selection bias

Biased labels

- Annotations in your dataset will reflect the worldviews of annotators

Confirmation bias

- The tendency to search for, interpret, favour, recall information in a way that confirms preexisting beliefs

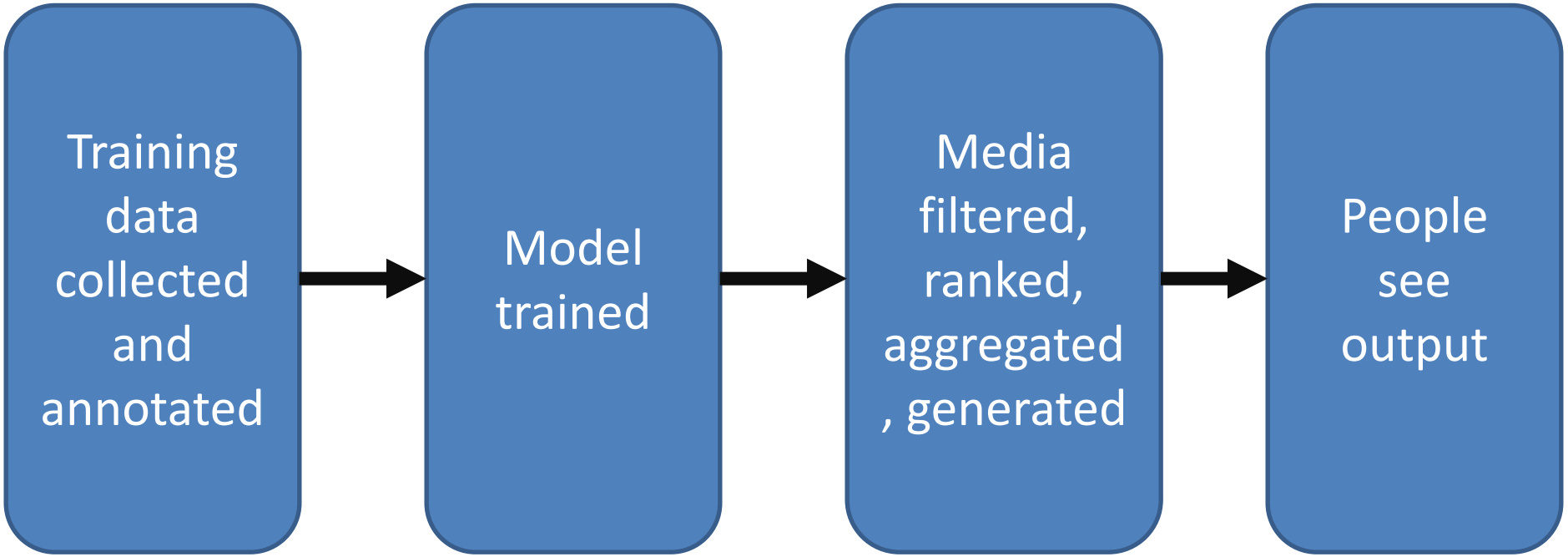


Overgeneralisation

- Coming to conclusion based on information that is too general and/or not specific enough (related: overfitting)

Automation bias

- Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation



Biased data created from process becomes new training data!

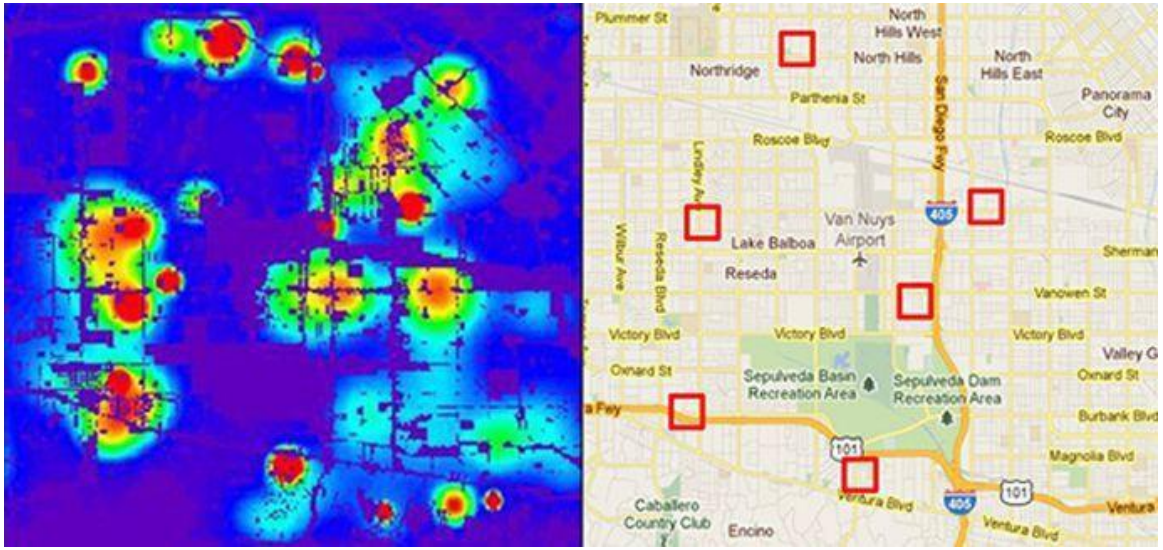
Algorithmic bias

- Unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making

NEAR TERM PROBLEMS

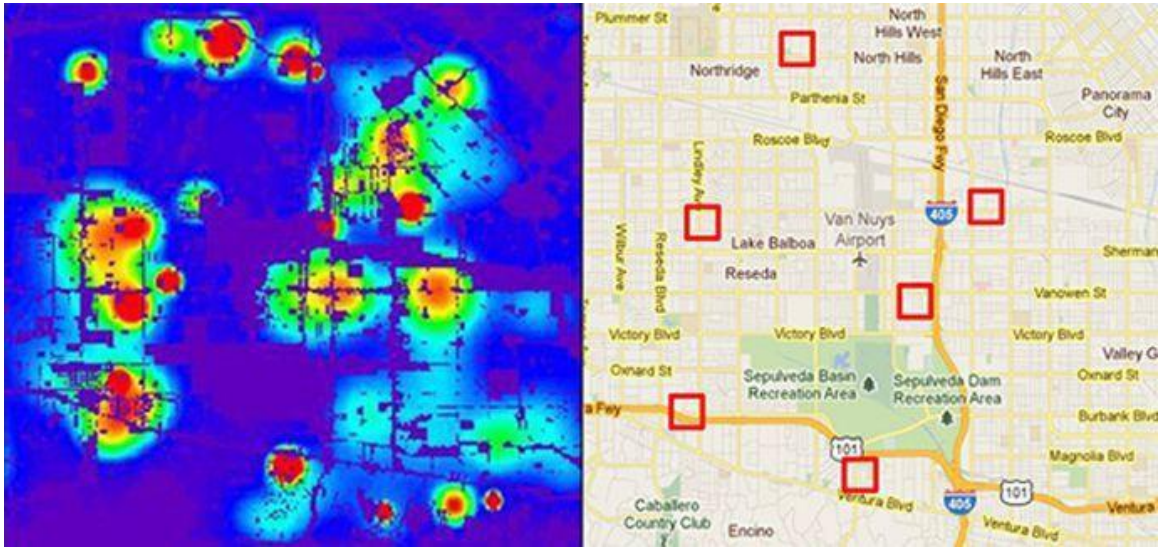
Predictive policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it is known to have occurred
- Predicts future events from past



Predictive policing

- Automation bias in face of:
  - Overgeneralization
  - Feedback Loops
  - Correlation Fallacy



Predicting criminality

- “Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial image.”

- Offering specialized engines for recognizing “High IQ”, “White-Collar Offender”, “Pedophile”, and “Terrorist” from a face image.

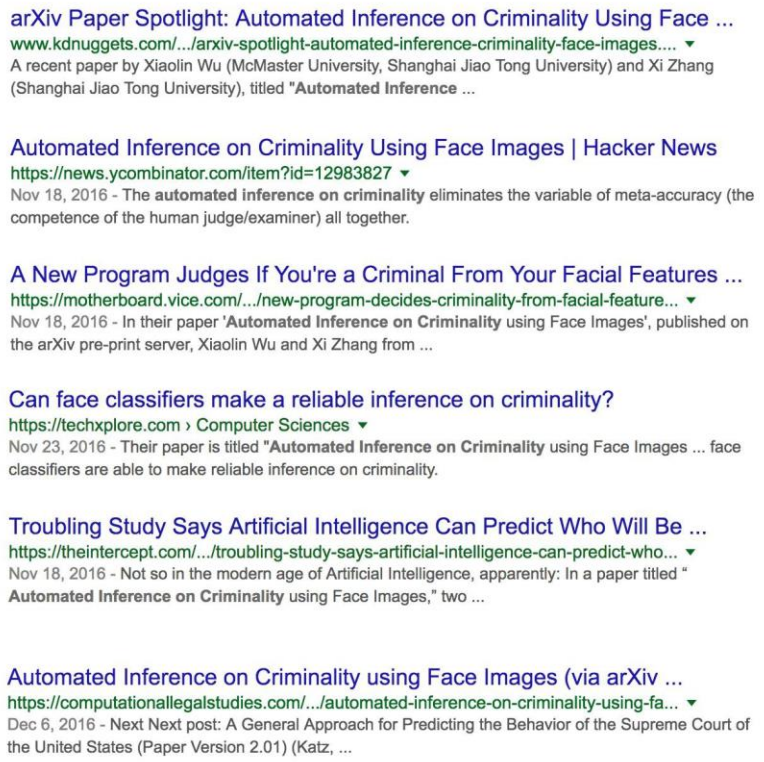
- Main clients are in homeland security and public safety

Predicting criminality

- *"Automated Inference on Criminality using Face Images" Wu and Zhang, 2016. arXiv*
- 1,856 closely cropped images of faces; Includes "wanted suspect" ID pictures from specific re
- *"[…] angle ϑ … is on average 19.6% smalle … ls …"*

Predicting criminality

- Selection Bias
- Experimenter's Bias
- Confirmation Bias
- Correlation Fallacy
- Feedback Loops



arXiv Paper Spotlight: Automated Inference on Criminality Using Face ...
www.kdnuggets.com/.../arxiv-spotlight-automated-inference-criminality-face-images....
A recent paper by Xiaolin Wu (McMaster University, Shanghai Jiao Tong University) and Xi Zhang (Shanghai Jiao Tong University), titled **"Automated Inference** ...

Automated Inference on Criminality Using Face Images | Hacker News
https://news.ycombinator.com/item?id=12983827
Nov 18, 2016 - The **automated inference on criminality** eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.

A New Program Judges If You're a Criminal From Your Facial Features ...
https://motherboard.vice.com/.../new-program-decides-criminality-from-facial-feature...
Nov 18, 2016 - In their paper '**Automated Inference on Criminality** using Face Images', published on the arXiv pre-print server, Xiaolin Wu and Xi Zhang from ...

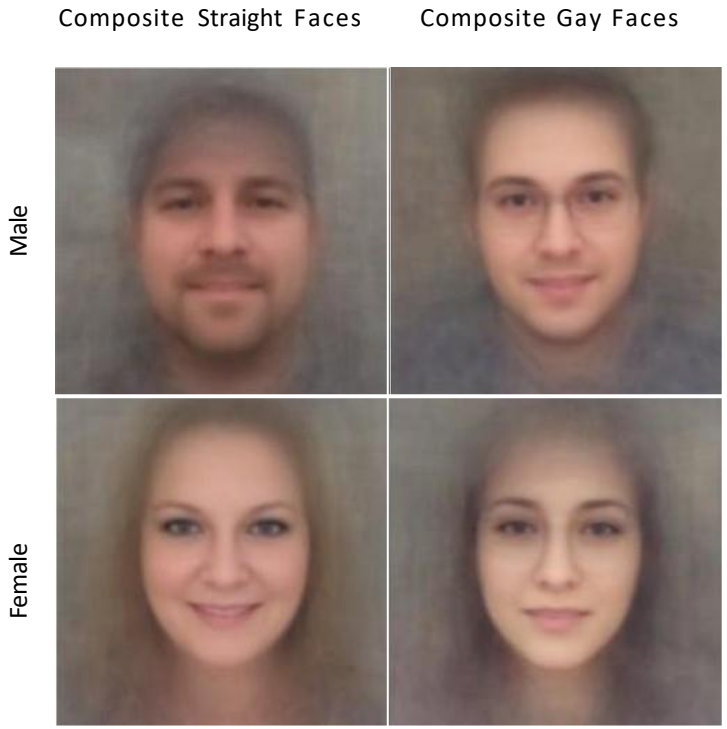
Can face classifiers make a reliable inference on criminality?
https://techxplore.com › Computer Sciences
Nov 23, 2016 - Their paper is titled **"Automated Inference on Criminality** using Face Images ... face classifiers are able to make reliable inference on criminality.

Troubling Study Says Artificial Intelligence Can Predict Who Will Be ...
https://theintercept.com/.../troubling-study-says-artificial-intelligence-can-predict-who...
Nov 18, 2016 - Not so in the modern age of Artificial Intelligence, apparently: In a paper titled " **Automated Inference on Criminality** using Face Images," two ...

Automated Inference on Criminality using Face Images (via arXiv ...
https://computationallegalstudies.com/.../automated-inference-on-criminality-using-fa...
Dec 6, 2016 - Next Next post: A General Approach for Predicting the Behavior of the Supreme Court of the United States (Paper Version 2.01) (Katz, ...

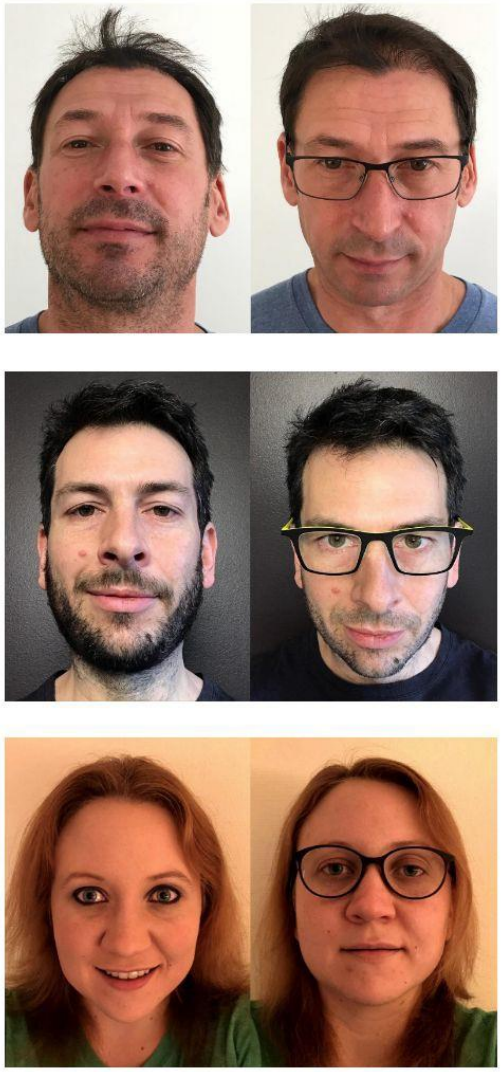
Predicting Homosexuality

- *Wang and Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, 2017.*
- "Sexual orientation detector" using 35,326 images from public profiles on a US dating website.
- "Consistent with the prenatal hormone theory [PHT] of sexual orientation, gay men and women tended to have gender-atypical facial morphology."



Predicting Homosexuality

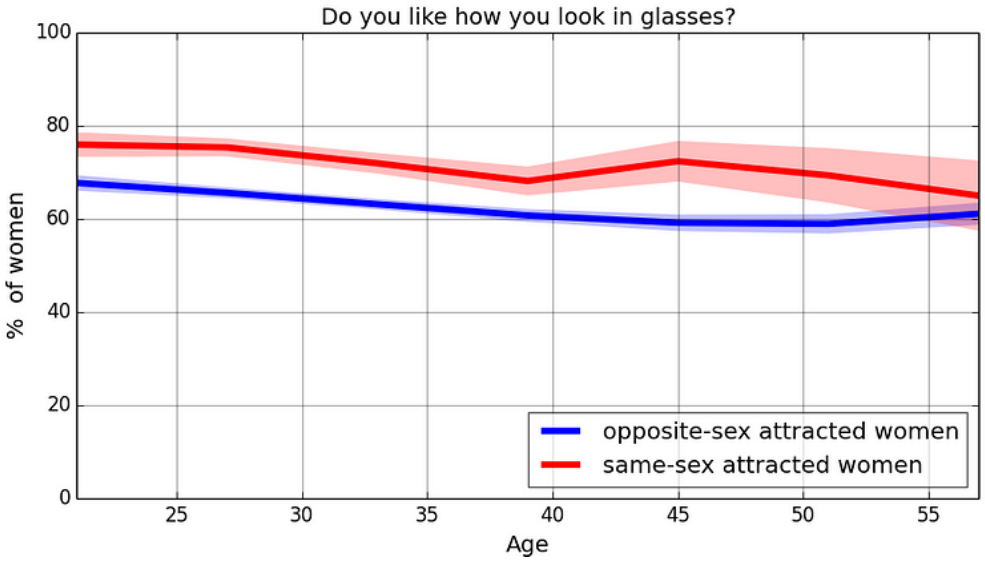
- *Differences between lesbian or gay and straight faces in selfies relate to grooming, presentation, and lifestyle — that is, differences in culture, not in facial structure.*



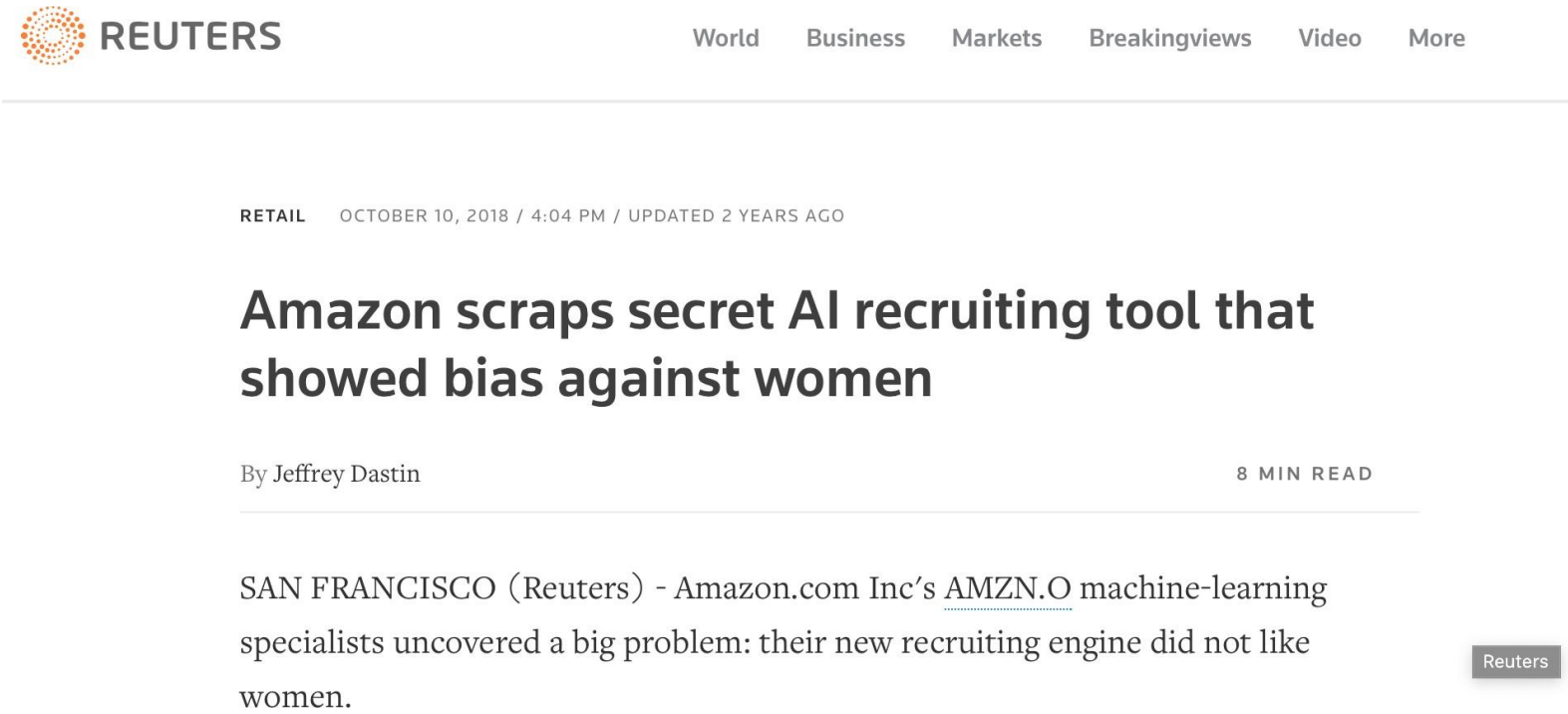
https://medium.com/%40blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Predicting Homosexuality

- Selection Bias
- Experimenter's Bias
- Correlation Fallacy



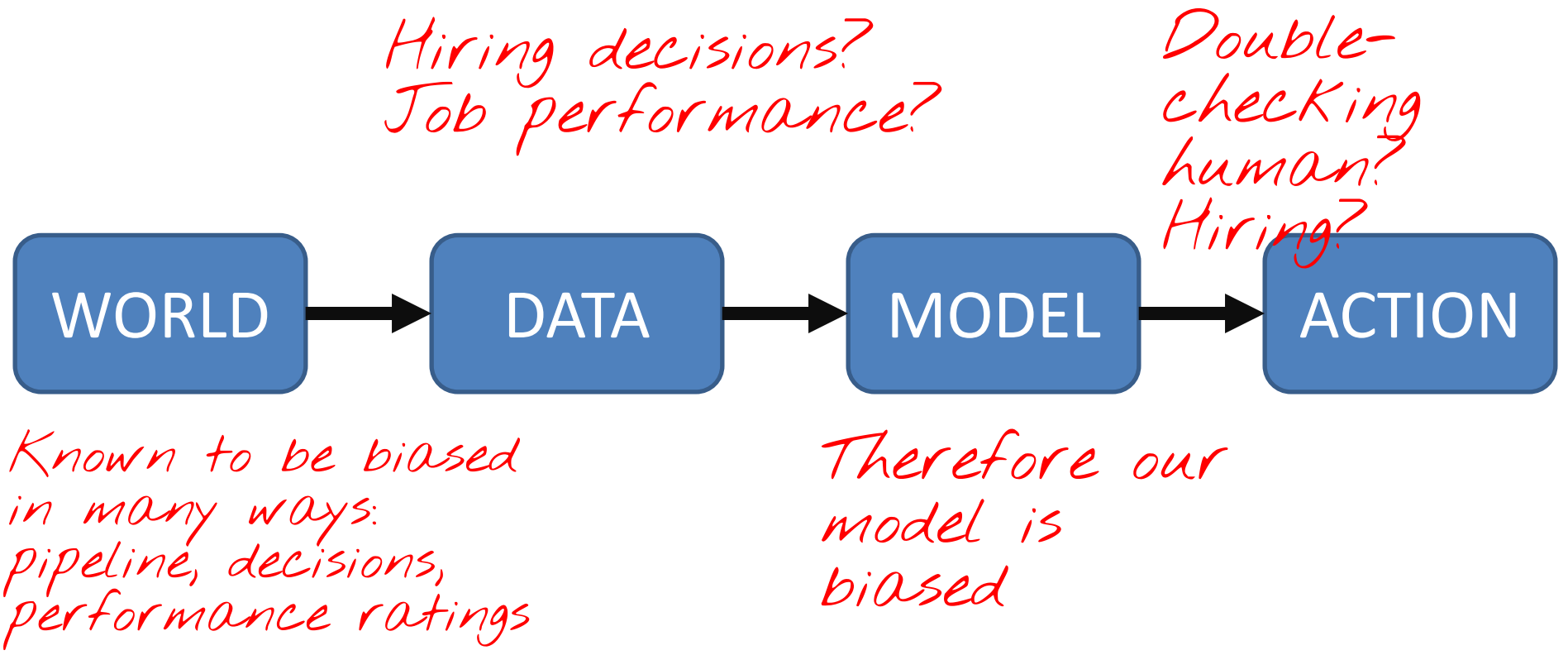
Hiring



https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias- against-women-idUSKCN1MK08G

Hiring

- ML model to predict hiring decision given CV



Amplifying existing biases is not aligned with our goals and values!

COMPAS

- Goal: predict recidivism, such that judges can consult 1-10 score in pre-trial sentencing decisions
- Motivation: be less biased than humans
- Solution:
  - Gather data
  - Exclude protected class attributes (race, etc)
  - Ensure that our model's score corresponds to same probability of recidivism across all groups

Outcome of COMPAS

Machine Bias

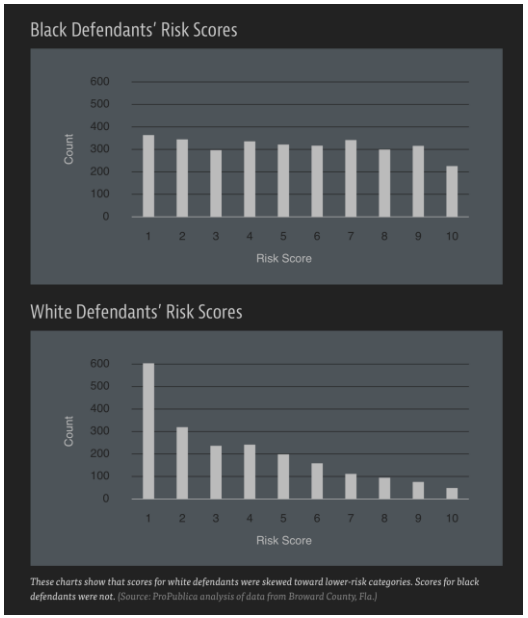
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

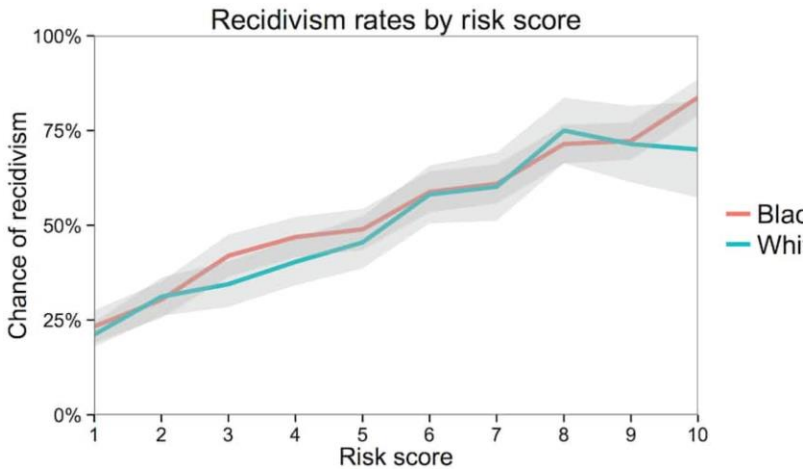


https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Statistical bias

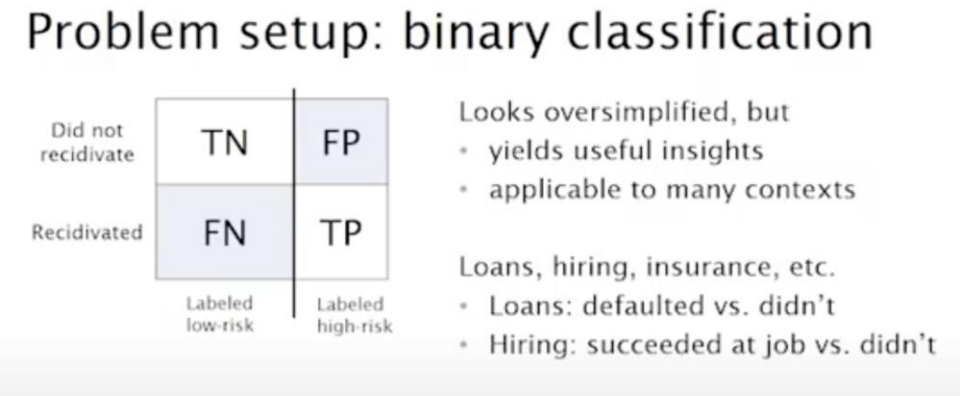
- Statistical bias: difference between estimator's expected value and the true value
- In this sense, COMPAS scores are not biased, w.r.t re-arrest
- But is it an adequate fairness criterion? Is it aligned with our values?

https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious- than-propublicas/



Fairness

- What do different stakeholders want from the classifier?
- Decision-maker: of those I've labeled high risk, how many recidivated?
  - Predictive value: TP / (TP + FP)
- Defendant: what's the probability I'll be incorrectly classified as high risk?
  - False positive rate: FP / (FP + TN)
- Society: is the selected set demographically balanced?
  - Demographic parity



Group fairness

- Do outcomes differ between groups (e.g. demographic), which we have no reason to believe are actually different?

Group fairness: impossibility theorem

if an instrument satisfies **predictive parity** … but the prevalence differs between groups, the instrument cannot achieve **equal false positive [rates]** and **[equal] false negative rates** across those groups.

Chouldechova. *Fair Prediction with Disparate Impact: A study of bias in recidivism prediction instruments.*

False positive
False negative

Prediction Fails Differently for Black Defendants

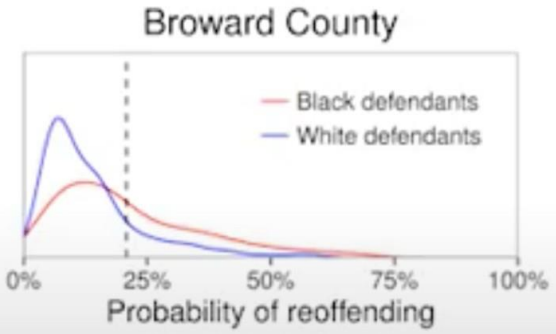
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Solution attempts

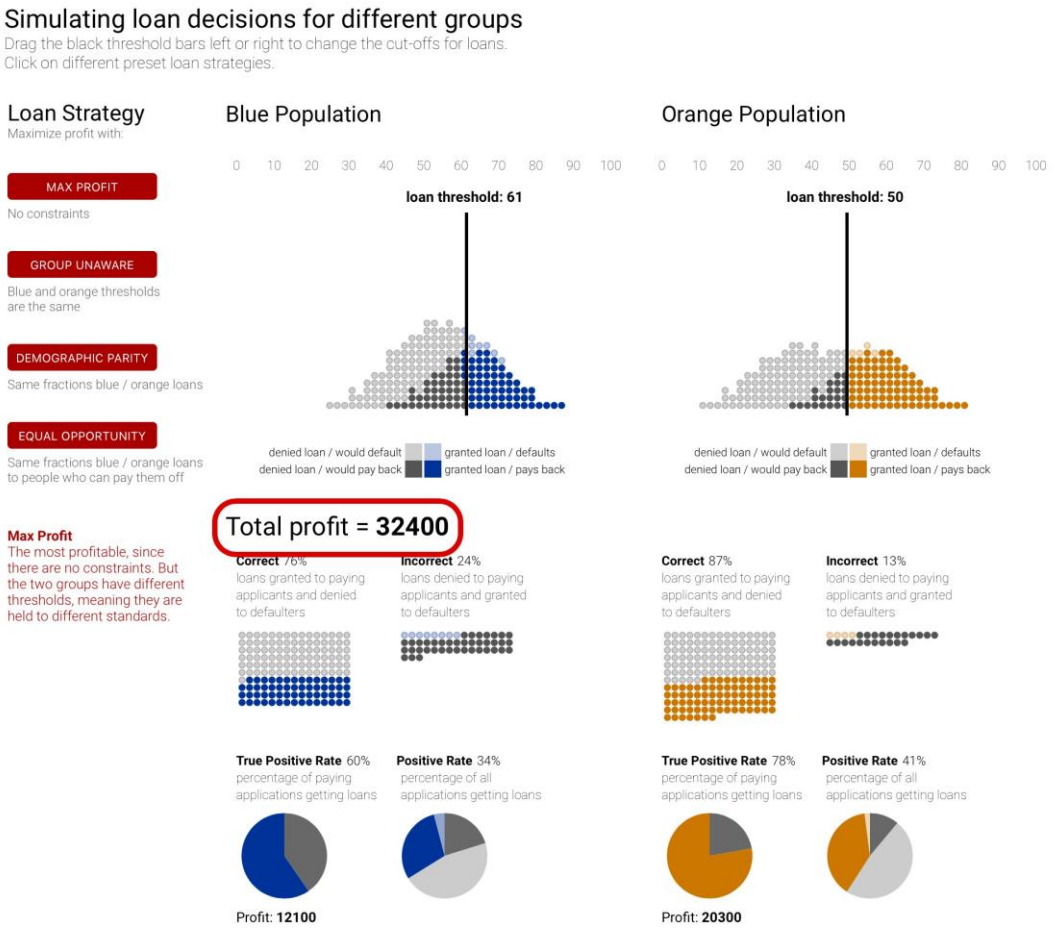
- Try win on group fairness:
  - e.g. Pick most important **two** metrics (FPR and FNR) and allow the model to use protected class attributes.
    - Now we fail individual fairness

Generally impossible to pick a single threshold for all groups that equalizes both FPR & FNR



*Assuming the scores are calibrated

Interactive example

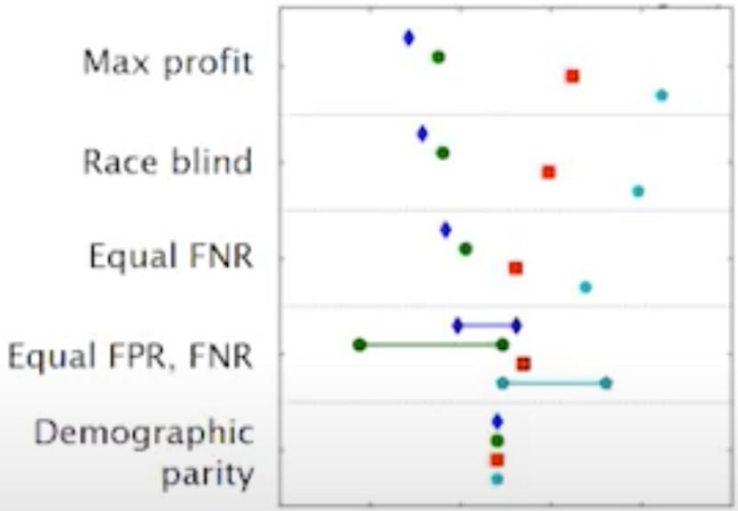


https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Masking protected attributes

Provocation

Ineffectiveness of blindness, quantified



Success rate of loan applications for 4 racial groups

Why is blindness in ML so ineffective compared to human decision-making?

1. ML bias is "just" a side effect of maximizing accuracy
2. ML is much better at picking up on proxies in the data

Hardt et al. *Equality of Opportunity in Supervised Learning.*

Representation

Nothing new

How Kodak's Shirley Cards Set Photography's Skin-Tone Standard

November 13, 2014 · 3:45 AM ET
Heard on Morning Edition

MANDALIT DEL BARCO

6-Minute Listen + PLAYLIST

Jersson Garcia works at Richard Photo Lab in Hollywood. He's 31 years old, and he's got a total crush on Shirley.

"Beautiful skin tones, beautiful eyes, great hair," he sighs. "She's gorgeous."

Garcia is holding a 4-by-6-inch photo of an ivory-faced brunette wearing a lacy, white, off-the-shoulders top. She has red lipstick and silver earrings, and the photo appears to have been taken sometime in the 1970s or '80s.

For many years, this "Shirley" card — named for the original model, who was an employee of Kodak — was used by photo labs to calibrate skin tones, shadows and light during the printing process.

For decades, Kodak's Shirley cards, like this one, featured only white models.

Kodak

Not just in our field

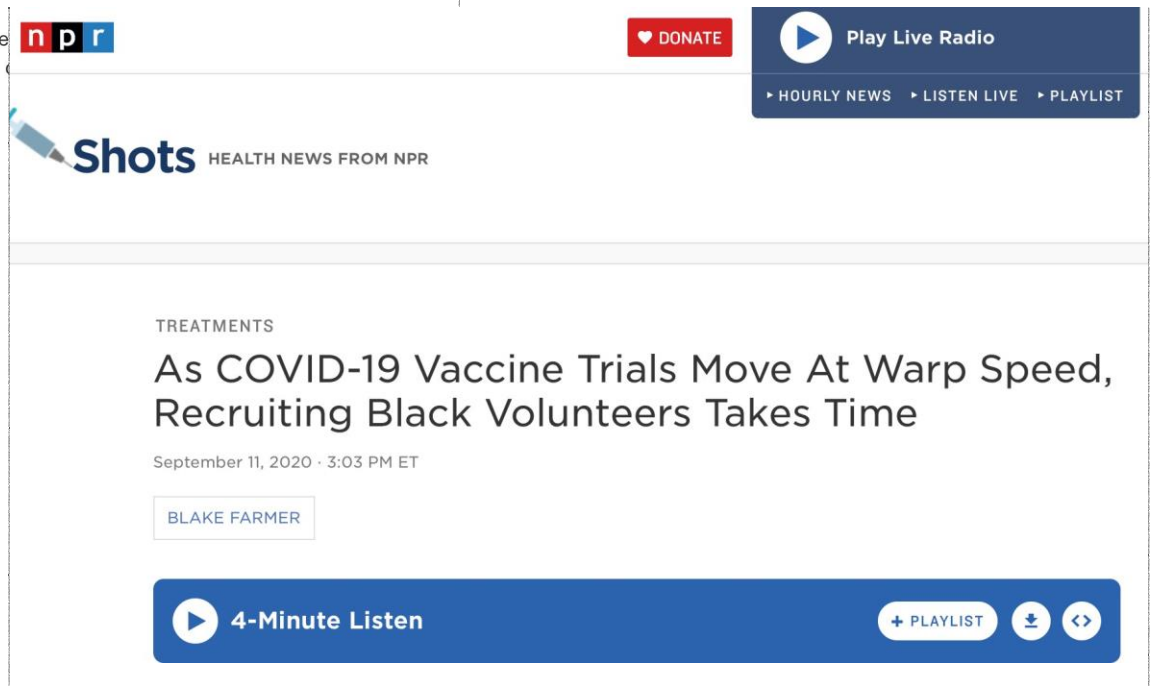
Lack of females in drug dose trials leads to overmedicated women

Gender gap leaves women experiencing adverse drug reactions nearly twice as often as men, study shows

Date: August 12, 2020

Source: University of California - Berkeley

Summary: Women are more likely than men to suffer adve
drug dosages have historically been based on
gests new research.



Facial recognition

- Old context: no expectation of privacy in public
- New context: "in public" now means "on any street in any city, or on any website on the internet"
- Is it ethical to work on this?
- Is it a problem if it does not work as well on some ethnicities?

At the same time, debating the merits of these technologies on the basis of their likely accuracy for different groups may distract from a more fundamental question: should we ever deploy such systems, even if they perform equally well for everyone? We may want to regulate the police's access to such tools, even if the tools are perfectly accurate. Our civil rights—freedom or movement and association—are equally threatened by these technologies when they fail and when they work well.

https://fairmlbook.org/introduction.html

SOUTH AFRICA

Vumacam captures around 4,500 suspicious vehicles and activities in Joburg daily

15 March 2023 - 06:53

Phathu Luvhengo

Journalist

GLOBAL

China's Surveillance State Should Scare Everyone

The country is perfecting a vast network of digital espionage as a means of social control—with implications for democracies worldwide.

By Anna Mitchell and Larry Diamond

https://www.theatlantic.com/international/archive/2018/02/china-surveillance/552203/

Suggestions

- Ethical risk sweeping
  - treat like cybersecurity penetration testing
- Expanding the ethical circle
  - whose interests, desires, experiences, values have we just assumed instead of consulted?
- Think about the terrible people
  - Who might abuse, steal, weaponize what we build? What incentives are we creating?
- Closing the loop
  - Remember that this is not a process to complete and forget. Set up ways to keep improving.