

KAIST Fall 2021

CS489: Computer Ethics and Social Issues

Fairness, Accountability, and Transparency

2021.09.27

Juho Kim

Who am I?

Why am I giving this guest lecture?

- Associate Professor in SoC
- Research area: Human-Computer Interaction
 - Building useful/usable software
 - Thinking a lot about the “human” side of computing
 - Understanding use and misuse of computing technology
 - Studying user perception, task improvements, & incentives
 - Supporting collaboration and social interaction
- Research interests: Interaction at scale, Social computing, Human-AI Interaction

Everything about Algorithmic Bias in One Tweet...

- Zoom Virtual Background
- Twitter Image Cropping Algorithm
- Racial bias in facial recognition / saliency detection
- Image dataset
- User control & agency
- ...

<https://twitter.com/colinmadland/status/1307111816250748933>

Colin Madland @colinmadland

A faculty member has been asking how to stop Zoom from removing his head when he uses a virtual background. We suggested the usual plain background, good lighting etc, but it didn't work. I was in a meeting with him today when I realized why it was happening.

9:18 AM · Sep 19, 2020 · Twitter Web App

20K Retweets 4.1K Quote Tweets 46.5K Likes

Colin Madland @colinmadland · Sep 19
Replying to @colinmadland
any guesses?

50 943 5.8K

25 538 4.8K

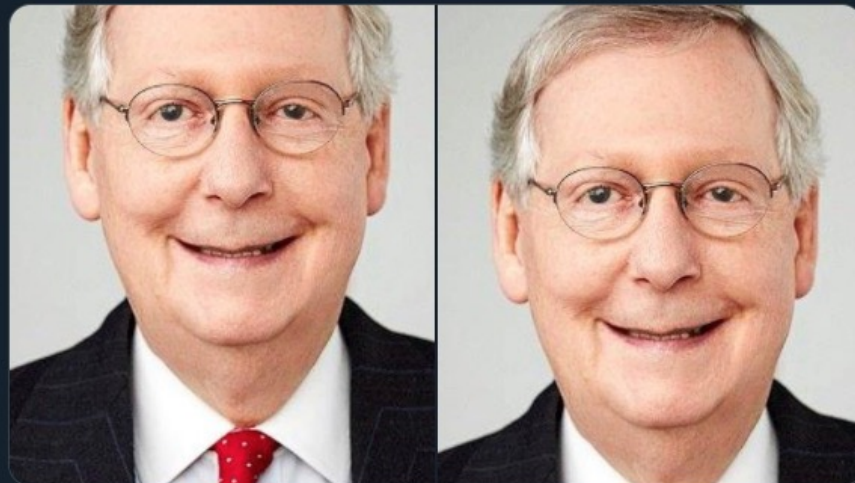




Tony "Abolish (Pol)ICE" Arcieri 🇺🇸
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?



7:05 AM · Sep 20, 2020 · Twitter Web App

51.9K Retweets 12.9K Quote Tweets 160.4K Likes



Zehan Wang @ZehanWang · 19h

We'll look into this. The algorithm does not do face detection at all (it actually replaced a previous algorithm which did). We conducted some bias studies before release back in 2017. At the time we found that there was no significant bias between ethnicities (or genders).



Dantley @dantley · 22h

Replying to @colinmadland and @Twitter

Based on some experiments I tried, I think @colinmadland's facial hair is affecting the model because of the contrast with his skin. I removed his facial hair and the Black man shows in the preview for me. Our team did test for racial bias before shipping the model.



8

185

363



Zehan Wang @ZehanWang · 19h

We purposefully constructed pairs of images of faces from different ethnic background as well as gender and ran them through the saliency detection model, checking for differences in saliency scores. No significant bias found. We'll review this study to see if we need to extend

6

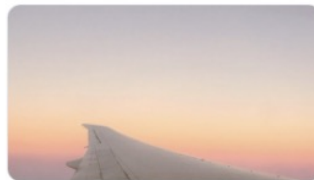
19

82



Infrastructure

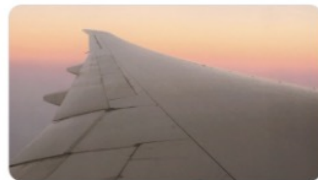
Speedy Neural Networks for Smart Auto-Cropping of Images



1



10



1



10



Eleanor Harding @t... · 08/01/2018

Couple of weeks at home in SA and my gallery is now 80% cat photos. She's called Phish. She likes to mlem.



2



40



Eleanor Harding @t... · 08/01/2018

Couple of weeks at home in SA and my gallery is now 80% cat photos. She's called Phish. She likes to mlem.



2



40



Left: Our face detector is unable to detect cats and other objects. Right: Our new cropping mechanism is able to focus on the most interesting part of the image.

<https://twitter.com/ZehanWang/status/1307461285811032066>

https://blog.twitter.com/engineering/en_us/topics/infrastructure/2018/Smart-Auto-Cropping-of-Images.html

Twitter Comms

Twitter CDO



Dantley 
@dantley

Replying to @TheNotoriousRBF @adrian_cadem and 6 others

It's 100% our fault. No one should say otherwise. Now the next step is fixing it.

6:32 AM · Sep 20, 2020 · Twitter for iPhone

219 Retweets **143** Quote Tweets **1.7K** Likes

<https://twitter.com/dantley/status/1307432466441859072>



 Cynthia Lee Retweeted

liz kelley @lizkelley · 57m

thanks to everyone who raised this. we tested for bias before shipping the model and didn't find evidence of racial or gender bias in our testing, but it's clear that we've got more analysis to do. we'll open source our work so others can review and replicate.

Tony "Abolish (Po)ICE" Arcieri  @bascule · 20h

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?

[Show this thread](#)



3 34 100

<https://twitter.com/lizkelley/status/1307742267193532416>



Amy X Zhang @amyxzh · 16h



The **easiest** fix for that biased cropping AI? No it's not to build another AI - it's to give people the power to select crop boundaries when posting a photo.

 17

 139

 854



5 mins on Twitter
points us to
all sorts of
algorithmic bias &
ethics issues.

ImageNet

- Bedrock of many modern AI systems
- Publicly available image dataset
- 14M images and 22K visual categories

Geological formation, formation

(geology) the geological features of the earth

1808 pictures

86.24% Popularity Percentile

Wordnet IDs

Numbers in brackets: (the number of synsets in the subtree).

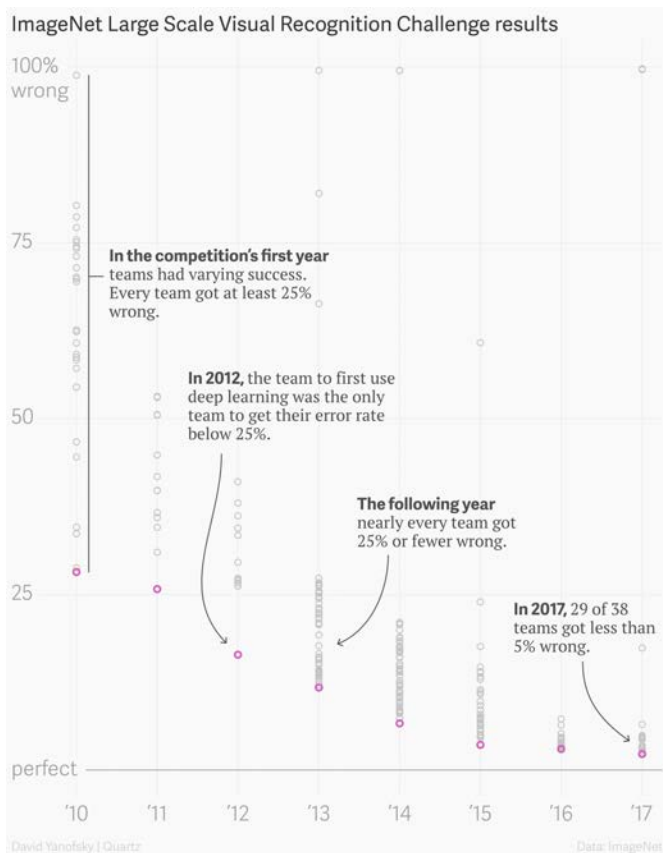
ImageNet 2011 Fall Release (32326)
- plant, flora, plant life (4486)
- geological formation, formation (1808)
- aquifer (0)
- beach (1)
- cave (3)
- cliff, drop, drop-off (2)
- delta (0)
- diapir (0)
- folium (0)
- foreshore (0)
- ice mass (10)
- lakefront (0)
- massif (0)
- monocline (0)
- mouth (0)
- natural depression, depression (0)
- natural elevation, elevation (41)
- oceanfront (0)
- range, mountain range, range of mountains (0)
- relict (0)
- ridge, ridgeline (2)
- ridge (0)
- shore (7)
- slope, incline, side (17)
- spring, fountain, outflow, outpouring (0)
- talus, scree (0)
- vein, mineral vein (1)
- volcanic crater, crater (2)
- wall (0)

Treemap Visualization Images of the Synset Downloads

ImageNet 2011 Fall Release Geological formation, formation

Natural	Slope	Shore
Natural	Ice	Water
Natural	Vein	Delta
Natural	Foreshore	Massif
Natural	Talus	Volcanic
Natural	Beach	Mouth
Natural	Lakefront	Range
Natural	Diapir	Cliff
Natural	Wall	Oceanfront
Natural	Aquifer	Cave
Natural	Spring	Monocline
Natural	Ridge	

ImageNet Visual Recognition Challenge



Driving Force for Deep Learning Revolution

[\[PDF\] Imagenet: A large-scale hierarchical image database](#)

[J Deng](#), [W Dong](#), [R Socher](#), [LJ Li](#), [K Li](#)... - 2009 IEEE conference ..., 2009 - researchgate.net

The explosion of image data on the Internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. But exactly how such data can be harnessed and organized remains a ...

☆ [🔗](#) Cited by 13027 [Related articles](#) [All 26 versions](#) [🔗](#)

[Imagenet classification with deep convolutional neural networks](#)

[A Krizhevsky](#), [I Sutskever](#), [GE Hinton](#) - Advances in neural ..., 2012 - papers.nips.cc

We trained a large, deep convolutional neural network to classify the 1.3 million high-resolution images in the LSVRC-2010 **ImageNet** training set into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 39.7% and 18.9% which is ...

☆ [🔗](#) Cited by 46699 [Related articles](#) [All 95 versions](#) [🔗](#)

[Imagenet large scale visual recognition challenge](#)

[O Russakovsky](#), [J Deng](#), [H Su](#), [J Krause](#)... - International journal of ..., 2015 - Springer

Abstract The **ImageNet** Large Scale Visual Recognition Challenge is a benchmark in object category classification and detection on hundreds of object categories and millions of images. The challenge has been run annually from 2010 to present, attracting participation ...

☆ [🔗](#) Cited by 12034 [Related articles](#) [All 17 versions](#)

[Delving deep into rectifiers: Surpassing human-level performance on imagenet classification](#)

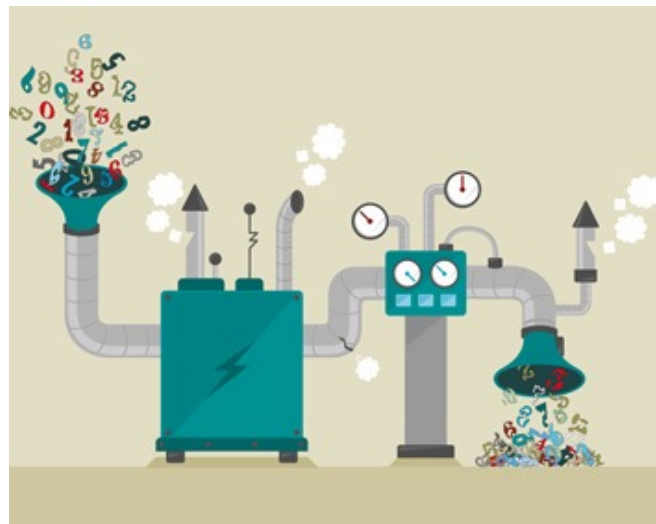
[K He](#), [X Zhang](#), [S Ren](#), [J Sun](#) - Proceedings of the IEEE ..., 2015 - cv-foundation.org

Rectified activation units (rectifiers) are essential for state-of-the-art neural networks. In this work, we study rectifier neural networks for image classification from two aspects. First, we propose a Parametric Rectified Linear Unit (PReLU) that generalizes the traditional rectified ...

☆ [🔗](#) Cited by 5685 [Related articles](#) [All 12 versions](#) [🔗](#)

Garbage in, garbage out

- Various sources of bias and political/sensitive decisions throughout the algorithmic pipeline
- Categories: where do they come from?
- Filtering of visual concepts: bias toward something visual
- Diversity of images: insufficient representation across dimensions



“programmer”

Original:



“programmer”

Original:



Balancing gender:



Balancing skin color:



Balancing age:



"We believe that ImageNet, as an influential research dataset, deserves to be critically examined, in order for the research community to design better collection methods and build better datasets."

Announced to scrub more than half of the 1.2 million pictures in the dataset's "people" category.

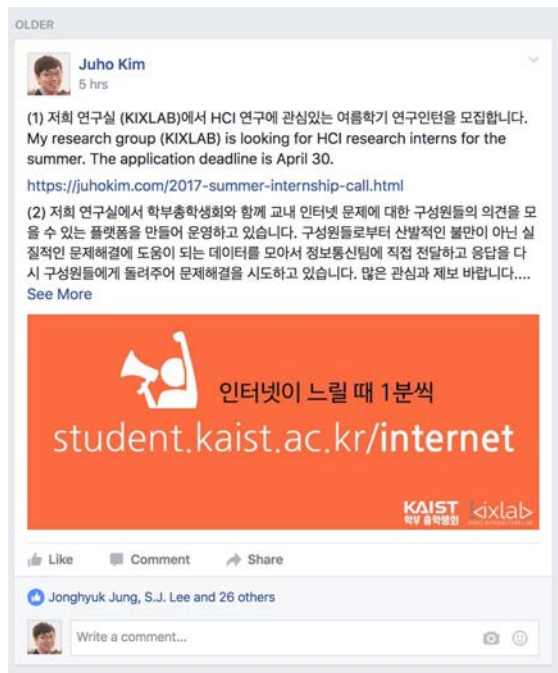
"There is no easy technical 'fix' by shifting demographics, deleting offensive terms, or seeking equal representation by skin tone,"

"The whole endeavor of collecting images, categorizing them, and labeling them is itself a form of politics, filled with questions about who gets to decide what images mean and what kinds of social and political work those representations perform."

**FAIRNESS, ACCOUNTABILITY, &
TRANSPARENCY**

Algorithms & humans interact closer than ever.

moderate



OLDER

Juho Kim
5 hrs

(1) 저희 연구실 (KIXLAB)에서 HCI 연구에 관심있는 여름학기 연구인턴을 모집합니다. My research group (KIXLAB) is looking for HCI research interns for the summer. The application deadline is April 30.
<https://juhokim.com/2017-summer-internship-call.html>

(2) 저희 연구실에서 학부총학생회와 함께 교내 인터넷 문제에 대한 구성원들의 의견을 모을 수 있는 플랫폼을 만들어 운영하고 있습니다. 구성원들로부터 산발적인 불만이 아닌 실질적인 문제해결에 도움이 되는 데이터를 모아서 정보통신팀에 직접 전달하고 응답을 다시 구성원들에게 돌려주어 문제해결을 시도하고 있습니다. 많은 관심과 제보 바랍니다....
See More

 인터넷이 느릴 때 1분씩
student.kaist.ac.kr/internet

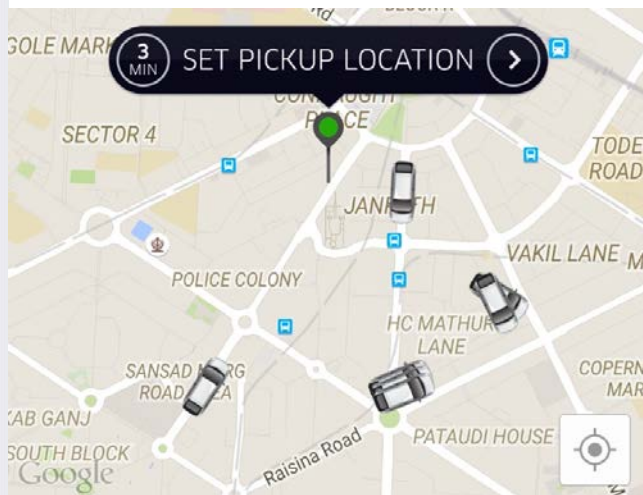
KAIST
학부 총학생회 <xlab>

Like Comment Share

Jonghyuk Jung, S.J. Lee and 26 others

Write a comment...

manage



compete



More decisions are made by algorithms

- Predictive criminal assessment
- Government resource allocation
- Loan / credit assessment
- Hiring
- Crime suspect identification with facial recognition

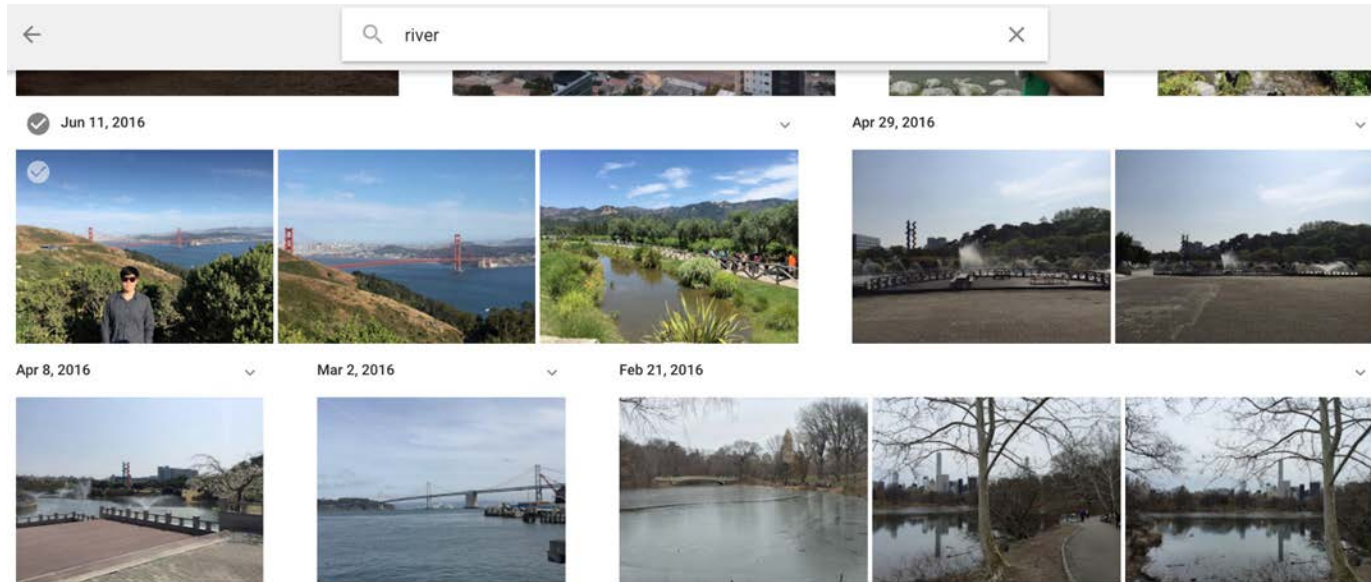
- Why is algorithmic decision-making a good idea?
- What could possibly go wrong?

Principle 1: Fairness

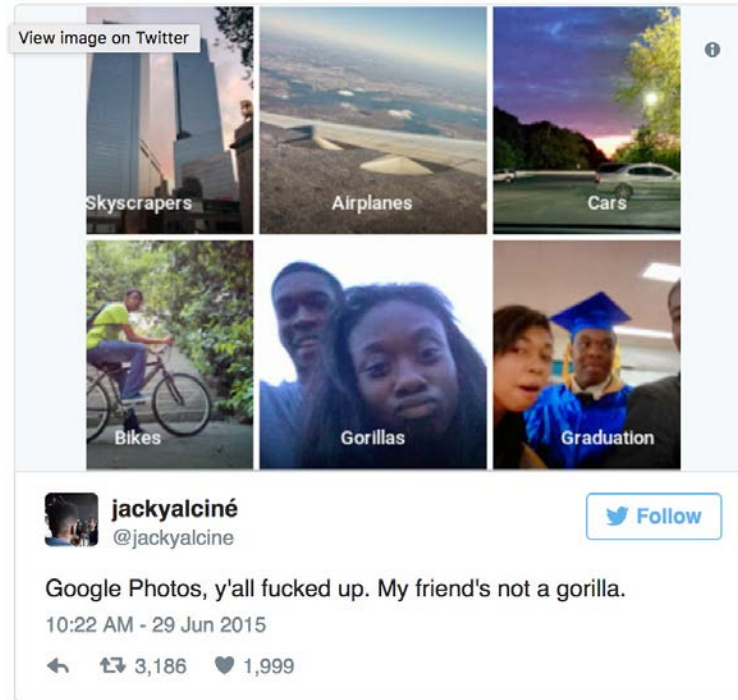
- “Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics” [fatml.org]
- **Treatment parity:** a classifier should be blind to a given protected characteristic. Also called anti-classification in [Corb2018], or “fairness through unawareness.”
- **Impact parity:** the fraction of people given a positive decision should be equal across different groups. This is also called demographic parity, statistical parity, or independence of the protected class and the score [Fair2018].

Intelligent Search

- Google Photos uses automated object recognition and tagging in their search interface.
















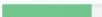




Implicit Racial Bias

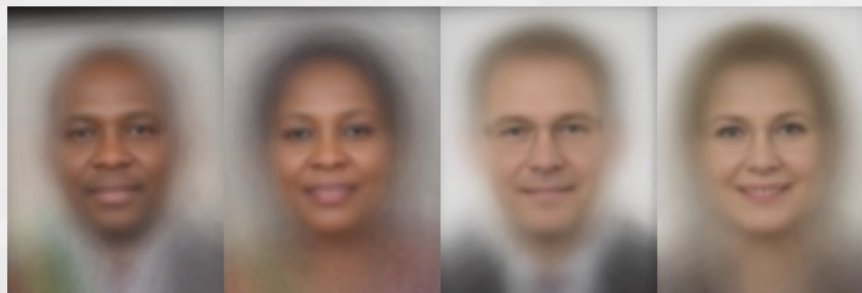


Gender Shades

When we analyze the results by intersectional subgroups - darker males, darker females, lighter males, lighter females - we see that all companies perform worst on darker females.

IBM and Microsoft perform best on lighter males. Face++ performs best on darker males.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

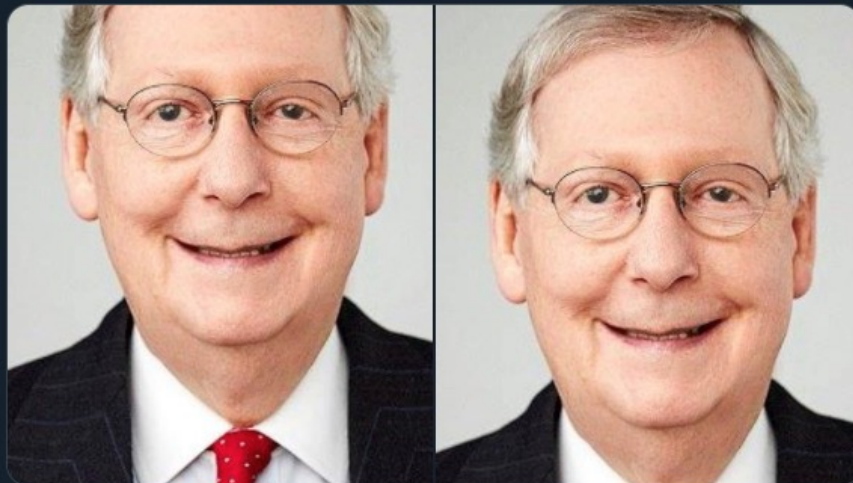




Tony "Abolish (Pol)ICE" Arcieri 🇺🇸
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?



7:05 AM · Sep 20, 2020 · Twitter Web App

51.9K Retweets 12.9K Quote Tweets 160.4K Likes

Emily/Brendan vs Lakisha/Jamal?

- If Emily and Brendan needed to send out 10 resumes on average to get one response, how many resumes did Lakisha and Jamal need to send?
- Emily and Brendan got 30% more callbacks when they sent out resumes listing high qualifications compared to when they sent out resumes with low qualifications. How much did Lakisha and Jamal benefit from getting higher qualifications?

15

9%

Learning from Biased Data in Word Embedding

$$\vec{\text{man}} - \vec{\text{woman}} \approx \vec{\text{king}} - \vec{\text{queen}}$$

$$\vec{\text{man}} - \vec{\text{woman}} \approx \vec{\text{computer programmer}} - \vec{\text{homemaker}}.$$

Implicit 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. fighter pilot | 12. boss |

DISCUSSION: Fair Hiring Algorithm

- Input: thousands of resumes from applicants
- Output: binary decision (yes: consider further, no: no hire)
- Let's discuss fairness dimensions to consider:
 - Race, Gender, Gender identity, Ability status, Socio-economic status, Education level, Religion, Country of origin, Way people look and dress
 - What are some other features w/ high likely correlations?
 - What dimensions should NOT be considered for fairness?
 - Data issues?
 - How might we design a "fair" hiring algorithm? How might we ensure getting good people in a fair manner?

Principle 2: Accountability

- Who is responsible if users are harmed by this product?
- Who will have the power to decide on necessary changes to the algorithmic system during design stage, pre-launch, and post-launch? [fatml.org]

Inadvertent Algorithmic Cruelty

- *“Yes, my year looked like that. True enough. My year looked like the now-absent face of my little girl. It was still unkind to remind me so forcefully,”*

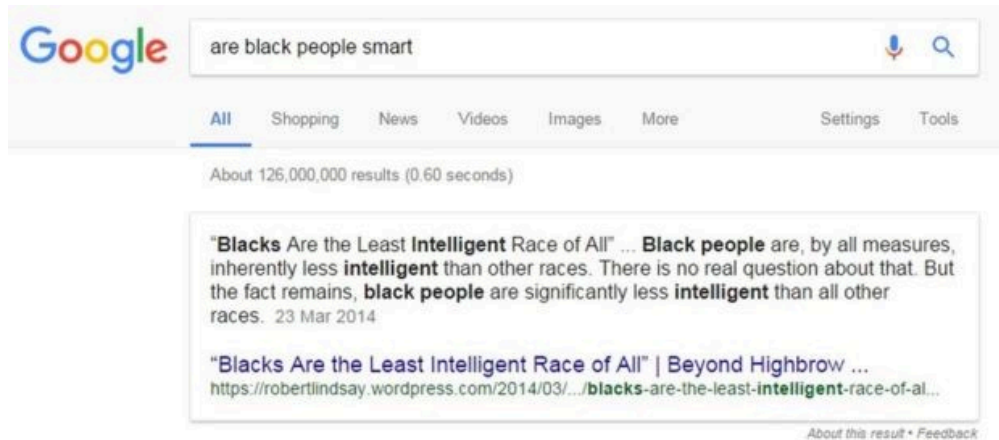


Who is accountable for algorithmic mistakes?

- Users exposed to only one political perspective for years
- Self-driving car killing pedestrians
- Facebook's "x years ago today" showing sad incidents
- Google photos classifying a human as a gorilla
- FB Newsfeed *showing* more content with negative emotion induces users to *write* more content with negative emotion (i.e., emotion contagion)

Responsible Actions: Are they effective?

- ImageNet team scrubbing potentially problematic images and applying debiasing techniques
- Google changing top search results that show false, questionable information



Twitter Comms

Twitter CDO



Dantley 
@dantley

Replying to @TheNotoriousRBF @adrian_cadem and 6 others

It's 100% our fault. No one should say otherwise. Now the next step is fixing it.

6:32 AM · Sep 20, 2020 · Twitter for iPhone

219 Retweets 143 Quote Tweets 1.7K Likes

<https://twitter.com/dantley/status/1307432466441859072>



Cynthia Lee Retweeted

liz kelley @lizkelley · 57m

thanks to everyone who raised this. we tested for bias before shipping the model and didn't find evidence of racial or gender bias in our testing, but it's clear that we've got more analysis to do. we'll open source our work so others can review and replicate.

Tony "Abolish (Po)ICE" Arcieri  @bascule · 20h

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?

[Show this thread](#)



3 34 100

<https://twitter.com/lizkelley/status/1307742267193532416>

Algorithmic Auditing

- “Enable interested **third parties to probe, understand, and review** the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of **detailed documentation, technically suitable APIs, and permissive terms of use.**” [fatml]



Thread



Casey Fiesler, PhD, JD, geekD

@cfiesler



One fascinating thing about the Twitter Photo Cropping Scandal of the Weekend was that it was essentially a collective algorithmic audit. I love that. I can see ways we could leverage crowdsourcing to conduct large-scale algorithmic audits. (Has this been done?)

10:41 PM · Sep 22, 2020 · Twitter Web App

76 Retweets **6** Quote Tweets **437** Likes

Algorithmic Auditing



Principle 3: Transparency

- **Explainability:** Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms. [fatml.org]
- **Accuracy:** Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures. [fatml.org]

Why is it hard?

- The benefit of transparency:
 - End Users: can understand the decision / result made by AI
 - Developers: can debug, tune, and optimize ML models
 - Companies: can generalize usage of ML by understanding the scope
- *“The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.”*

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Transparent Explanation Examples

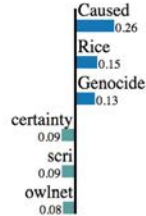
LIME: automatically generated explanations

Prediction probabilities

atheism	0.50
christian	0.43
religion.misc	0.05
mideast	0.02
Other	0.00

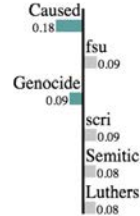
NOT atheism

atheism

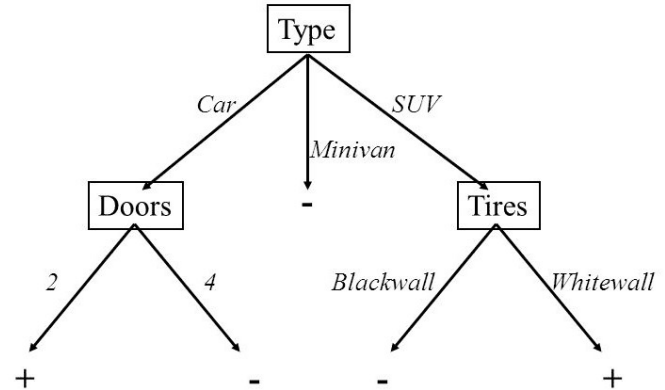


NOT christian

christian



Decision tree: inherently interpretable



Exercise: Transparent Hiring Algorithm

- When our “fair” hiring algorithm makes a decision, we need to tell applicants about the decision. In groups of 4, design a text-based explanation that will be provided to a candidate in a decision email that will be sent to them.
- Some food for thought:
 - Include both the procedures followed by the algorithm and the specific decisions that are made.
 - How much to “open up” about the algorithmic process & source of data?
 - Provide any channel for candidate feedback and appeal?
 - Use plain and intuitive language.

https://docs.google.com/forms/d/e/1FAIpQLSd27b5k1EMAf_40v6tpbncWoGrqX8LjqCUbLz9y54W93YPr2A/viewform?usp=sf_link

http://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

“The two hardest problems in computer science are: (i) people, (ii), convincing computer scientists that the hardest problem in computer science is people.

- Jeff Bigham (CMU)

Take-Home Messages

- Algorithms are inherently human.
- FAccT issues critically determine success of systems. They are not simple “cost” to pay and “nice” things to have.
- Human-machine collaboration could help.
- FAccT issues span all areas of computer science: theoretical guarantees, system performance, security & privacy, verification & testing, explainable AI, data processing techniques, usability & human-AI interaction

Useful Resources

- <https://cdt.org/issue/privacy-data/digital-decisions/>
- <https://www.fatml.org/resources/principles-for-accountable-algorithms>
- http://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf