A specialist energy consultancy

Bias Traps in AI

Dr Diptargha Chakravorty

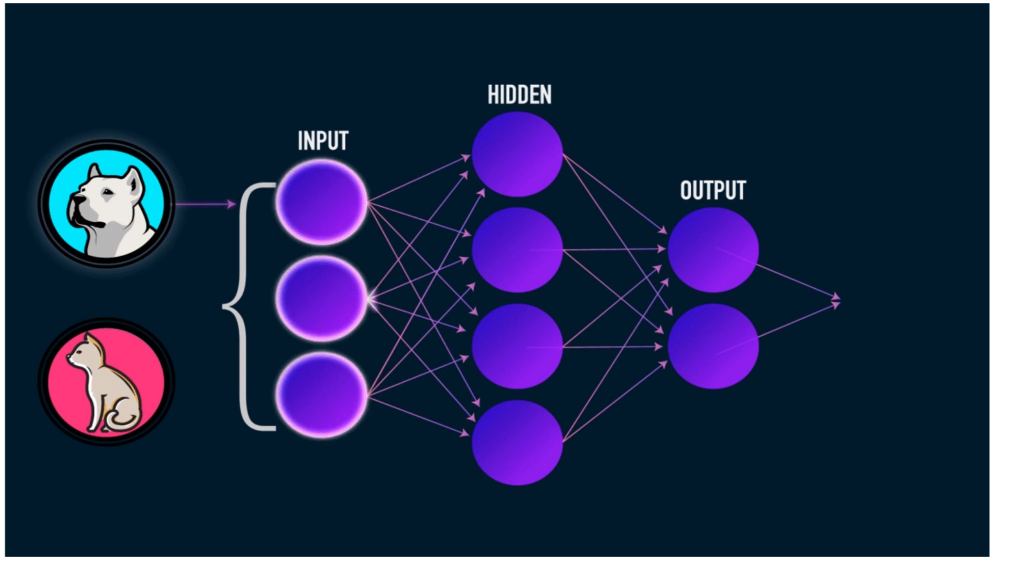
Head of Innovation

21st May 2024

tneigroup.com

What do we mean by bias in AI?

- It refers to the presence of unfair or prejudiced outcomes in algorithms and models
- This can manifest in different ways such as
  - **Data bias:** If historical data reflects some prejudices, then AI can learn and perpetuate those biases
  - **Algorithmic bias:** Inherent biases due to the way they are designed or trained such as data preprocessing. These can be unintentional but still result in biased outcomes.
  - **User bias:** Users interacting with AI systems can introduce bias such as search engines or user feedback systems can reinforce existing bias in recommendation systems.



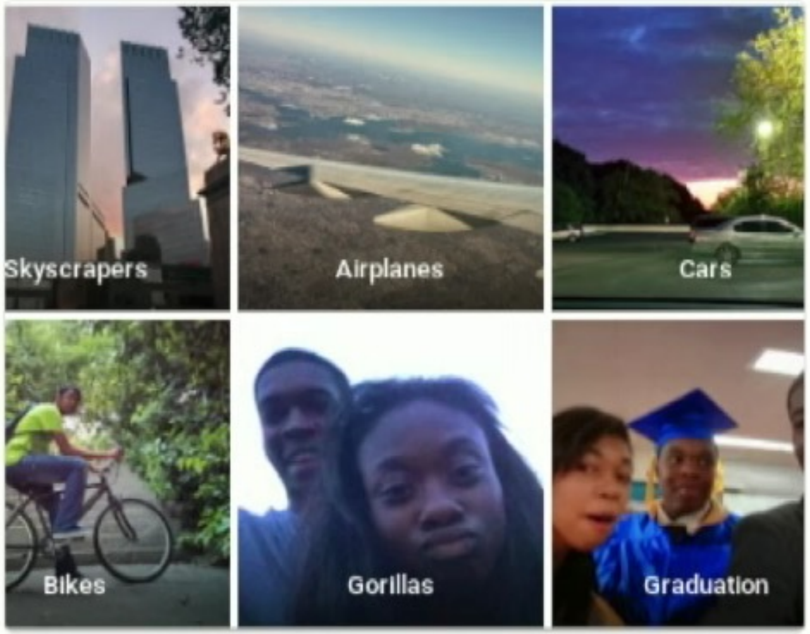
Ethical Considerations

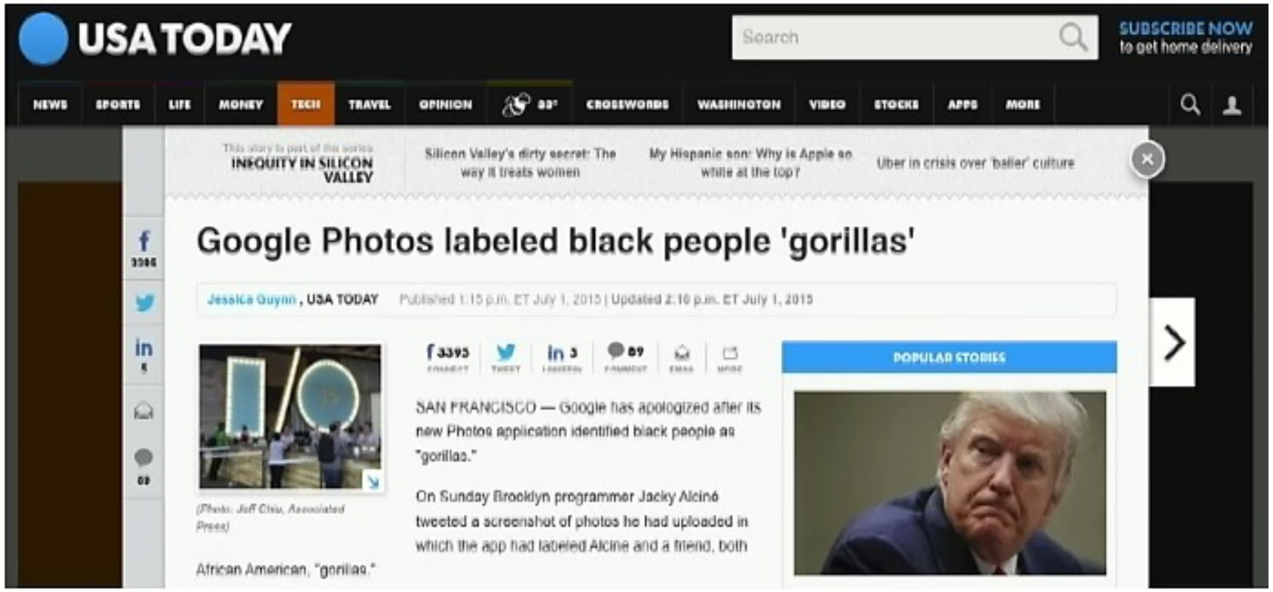
- Closely linked to biases are the ethical considerations of using an AI model. Some of the essential considerations are –
  - **Fairness:** To ensure equitable outcomes for all individuals regardless of their demographic characteristics
  - **Accountability:** Mechanisms for accountability in AI development, maintenance and deployment to address bias issues
  - **Transparency:** A clear explanation of how they have arrived at a decision, especially for critical systems such as finance, healthcare and criminal justice
  - **Bias awareness:** People developing AI models should be aware of their own biases and how these biases can influence algorithm design and decision making

Sources: [1] Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML
[2] Centre for Data Ethics and Innovation - GOV.UK (www.gov.uk)

Bias in Photo Labelling

Google Photos, y'all f ***** up. My friend's not a gorilla.





This story is part of the series INEQUITY IN SILICON VALLEY

Silicon Valley's dirty secret: The way it treats women

My Hispanic son: Why is Apple so white at the top?

Uber in crisis over 'baller' culture

Google Photos labeled black people 'gorillas'

Jessica Guynn, USA TODAY Published 1:15 p.m. ET July 1, 2015 | Updated 2:10 p.m. ET July 1, 2015

(Photo: Jeff Chiu, Associated Press)

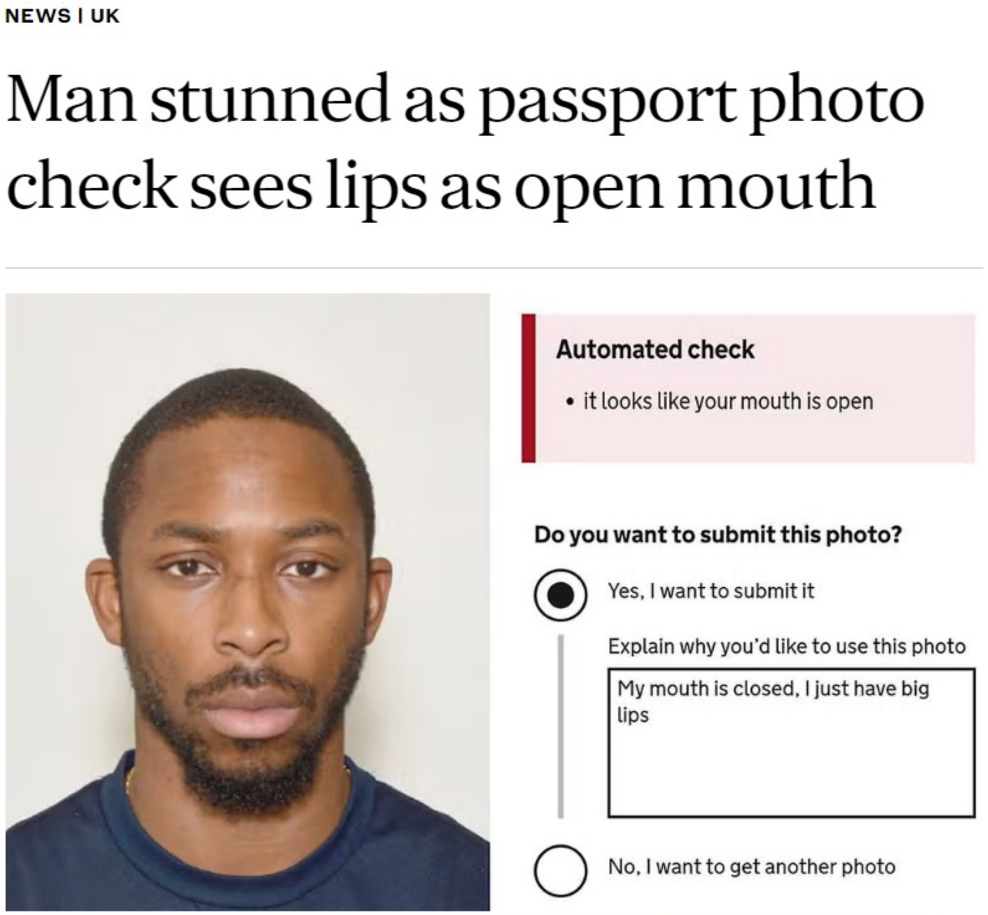
SAN FRANCISCO — Google has apologized after its new Photos application identified black people as "gorillas."

On Sunday Brooklyn programmer Jacky Alciné tweeted a screenshot of photos he had uploaded in which the app had labeled Alcine and a friend, both African American, "gorillas."

- Google's Photos app mistakenly labelled a black couple as being "gorillas"

Source: https://www.bbc.co.uk/news/technology-33347866

Bias in Photo Checking



NEWS | UK

Man stunned as passport photo check sees lips as open mouth

Automated check

- it looks like your mouth is open

Do you want to submit this photo?

● Yes, I want to submit it

Explain why you'd like to use this photo

My mouth is closed, I just have big lips

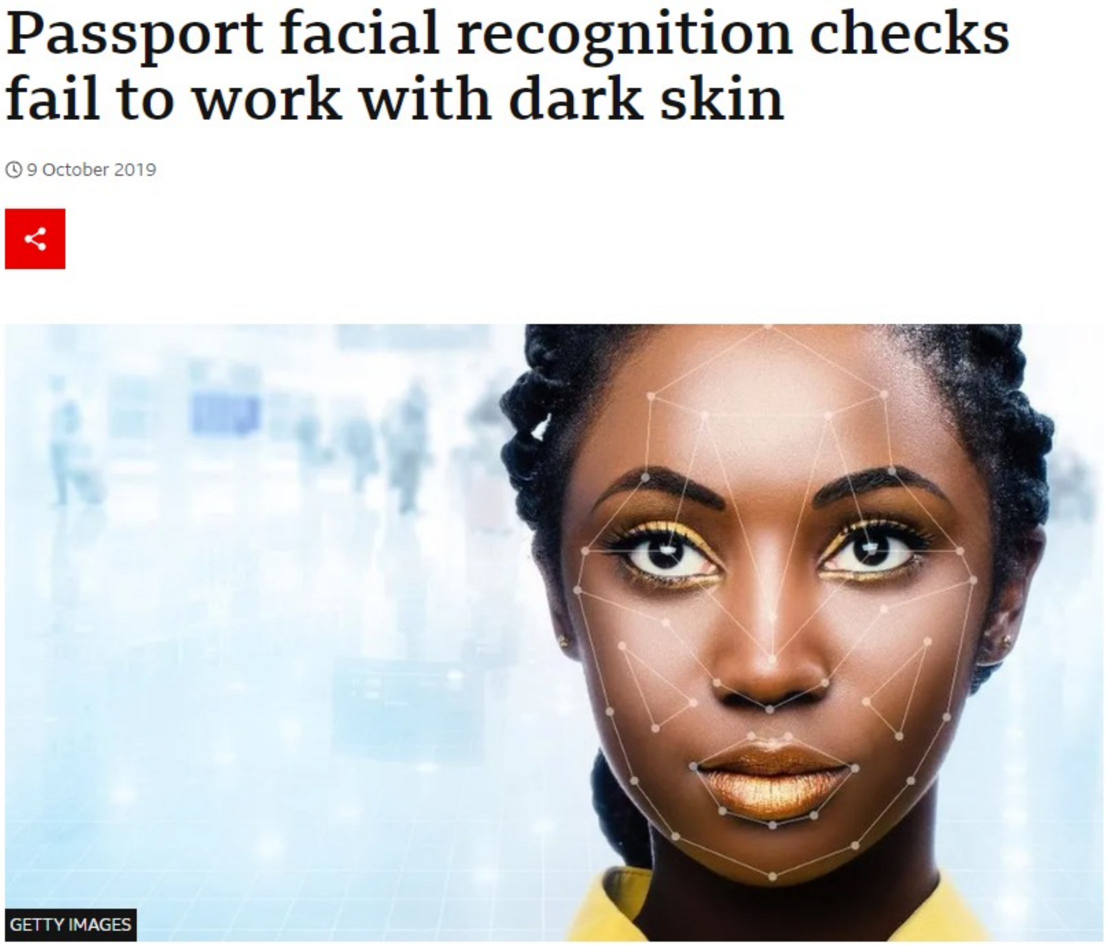
○ No, I want to get another photo

"ANNOYED": JOSHUA BADA USED A HIGH-QUALITY IMAGE FOR AN ONLINE APPLICATION TO RENEW HIS PASSPORT

- Automated photo checker for online passport renewal mistook his lips for an open mouth
- Noel Sharkey, professor of artificial intelligence and robotics at the University of Sheffield, believes an unrepresentative sample of black people is one possible reason for the error.

Source: Man stunned as passport photo check sees lips as open mouth | London Evening Standard | Evening Standard

Bias in Facial Recognition System



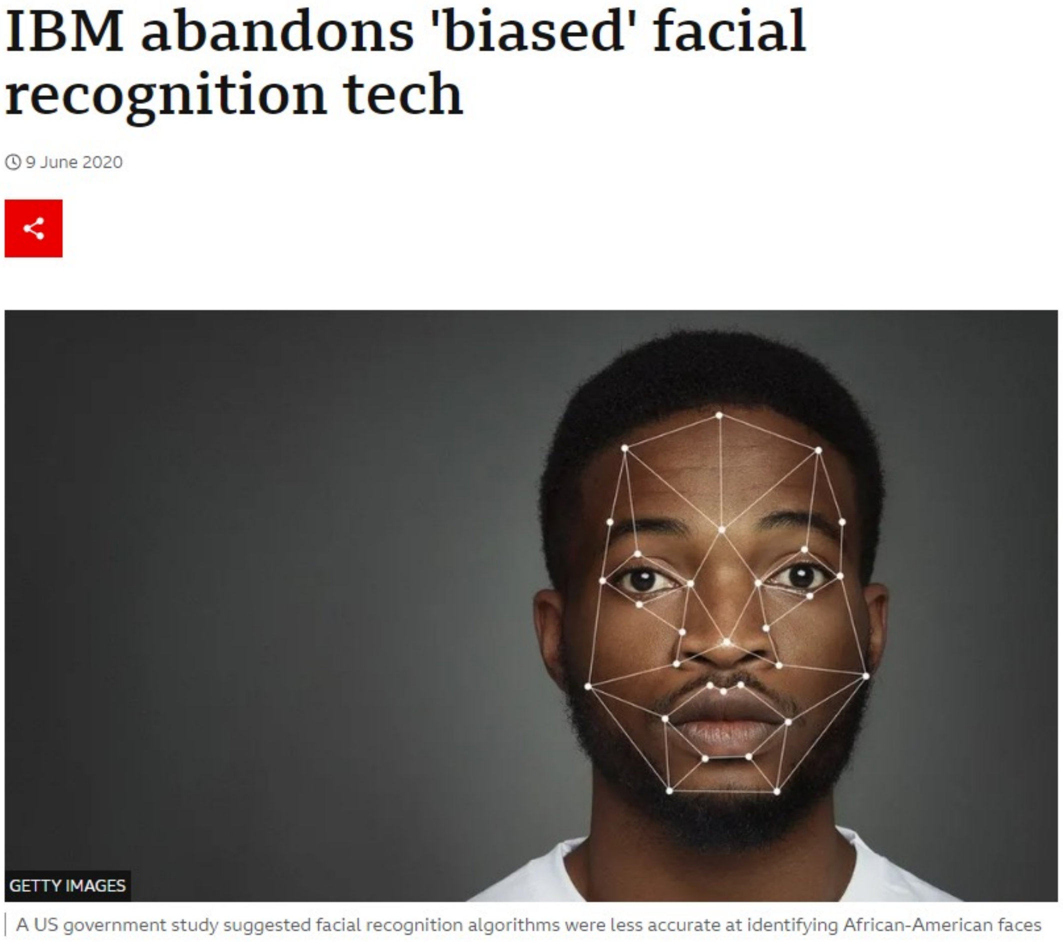
Passport facial recognition checks fail to work with dark skin

9 October 2019

- A passport checking service launched by the Home Office had trouble handling some shades of skin.
- A 2019 study conducted by the Massachusetts Institute of Technology found that none of the facial recognition tools from Microsoft, Amazon and IBM were 100% accurate when it came to recognising men and women with dark skin.

Source: Passport facial recognition checks fail to work with dark skin - BBC News

Bias in Facial Recognition System



IBM abandons 'biased' facial recognition tech

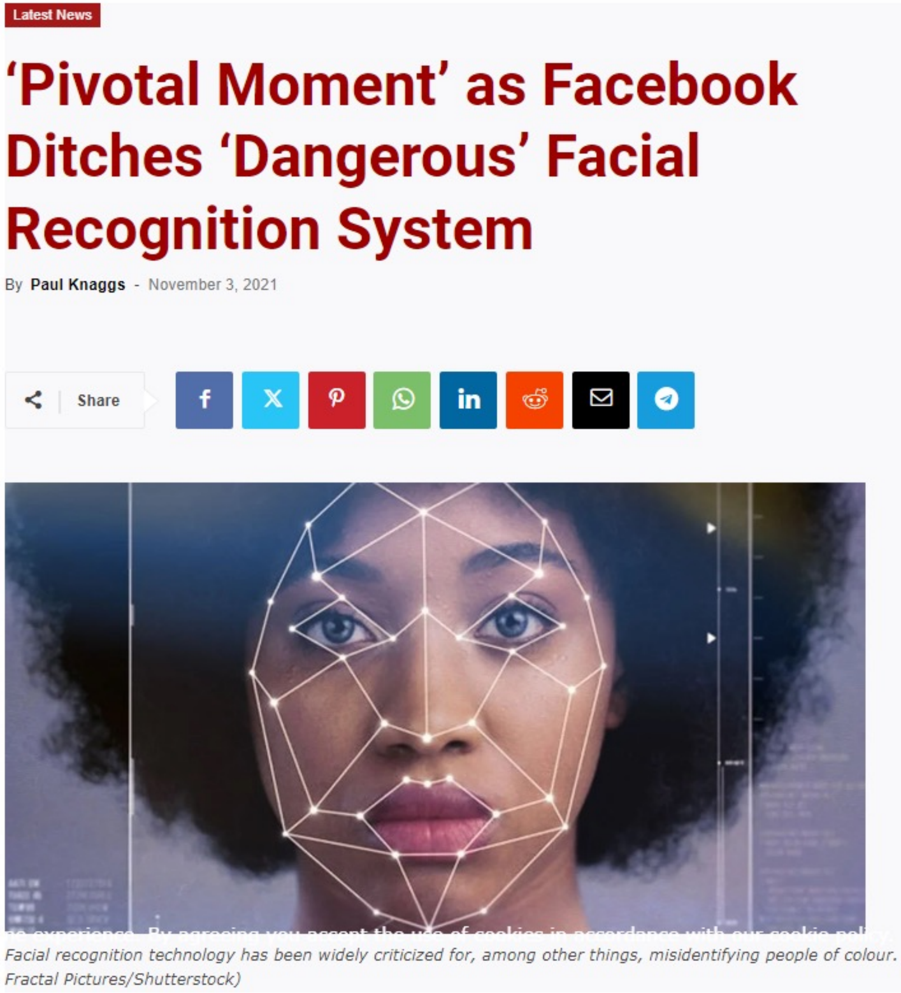
9 June 2020

A US government study suggested facial recognition algorithms were less accurate at identifying African-American faces

- IBM stopped offering their facial recognition software for “mass surveillance or racial profiling”
- IBM said AI systems used in law enforcement needed testing for ‘bias’

Source: IBM abandons 'biased' facial recognition tech - BBC News

Bias in Facial Recognition System



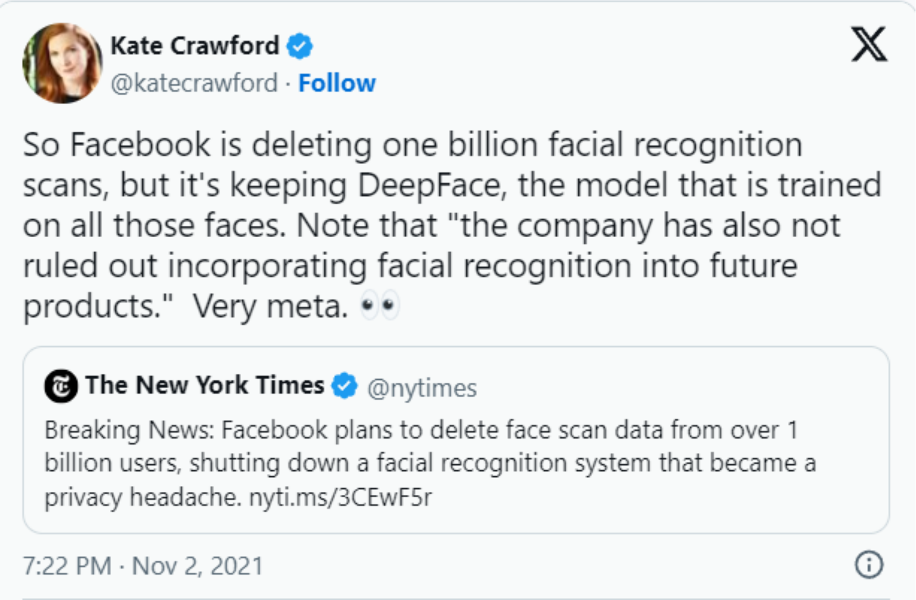
Latest News

‘Pivotal Moment’ as Facebook Ditches ‘Dangerous’ Facial Recognition System

By **Paul Knaggs** - November 3, 2021

Share

Facial recognition technology has been widely criticized for, among other things, misidentifying people of colour. Fractal Pictures/Shutterstock)



Kate Crawford
@katecrawford · **Follow**

So Facebook is deleting one billion facial recognition scans, but it's keeping DeepFace, the model that is trained on all those faces. Note that "the company has also not ruled out incorporating facial recognition into future products." Very meta. 👀

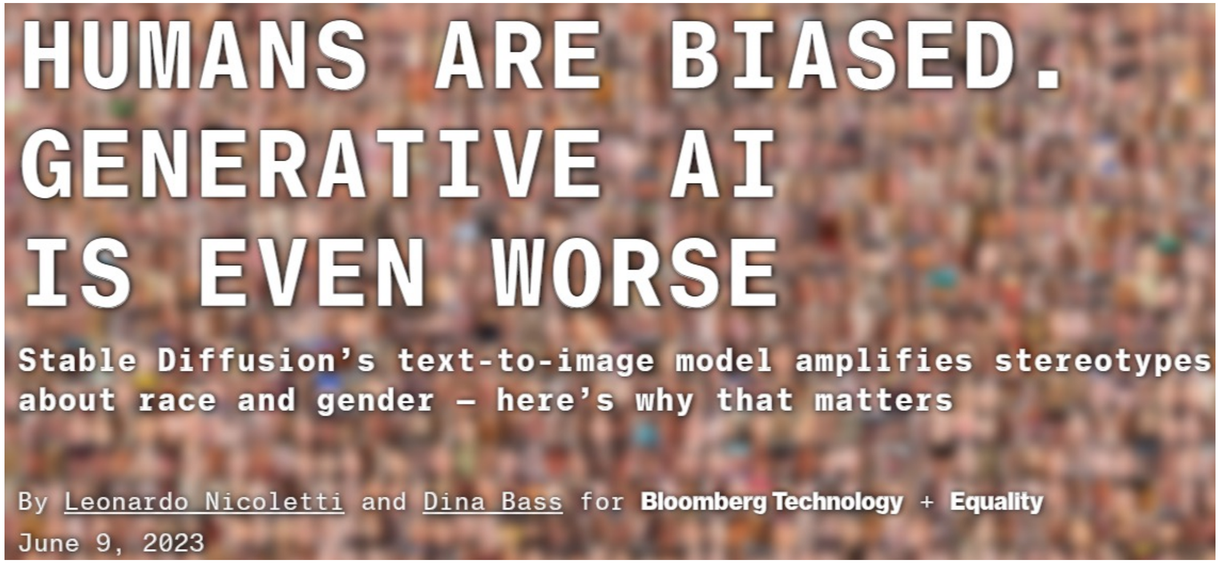
The New York Times @nytimes
Breaking News: Facebook plans to delete face scan data from over 1 billion users, shutting down a facial recognition system that became a privacy headache. nyti.ms/3CEwF5r

7:22 PM · Nov 2, 2021

“Last year, the company also agreed to pay $650 million to settle a class-action lawsuit in Illinois that accused Facebook of violating a state law that requires residents’ consent to use their biometric information, including their “face geometry.”

Source: [1] ‘Pivotal Moment’ As Facebook Ditches ‘Dangerous’ Facial Recognition System - Labour Heartlands
[2] Facebook Plans to Shut Down Its Facial Recognition System - The New York Times (nytimes.com)

Bias in Generative AI



HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE

Stable Diffusion's text-to-image model amplifies stereotypes about race and gender — here's why that matters

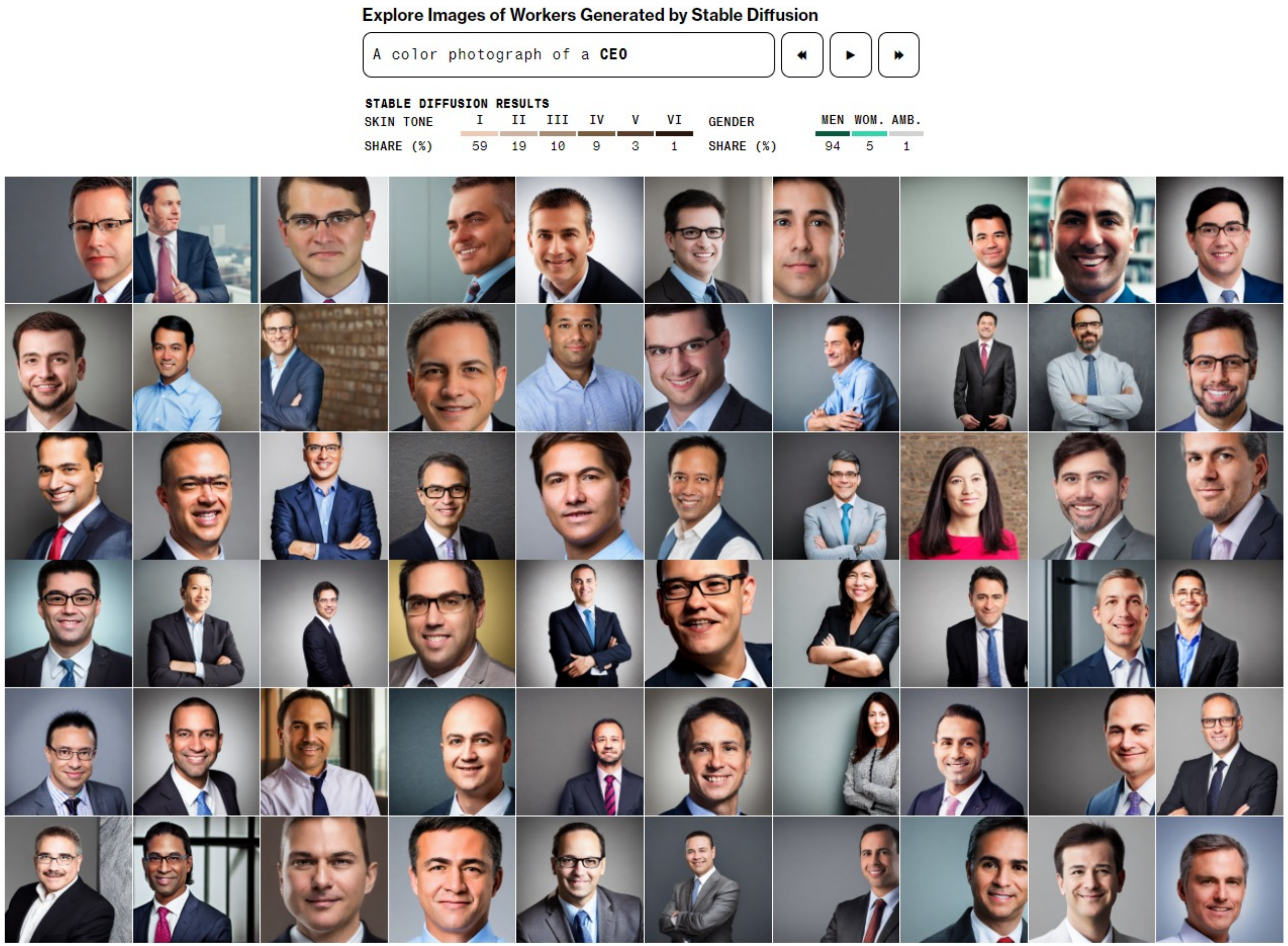
By Leonardo Nicoletti and Dina Bass for Bloomberg Technology + Equality

June 9, 2023

The world according to Stable Diffusion is run by White male CEOs. Women are rarely doctors, lawyers or judges. Men with dark skin commit crimes, while women with dark skin flip burgers.

https: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Bias in Generative AI



Explore Images of Workers Generated by Stable Diffusion

A color photograph of a **CEO**

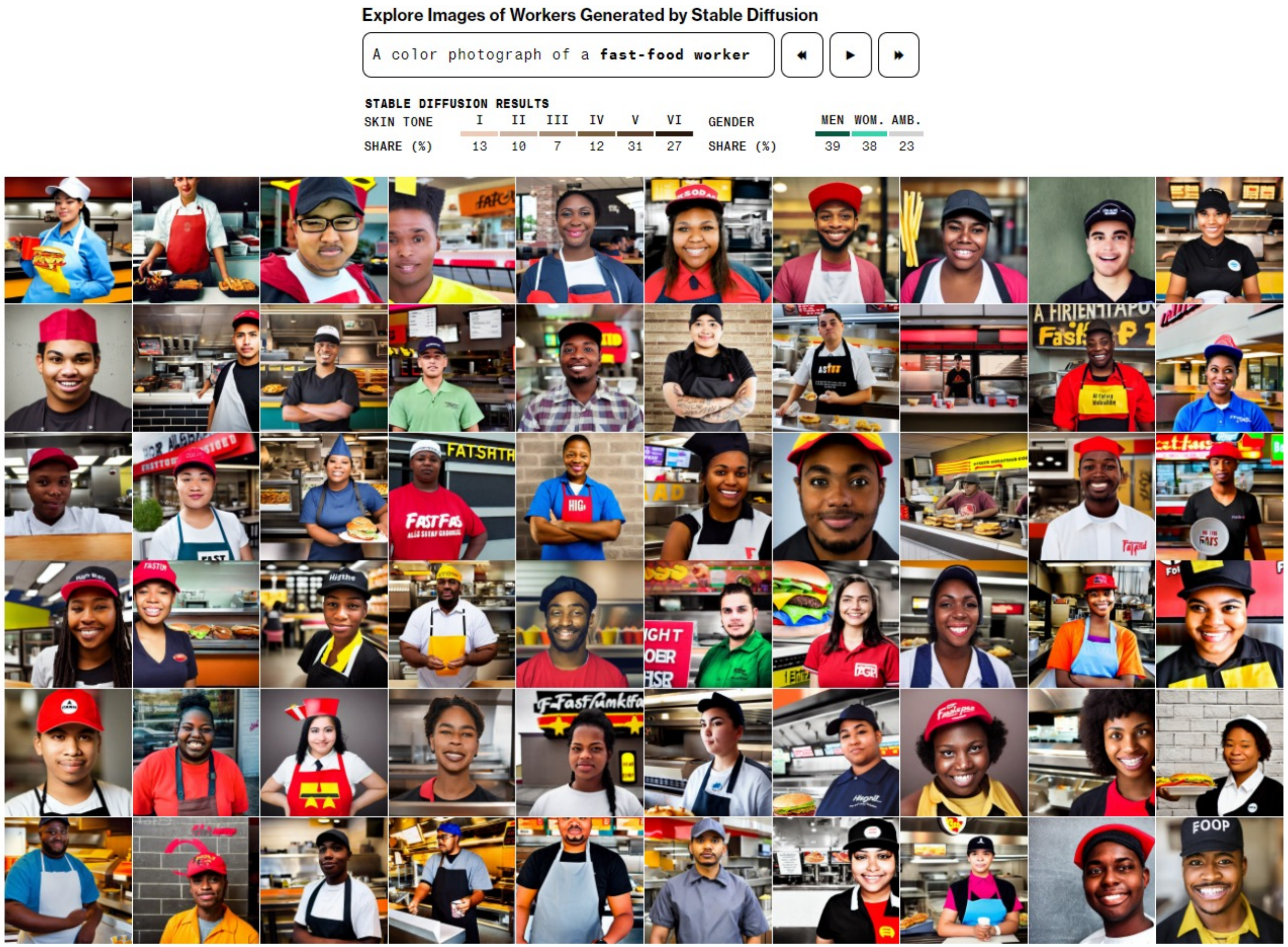
STABLE DIFFUSION RESULTS

SKIN TONE	I	II	III	IV	V	VI	GENDER	MEN	WOM.	AMB.
SHARE (%)	59	19	10	9	3	1	SHARE (%)	94	5	1

Image sets generated for every high-paying jobs were dominated by subjects with lighter skin tones, while subjects with darker skin tones were more commonly generated by prompts like “fast-food worker” and “social worker.”

https: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Bias in Generative AI



Explore Images of Workers Generated by Stable Diffusion

A color photograph of a **fast-food worker**

STABLE DIFFUSION RESULTS

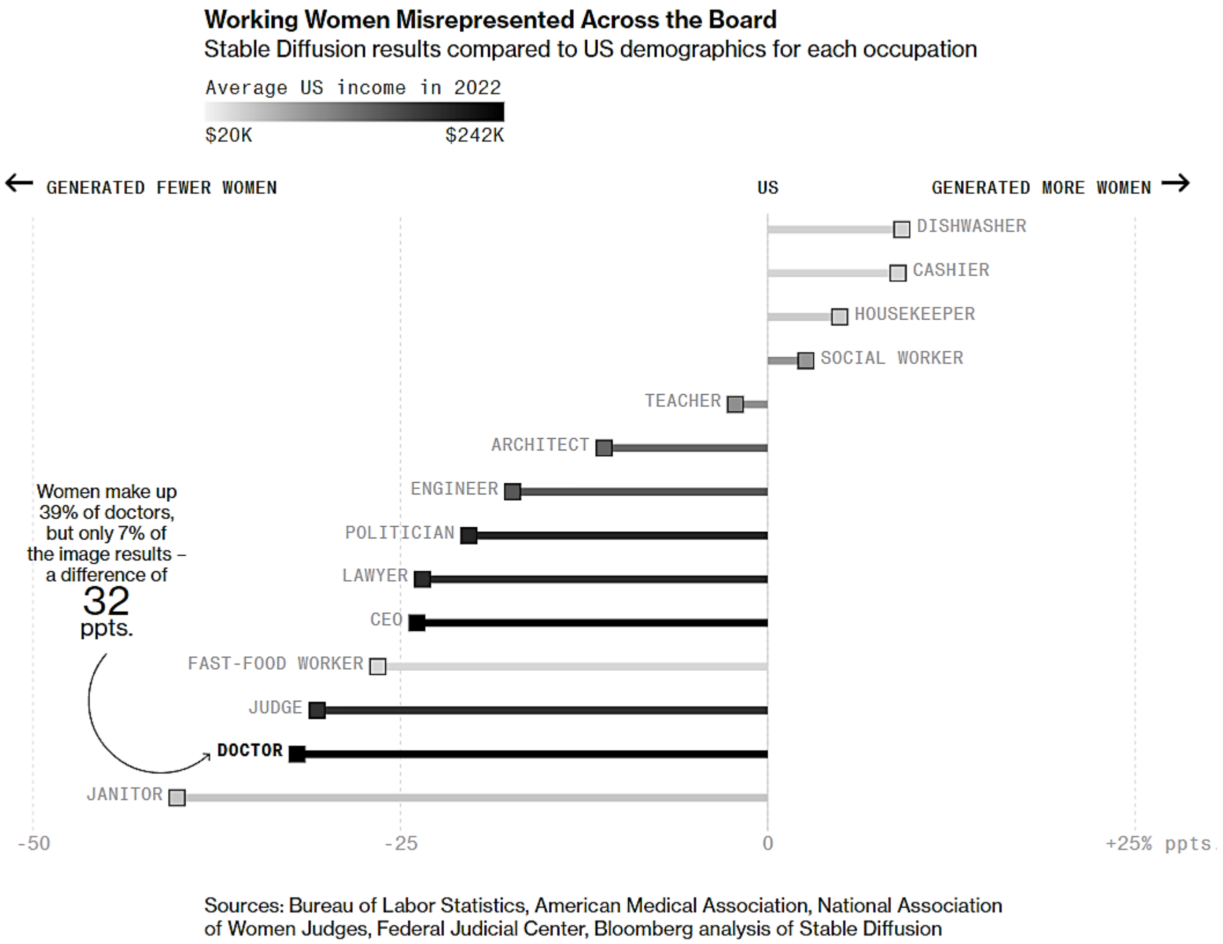
SKIN TONE	I	II	III	IV	V	VI
SHARE (%)	13	10	7	12	31	27

GENDER	MEN	WOM.	AMB.
SHARE (%)	39	38	23

Image sets generated for every high-paying jobs were dominated by subjects with lighter skin tones, while subjects with darker skin tones were more commonly generated by prompts like “fast-food worker” and “social worker.”

https: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Bias in Generative AI



Working Women Misrepresented Across the Board

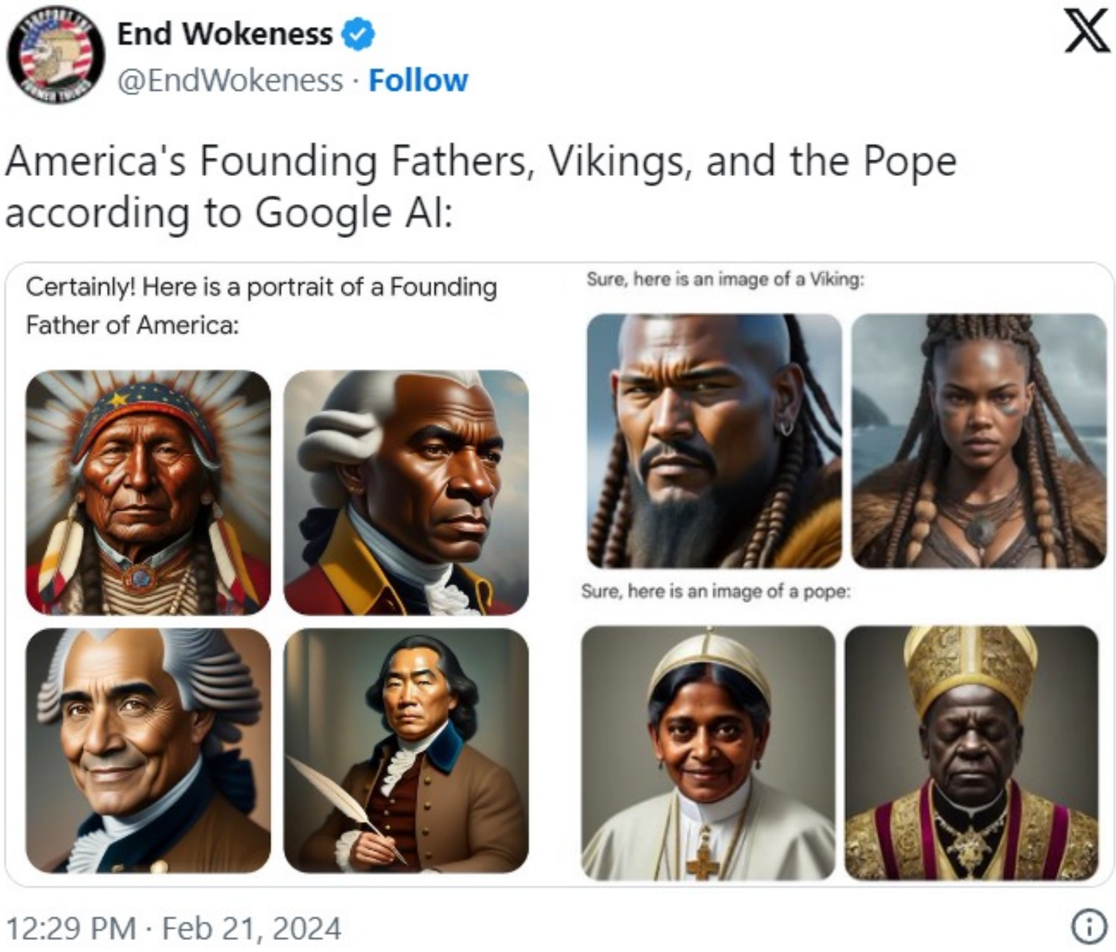
Stable Diffusion results compared to US demographics for each occupation

Sources: Bureau of Labor Statistics, American Medical Association, National Association of Women Judges, Federal Judicial Center, Bloomberg analysis of Stable Diffusion

- Results generated by Stable Diffusion were compared with the US Bureau of Labour Statistics
- For the keyword ‘judge’ results from stable diffusion show only 3% are women when in reality 34% of US judges are women.

https: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Is Google Gemini Racist?



End Wokeness
@EndWokeness · Follow

America's Founding Fathers, Vikings, and the Pope according to Google AI:

Certainly! Here is a portrait of a Founding Father of America:

Sure, here is an image of a Viking:

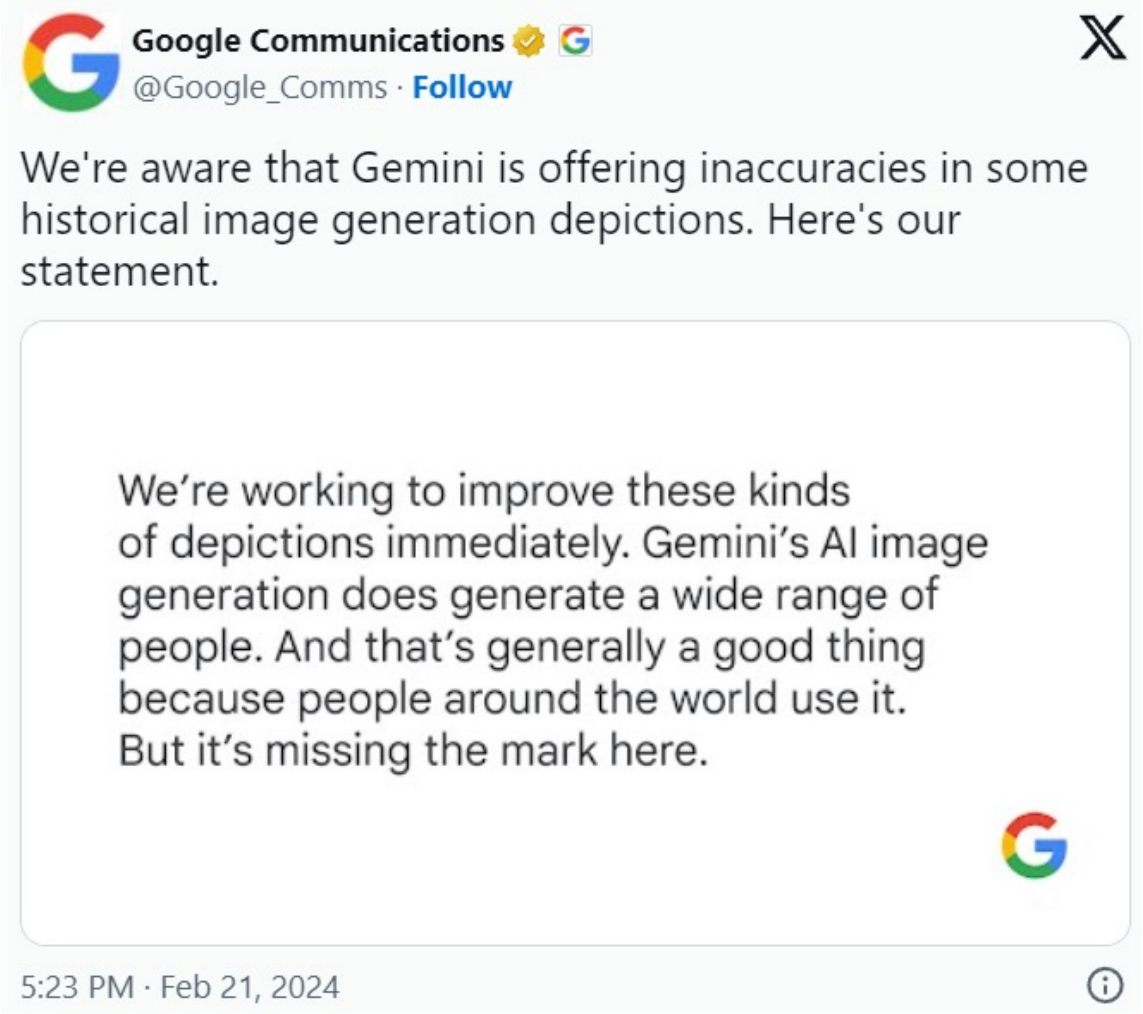
Sure, here is an image of a pope:

12:29 PM · Feb 21, 2024

- Gemini when asked to generate images, would modify the race of historical figures who were white.
- Gemini even refused to generate images in response to prompts like 'show a picture of a white person'
- This led to accusations that Gemini is racist due to 'anti-white bias.'

Source: https://em360tech.com/tech-article/is-gemini-racist

Is Google Gemini Racist?



Google Communications
@Google_Comms · Follow

We're aware that Gemini is offering inaccuracies in some historical image generation depictions. Here's our statement.

We're working to improve these kinds of depictions immediately. Gemini's AI image generation does generate a wide range of people. And that's generally a good thing because people around the world use it. But it's missing the mark here.

5:23 PM · Feb 21, 2024

- Google has paused the image generation feature of Gemini
- AI bias is an issue across most image generation platforms such as 'Dall-E' from OpenAI showed biases such as only white men for 'CEO' or only white women for 'nurse'

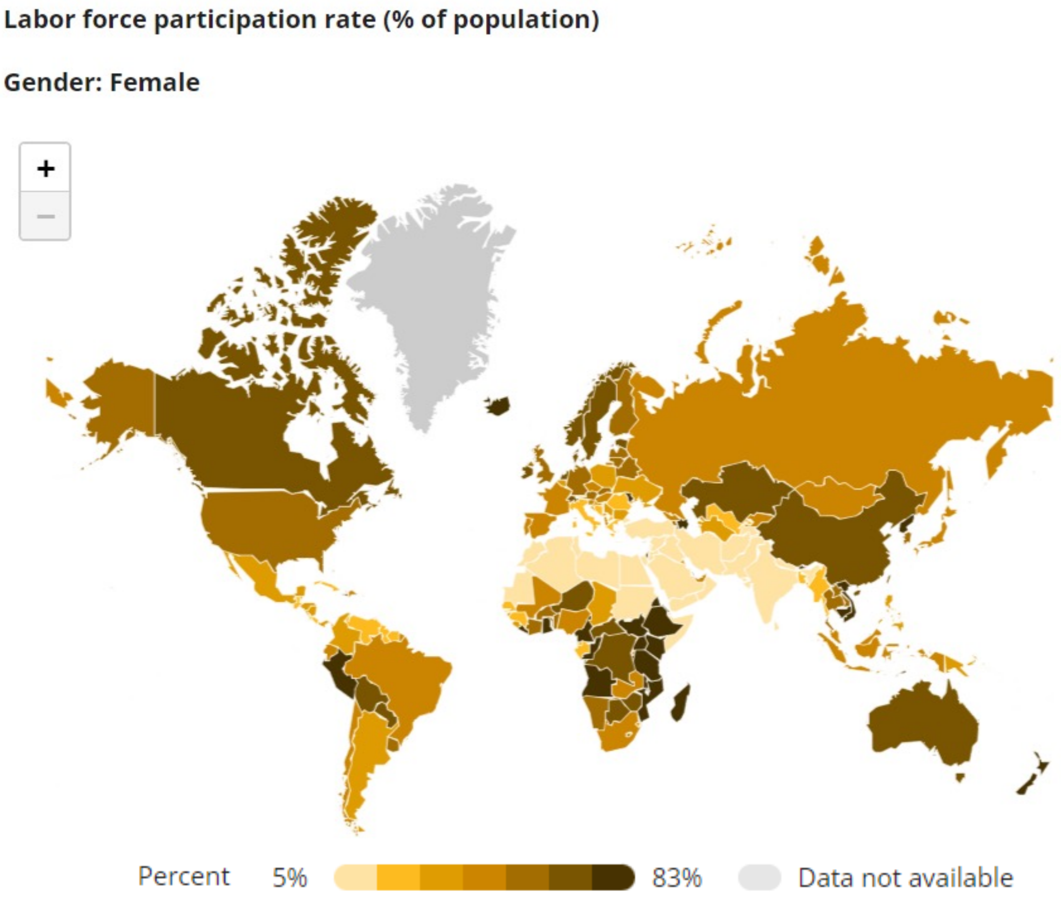
Source: https://em360tech.com/tech-article/is-gemini-racist

Datasphere Initiative

- We need to ensure that data used for policy and decision-making provides an accurate picture of all human lives building a more equitable and inclusive digital society
- The author of the book '*Invisible Women: Exposing Data Bias in a World Designed for Men*' has referred to the "gender data gap" as a "phenomenon whereby the vast majority of information that we have collected globally and continue to collect — everything from economic data to urban planning data to medical data — have been collected on men"

Source: Breaking gender bias for a more diverse and equitable Datasphere - The Datasphere Initiative

Gender Data Portal – The World Bank



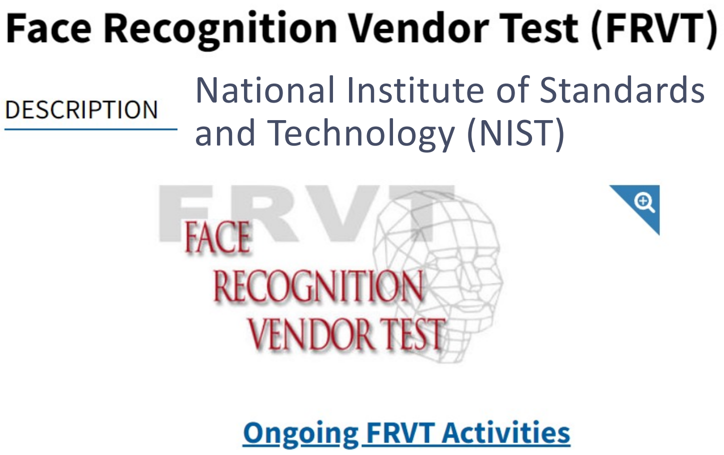
Labor force participation rate (% of population)

Gender: Female

- Labour force participation rate in different countries
- The portal provides latest gender statistics in several areas to inform policy choices

Source: [Home | Gender Data Portal | World Bank Gender Data Portal]

Few other resources

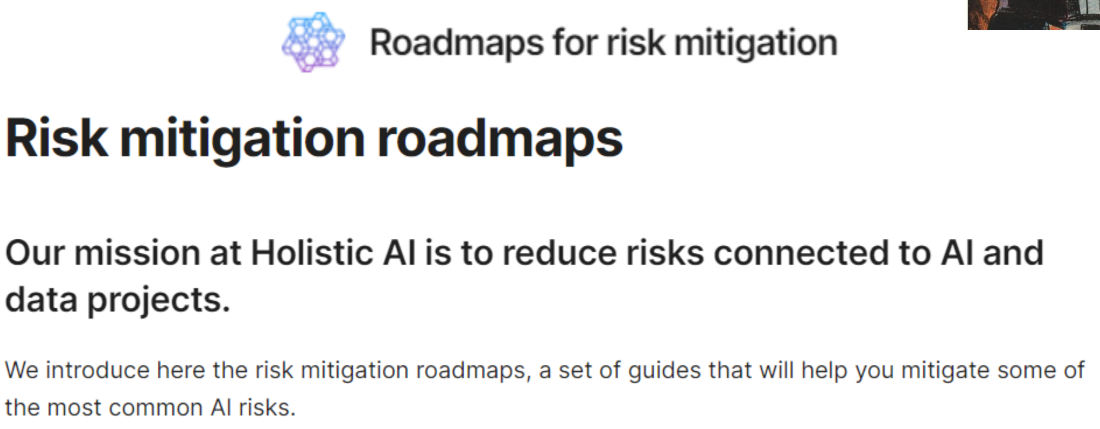


Face Recognition Vendor Test (FRVT)

DESCRIPTION National Institute of Standards and Technology (NIST)

Ongoing FRVT Activities

Source: Face Recognition Vendor Test (FRVT) | NIST



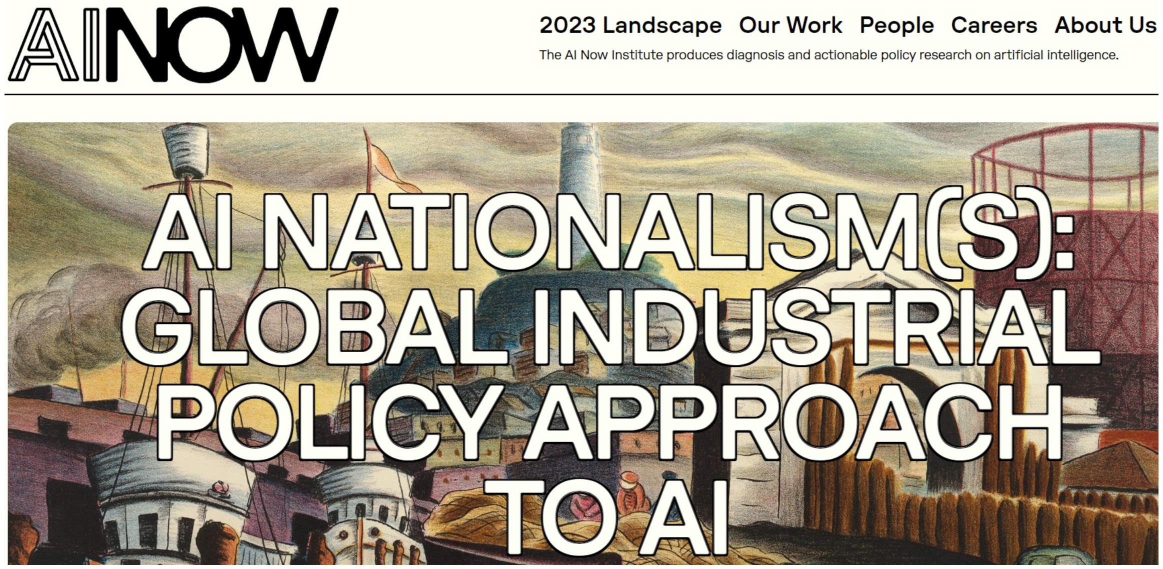
Roadmaps for risk mitigation

Risk mitigation roadmaps

Our mission at Holistic AI is to reduce risks connected to AI and data projects.

We introduce here the risk mitigation roadmaps, a set of guides that will help you mitigate some of the most common AI risks.

Source: Risk mitigation roadmaps | Roadmaps for risk mitigation (gitbook.io)



AINOW

2023 Landscape Our Work People Careers About Us

The AI Now Institute produces diagnosis and actionable policy research on artificial intelligence.

AI NATIONALISM(S): GLOBAL INDUSTRIAL POLICY APPROACH TO AI

The AI Now Institute produces diagnosis and actionable policy research on artificial intelligence.

Source: Home - AI Now Institute

A specialist energy consultancy

Thank you for listening!

For any questions, please contact me at diptargha@tneigroup.com