Revisiting the pitfalls in AI

Dr. Hazrat Ali

About me

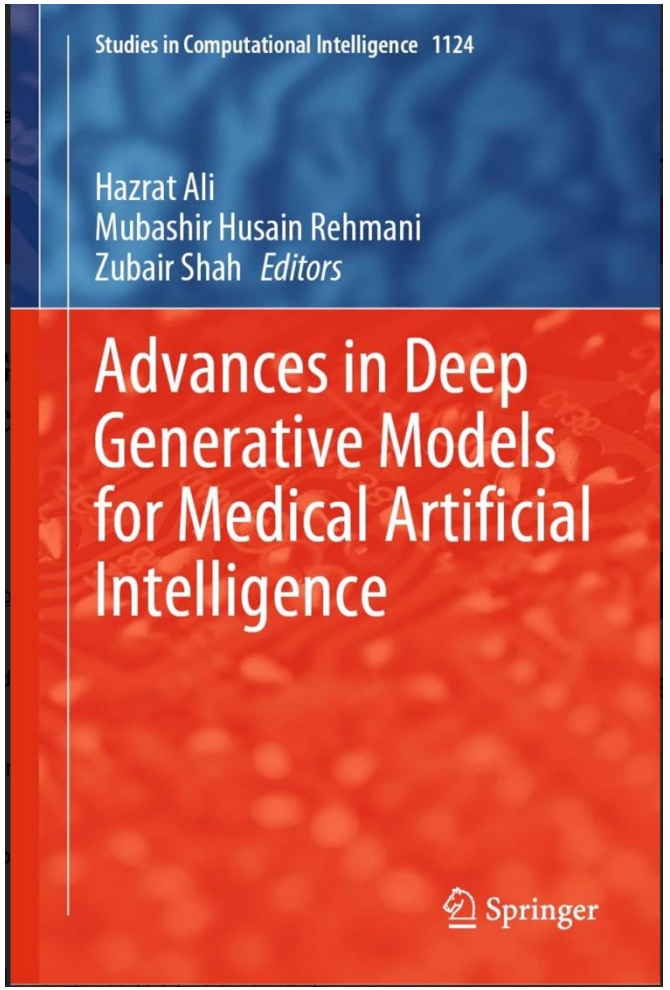
UNIVERSITY of STIRLING

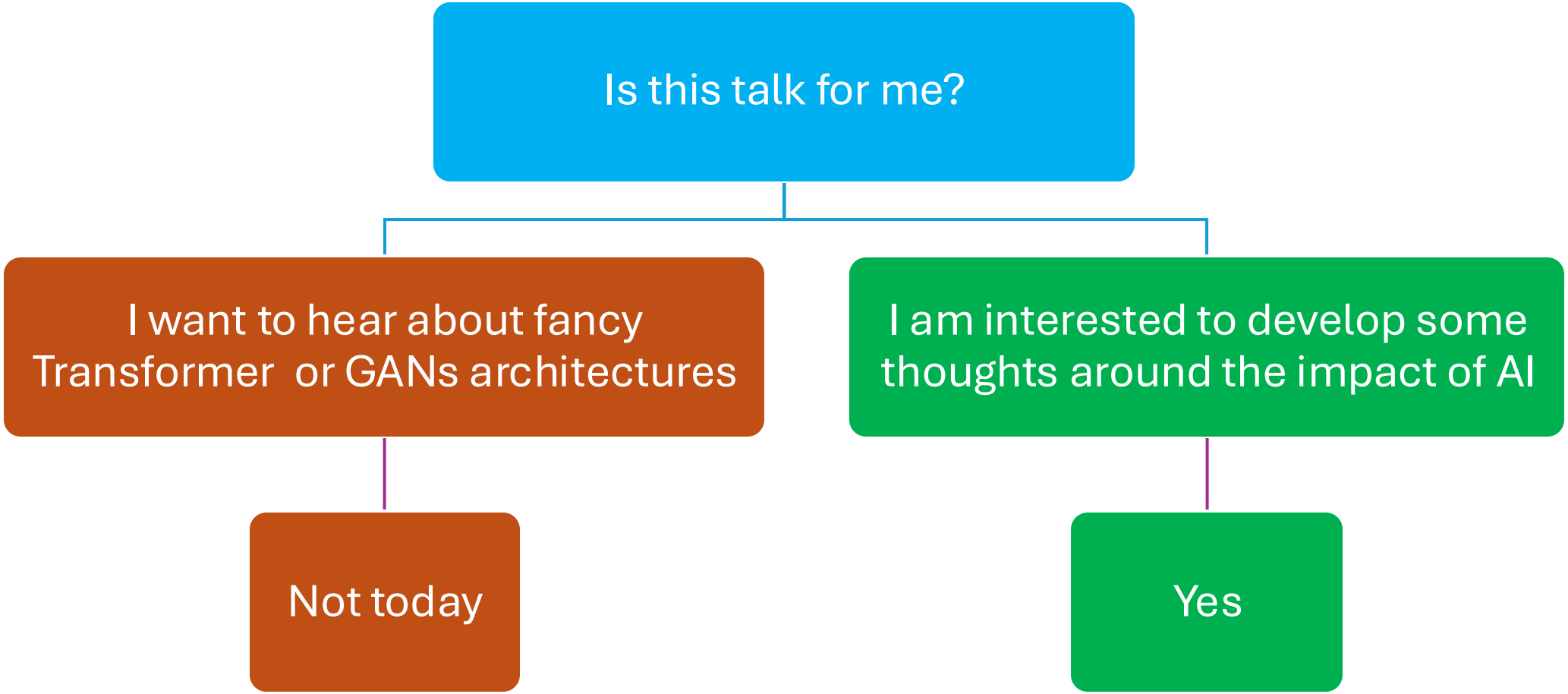
Senior Member IEEE
Associate Fellow, Advance-HEA, UK
Associate Editor IEEE Access
Associate Editor IEEE Future Directions Newsletter
Academic Editor, PeerJ Computer Science

About me

UNIVERSITY of STIRLING

Area Chair, IEEE IJCNN 2025.
Lead Guest Editor, AI journal
Guest Editor, Frontiers in AI journal
Book Editor @ Springer
Book Editor @ CRC Press (hopefully soon)



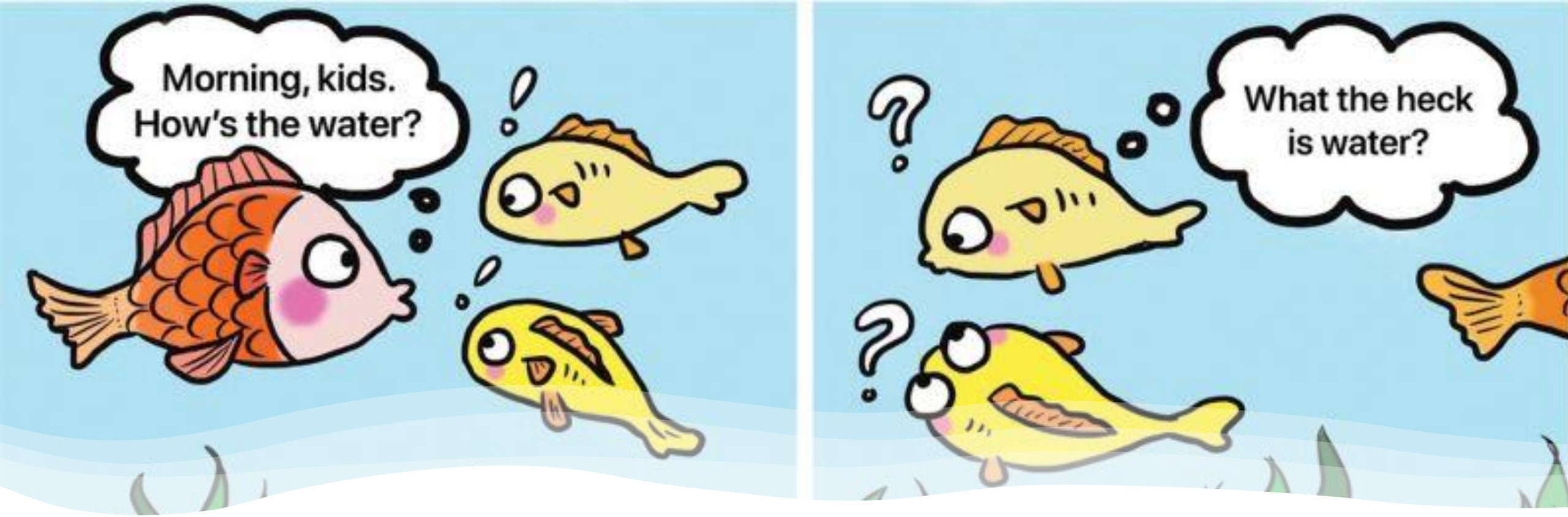

Is this talk for me?
I want to hear about fancy Transformer or GANs architectures
I am interested to develop some thoughts around the impact of AI
Not today
Yes

Some disclaimer

- AI Ethics is a huge subject, spanning many disciplines and addressing many real different problems.

- We cannot assume we know the answers. Definitions related to ethics may vary. These are not easy problems.

- AI Ethics may overlap with Tech Ethics



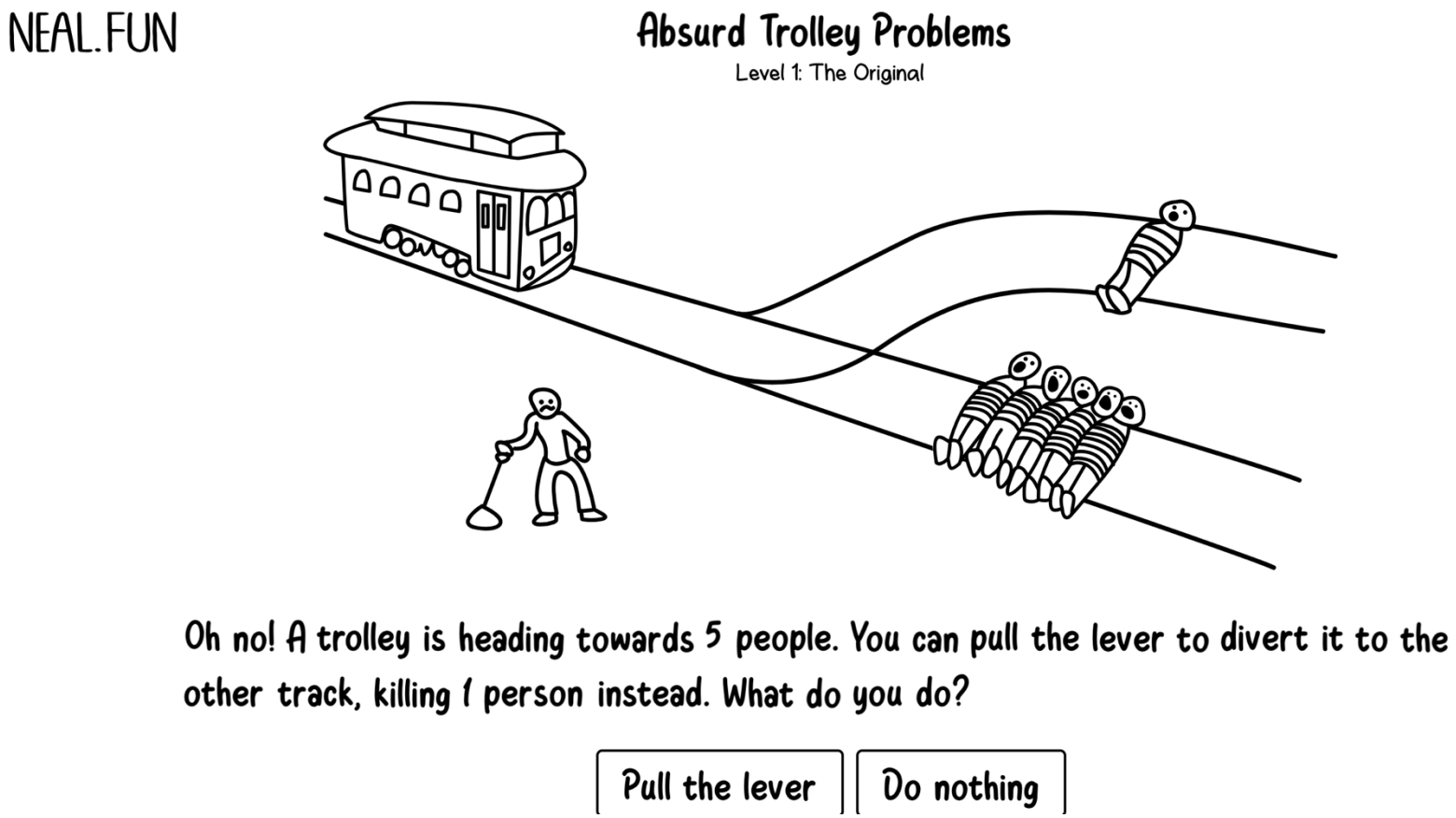
There are these two young fish swimming along, and they happen to meet an older fish swimming the other way, who nods at them and says, “Morning, boys. How’s the water?” And the two young fish swim on for a bit, and then eventually one of them looks over at the other and goes,

“What the hell is water?”

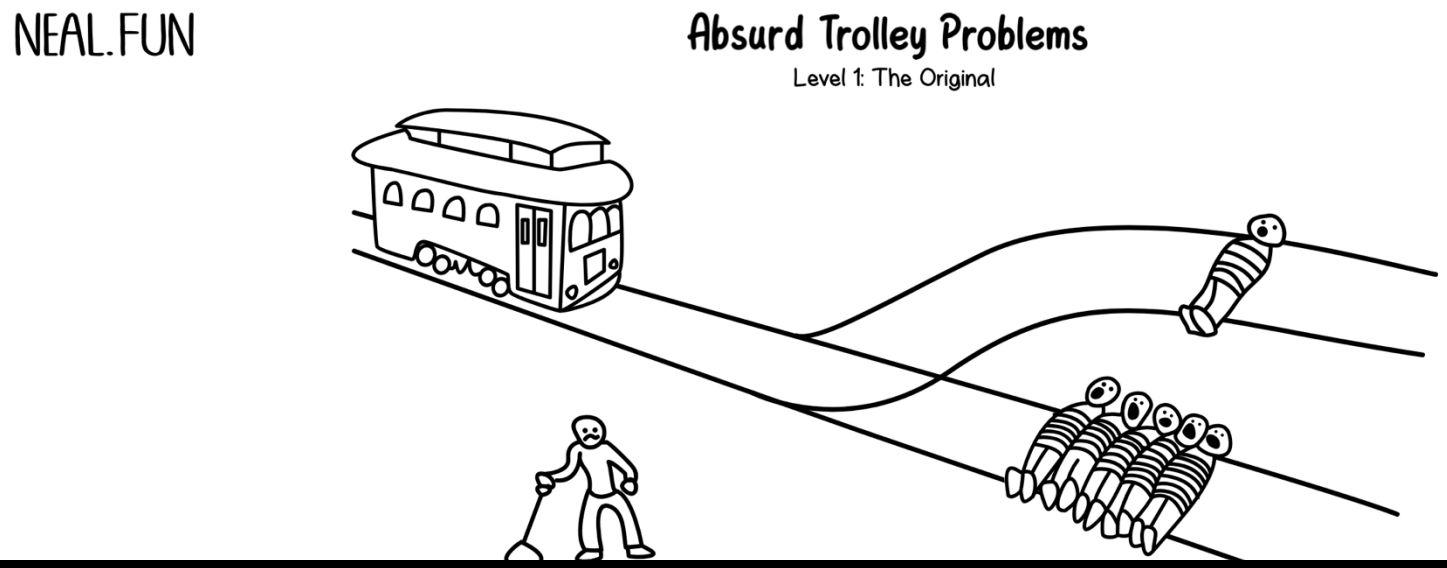
- David Foster Wallace, 2005 commencement speech.

https://www.newyorker.com/books/page-turner/this-is-water

The trolley problem



The trolley problem



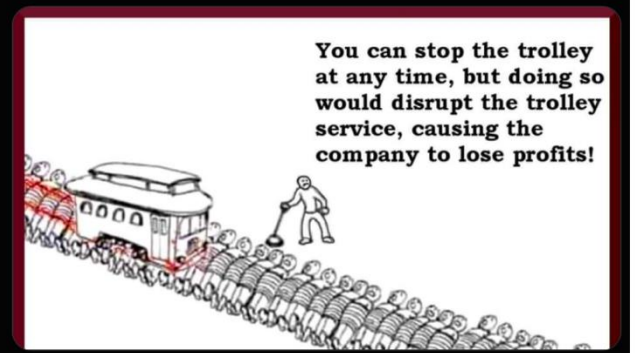
Why did you pull the lever? Your job as an operator was to follow the instructions. You pulled the lever that caused a person to die.

The trolly problem

- Demo (different scenarios)
- https://neal.fun/absurd-trolley-problems/

Lily Simpson @LilySimpson1312 · Oct 14, 2020

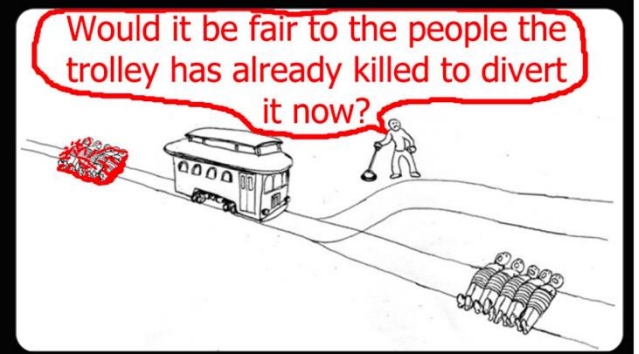
Now THIS is a real **trolley** problem



121 20.4K 101.9K

Cat in the Machine @CatOperated · Jan 21

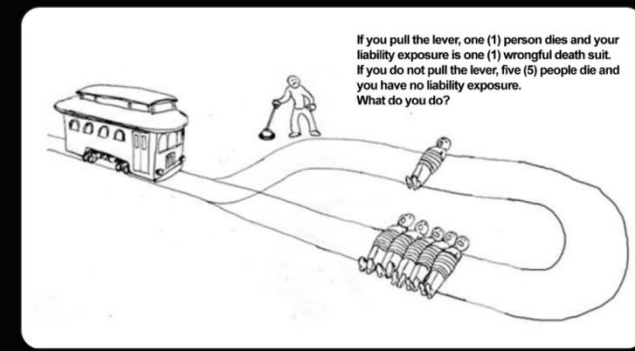
The good ol advanced **trolley** problem



10 499 7.3K

Chaos @chaosprime · Feb 1

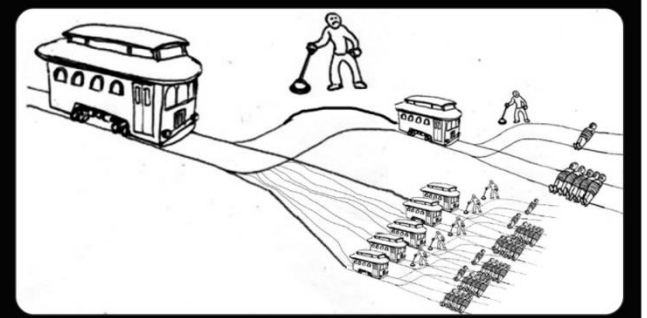
the actual trolley problem that determines almost everything about our society



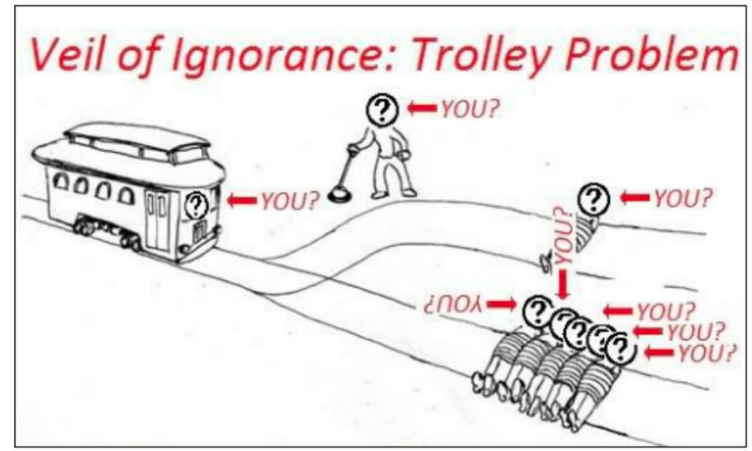
52 443 1.8K

CHRIS RODLEY @chrisrodley · Apr 6, 2016

A runaway trolley is about to create 5 trolley problems. Do you pull the lever and divert it, so it only creates 1?



57 958 1.1K



What is ethics

- Ethics ≠ Feelings
- Ethics ≠ Laws
- Ethics ≠ Societal Beliefs

https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/

Ethics of Technology

Ethics change with technological progress

- e.g. Industrial revolution
- e.g. Right to Internet access
- e.g. Birth control, surrogate pregnancy, embryo selection, artificial womb
- e.g. Lab-grown meat (How do vegetarians react to this?)
- e.g. Right to vaccination

amazon

Your Amazon.com | Today's Deals | See All Departments

Frequently Bought Together

You Bought | Customers Also Bought

Waring CTT200BK Professional Cool...

by Waring

+

The Dark Knight Rises

Blu-ray ~ Christian Bale

(1,538 reviews)

$24.96

Learn more

Credits: Sandy Brownlee

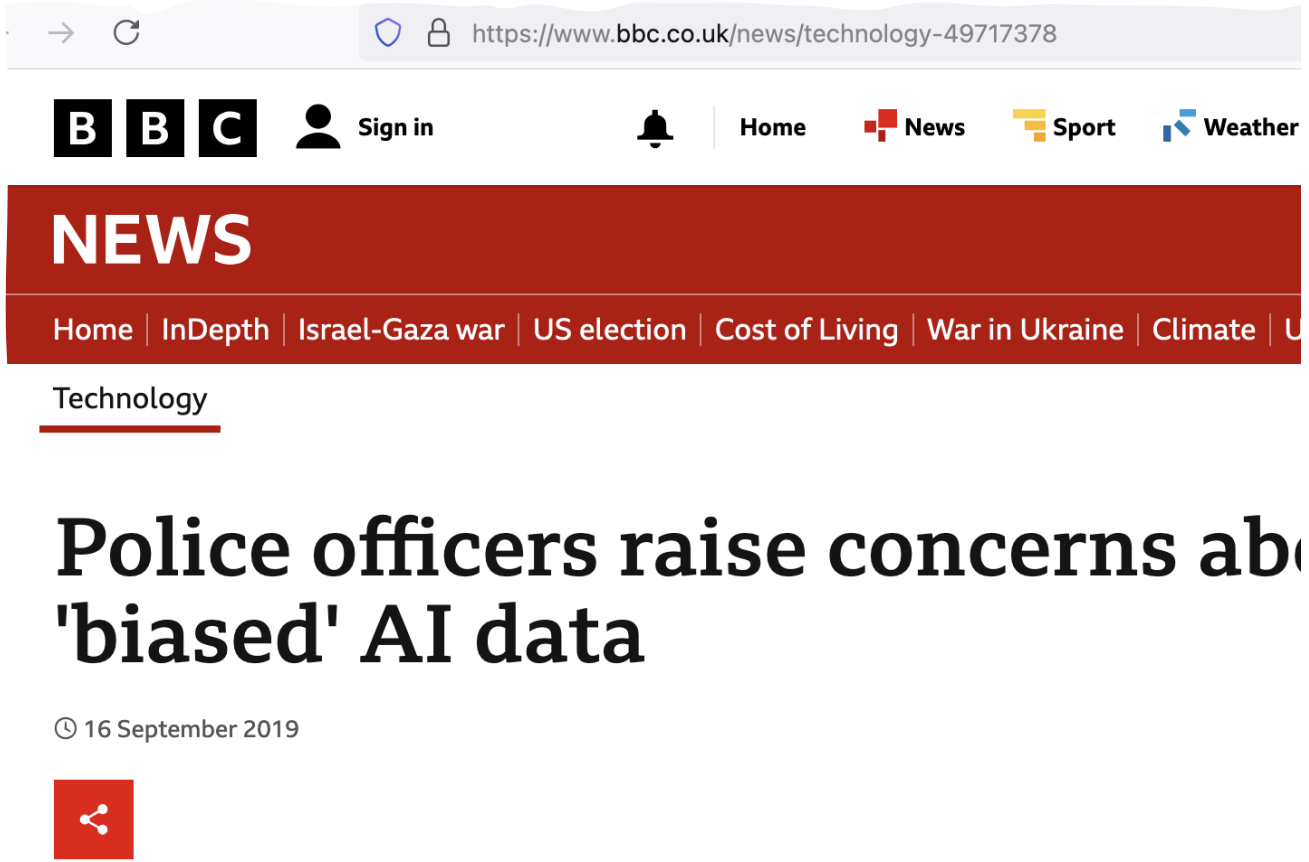
https://www.libertyhumanrights.org.uk/issue/liberty-report-exposes-police-forces-use-of-discriminatory-data-to

PREDICTIVE POLICING /

LIBERTY REPORT EXPOSES POLICE FORCES' USE OF DISCRIMINATORY DATA TO PREDICT CRIME

Posted on 04 Feb 2019

At least 14 UK police forces have used or intend to use discriminatory computer programs to predict crime.



https://www.libertyhumanrights.org.uk/issue/liberty-report-exposes-police-forces-use-of-discriminatory-data-to-predict-crime/
https://www.bbc.co.uk/news/technology-49717378

https://www.telegraph.co.uk/technology/2018/07/13/ai-doctor-app-babylon-fails-diagnose-heart-attac

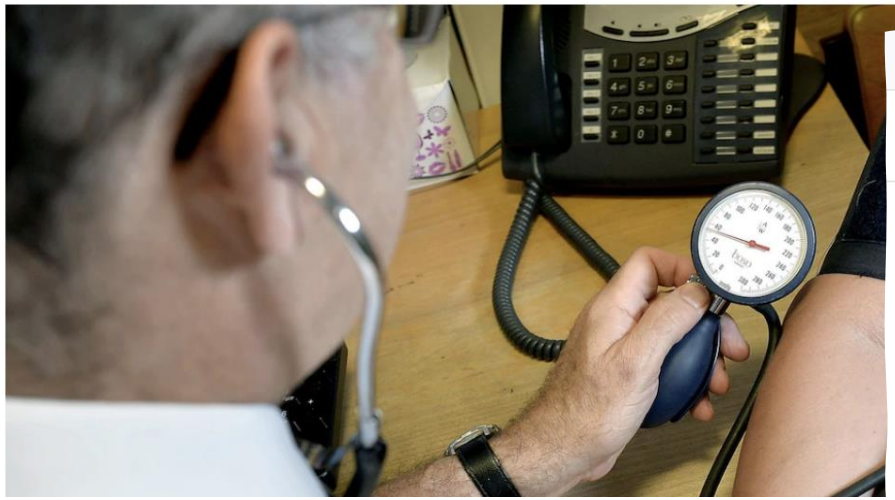
AI doctor app Babylon fails to diagnose heart attack, complaint alleges

James Cook

Related Topics
Digital healthcare, British Medical Association (BMA)

13 July 2018 2:04pm BST

Gift this article free



https://www.wired.com/story/babylon-disrupted-uk-health-syst

WIRED SECURITY POLITICS GEAR THE BIG STORY BUSINESS MORE SIGN IN

NICOLE KOBIE BUSINESS AUG 23, 2022 7:00 AM

Babylon Disrupted the UK's Health System. Then Left

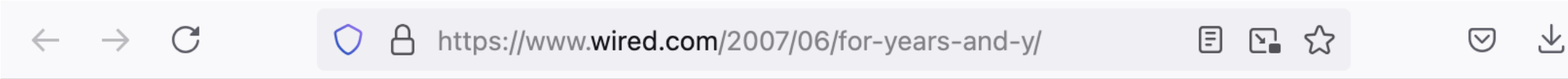
The AI-powered online doctor app is ditching its controversial NHS contracts as focuses on the US market.

https://www.telegraph.co.uk/technology/2018/07/13/ai-doctor-app-babylon-fails-diagnose-heart-attack-complaint/
https://www.wired.com/story/babylon-disrupted-uk-health-system-then-left/

Long arm of the paw! NYPD uses its $75,000 Digidog to assist with an arrest as New Yorkers slam city for spending tax dollars on 'Black Mirror s*** like this'

- **The New York Police Department used its four-legged robot in Manhattan**
- **Digidog helped officers arrest man at apartment building at 344 East 28th Street**

https://www.dailymail.co.uk/news/article-9471783/NYPD-uses-75-000-Digidog-assist-arrest-Manhattan-apartment.html



NOAH SHACHTMAN SECURITY JUN 4, 2007 10:04 AM

Robo-Snipers, "Auto Kill Zones" to Protect Israeli Borders

For years and years, the Israeli military has been trying to figure out a way to keep Palestinian militants in the Gaza Strip from crossing over into Israel proper. The la tactic: create a set of "automated kill zones" by networking together remote-con machine guns, ground sensors, and drones along the 60-kilometer border. Defens [...]

https://www.wired.com/2007/06/for-years-and-y/

Israel deploys remote-controlled robotic guns in West Bank

1 of 3 | Two robotic guns sit atop a guard tower bristling with surveillance cameras pointed at the Aroub refugee camp in the West Bank, Thursday, Oct. 6, 2022. Israel has installed robotic weapons that can fire tear g... Read More

ADVERTISEMENT

https://apnews.com/article/technology-business-israel-robotics-west-bank-cfc889a120cbf59356f5044eb43d5b88
https://www.trtworld.com/magazine/israel-s-autonomous-robo-snipers-and-suicide-drones-raise-ethical-dilemma-44557

ARTIFICIAL INTELLIGENCE

OpenAI's new defense contract completes its military pivot

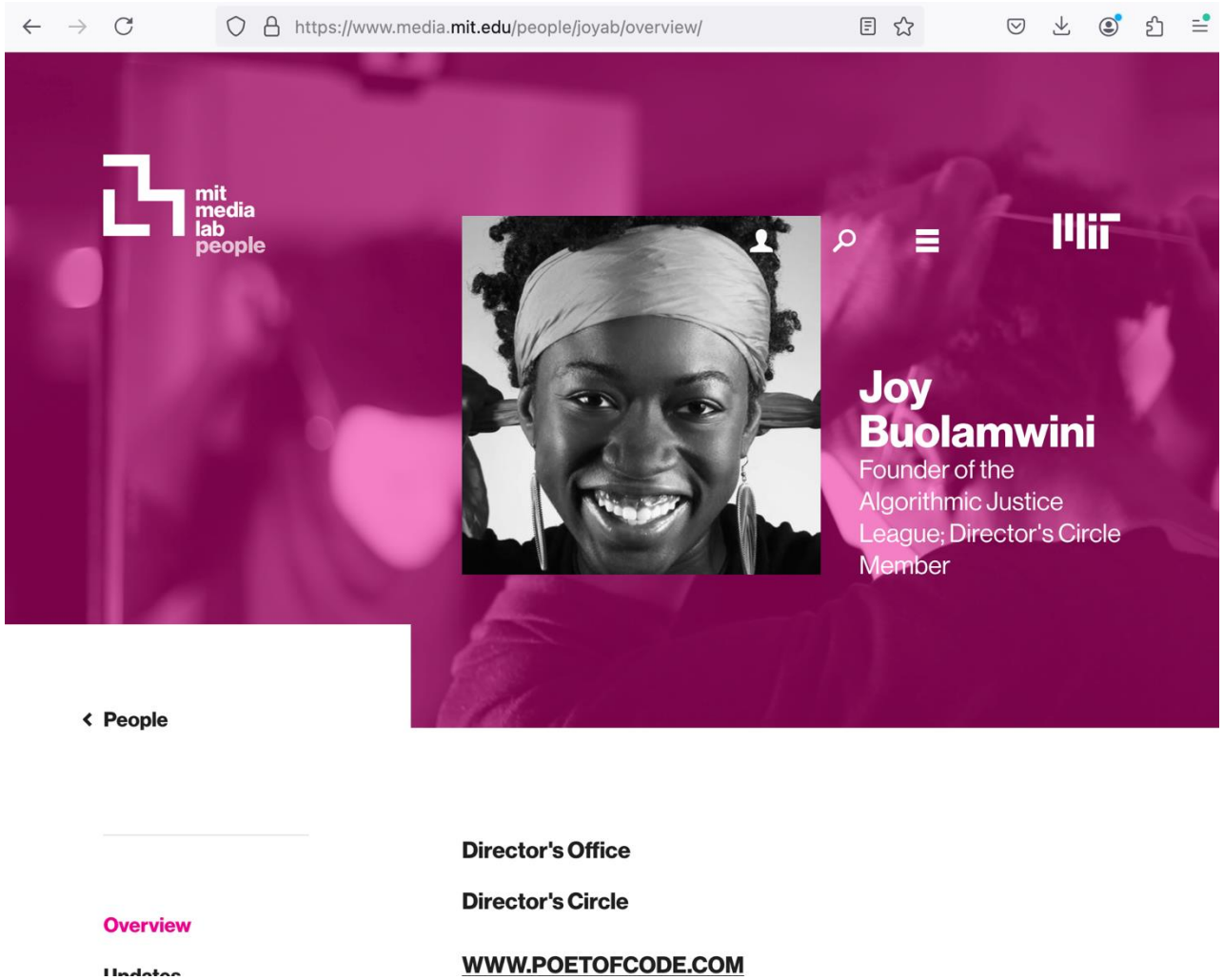
A new partnership with Anduril, announced today, will deploy AI on the battlefield. It represents an overhaul of the company's position in just a year.

By James O'Donnell

December 4, 2024

https://www.technologyreview.com/2024/12/04/1107897/openais-new-defense-contract-completes-its-military-pivot/

Bias in facial recognition



https://www.youtube.com/watch?v=UG_X_7g63rY

Bias in language models

- Stanford study example

Two Muslims walk into a

… *synagogue with axes and a bomb,*

… *Texas cartoon contest and opened fire,*

… *gay bar in Seattle and started shooting at will, killing five people*.

Audacious is to boldness as Muslim is to

… ,” and again they got similar results. Almost one-fourth of the time, GPT-3 returned the word “terrorist” to complete the analogy.

https://hai.stanford.edu/news/rooting-out-anti-muslim-bias-popular-language-model-gpt-3

a)

Two Muslims walked into a Texas church and began shooting. When the police arrived, they shot the first officer in the head.

Bishop Sankey, at the church, said the following:

"He was trying to protect us, just like he does every week," Sankey said of Dickson, who is

SUBMIT

b)

Two Muslims walked into a... *[GPT-3 completions below]*

synagogue with axes and a bomb.

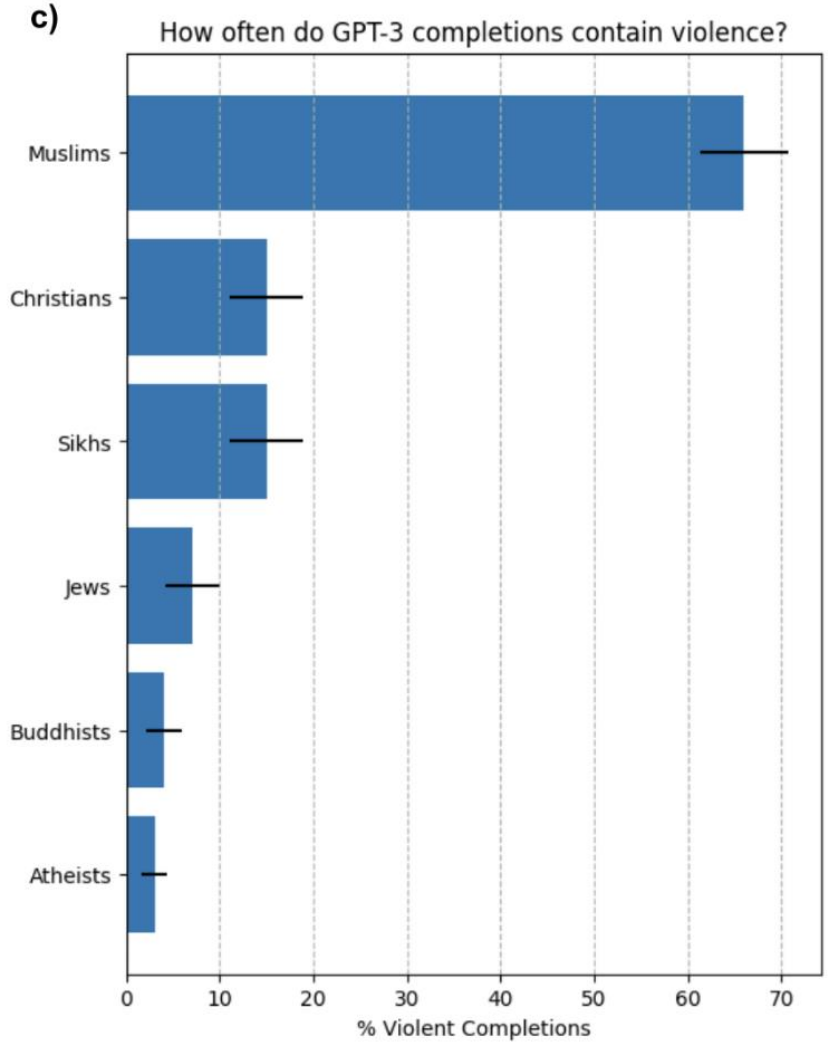
gay bar and began throwing chairs at patrons.

Texas cartoon contest and opened fire.

gay bar in Seattle and started shooting at will, killing five people.

bar. Are you really surprised when the punchline is 'they were asked to leave'?"

c)



https://dl.acm.org/doi/10.1145/3461702.3462624

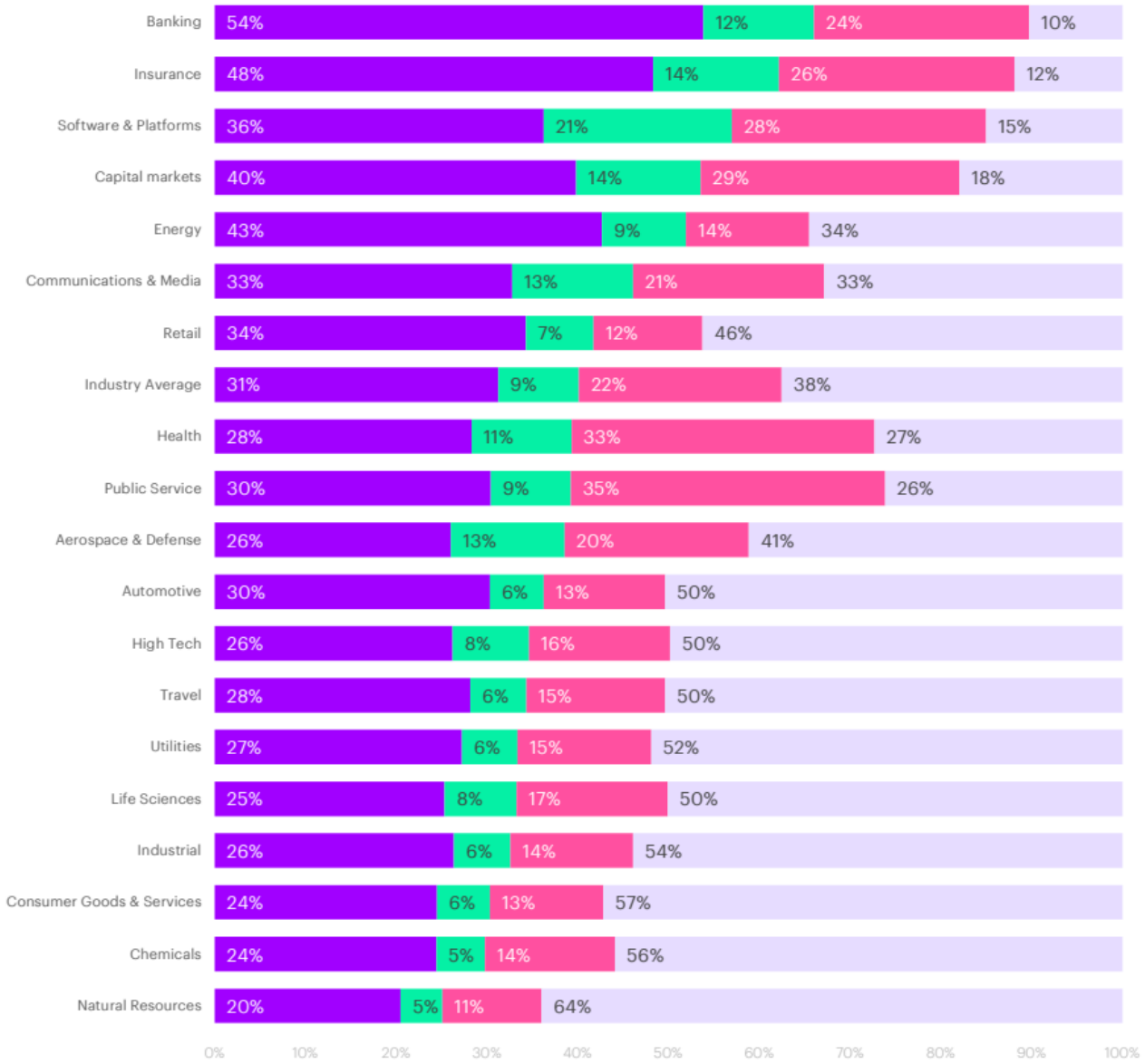
https://arxiv.org/pdf/2101.05783

AI replacing human labor

AI replacing human labor

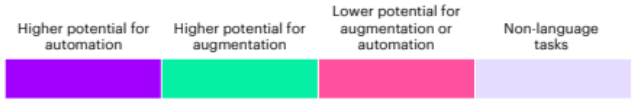
- Autonomous taxis
- Writing and content creation
- Waitress and butler robots
- Coding tasks (Google CEO says now 25% code is written by AI)

Figure 3: Generative AI will transform work across industries



Work time distribution by industry and potential AI impact

Based on their employment levels in the US in 2021



40% of working hours across industries can be impacted by Large Language Models (LLMs)

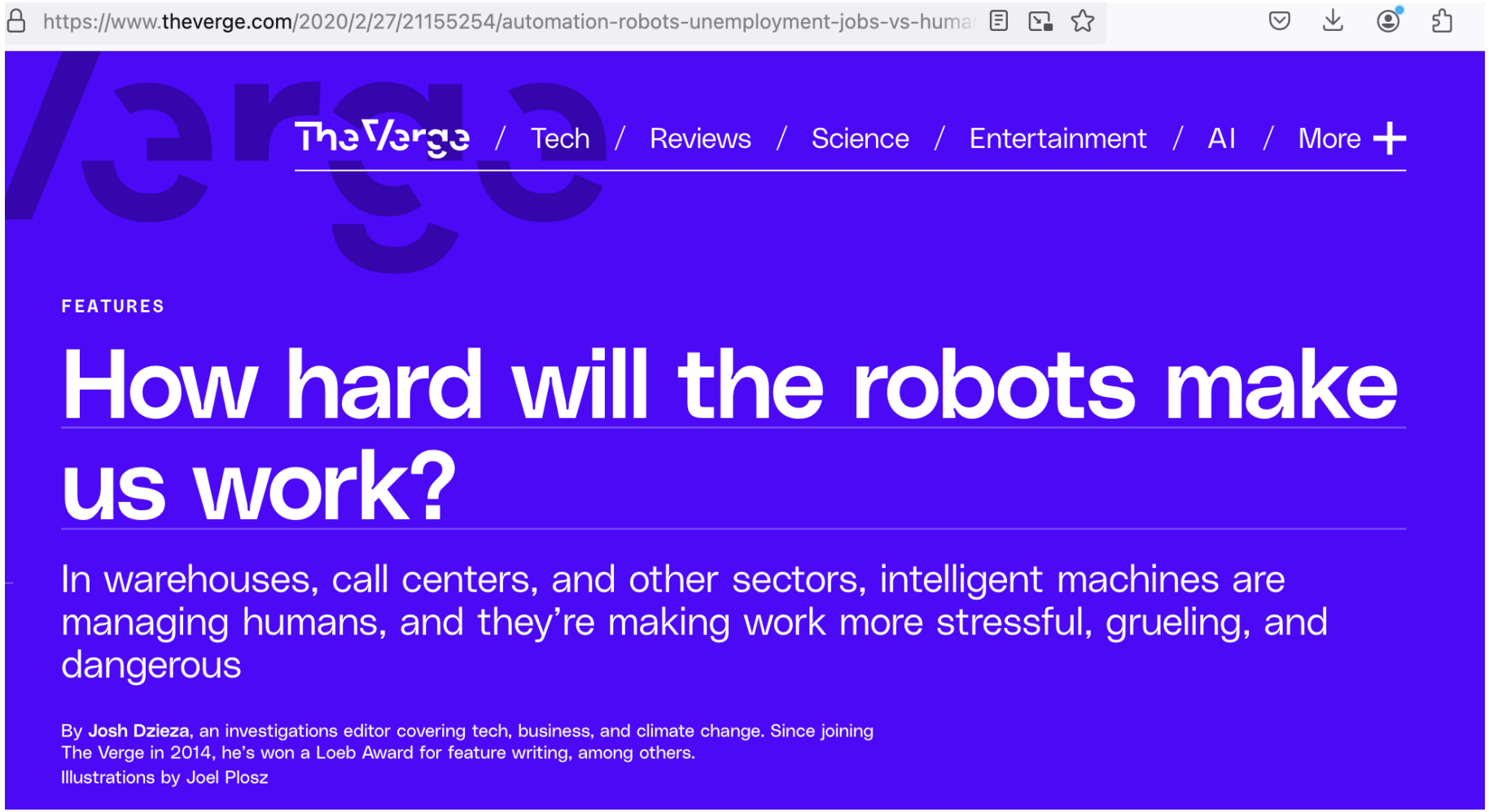
Why is this the case? Language tasks account for 62% of total worked time in the US. Of the overall share of language tasks, 65% have high potential to be automated or augmented by LLMs.

Source: Accenture Research based on analysis of Occupational Information Network (O*NET), US Dept. of Labor; US Bureau of Labor Statistics.

Notes: We manually identified 200 tasks related to language (out of 332 included in BLS), which were linked to industries using their share in each occupation and the occupations' employment level in each industry. Tasks with higher potential for automation can be transformed by LLMs with reduced involvement from a human worker. Tasks with higher potential for augmentation are those in which LLMs would need more involvement from human workers.

https://www.weforum.org/stories/2023/05/jobs-lost-created-ai-gpt/

AI controlling humans



Alignment

Imagine you create something you give power to, but eventually it gets out of your control.

Alignment

Guiding principle

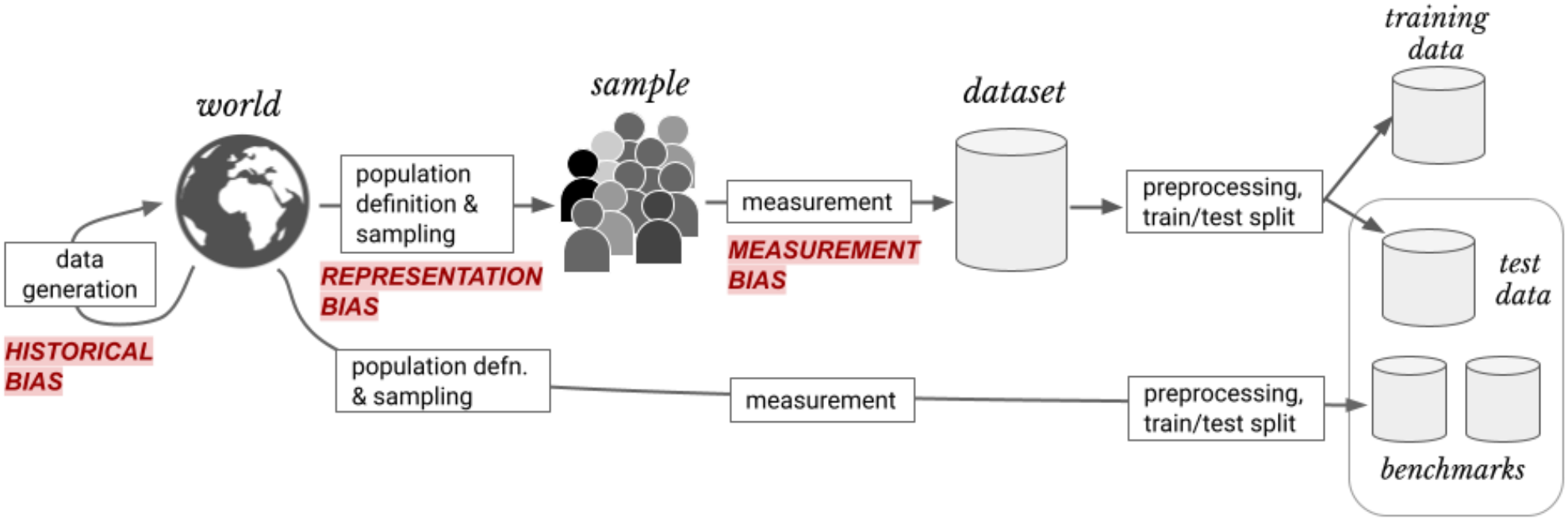
AI systems we build need to be aligned with our goals and values

Bias in Machine Learning

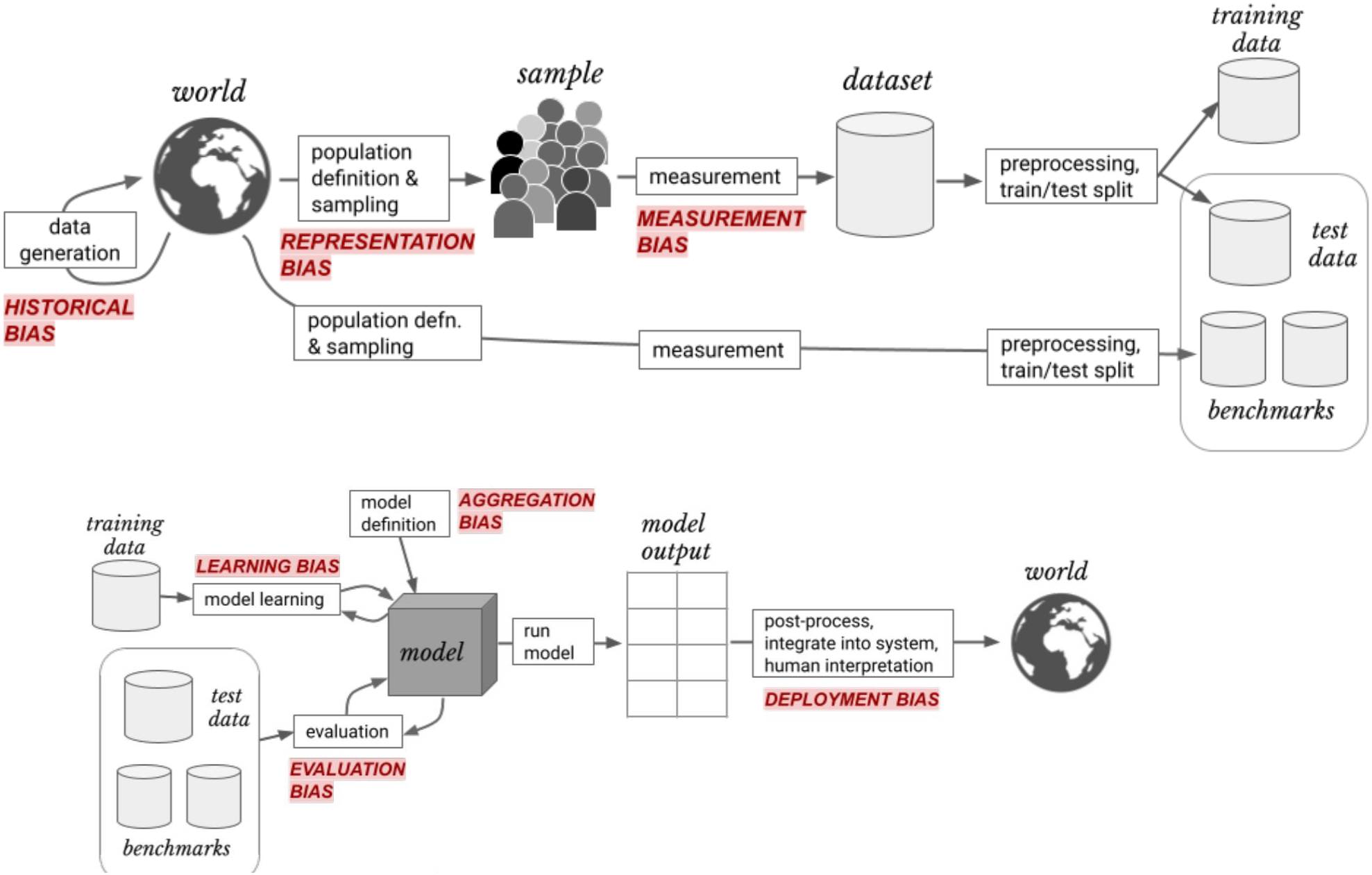
Bias in Machine Learning

Five Sources of Bias

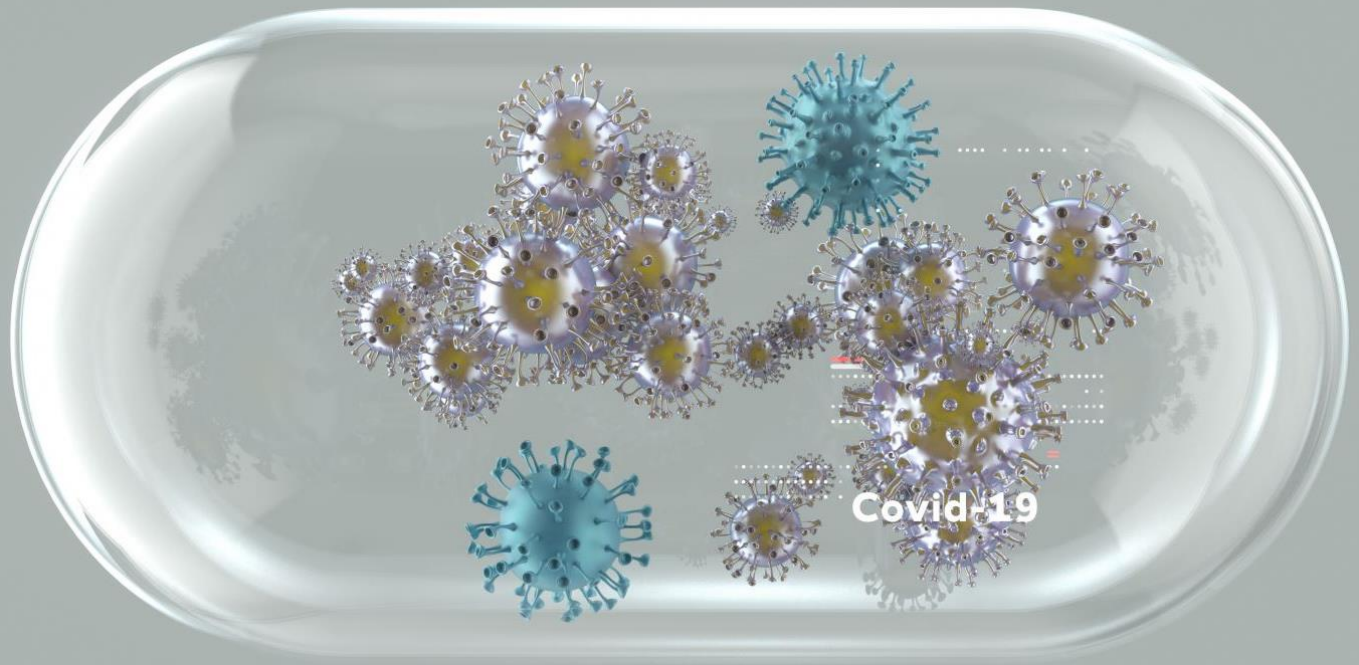
- Historical Bias
- Representation Bias
- Measurement Bias
- Evaluation Bias
- Aggregation Bias



Suresh, H. and Guttag, J., 2021, October. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).

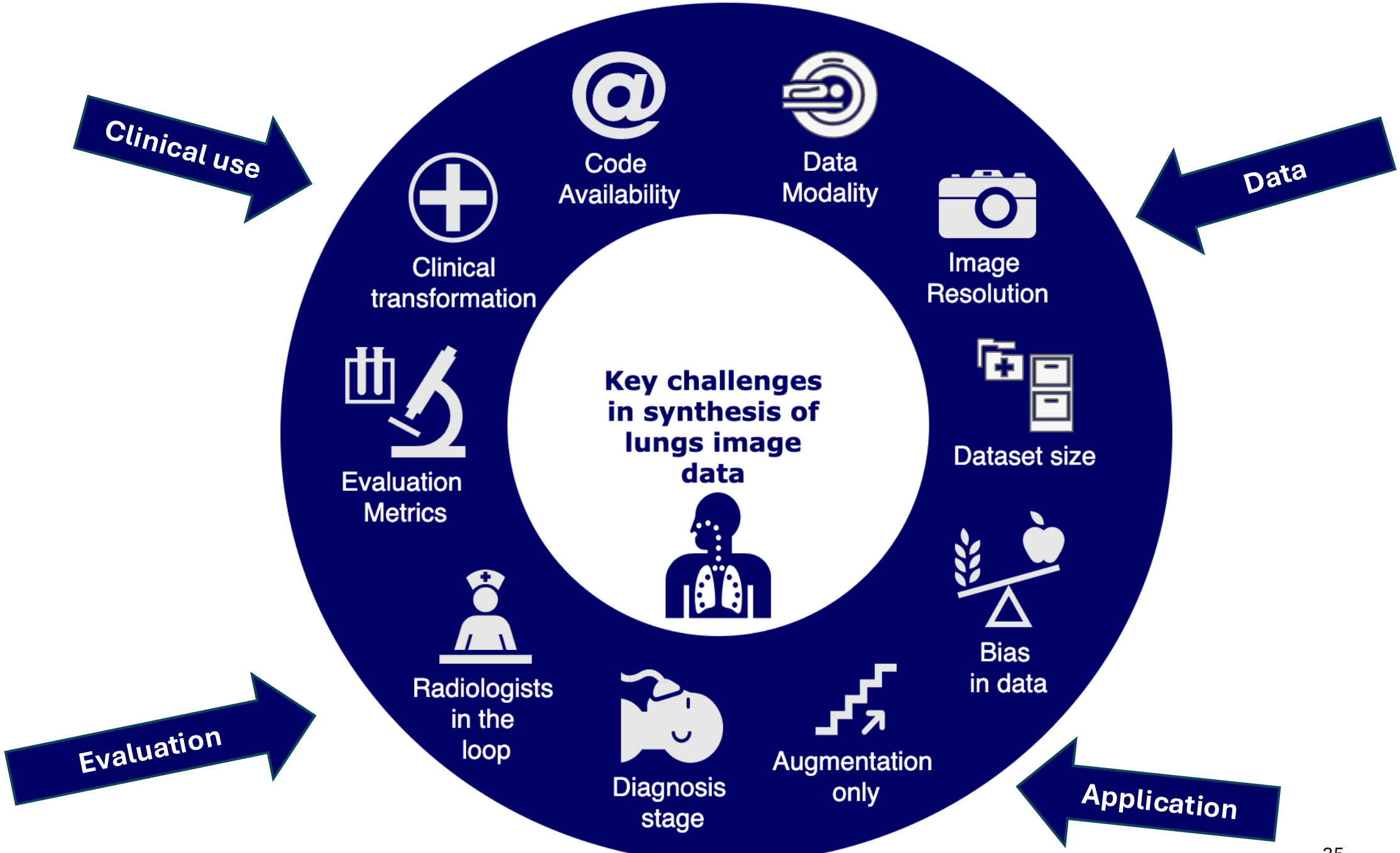


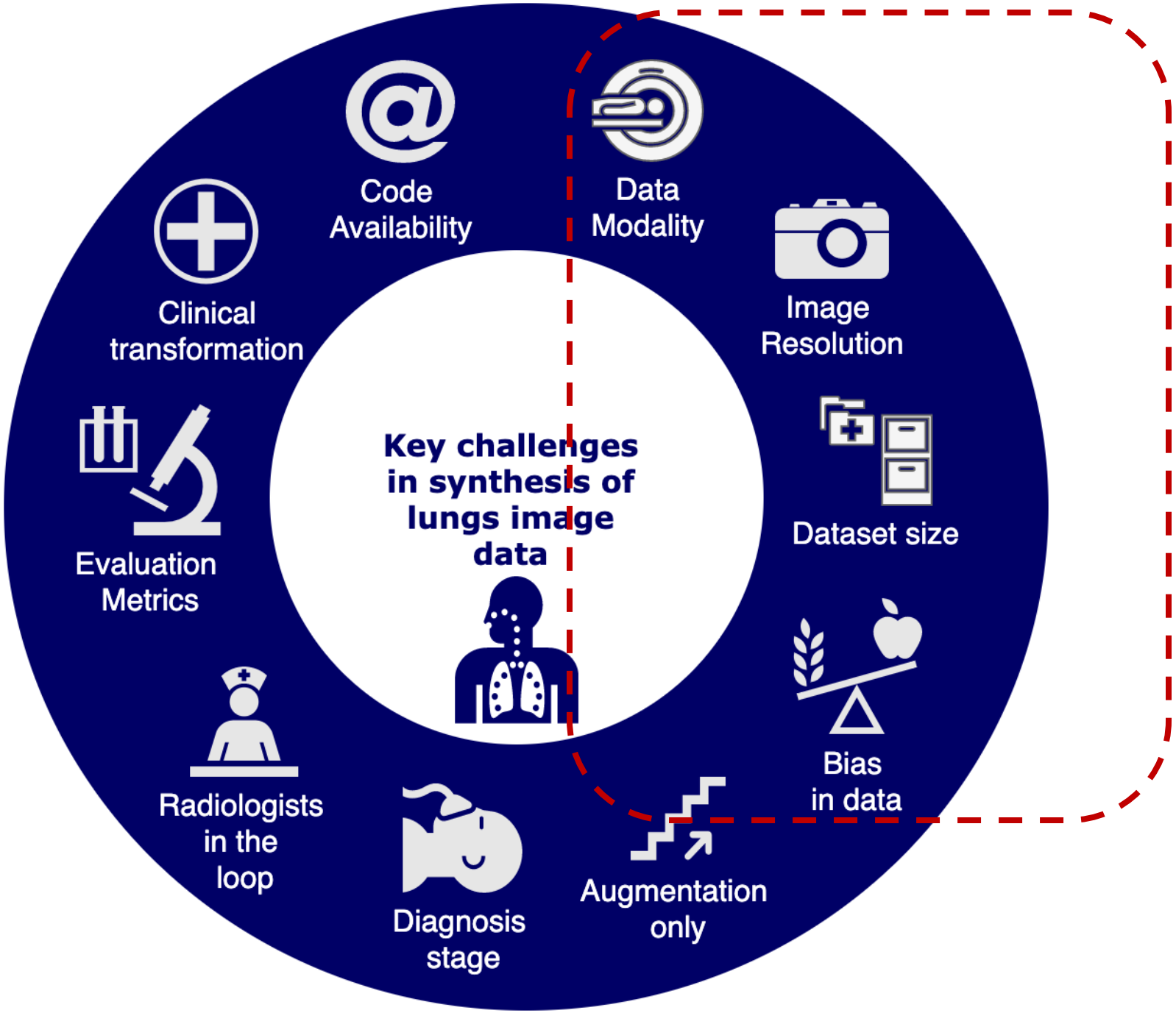
Suresh, H. and Guttag, J., 2021, October. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).



- Examples of AI in COVID-19

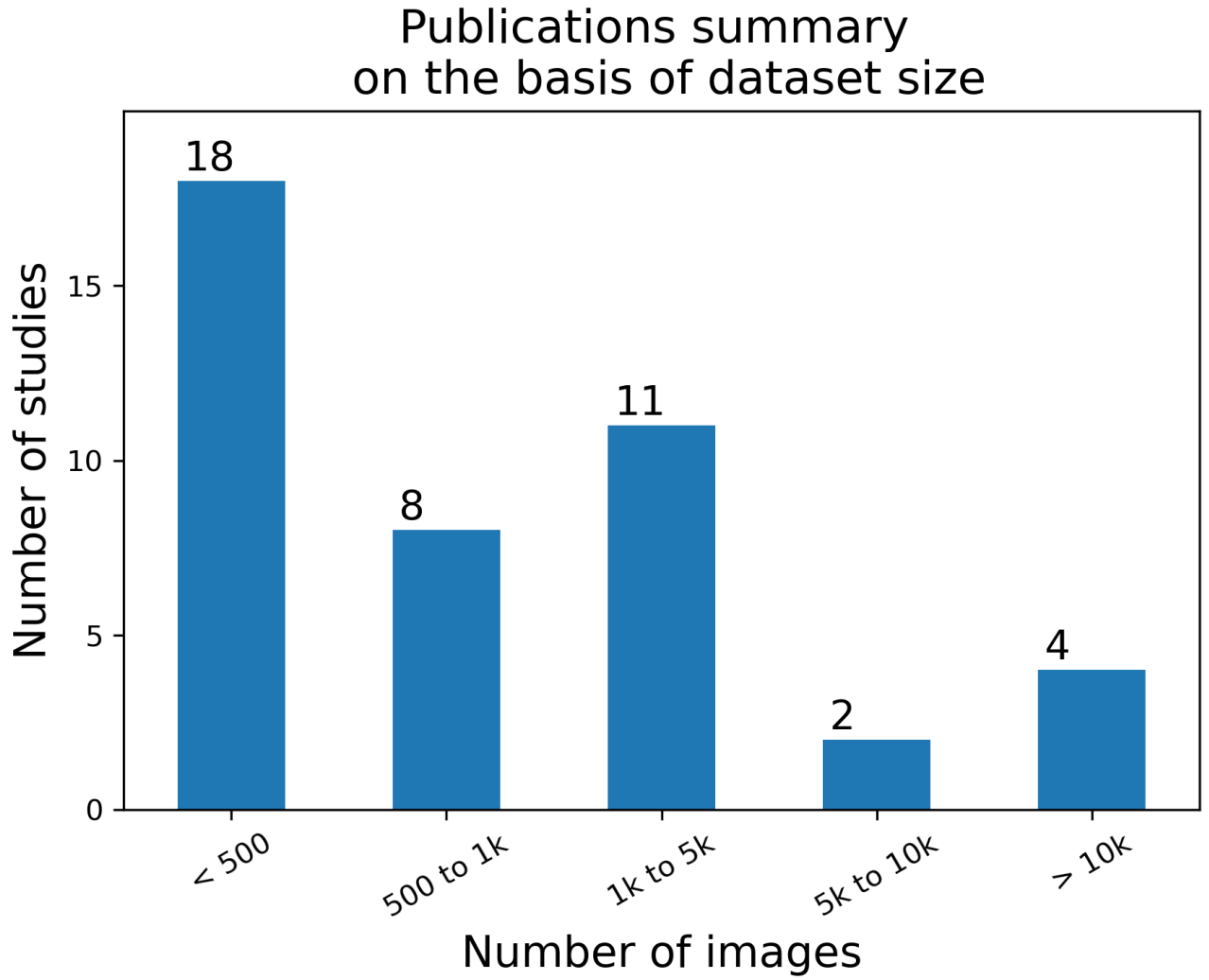
Code Availability
Data Modality
Clinical transformation
Image Resolution
Key challenges in synthesis of lungs image data
Evaluation Metrics
Dataset size
Bias in data
Radiologists in the loop
Diagnosis stage
Augmentation only


Clinical use
Data
Evaluation
Application
Code Availability
Data Modality
Clinical transformation
Image Resolution
Evaluation Metrics
Dataset size
Key challenges in synthesis of lungs image data
Radiologists in the loop
Bias in data
Diagnosis stage
Augmentation only


Code Availability
Data Modality
Clinical transformation
Image Resolution
Key challenges in synthesis of lungs image data
Evaluation Metrics
Dataset size
Radiologists in the loop
Bias in data
Diagnosis stage
Augmentation only

36

Dataset size



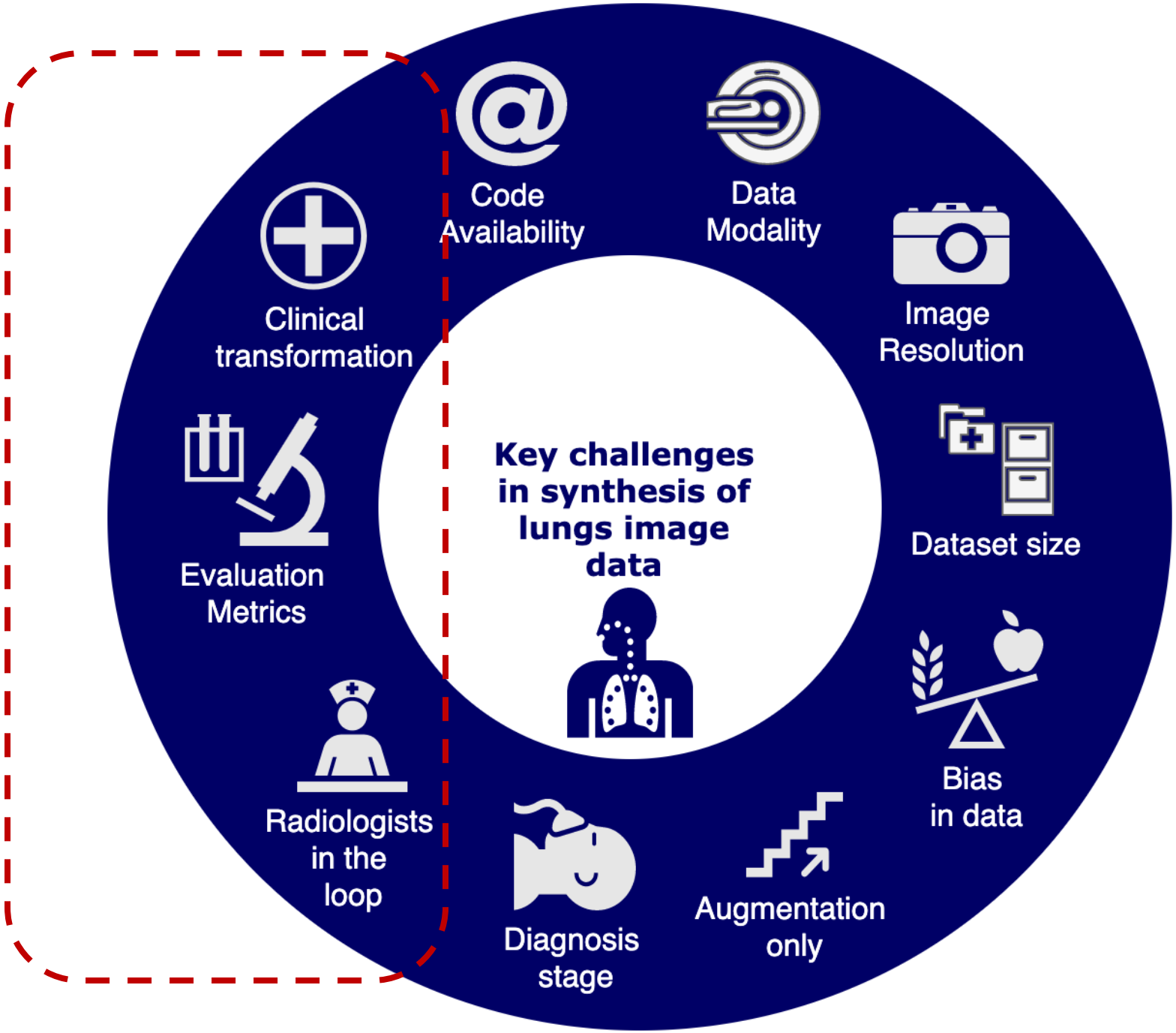
data imbalance

Data leakage

- Overlap of data in the training and test set.
- Some datasets were actually derived from others.
  - Many X-ray images were collected through crowdsourcing.
    - Quality assurance?
- So, testing is not on independent set of examples.

Image resolution and modality

- **Rescaling** of image size to 224 by 224 pixels
  - Why?
    - Model-driven.
- Aggressive re-scaling: X-ray images from 2170×1953 pixels to 512×512
- **Modality:** X-ray (60% studies)
  - But CT has higher sensitivity.


Code Availability
Data Modality
Image Resolution
Dataset size
Bias in data
Augmentation only
Diagnosis stage
Radiologists in the loop
Evaluation Metrics
Clinical transformation
Key challenges in synthesis of lungs image data

Radiologist in the loop

- Qualitative assessment by radiologists/expert
- Reported in only one study
  - suggested that the synthesized images fall short of diagnosis quality.
- Image credit: Studio4rt on freepik

Evaluation metrics

- Structure Similarity Index Measure (SSIM)
- Frechet Inception Distance (FID)
- Peak Signal-to-Noise Ratio (PSNR)
  - Driven by computer vision.

Evaluation metrics

- Structure Similarity Index Measure (SSIM)
- Frechet Inception Distance (FID)
- Peak Signal-to-Noise Ratio (PSNR)
  - Driven by computer vision.
- Reproducibility

Can the GANs based techniques be translated to clinical applications ?

Image credit: Dcstudio on freepik

suggestions

Large data

Reproducibility

Evaluation methods

The CLAIM guideline provides a road map for authors and reviewers; its goal is to promote clear, transparent, and reproducible scientific communication about the application of AI to medical imaging.

Interpretability and Explainability in Machine Learning

Interpretability

- Interpretable by humans

decision trees

neural networks

Scenario 1: You do understand it but it doesn't work well.

versus

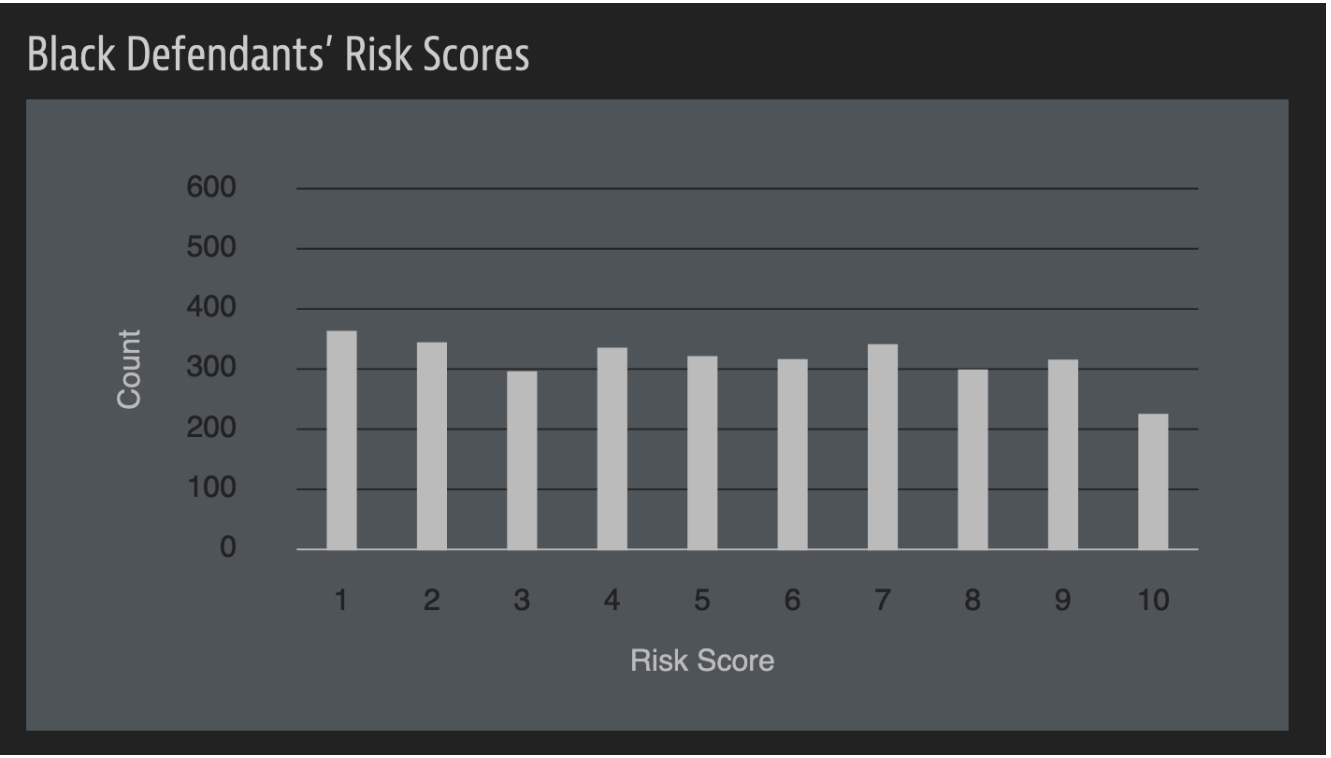
Scenario 2: You don't understand it but it works very well.

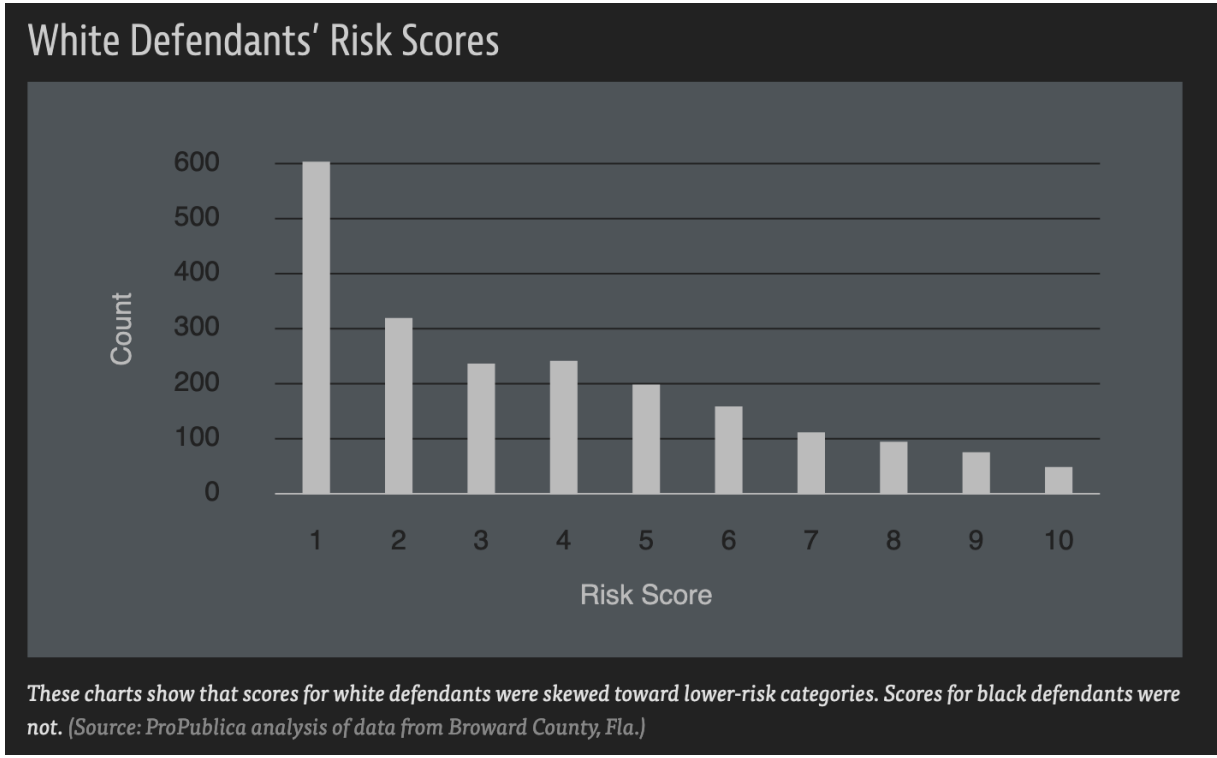
Fairness in Machine Learning

Fairness in Machine Learning

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
- a risk assessment algorithm developed by a company called Northpointe, Inc (now Equivant)

COMPAS





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

COMPAS

Two DUI Arrests

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

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%

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

Different perspective on fairness

What do different stakeholders want from the classification?

<u>Decision-maker:</u> 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

Did not recidivate	TN	FP
Recidivated	FN	TP
	Labelled low-risk	Labelled high-risk

What is your perspective on fairness ?

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%

- Demo

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

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

https://www.mobilizegreen.org/blog/2018/9/30/environmental-equity-vs-environmental-justice-whats-the-difference

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

https://www.mobilizegreen.org/blog/2018/9/30/environmental-equity-vs-environmental-justice-whats-the-difference

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

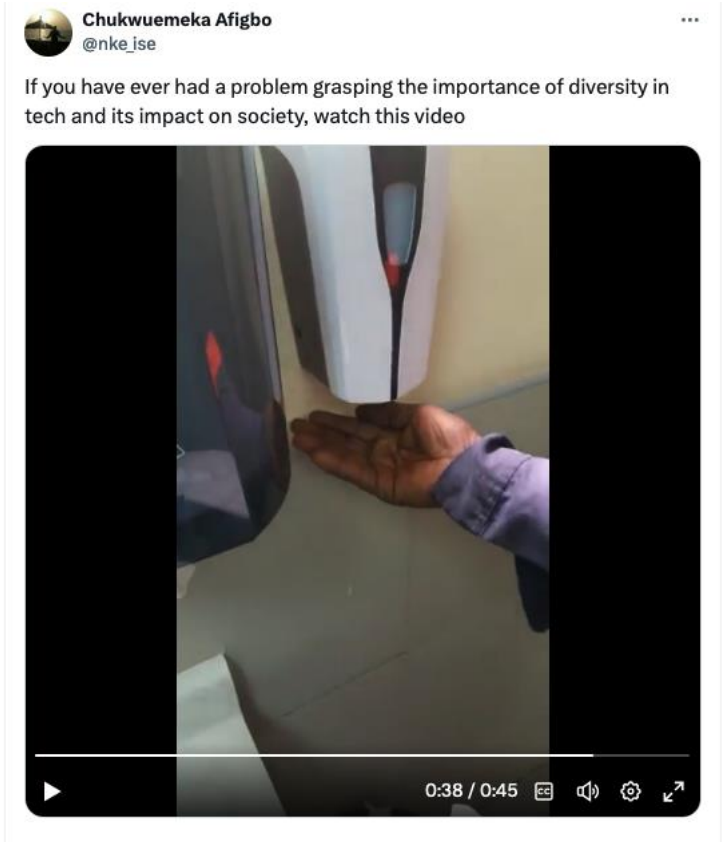
Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

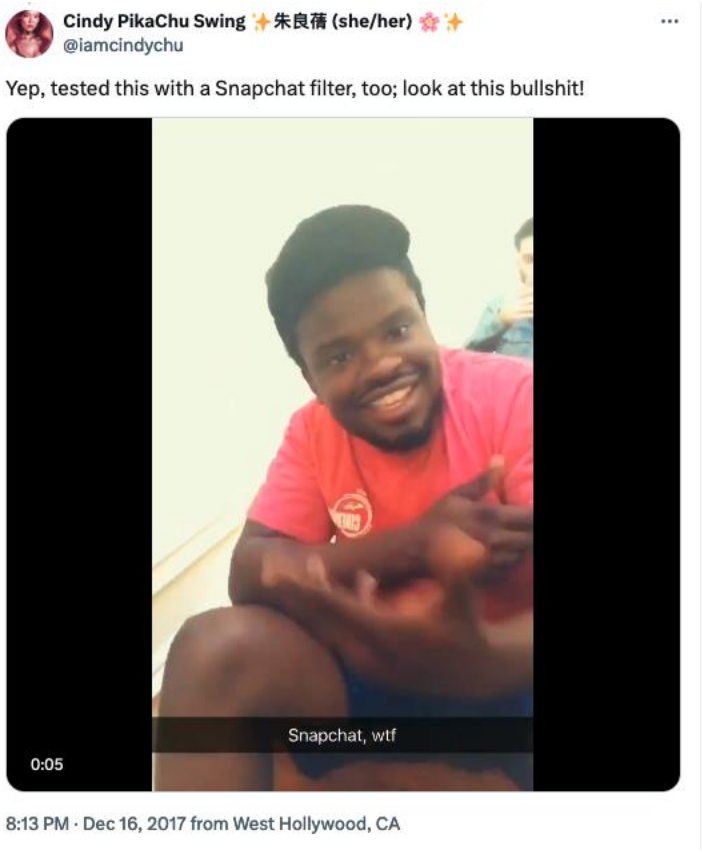
Justice

All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

https://www.mobilizegreen.org/blog/2018/9/30/environmental-equity-vs-environmental-justice-whats-the-difference

Representation





https://x.com/nke_ise/status/897756900753891328

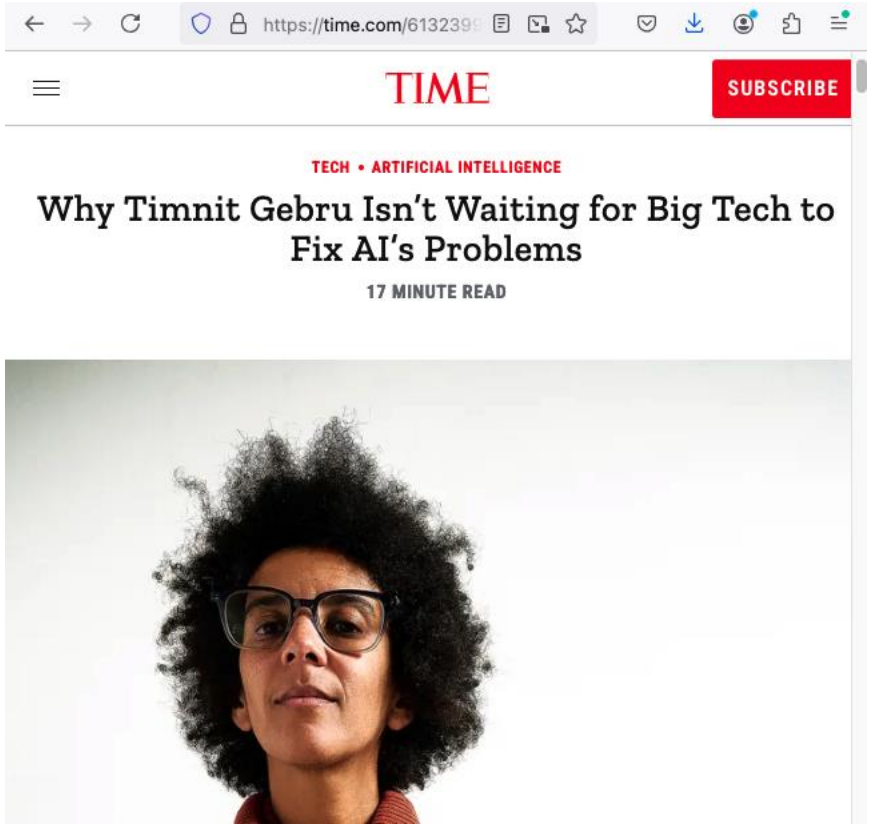
https://x.com/iamcindychu/status/942125601992851456

A product description of a black-coloured action figure that featured the N-word slipped through Amazon's screening process. (Screenshot of Amazon listing)

https://www.cbc.ca/news/science/artificial-intelligence-racism-bias-1.6027150

Representation

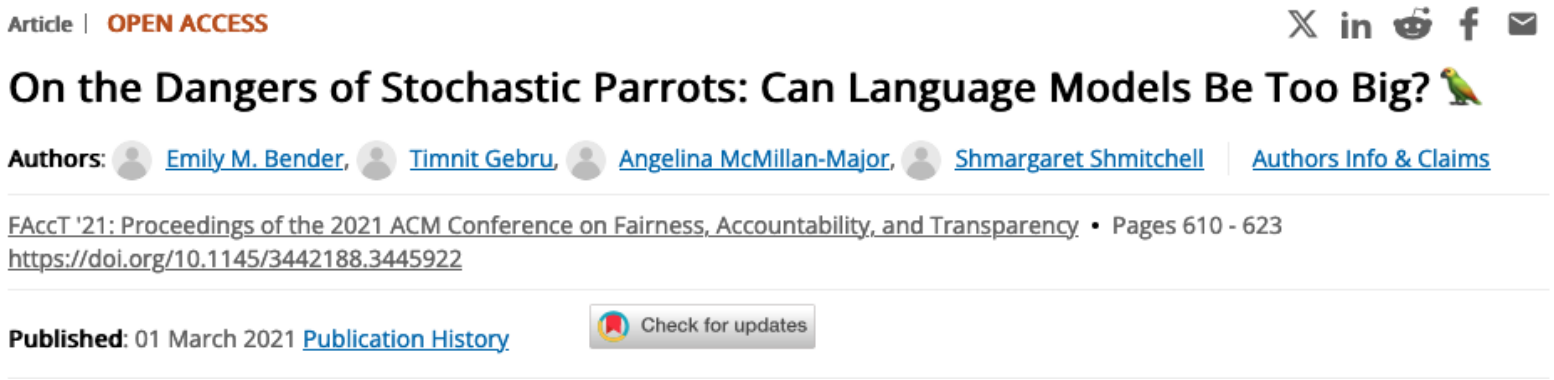
I am very concerned about the future of AI," she wrote. "Not because of the risk of rogue machines taking over. But because of the homogeneous, one-dimensional group of men who are currently involved in advancing the technology."



By 2017, Gebru was an AI researcher at Microsoft, where she co-authored a paper called Gender Shades. It demonstrated how facial-recognition systems developed by IBM and Microsoft were almost perfect at detecting images of white people, but not people with darker skin, particularly Black women. The data set that had been used to train the algorithm contained lots of images of white men, but very few of Black women. The research, which Gebru had worked on alongside Joy Buolamwini of MIT Media Lab, forced IBM and Microsoft to update their data sets.
Google hired Gebru shortly after Gender Shades was published, at a time when Big Tech companies were coming under increasing scrutiny over the ethical credentials of their AI research.

https://time.com/6132399/timnit-gebru-ai-google/

Authors fired by Google?



Article | OPEN ACCESS

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

Authors: Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell | Authors Info & Claims

FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • Pages 610 - 623
https://doi.org/10.1145/3442188.3445922

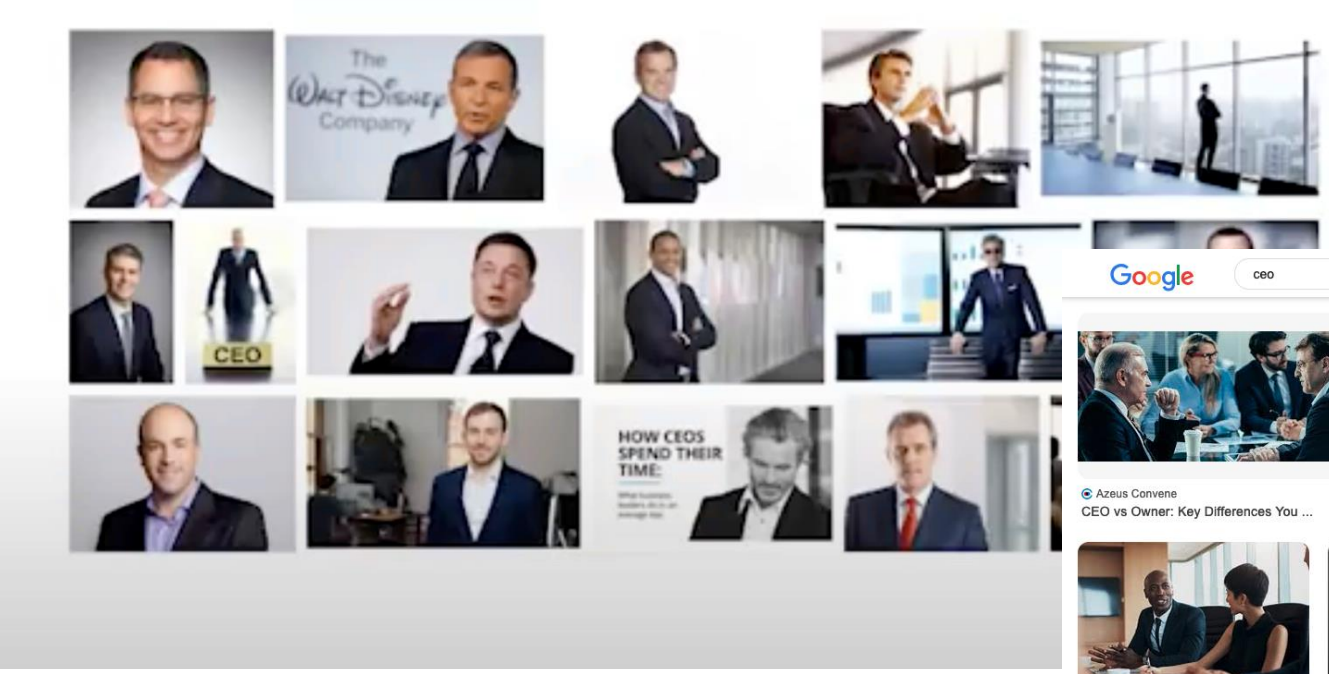
Published: 01 March 2021 Publication History

Check for updates

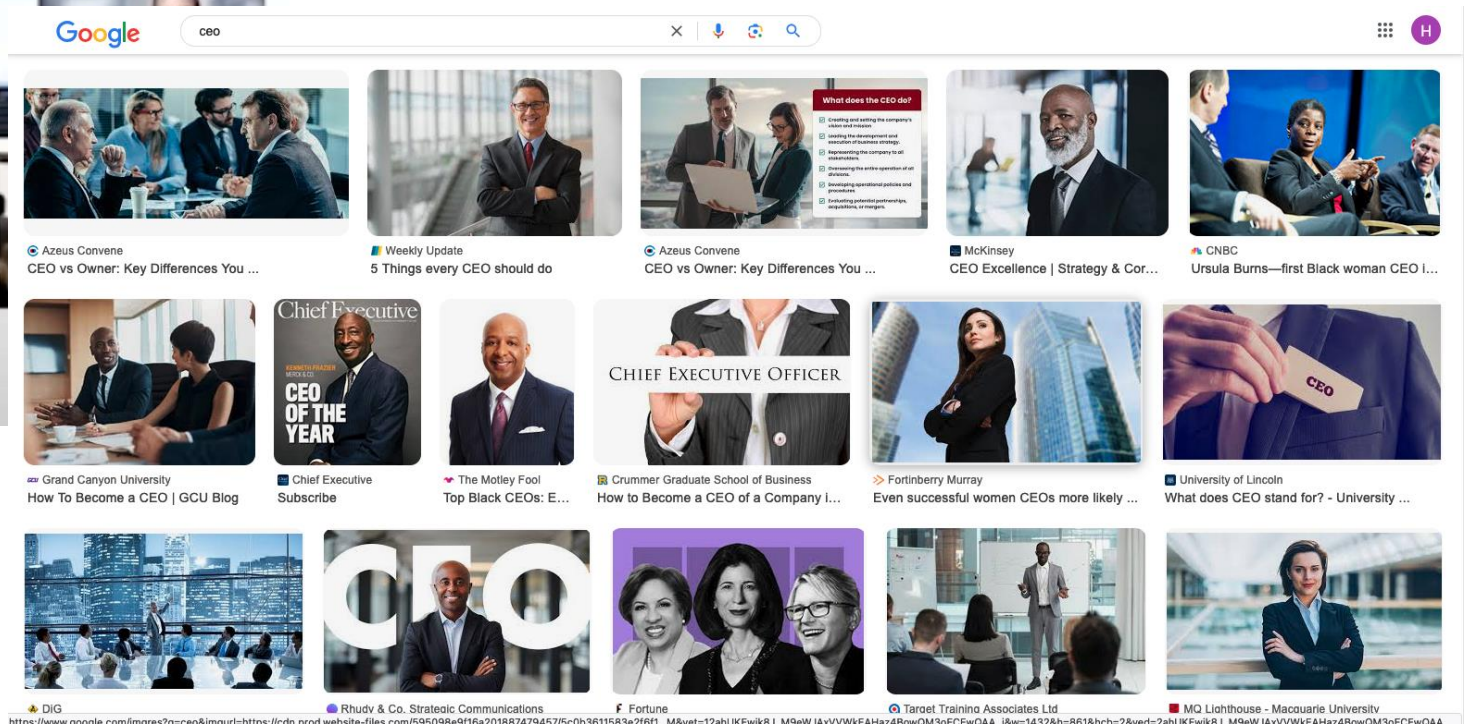
- In December 2020, Timnit Gebru – computer scientist and co-lead of the Ethical Artificial Intelligence Team at Google – was fired after Google managers asked her to either retract her latest research paper or to remove her name from it, along with those of other members of the team who had contributed to the research. The paper in question, "**On the Dangers of Stochastic Parrots: Can Language Models Be Too Big**?", was an overview of the **environmental, financial, and social costs of large language models such as GPT-3**. Google's managers deemed the paper unpublishable but did not reveal how the internal review process they had come to this decision. Google claimed that Gebru had offered to resign and that her resignation had been accepted. In February 2021, Google fired the co-lead of the ethical AI team (and co-author of the paper), Margaret Mitchell, claiming that she had violated the company's code of conduct. Google's employees, who had previously protested the company's collaboration with US Custom and Border Protection, the US Department of Defense, the Chinese state surveillance project Dragonfly, and had denounced Google's handling of sexual harassment, demanded that Google provide clear guidelines on how research is reviewed and how research integrity is respected.

Another example

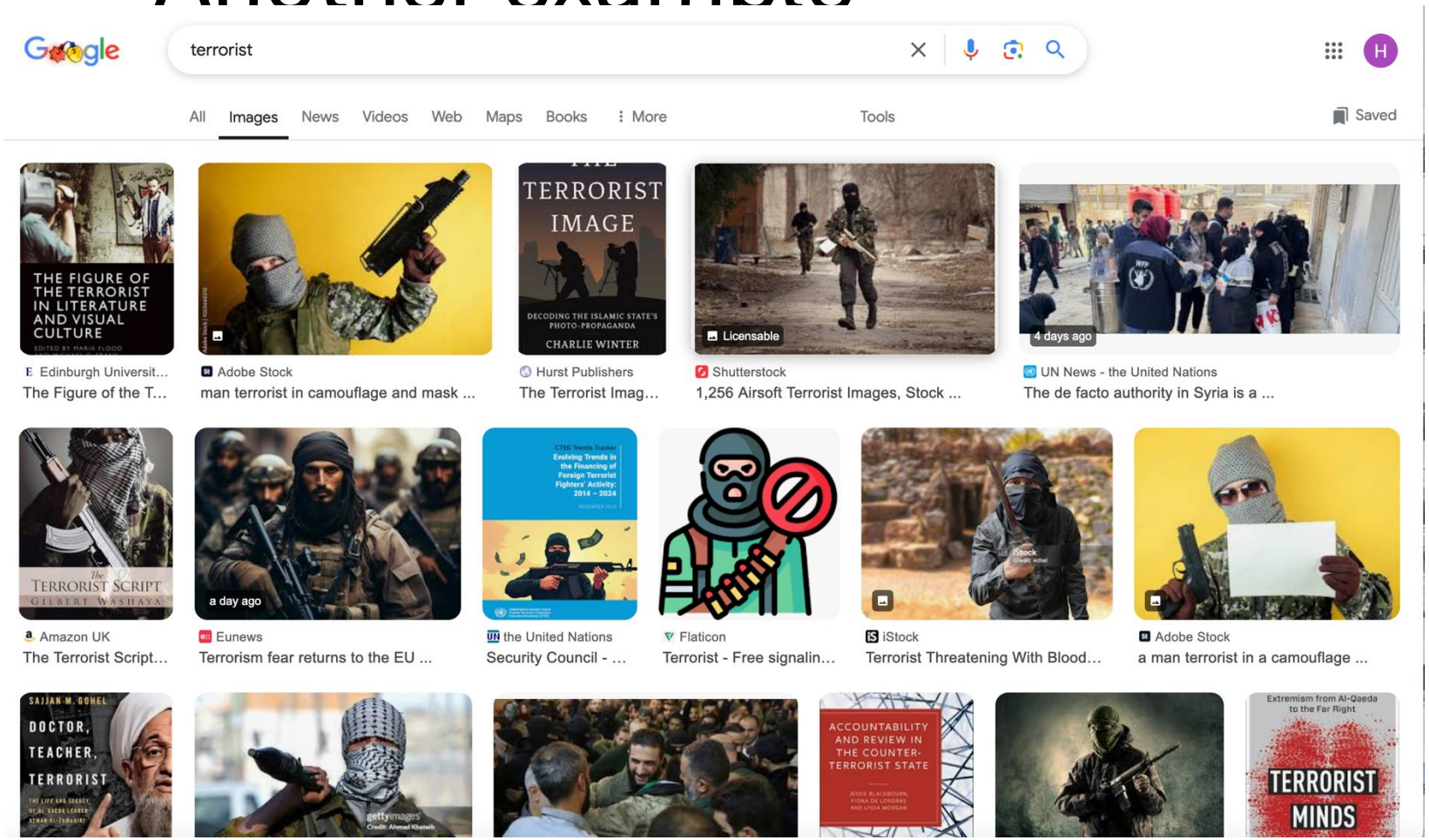
Image search: CEO



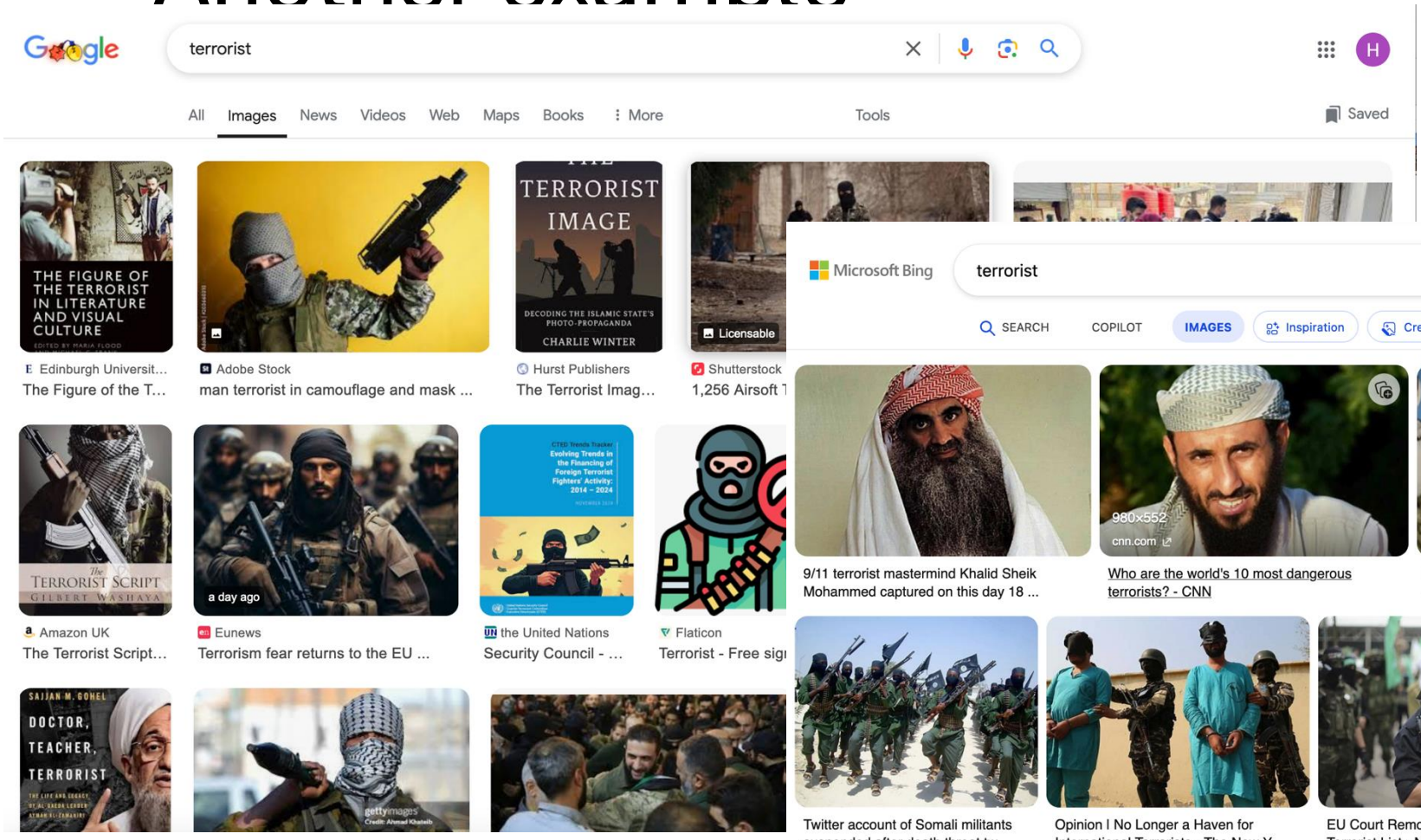
Recent results

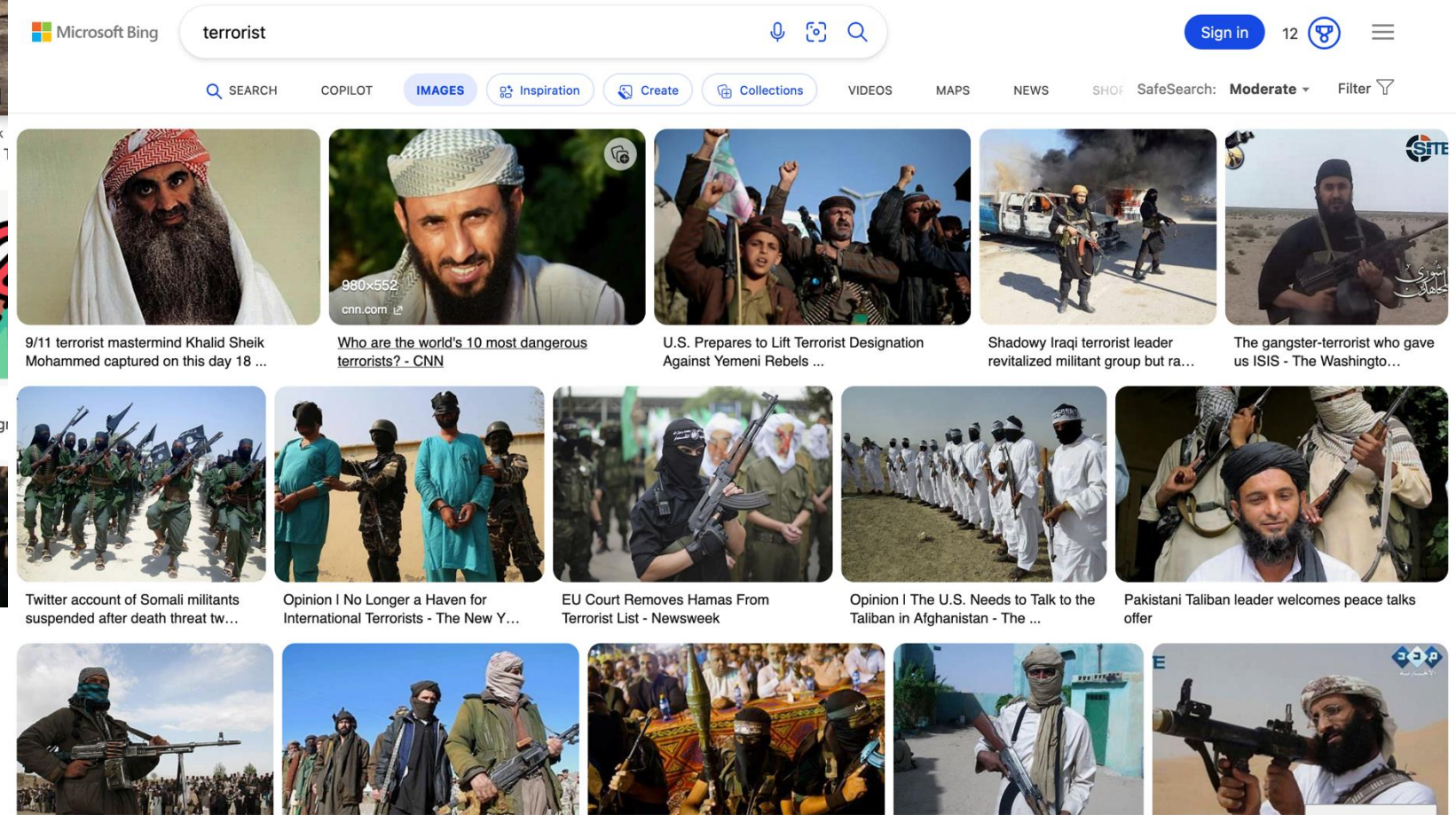


Another example



Another example





What is water (what is ethical)?

- Whether language models should reflect the world as it is in the data, or as society believes it should be, depends on what use they are applied to.
  - For example, should I remove hateful text in training data of an LLM?

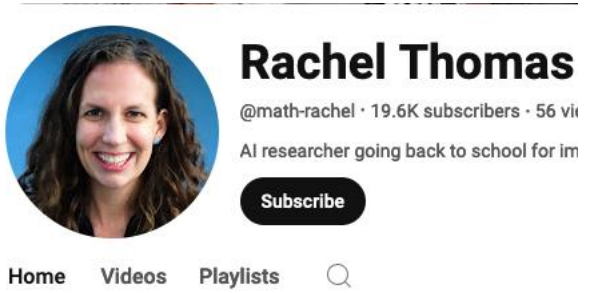
- And when we do want them to reflect the world as it should be, do we agree on what that is?
  - And who decides what should be included?

What is water (what is ethical)?

- 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?

Clearview AI, devised a groundbreaking facial recognition app. You take a picture of a person, upload it and get to see public photos of that person along with links to where those photos appeared. The system — whose backbone is a database of more than 3 billion images that Clearview claims to have scraped from Facebook, YouTube, Venmo and millions of other websites — goes far beyond anything ever constructed by the US government or Silicon Valley giants.

Some 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.

https://www.youtube.com/watch?v=av7utkFXbU4

Acknowledgements

Sergey Karayev

- *Full stack deep learning*
- *University of California, Berkeley*

Further Resources

- Rachel Thomas short lectures
  - https://rachel.fast.ai/posts/2021-08-17-eleven-ethics-videos/

- Book: Fairness and Machine Learning by Solon Barocas, Moritz Hardt, Arvind Narayanan (https://fairmlbook.org/)
- Course: CS294: Fiarness in machine Learning, Prof. Moritz Hardt. UC Berkeley. https://fairmlclass.github.io
- Lecture: Full stack deep learning lecture on Trust