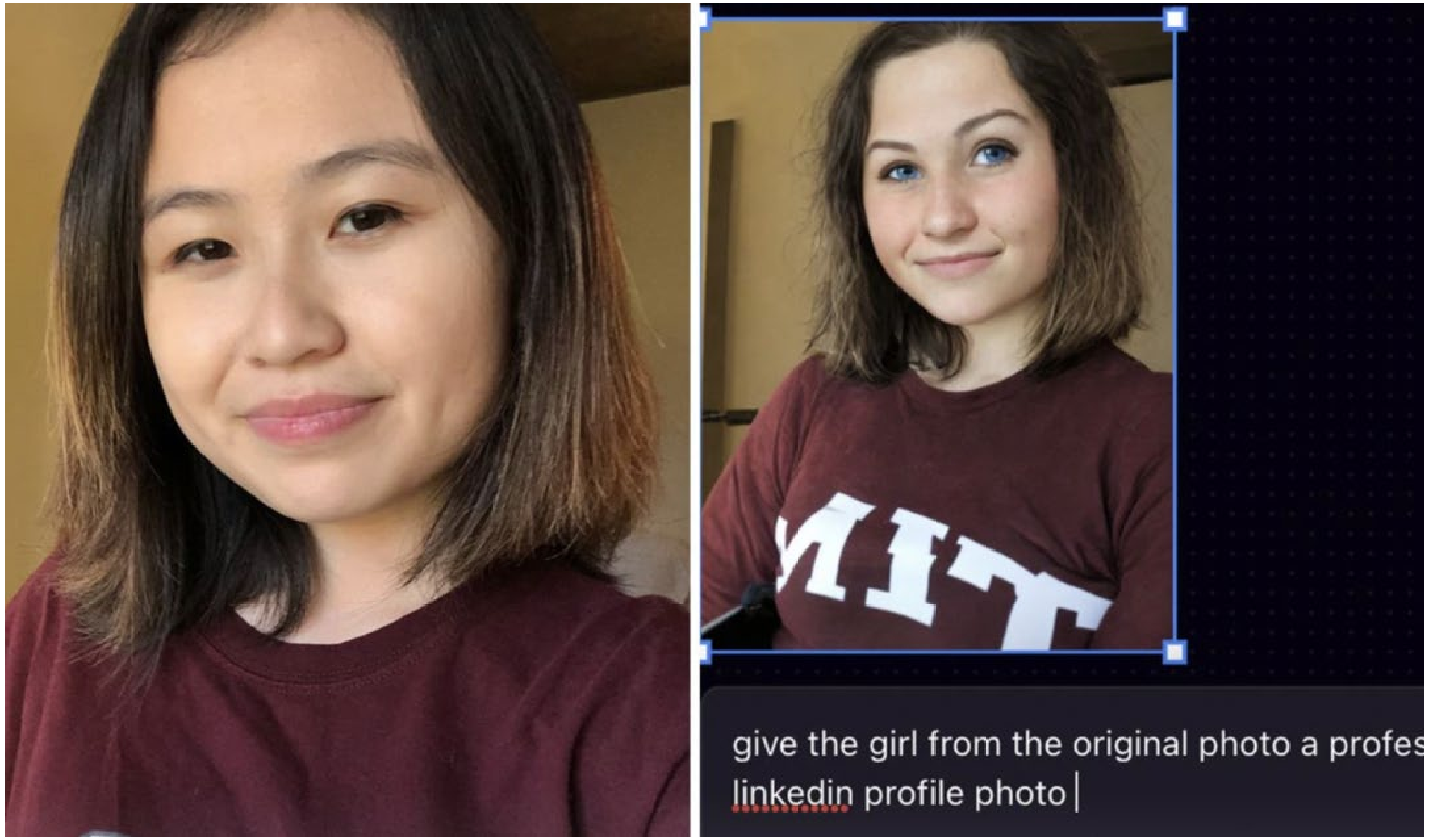
CS 640 Lecture 3:

Ethical and Societal Concerns in AI

Bias in AI Image Generation: MIT Graduate Asked AI Image Generating App "Playground AI" to Make Her Headshot More Professional -- It "Whitewashed" Her Instead



August 9, 2023

Credit:
Peopleofcolorintech.com

Joy Buolamwini, MIT Media Lab, 2017

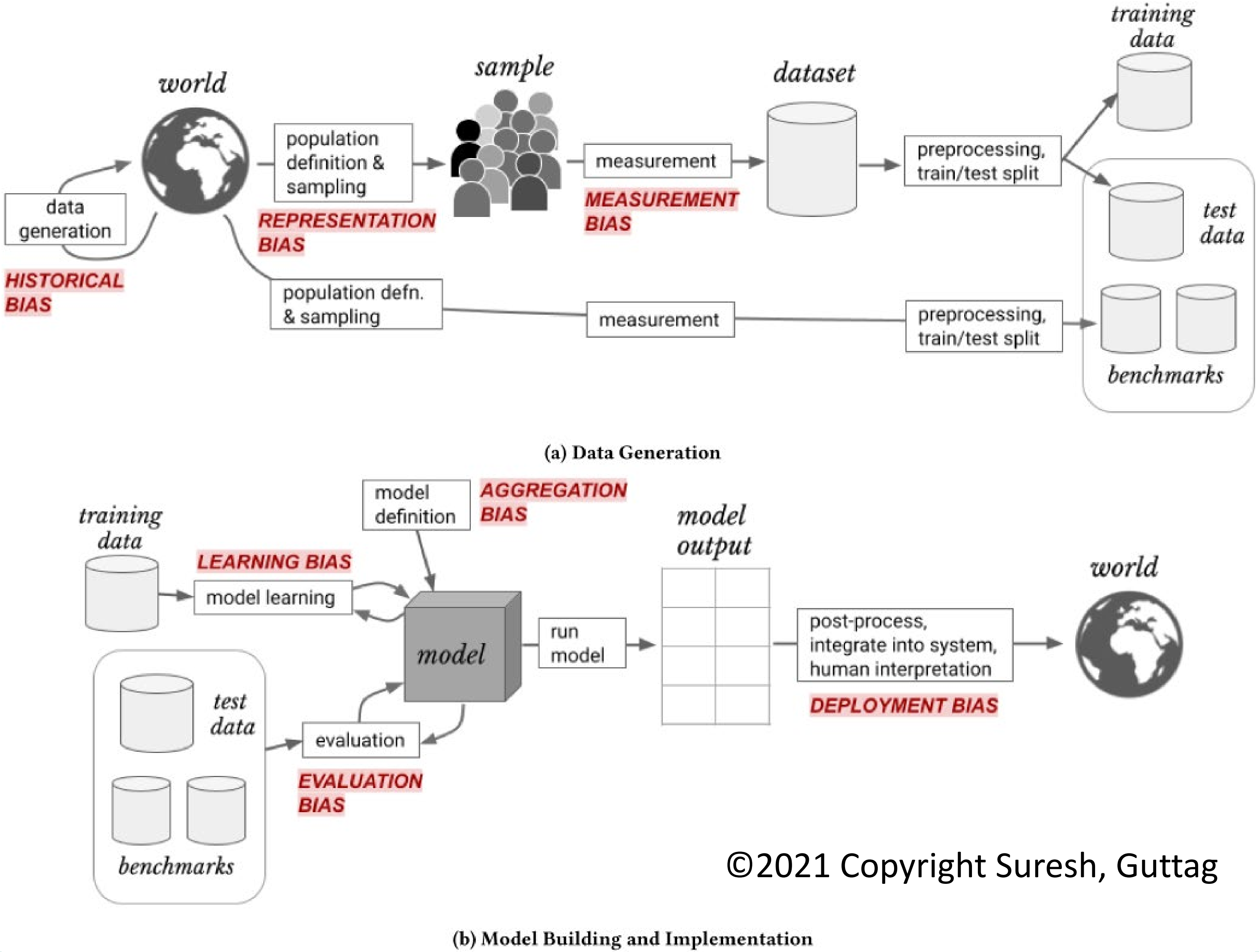
http://gendershades.org https://youtu.be/TWWsW1w-BVo

Racial and Gender Bias
in AI-based Face Detection

Joy asks for transparency and accountability

A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

[The 7 Sources of Harm in ML]



Suresh and Guttag

ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2021

Deployment and Representation Bias

"Facial recognition technology can expose political orientation from naturalistic facial images" by Michal Kosinski, 2021

- "We are aiming to study existing privacy threats, rather than develop new privacy-invading tools"
- Algorithm: Input: 224x224 cropped face. Converted by VGGFace2 to a 2048-dim feature vector, which is then compared to the average feature vector of liberals or conservatives.
- Dating website sample: 1,085,795. But preselection: 27% conservative, 23% liberal. 50% data not included.
- Argues: Even if one knew which transient facial features reveal political orientation and changed them, AI would circumvent this. "An arms race that humans are unlikely to win."

Representation Bias

Dating site studies – Extrapolation to other data?

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images by Y. Wang and M. Kosinski, 2018

- AI could correctly distinguish between gay and heterosexual men in 81% of cases, and in 71% of cases for women (human 61% vs. 54%).

Presentation in self-posted facial images can expose sexual orientation: Implications for research and privacy by Dawei Wang, 2022

- Differentiating features: Eyeglasses, brightness, background, only 3 values/color channel. Masking, blurring not sufficient.

AI may have picked up on a certain combination of features (glasses)

Measurement and Representation Biases

Automated Inference on Criminality using Face Images by Wu, Zhang, 2016

Data: 1856 ID photos. "Non-criminals" from internet photos, "Criminals" from police departments

Measurement bias: Police custody may cause facial expressions or damage.

Representation bias: Faces of people in custody are not representative of crime, but include criminals that have been caught, jailed, and photographed.

AI & Physiognomy? Assessing a person's criminality, character, or personality from the appearance of their face?

When Machine Learning Is Facially Invalid

Frank Pasquale, SEPTEMBER 2018 | VOL. 61 | NO. 9 | COMMUNICATIONS OF THE ACM 25

Goal of the Article:

Explore whether the ML research community should improve certain facial inference work or shun it: ML systems to

- detect a person’s sexual orientation & intelligence,
- infer a person’s political leaning,
- stereotype facial features of criminals.

When it comes to criminal law, extreme caution should be exercised with respect to the new physiognomy.

Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

Goal of the Study:

Evaluate potential racial bias of AI systems that recognize emotions by analyzing facial expressions in images

Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

Commercial AI Systems tested:

Face++: https://www.faceplusplus.com

Microsoft Face API:

https://azure.microsoft.com/en-us/services/cognitive-services/face

Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

Study data:

- Professional photos of 400 basketball players from the 2016 to 2017 NBA season
- Players appear similar in their clothing, athleticism, and age
- Players look at the camera in the picture

Emotion-reading tech fails the racial bias test

The Conversation, [Lauren Rhue](), January 3, 2019 6.23am

Example of study data:

Darren Collison and Gordon Hayward

Face++ detects:

Both players are smiling. Similar smile scores: 48.7 and 48.1 out of 100

Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

	Darren	Gordon
Smile Scores:	48.7	48.1
Emotions		
Happy	39	60
Angry	27	0.1

Face++

Face++ rated the emotions on facial expressions of basketball players out of 100. Black faces were, on average, rated as angrier and unhappier than white faces.

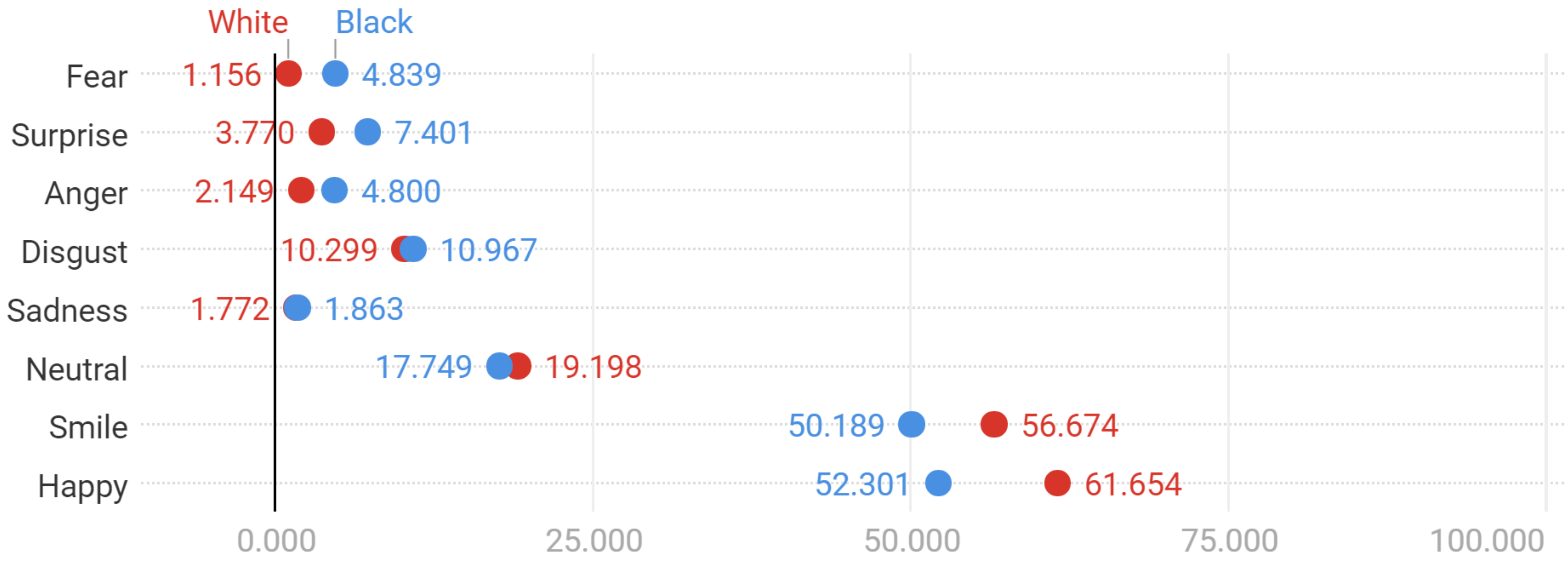


Chart: The Conversation, CC-BY-ND • Source: SSRN (2018) • Get the data

Face API

Face API rated the emotions on facial expressions of basketball players out of 100. White faces were seen, on average, as happier than black faces.

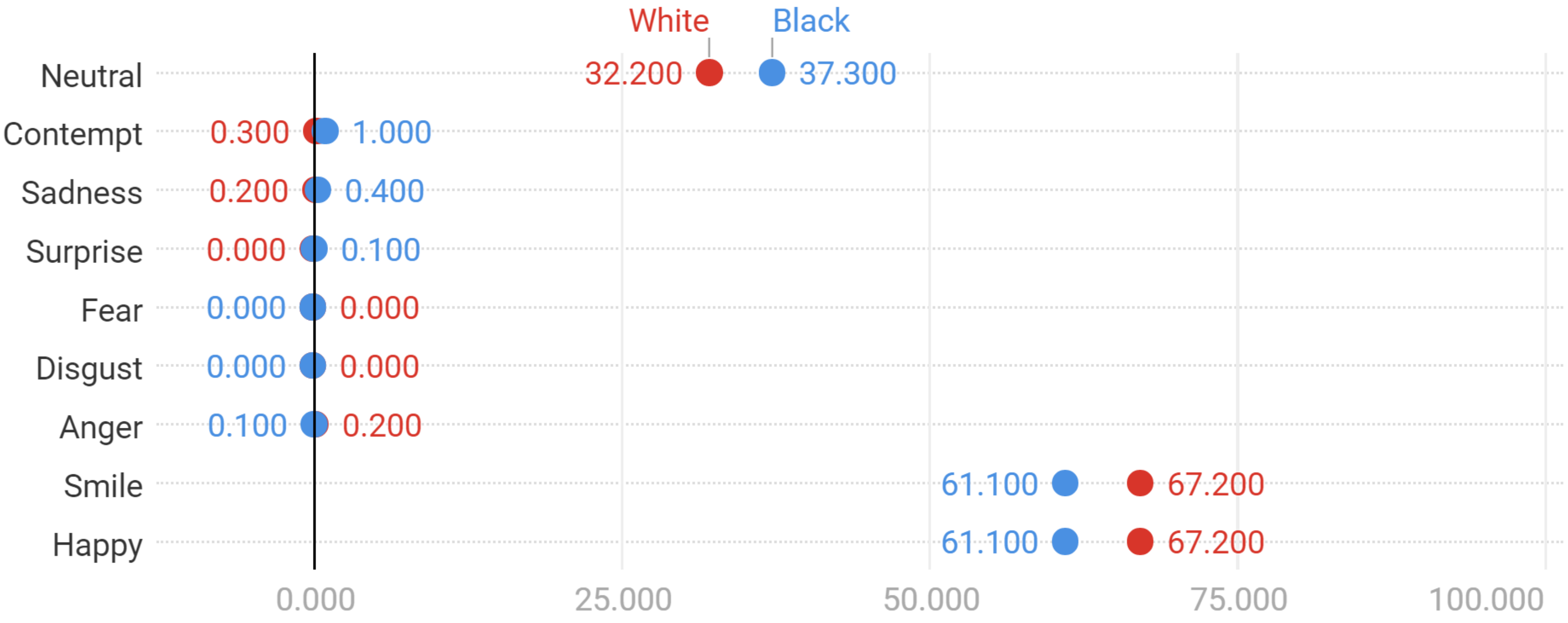


Chart: The Conversation, CC-BY-ND • Source: SSRN (2018) • Get the data

Lauren Rhue’s Analysis of her Study Results:

- Some research suggests that black professionals must amplify positive emotions to receive parity in their workplace performance evaluations.
- Some researchers argue that facial recognition technology is more objective than humans.
- Rhue’s study suggests that facial recognition reflects the same biases that people have.
- Black men’s facial expressions are scored with emotions associated with threatening behaviors more often than white men, even when they are smiling.
- The use of facial-analysis systems could formalize preexisting stereotypes into widely-used AI, automatically embedding them into everyday life.

Lauren Rhue’s Analysis of her Study Results:

Applications of commercial face analysis systems:

- Help companies with interviewing and hiring decisions.
- Scan faces in crowds to identify threats to public safety.

Until AI systems assess black and white faces similarly, black people may need to exaggerate their positive facial expressions – essentially smile more – to reduce ambiguity and potentially negative interpretations by the technology.

Lauren Rhue’s Analysis of her Study Results:

Although innovative, artificial intelligence can perpetrate and exacerbate existing power dynamics, leading to disparate impact across racial/ethnic groups.

Some societal accountability is necessary to ensure fairness to all groups because facial recognition, like most artificial intelligence, is often invisible to the people most affected by its decisions.

Evolution of LMs

- Larger is better? https://huggingface.co/blog/large-language-models

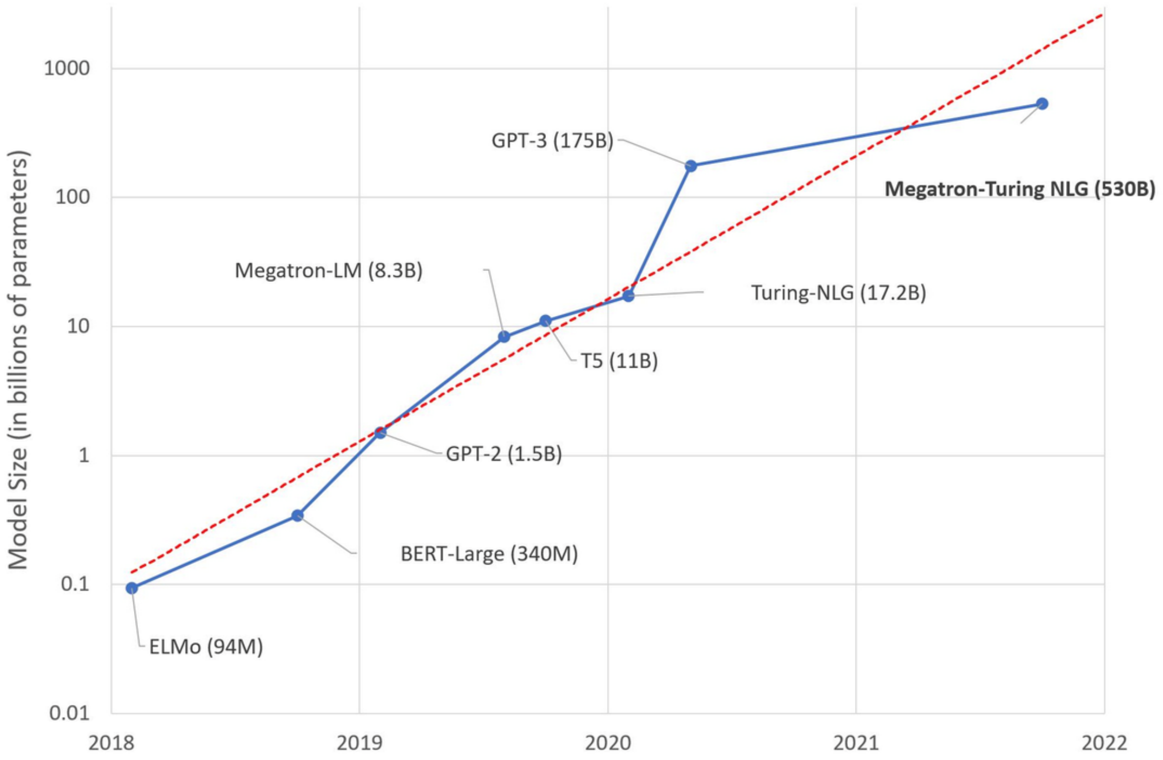
Published October 26, 2021.

Update on GitHub

juliensimon
Julien Simon Opinion piece

A few days ago, Microsoft and NVIDIA introduced Megatron-Turing NLG 530B, a Transformer-based model hailed as "*the world's largest and most powerful generative language model*."

This is an impressive show of Machine Learning engineering, no doubt about it. Yet, should we be excited about this mega-model trend? I, for one, am not. Here's why.



BOSTON UNIVERSITY

Bigger is better?

- Brain
  - Average: 86 billion neurons, 100 trillion synapses (not all about language)
  - GPT-4 is estimated to have 1.76 trillion parameters
- Megatron
  - Cost: 530 billion parameters, hundreds of DGX A100 multi-GPU servers (each cost $200k) + network + hosting + ... = total of $100 million dollars
  - Training cost: each DGX server can consume up to 6.5 kilowatts + cooling power = carbon footprint
  - BERT-base: 110 million parameters → carbon footprint = NY-SF flight

Carbon Footprint

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Carbon Footprint

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Carbon footprint of GPT-3?
552 metric tons of carbon emissions, equivalent to driving a passenger vehicle 1.24 million miles (2 million kilometers)

Carbon footprint of GPT-4?
Between 12,456 and 14,994 metric tons CO_2e

BOSTON UNIVERSITY

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Bender et al., 2021

What are the possible risks associated with LLMs and what paths are available for mitigating those risks?

Recommendations:

- weigh the environmental and financial costs first,
- invest resources into curating and carefully documenting datasets rather than ingesting everything on the web,
- carry out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values,
- encourage research directions beyond ever larger language models.

AI's Influence on Elections 2024 ?

The Economist, September 20203

- 4 billion people will vote in Britain, India, Indonesia, Mexico, Taiwan, USA in 2024
- FB on Election 2016: Russian government's 80,000 posts reached 126 million Americans, ~half the electorate
- Micah Musser, Georgetown U.: AI could save $3m of content generating in a $10m campaign
- Meta and X/Twitter cut safety teams
- Quiller helps campaigns write better fundraising emails

- Concerned:
  - Eric Schmidt, former Google CEO: "the 2024 elections are going to be a mess because social media is not protecting us from false generative AI."
  - Sam Altman, CEO of Open AI: "nervous about the impact AI is going to have on future elections (at least until everyone gets used to it)"
- Not so concerned:
  - Jacky Chang, CTO Biden 2020: "The problem is more with demand than supply."
  - Brendan Nyhan, Dartmouth: "We still have not one convincing case of a deepfake making any difference whatsoever in politics."

Regulation of AI

European Union:

- 2021: “The Artificial Intelligence Act” proposed by the European Commission
- 2022: “General approach position” on the AI Act adopted by the European Council
- June 2023: Amendments adopted by the European Parliament
- Now: Negotiation between European Commission and member states

USA:

- June 2023: Hearings in US Congress on AI
- July 2023: Federal Trade Commission investigation into ChatGPT

China:

- “AI algorithms must be registered with a government body and somehow embody core socialist values” according to The Economist, Sep. 2023

Researchers are BU are helping with the process of AI regulation

National Telecommunications and Information Administration request for comments on AI accountability: "What policies can support the development of AI audits, assessments, certifications and other mechanisms to create earned trust in AI systems?"

Boston University & Chicago University researchers submitted: NTIA-2023-0005-1268

1. AI accountability must be implemented through the entire lifecycle of systems.
2. Accountability mechanisms must be both robust and broadly accessible.
3. Access and transparency are consistent with protecting privacy and intellectual property rights.
4. Accountability and transparency mechanisms are a necessary but not sufficient aspect of AI regulation.
5. AI regulation requires rules for both generalized and specific contexts; we recommend collaboration between specialized agencies and a meta-agency with AI-specific expertise.

BU Course on Responsible AI, Law, Ethics & Society

- Spring 2024
- https://learn.responsibly.ai
- CDS DS 682
- Counts toward “MS in AI” degree as an elective
- Taught together with UC Berkeley

  by Shlomi Hod and other: https://shlomi.hod.xyz/