Responsible AI

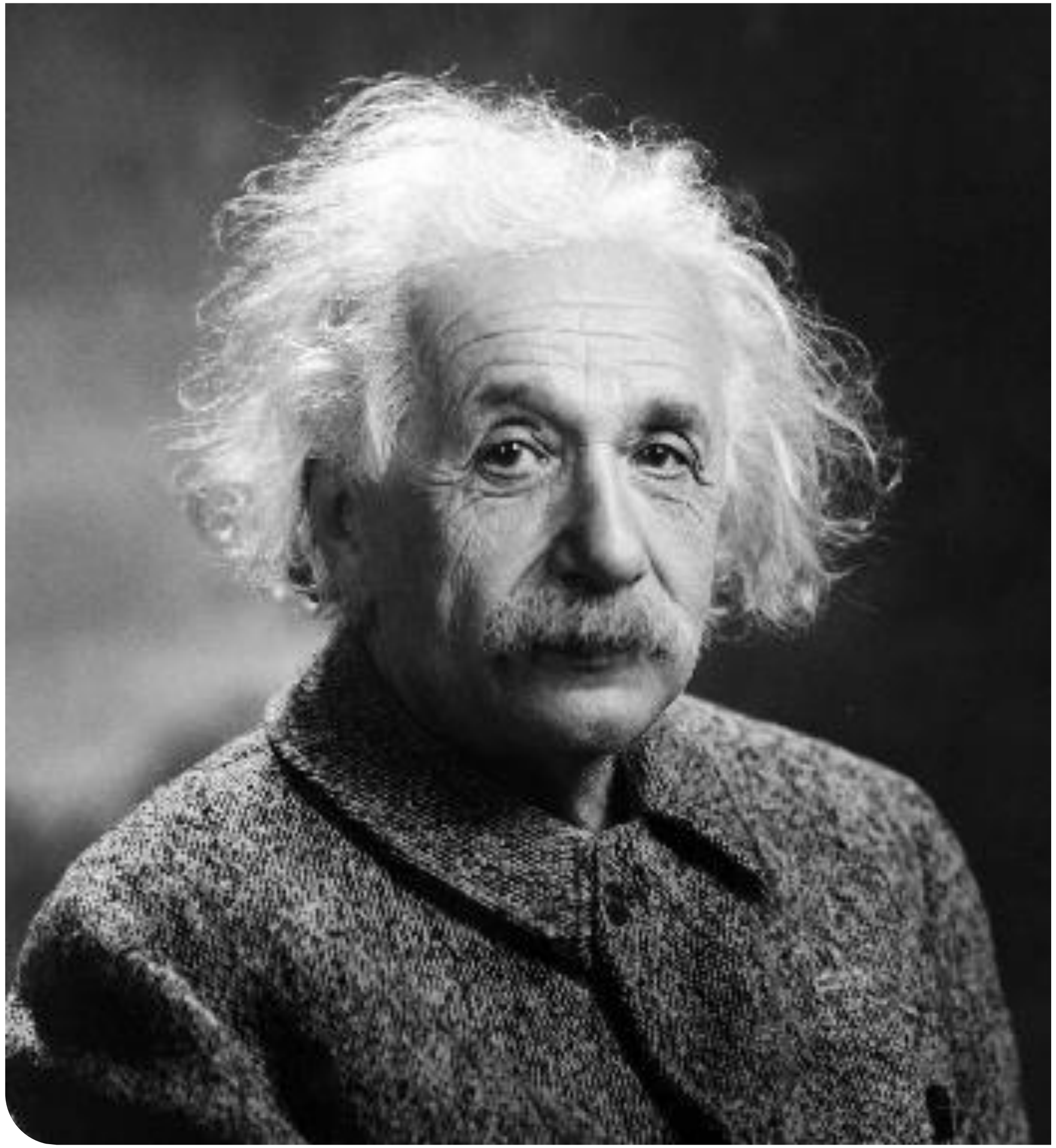
A short introduction for a general audience

Prof. David Martens

Co-director ACRAI
Antwerp Center for Responsible AI

Nov 2023

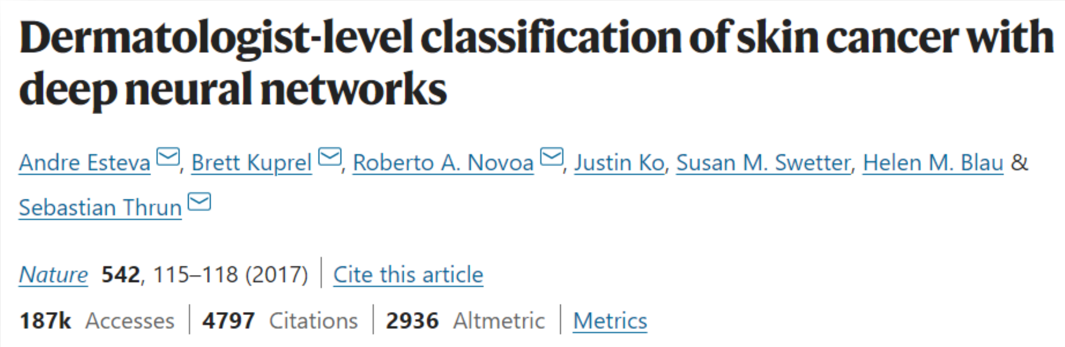
Universiteit Antwerpen



Einstein's Warning

"Technological progress is like an axe in the hand of a pathological criminal."

The age of AI

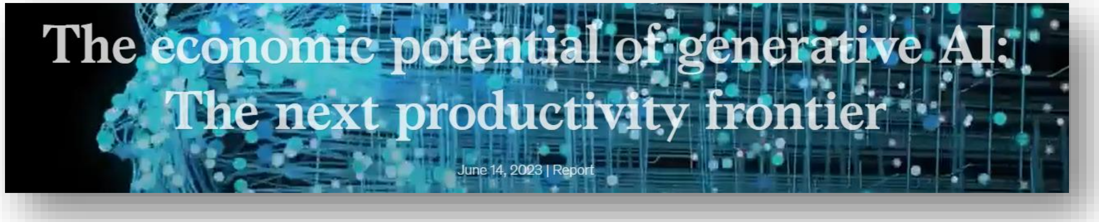


Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

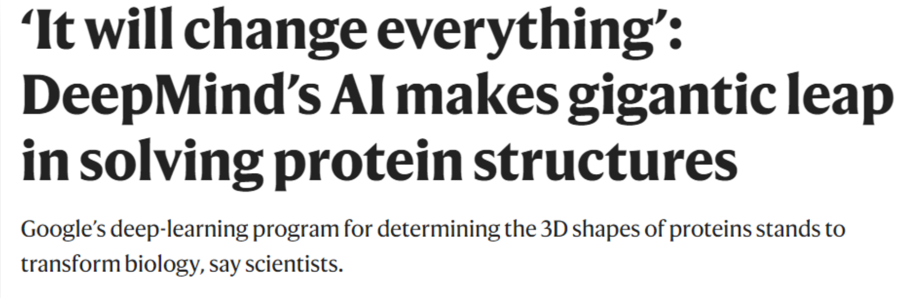
Nature **542**, 115–118 (2017) | Cite this article

187k Accesses | **4797** Citations | **2936** Altmetric | Metrics



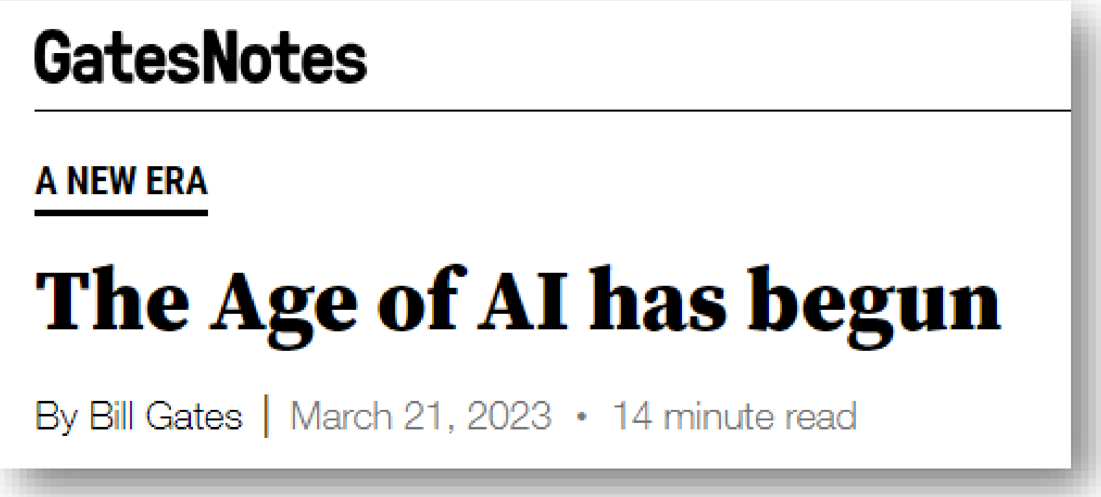
The economic potential of generative AI: The next productivity frontier

June 14, 2023 | Report



‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures

Google’s deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.



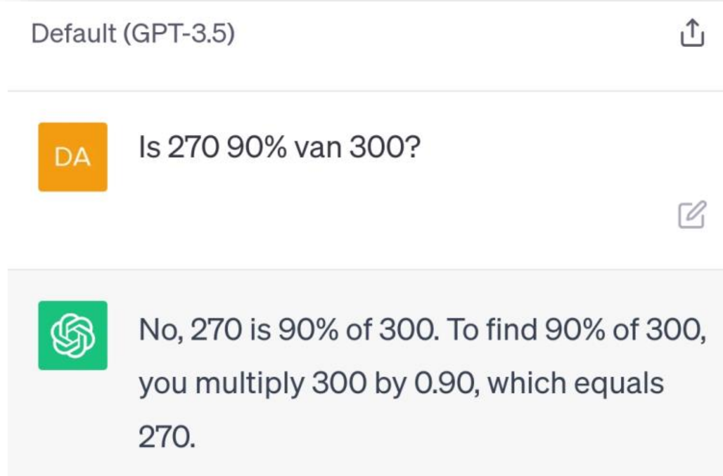
GatesNotes

A NEW ERA

The Age of AI has begun

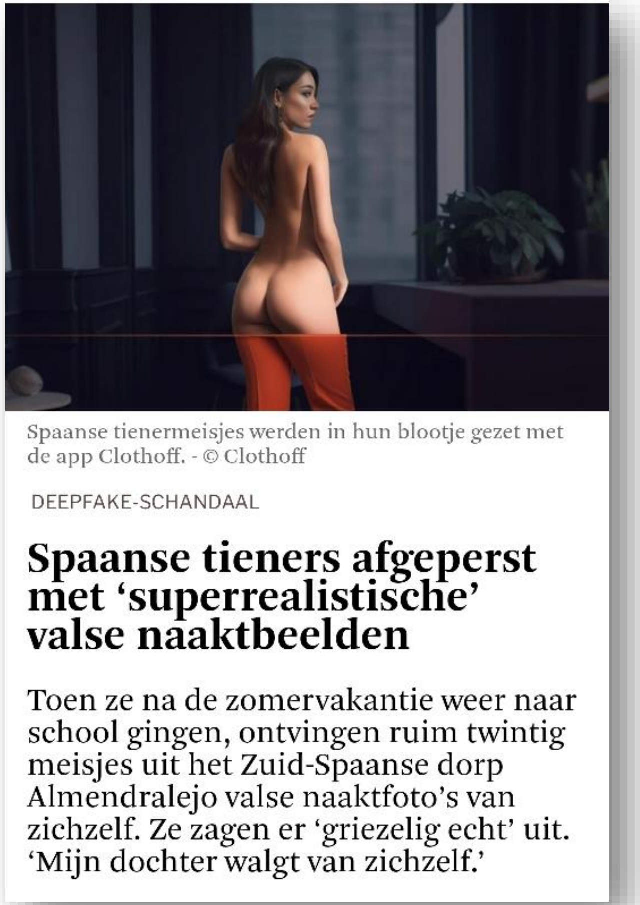
By Bill Gates | March 21, 2023 • 14 minute read

The age of AI?



Is 270 90% van 300?

No, 270 is 90% of 300. To find 90% of 300, you multiply 300 by 0.90, which equals 270.

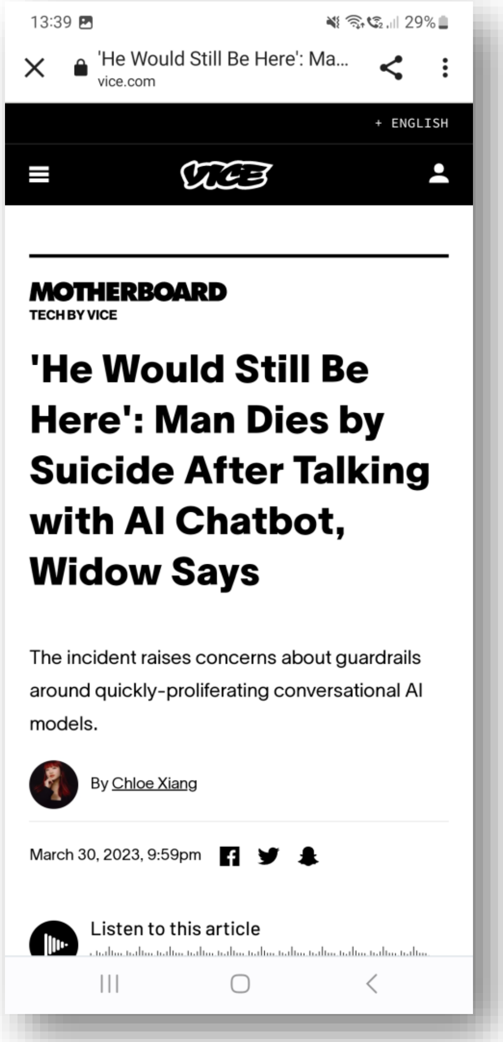


Spaanse tienermeisjes werden in hun blootje gezet met de app Clothoff. - © Clothoff

DEEPFAKE-SCHANDAAL

Spaanse tieners afgeperst met ‘superrealistische’ valse naaktbeelden

Toen ze na de zomervakantie weer naar school gingen, ontvingen ruim twintig meisjes uit het Zuid-Spaanse dorp Almendralejo valse naaktfoto’s van zichzelf. Ze zagen er ‘griezelig echt’ uit. ‘Mijn dochter walgt van zichzelf.’



'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says

The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

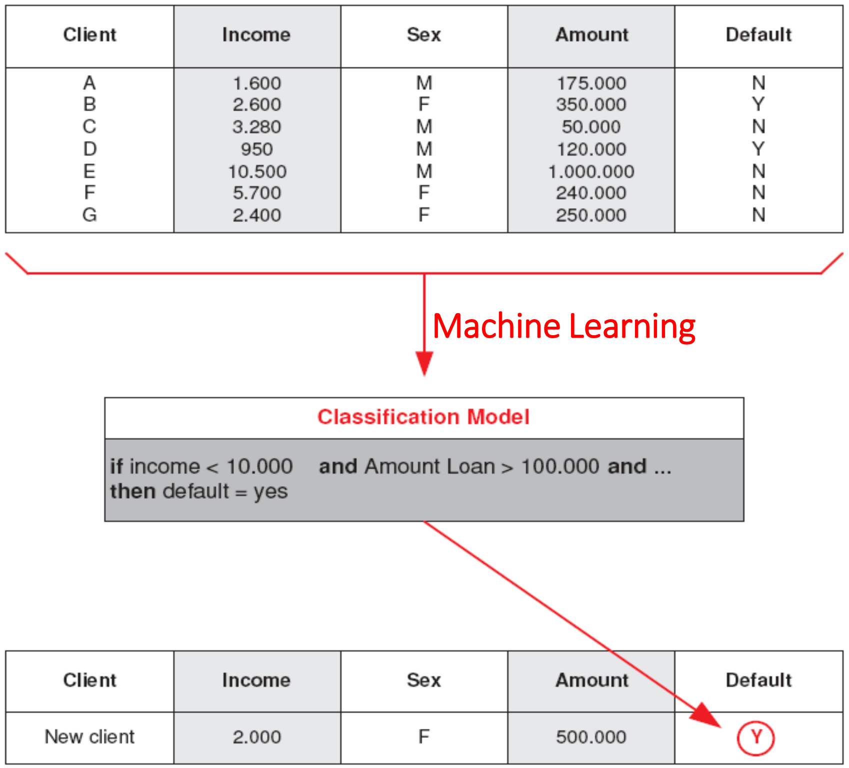
By Chloe Xiang

March 30, 2023, 9:59pm

Listen to this article

Machine Learning

- **Machine Learning:** automatic extraction of knowledge from data
- Setting the scene with credit scoring example



Client	Income	Sex	Amount	Default
A	1.600	M	175.000	N
B	2.600	F	350.000	Y
C	3.280	M	50.000	N
D	950	M	120.000	Y
E	10.500	M	1.000.000	N
F	5.700	F	240.000	N
G	2.400	F	250.000	N

Data

Machine Learning

Machine learning technique

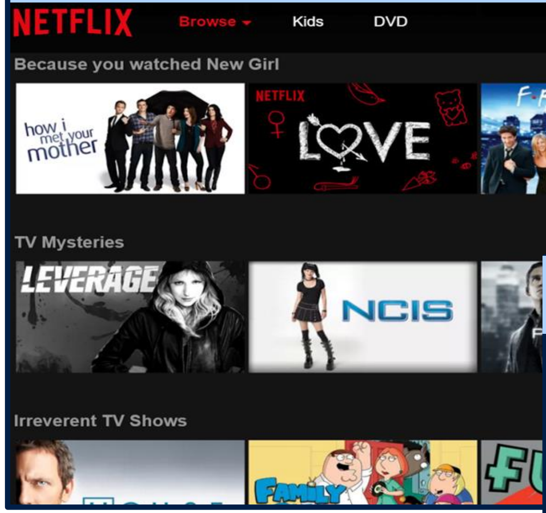
Classification Model

if income < 10.000 **and** Amount Loan > 100.000 **and** ...
then default = yes

Pattern

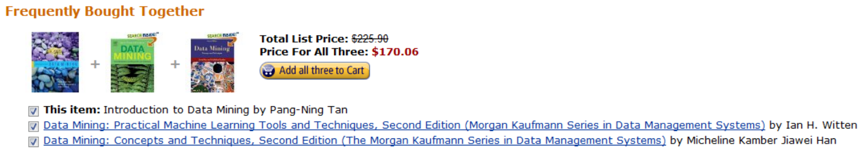
Client	Income	Sex	Amount	Default
New client	2.000	F	500.000	Y

Recommender Systems



Targeted Advertising

Many more …

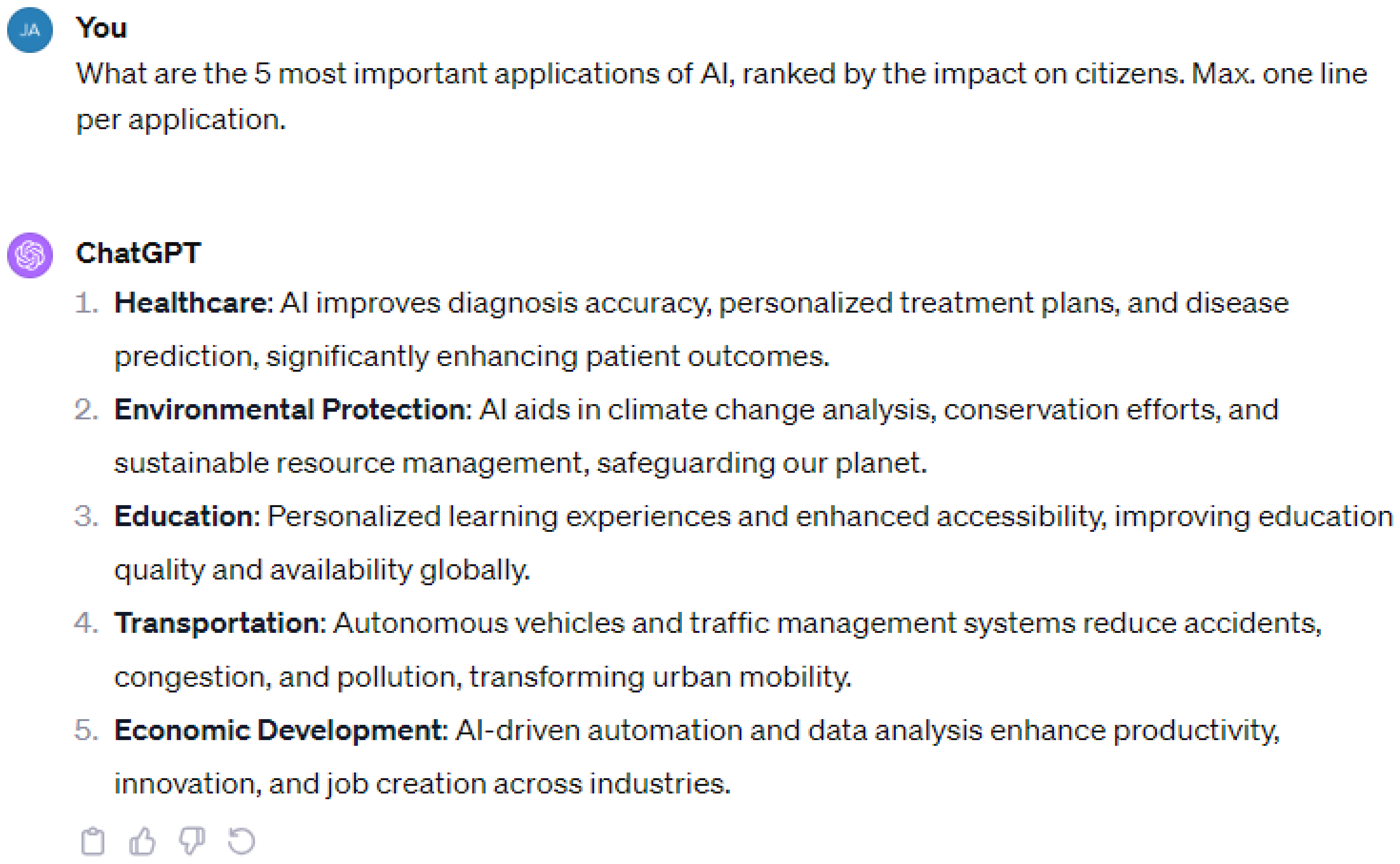

Frequently Bought Together

Total List Price: ~~$225.90~~
Price For All Three: $170.06

Add all three to Cart

☑ **This item:** Introduction to Data Mining by Pang-Ning Tan
☑ Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems) by Ian H. Witten
☑ Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems) by Micheline Kamber Jiawei Han

Fraud Detection

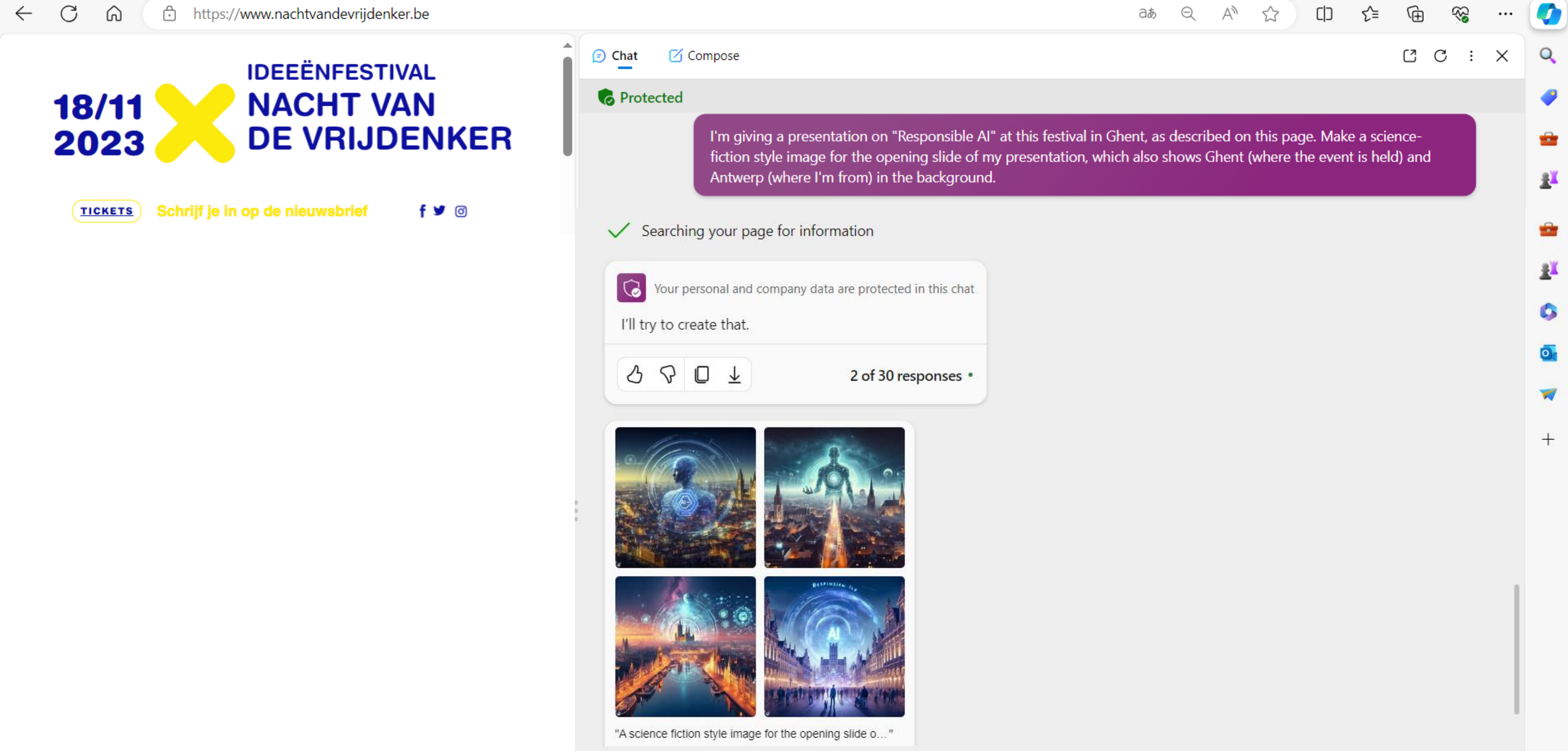


You

What are the 5 most important applications of AI, ranked by the impact on citizens. Max. one line per application.

ChatGPT

1. **Healthcare**: AI improves diagnosis accuracy, personalized treatment plans, and disease prediction, significantly enhancing patient outcomes.
2. **Environmental Protection**: AI aids in climate change analysis, conservation efforts, and sustainable resource management, safeguarding our planet.
3. **Education**: Personalized learning experiences and enhanced accessibility, improving education quality and availability globally.
4. **Transportation**: Autonomous vehicles and traffic management systems reduce accidents, congestion, and pollution, transforming urban mobility.
5. **Economic Development**: AI-driven automation and data analysis enhance productivity, innovation, and job creation across industries.



I'm giving a presentation on "Responsible AI" at this festival in Ghent, as described on this page. Make a science-fiction style image for the opening slide of my presentation, which also shows Ghent (where the event is held) and Antwerp (where I'm from) in the background.

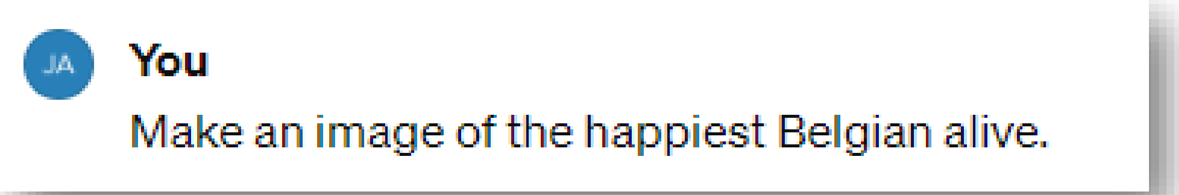
✓ Searching your page for information

Your personal and company data are protected in this chat

I'll try to create that.

2 of 30 responses

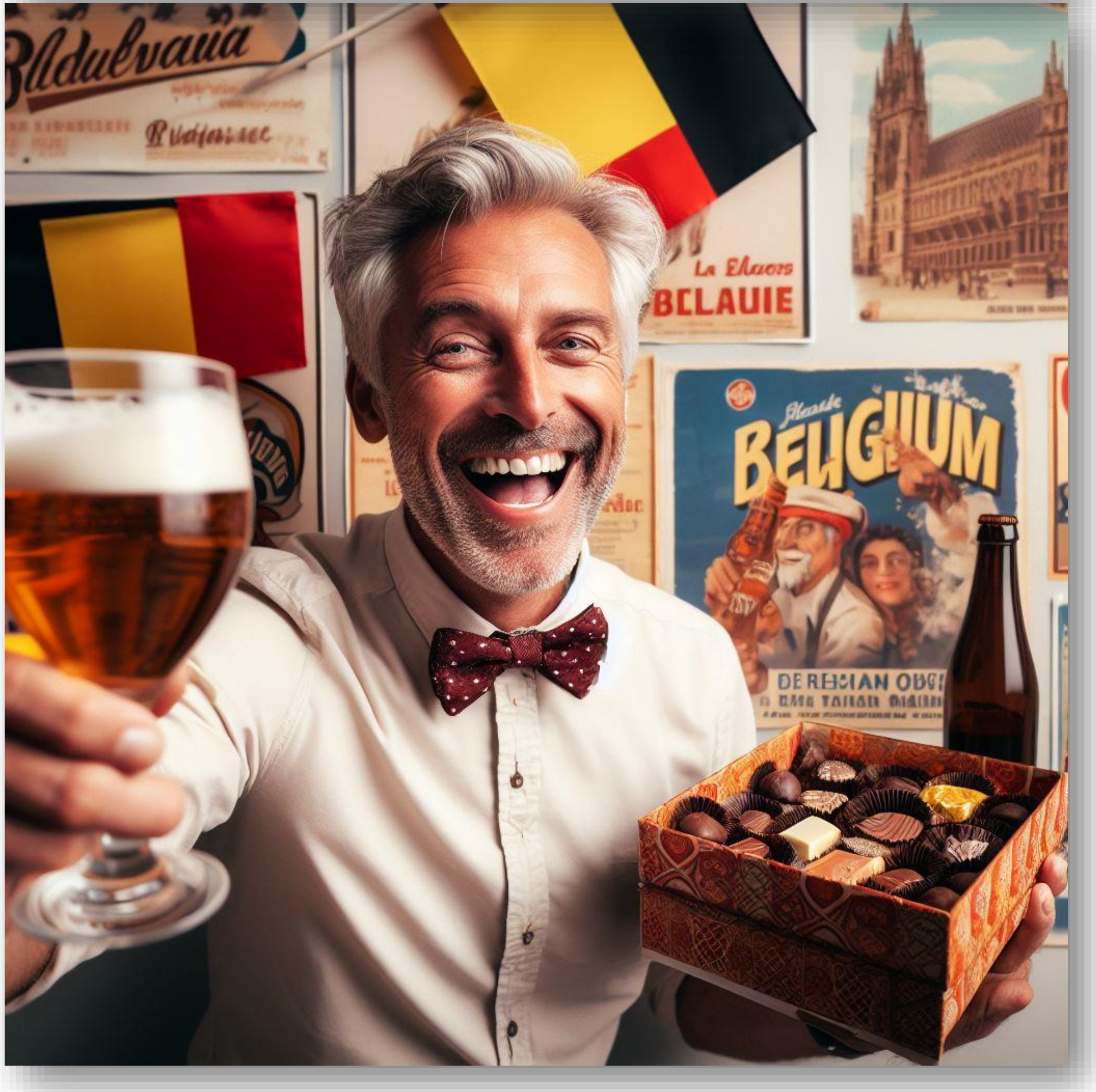
"A science fiction style image for the opening slide o..."

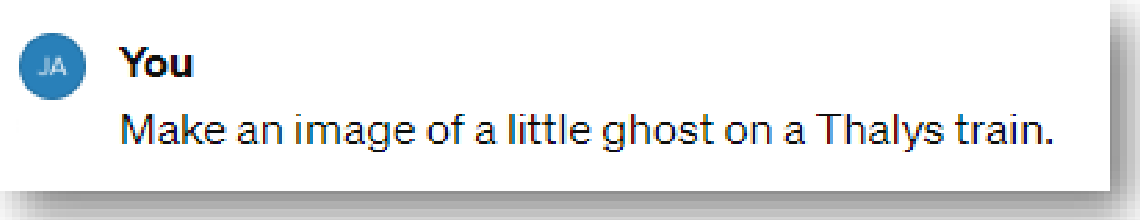


JA

You

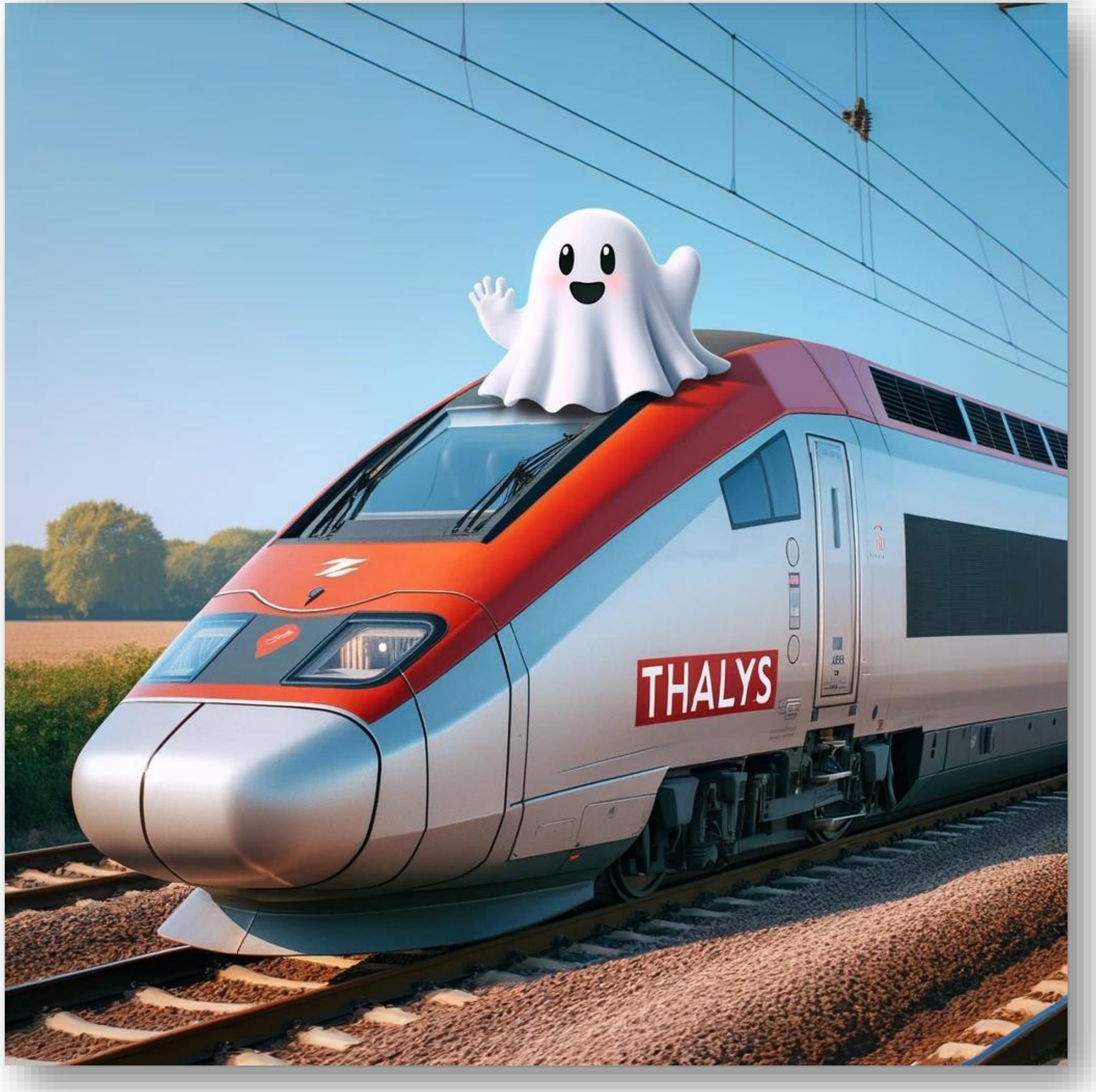
Make an image of the happiest Belgian alive.





JA **You**

Make an image of a little ghost on a Thalys train.



Responsible AI

The development and application of AI that is aligned with moral values of society

AI Risks

Immediate

Systemic

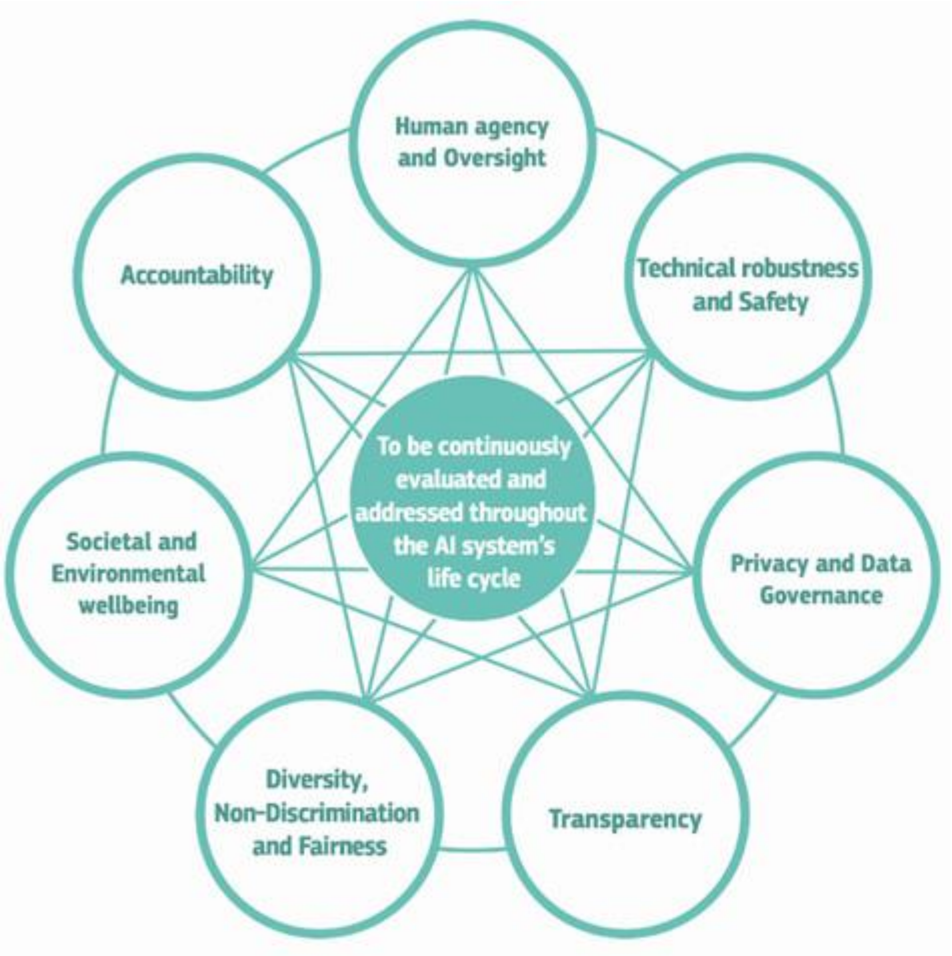
Long term

Images by Midjourney

AI Risks

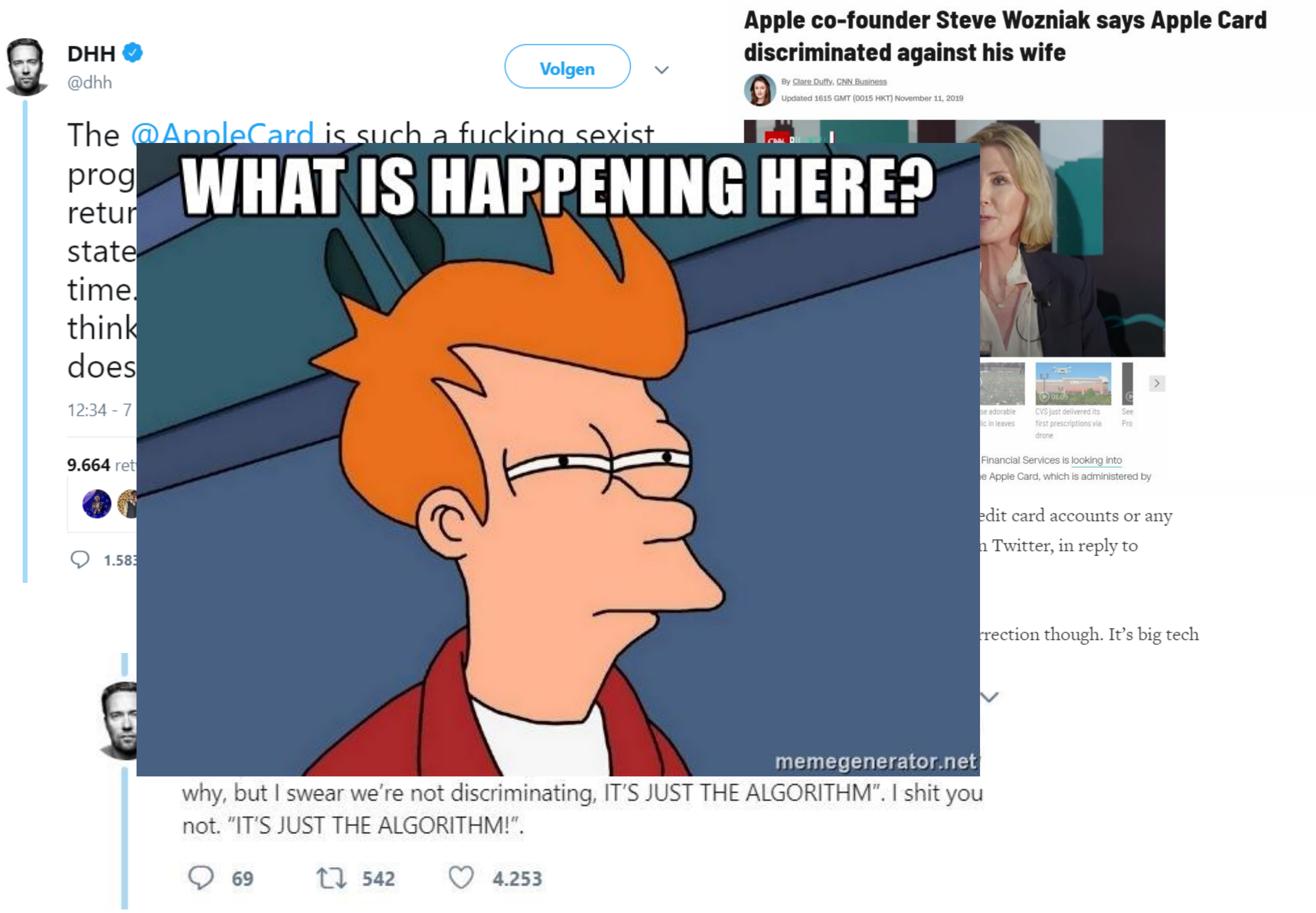
Immediate

A selection follows next.


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

Europe's Ethics Guidelines for Trustworthy AI (2019)

1. Explainable AI



DHH @dhh

Volgen

The @AppleCard is such a fucking sexist
prog
retur
state
time.
think
does

12:34 - 7

9.664 ret

1.58

Apple co-founder Steve Wozniak says Apple Card discriminated against his wife

By Clare Duffy, CNN Business

Updated 1615 GMT (0015 HKT) November 11, 2019

Financial Services is looking into
e Apple Card, which is administered by

edit card accounts or any
n Twitter, in reply to

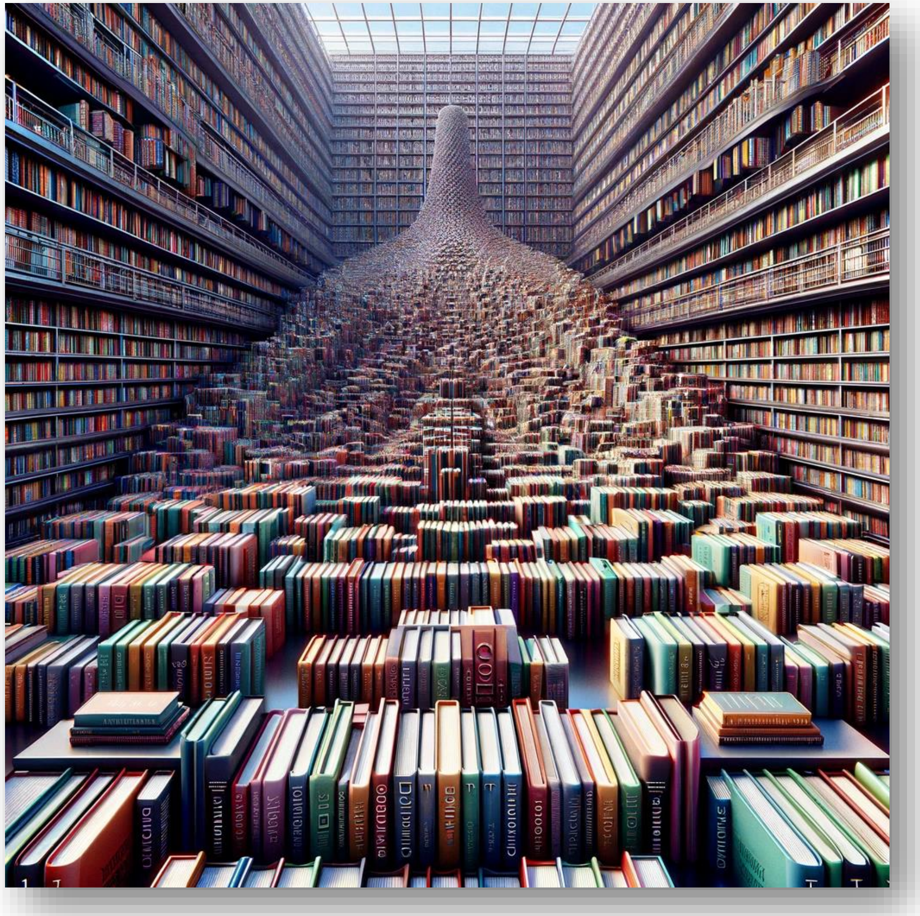
rrection though. It's big tech

why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM". I shit you not. "IT'S JUST THE ALGORITHM!".

69 542 4.253

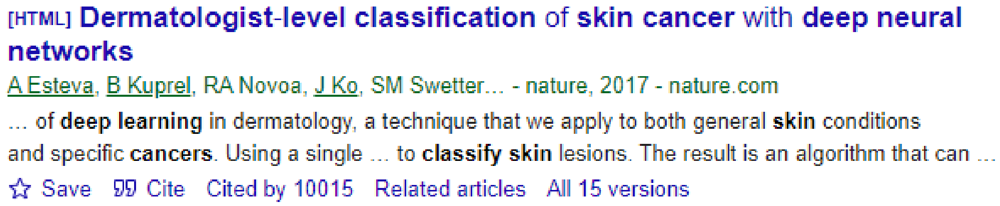
Black Box?

- Deep learning: large artificial neural network with massive number of parameters
  - MobileNetV2: 4.3 million parameters
  - GPT-4: >1 trillion parameters (an image of a printed version of the formula…)



Trust: lab-setting versus real-life

- Data: image of skin lesion
- Task: diagnose skin cancer
- High test accuracy, matching accuracy of 21 dermatologists



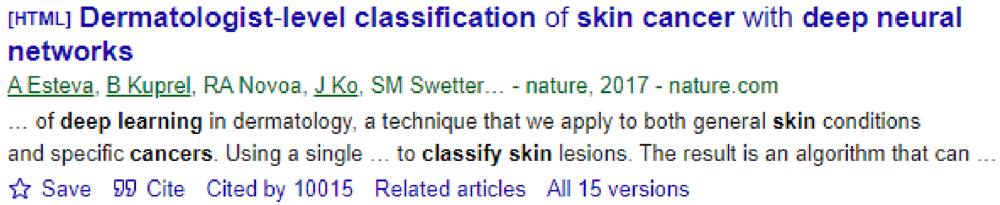
[HTML] **Dermatologist-level classification** of **skin cancer** with **deep neural networks**

A Esteva, B Kuprel, RA Novoa, J Ko, SM Swetter… - nature, 2017 - nature.com

… of **deep learning** in dermatology, a technique that we apply to both general **skin** conditions and specific **cancers**. Using a single … to **classify skin** lesions. The result is an algorithm that can …

https://towardsdatascience.com/is-the-medias-reluctance-to-admit-ai-s-weaknesses-putting-us-at-risk-c355728e9028

Trust: lab-setting versus real-life

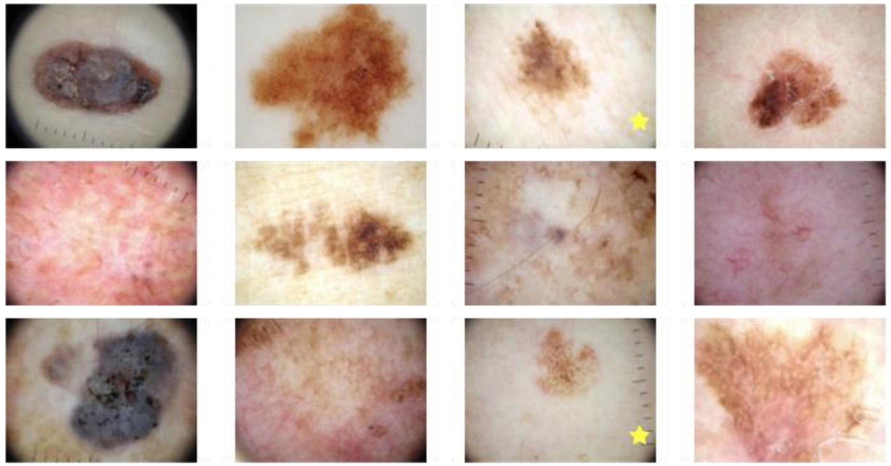


[HTML] **Dermatologist-level classification of skin cancer with deep neural networks**

A Esteva, B Kuprel, RA Novoa, J Ko, SM Swetter... - nature, 2017 - nature.com

... of **deep learning** in dermatology, a technique that we apply to both general **skin** conditions and specific **cancers**. Using a single ... to **classify skin** lesions. The result is an algorithm that can ...

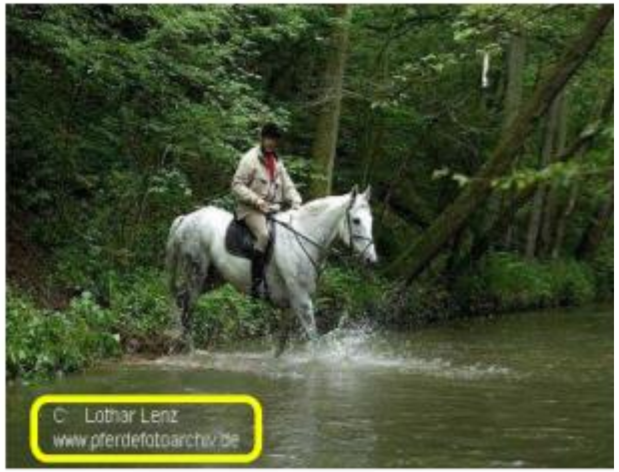
- Data: image of skin lesion
- Task: diagnose skin **cancer**
- High test accuracy, matching accuracy of 21 dermatologists
- Issue: when dermatologist is concerned about lesion, a ruler is placed next to it in the picture
- Pattern learnt?: if ruler then malignant
  (authors stated ruler had limited impact)



https://towardsdatascience.com/is-the-medias-reluctance-to-admit-ai-s-weaknesses-putting-us-at-risk-c355728e9028

Trust: lab-setting versus real-life

- Data: picture
- Task: predict if **horse** or not
- High test accuracy, low accuracy when deployed in the field
- Issue: horse pictures had watermark with copyright at bottom left
- Pattern learnt: if watermark then horse



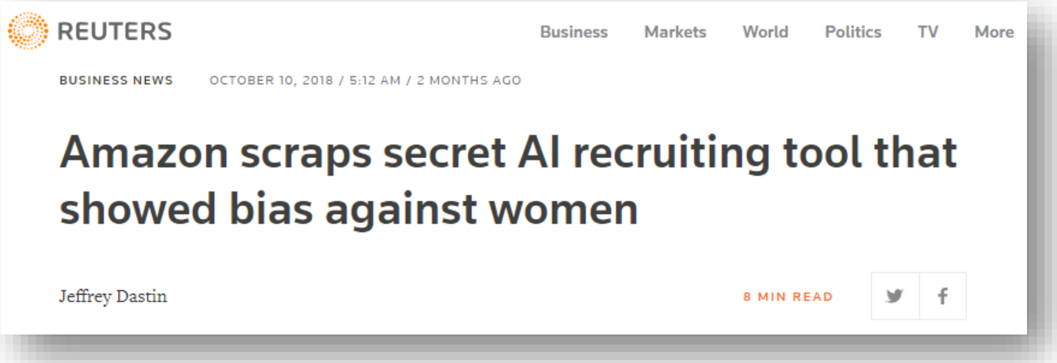
https://arxiv.org/pdf/1902.10178.pdf

2. Fairness

- Discrimination against sensitive groups
- Bias

2. Fairness

- Discrimination against sensitive groups
- Bias
  - HR Analytics, prediciton model to review job applicants’ resumes to automate the search for top talent
  - Trained on resumes from past (10 year period), biased data
  - Model trained to prefer male candidates, for example:
    - Penalized uses of word “woman’s” (eg woman’s chess club president)
    - Penalizes all-woman colleges

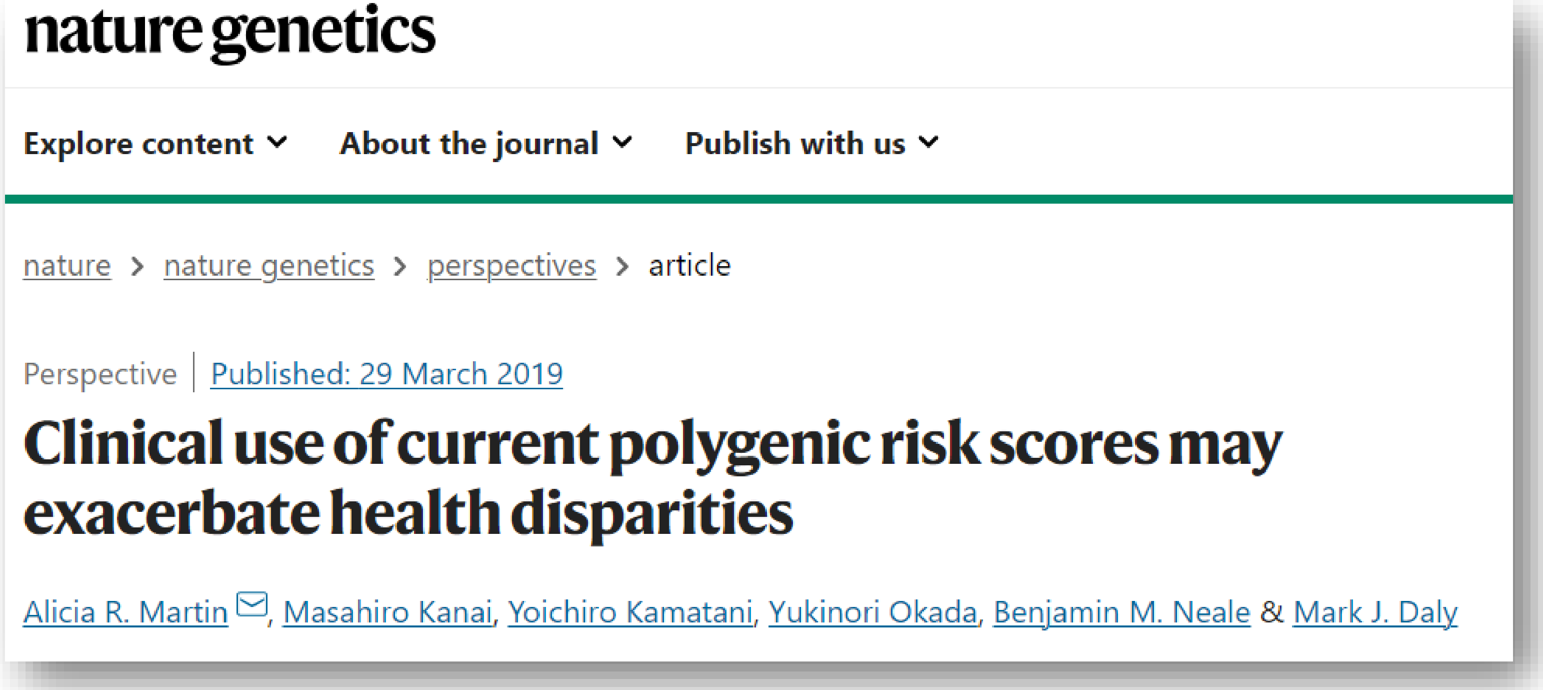


BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / 2 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ



nature genetics

nature > nature genetics > perspectives > article

Perspective | Published: 29 March 2019

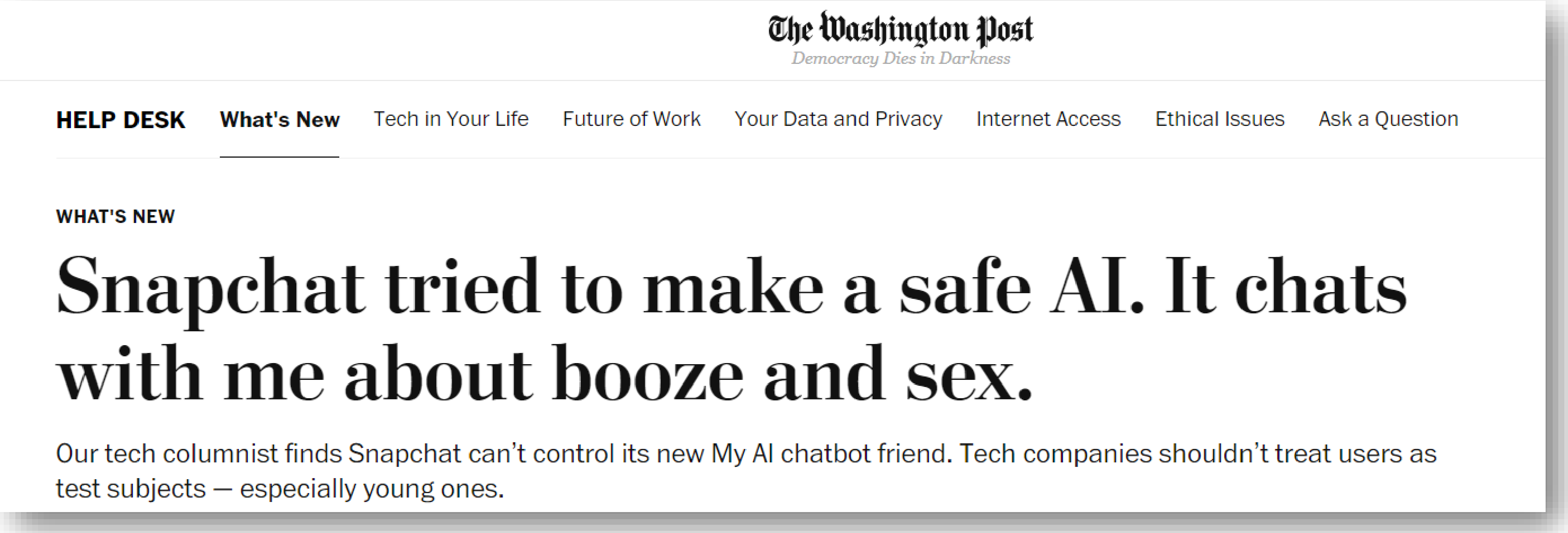
Clinical use of current polygenic risk scores may exacerbate health disparities

Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale & Mark J. Daly

“Although we share enthusiasm about the potential of PRS to improve health outcomes through their eventual routine implementation as clinical biomarkers, we consider the consistent observation that they currently have ***far greater predictive value*** *in individuals of* ***recent European descent*** *than of other ancestries to be the major ethical and scientific challenge surrounding clinical translation and, at present, the* ***most critical limitation to genetics in precision medicine.”***

3. Experimentation

- Ethics as an afterhtought…
- We should be mindful of experiments on our children


The Washington Post

Democracy Dies in Darkness

HELP DESK **What's New** Tech in Your Life Future of Work Your Data and Privacy Internet Access Ethical Issues Ask a Question

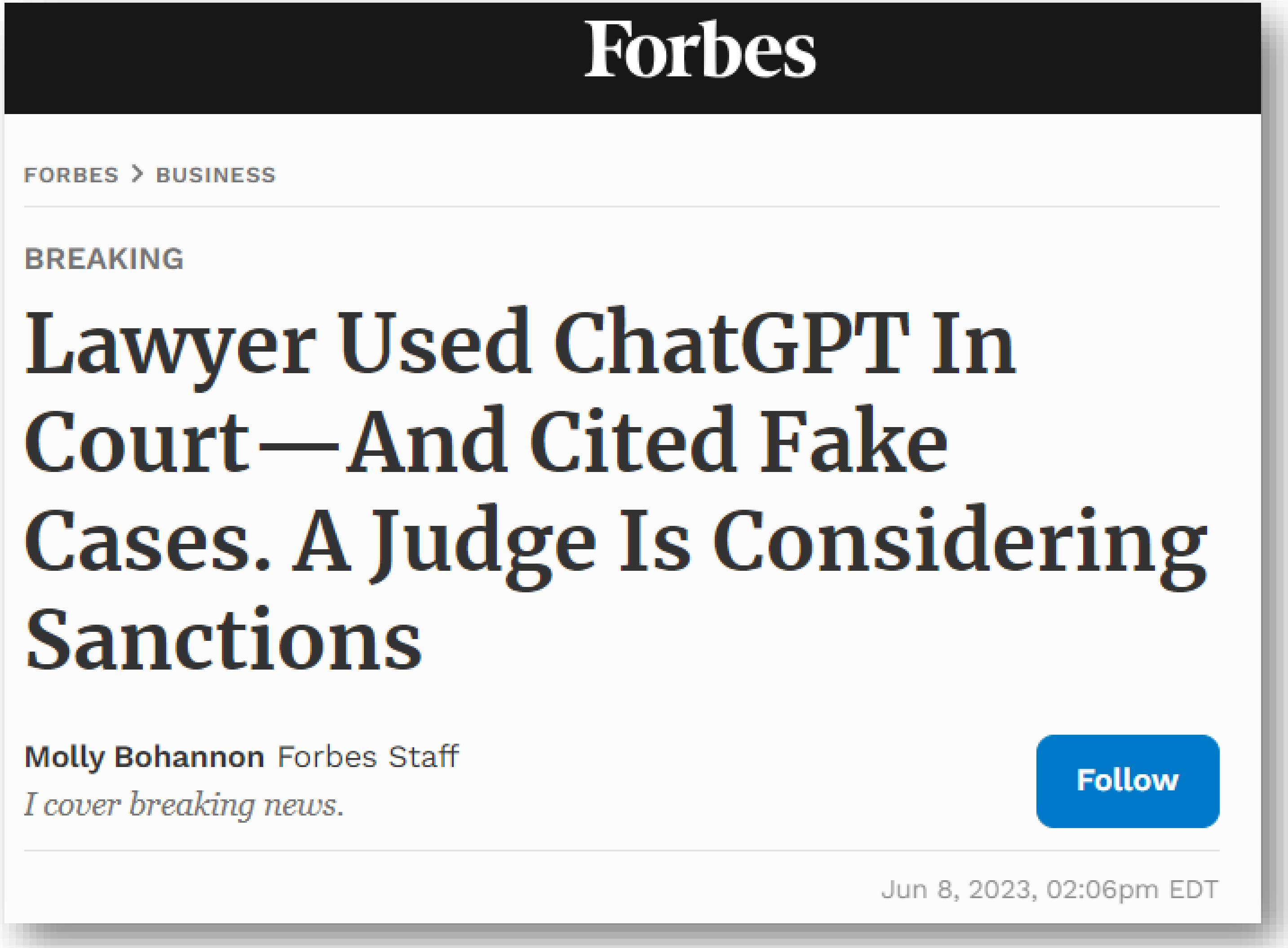
WHAT'S NEW

Snapchat tried to make a safe AI. It chats with me about booze and sex.

Our tech columnist finds Snapchat can't control its new My AI chatbot friend. Tech companies shouldn't treat users as test subjects — especially young ones.

4. Generative AI risks

- **Hallucinations**



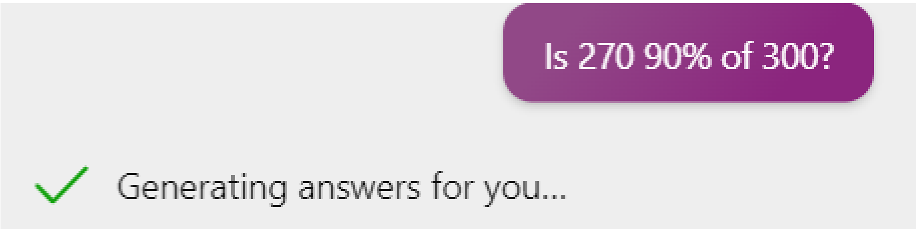
BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

Molly Bohannon Forbes Staff

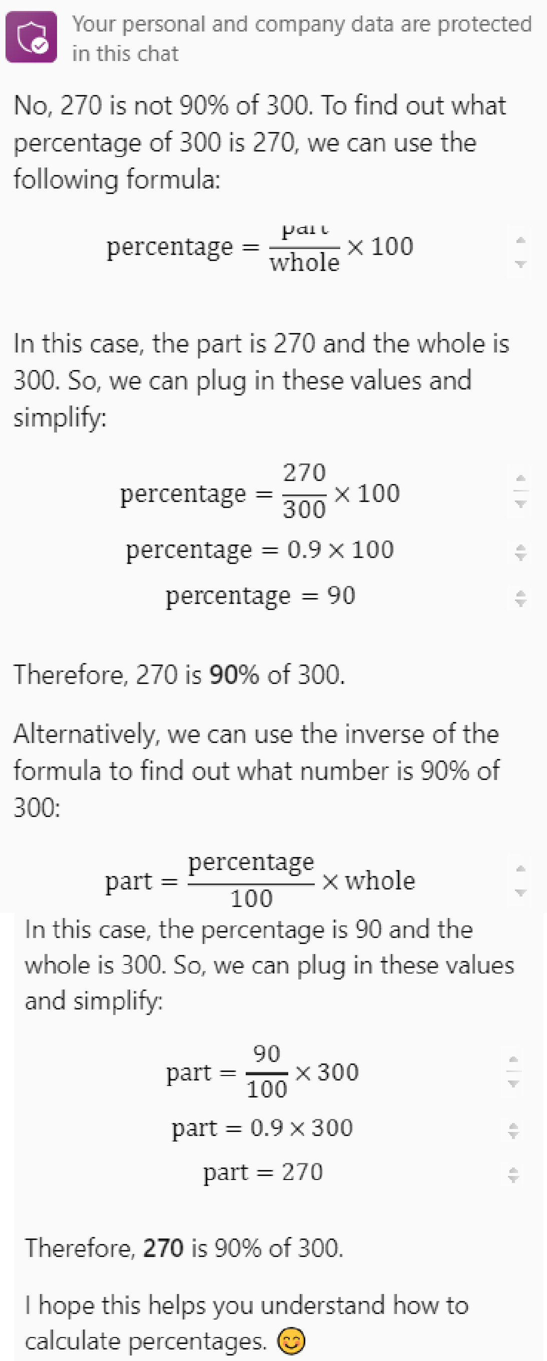
I cover breaking news.

Jun 8, 2023, 02:06pm EDT



Is 270 90% of 300?

✓ Generating answers for you...



No, 270 is not 90% of 300. To find out what percentage of 300 is 270, we can use the following formula:

$$\text{percentage} = \frac{\text{part}}{\text{whole}} \times 100$$

In this case, the part is 270 and the whole is 300. So, we can plug in these values and simplify:

$$\text{percentage} = \frac{270}{300} \times 100$$

$$\text{percentage} = 0.9 \times 100$$

$$\text{percentage} = 90$$

Therefore, 270 is **90%** of 300.

Alternatively, we can use the inverse of the formula to find out what number is 90% of 300:

$$\text{part} = \frac{\text{percentage}}{100} \times \text{whole}$$

In this case, the percentage is 90 and the whole is 300. So, we can plug in these values and simplify:

$$\text{part} = \frac{90}{100} \times 300$$

$$\text{part} = 0.9 \times 300$$

$$\text{part} = 270$$

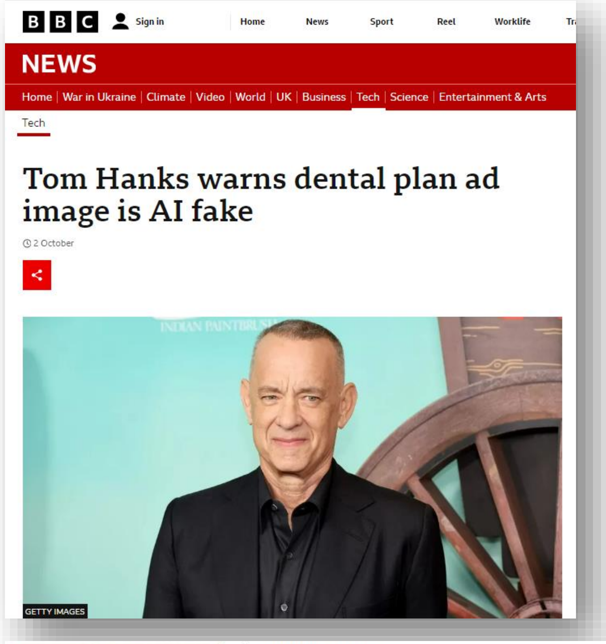
Therefore, **270** is 90% of 300.

I hope this helps you understand how to calculate percentages. 😊

Asked to Bing Chat on October 12th 2023.

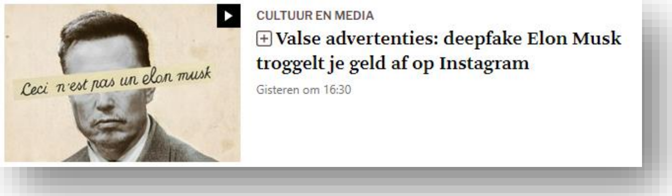
Generative AI risks

1. **Hallucinations**
2. **Misuse**


BBC
Sign in
Home
News
Sport
Reel
Worklife
NEWS
Home | War in Ukraine | Climate | Video | World | UK | Business | Tech | Science | Entertainment & Arts
Tech
INDIAN PAINTBRUSH
GETTY IMAGES

Tom Hanks warns dental plan ad image is AI fake

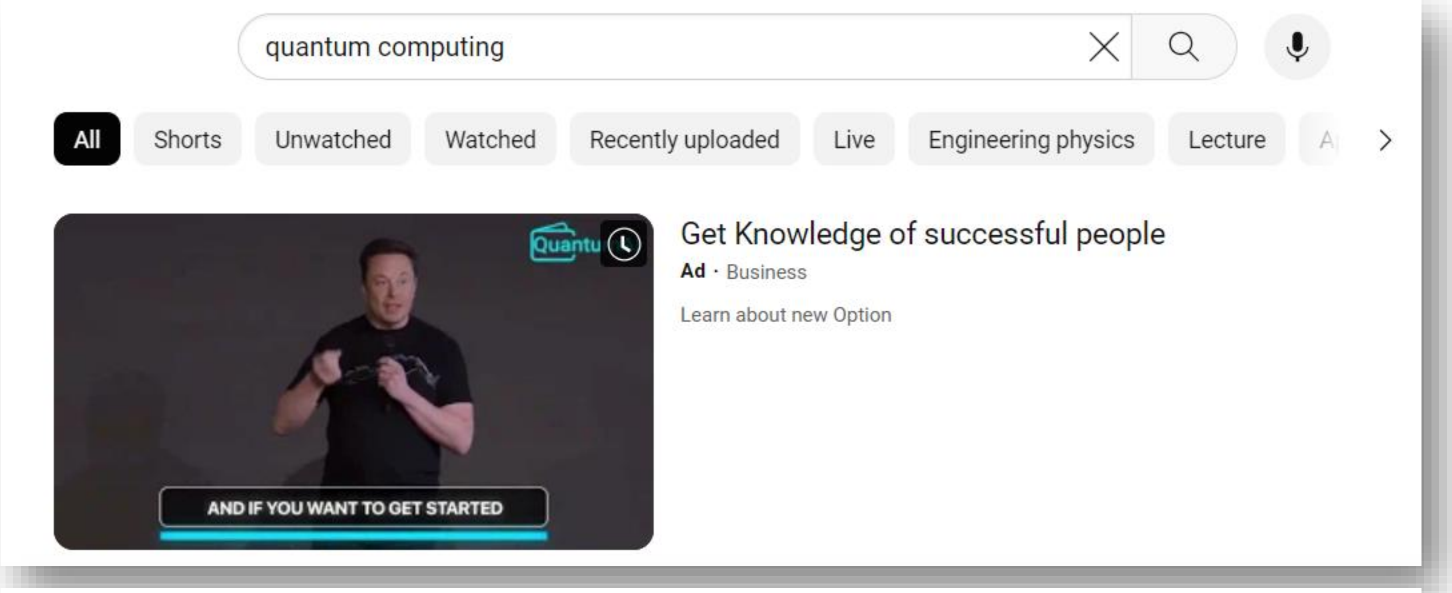
2 October



CULTUUR EN MEDIA

Valse advertenties: deepfake Elon Musk troggelt je geld af op Instagram

Gisteren om 16:30

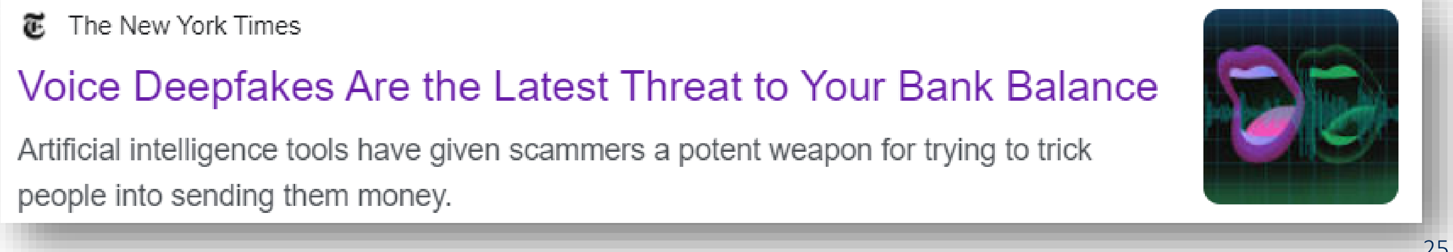


quantum computing

Get Knowledge of successful people

Ad · Business

Learn about new Option



The New York Times

Voice Deepfakes Are the Latest Threat to Your Bank Balance

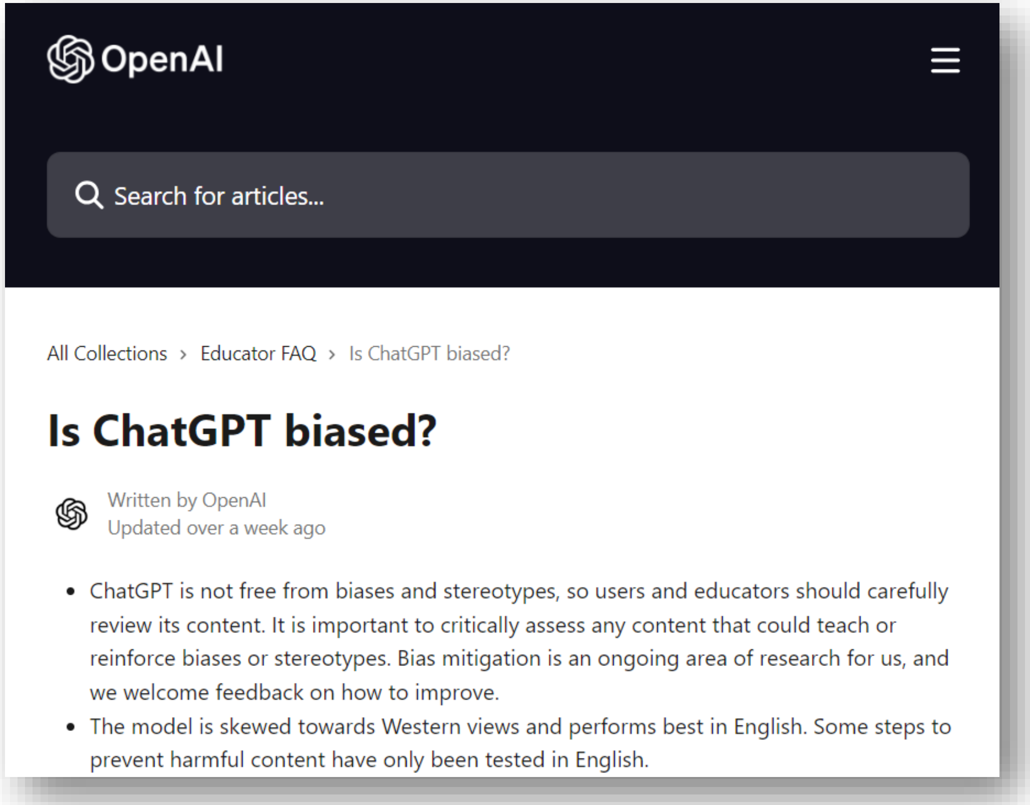
Artificial intelligence tools have given scammers a potent weapon for trying to trick people into sending them money.

Screenhots captured on October 11th, 2023.

Generative AI risks

- **Hallucinations**
- **Misuse**
- **Bias**

GPT-3: based on
- 60% from internet-crawled material,
- 22% from curated content from the internet,
- 16% from books,
- 3% from Wikipedia.



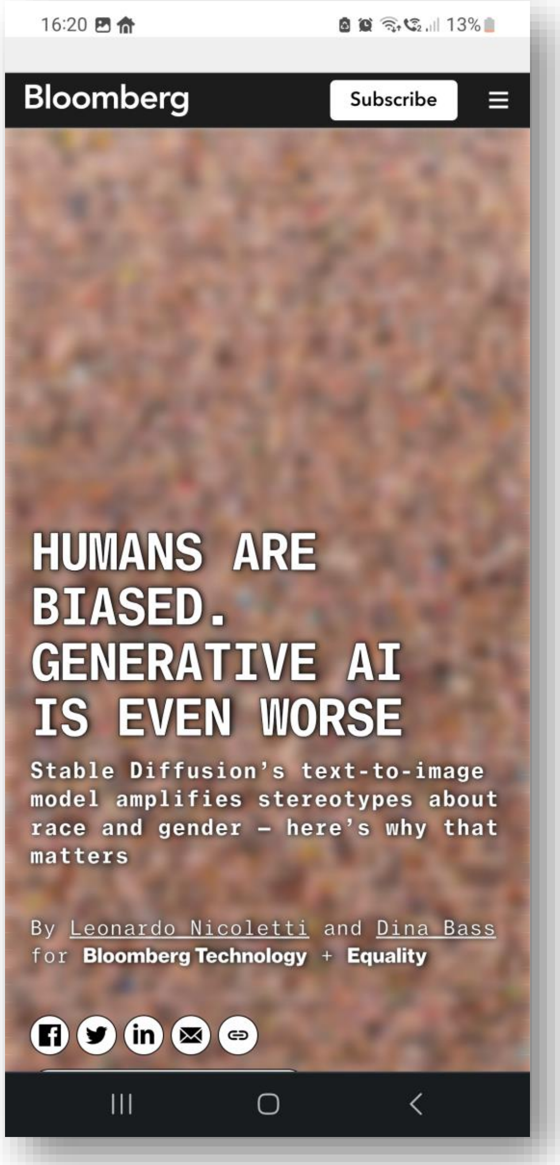
Is ChatGPT biased?

Written by OpenAI
Updated over a week ago

- ChatGPT is not free from biases and stereotypes, so users and educators should carefully review its content. It is important to critically assess any content that could teach or reinforce biases or stereotypes. Bias mitigation is an ongoing area of research for us, and we welcome feedback on how to improve.
- The model is skewed towards Western views and performs best in English. Some steps to prevent harmful content have only been tested in English.

https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/
https://arxiv.org/abs/2005.14165

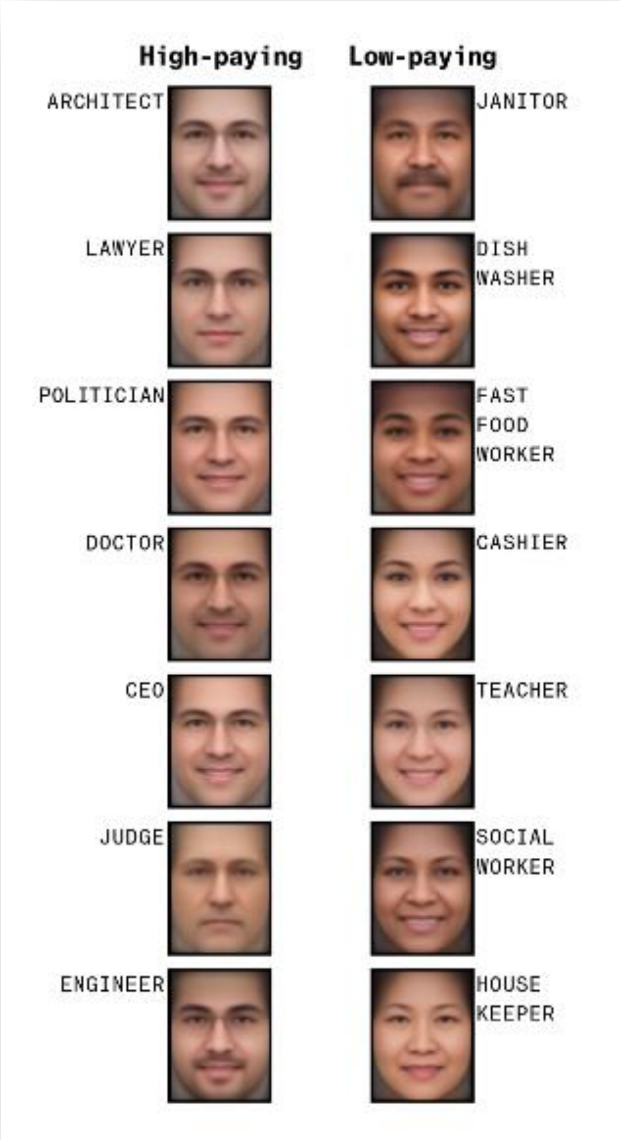
Generative AI risks



HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE

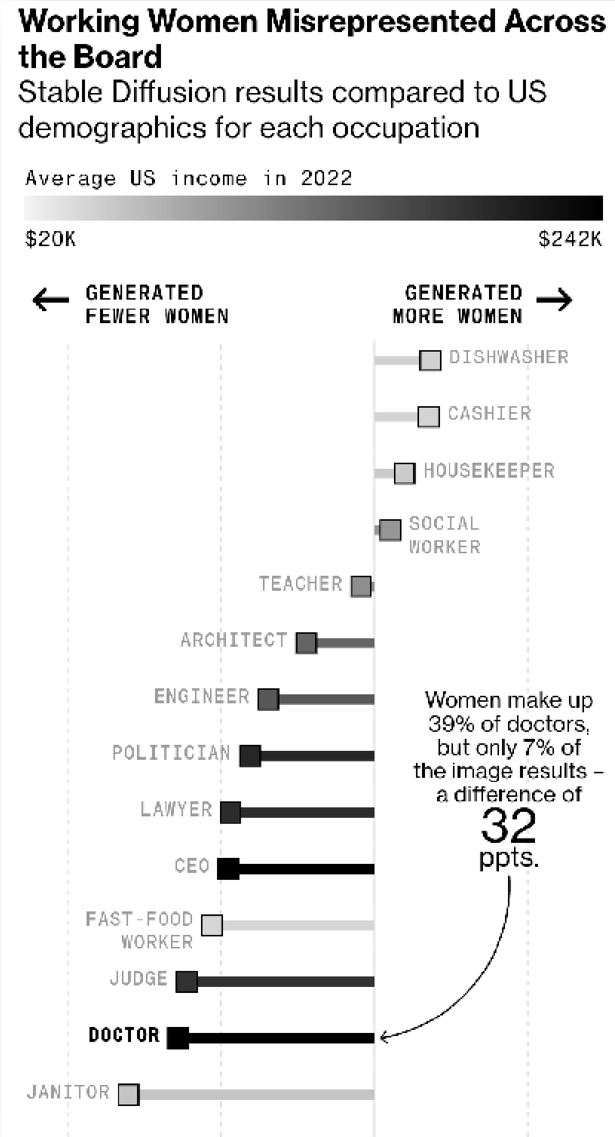
Stable Diffusion's text-to-image model amplifies stereotypes about race and gender – here's why that matters

By Leonardo Nicoletti and Dina Bass for **Bloomberg Technology** + **Equality**



Working Women Misrepresented Across the Board

Stable Diffusion results compared to US demographics for each occupation



Women make up 39% of doctors, but only 7% of the image results – a difference of 32 ppts.

AI Risks

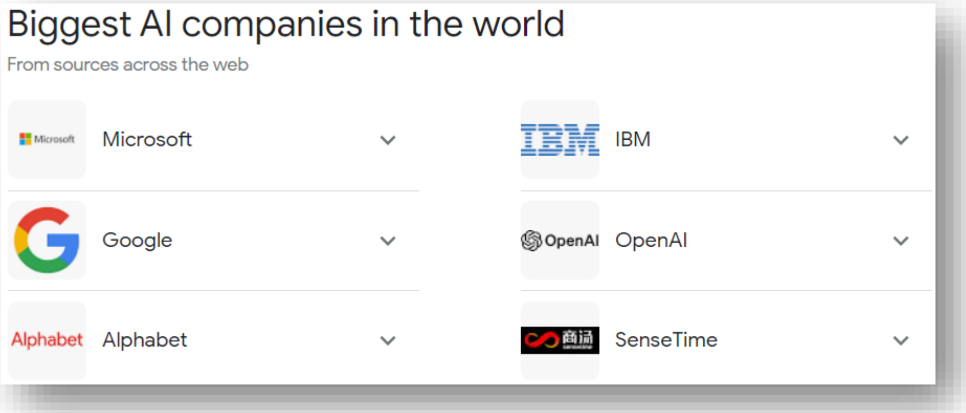
Immediate

Systemic

Long term

Systemic Risks

- **Concentration of power**: a few US and Asian giants dominating the field



Biggest AI companies in the world

From sources across the web

Microsoft

Google

Alphabet

IBM

OpenAI

SenseTime

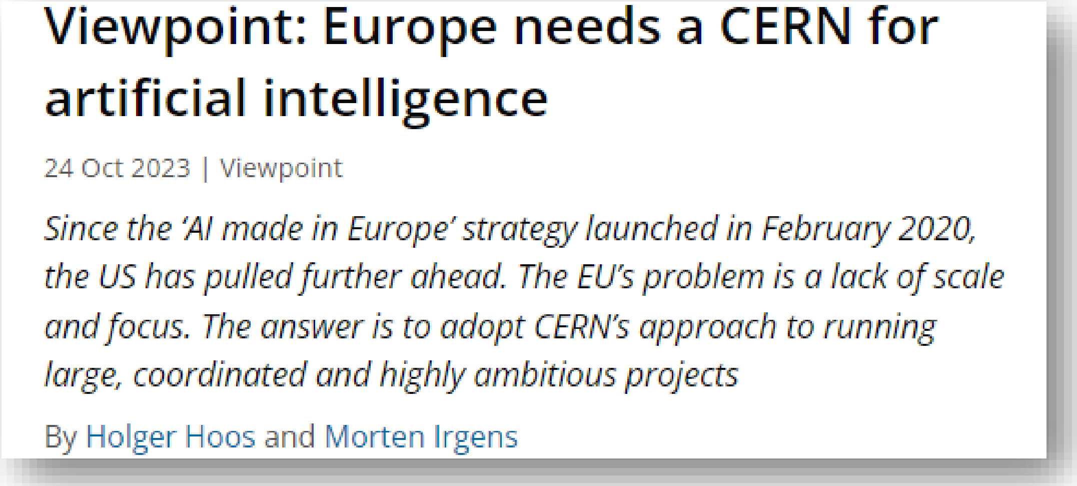
OCT. 9, 2023
THE 100 MOST INFLUENTIAL PEOPLE IN ARTIFICIAL INTELLIGENCE
TIME 100 AI
ELON MUSK'S FIGHT FOR THE FUTURE OF AI
By WALTER ISAACSON
time.com

About 10% academics

About 10% European

Systemic Risks

- **Concentration of power**: a few US and Asian giants dominating the field
- Why is that bad?
  - They understand the technology (cf. discussion on open source)
  - They control the technology (cf. Elon Musk and Starlink)
  - They drive the discussions and regulations on what is right and wrong (cf. closed door meeting in Washington)



Viewpoint: Europe needs a CERN for artificial intelligence

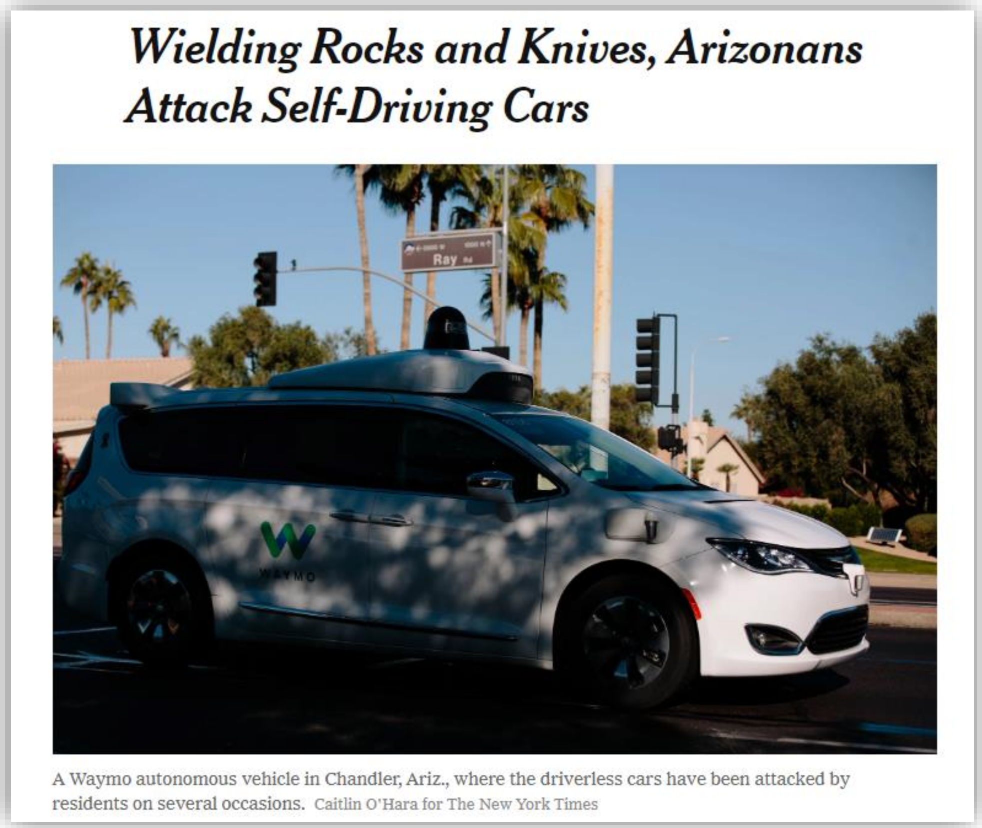
24 Oct 2023 | Viewpoint

Since the 'AI made in Europe' strategy launched in February 2020, the US has pulled further ahead. The EU's problem is a lack of scale and focus. The answer is to adopt CERN's approach to running large, coordinated and highly ambitious projects

By Holger Hoos and Morten Irgens

Systemic Risks

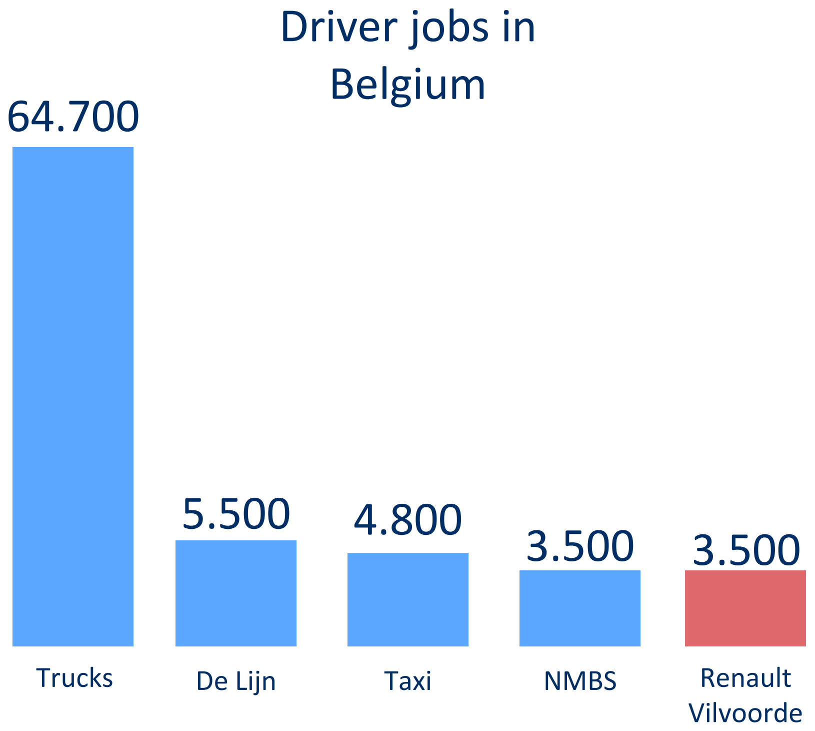
- **Jobs:** displacement and new opportunities



Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars

A Waymo autonomous vehicle in Chandler, Ariz., where the driverless cars have been attacked by residents on several occasions. Caitlin O'Hara for The New York Times

Systemic Risks



Sources:
https://nl.wikipedia.org/wiki/De_Lijn_(vervoermaatschappij)
https://www.taxi-info.be/bestanden/file/Jaarverslag%20Sociaal%20Fonds%20Taxi%20en%20VVB%202018.pdf
https://www.vrt.be/vrtnws/nl/2018/01/11/treinbestuurders-zijn-gegeerd-goed/
https://en.wikipedia.org/wiki/Vilvoorde_Renault_Factory

Impact of AI

- Will it impact jobs?
  - Consensus seems: 30-40% of jobs at risk due to automation over next decade(s) in advanced economies
  - OECD (2023) warns “AI likely to significantly impact jobs”
- Whose jobs?
  - Net gain job increase not evenly distributed
  - Which ones? Also non-routine jobs

AI

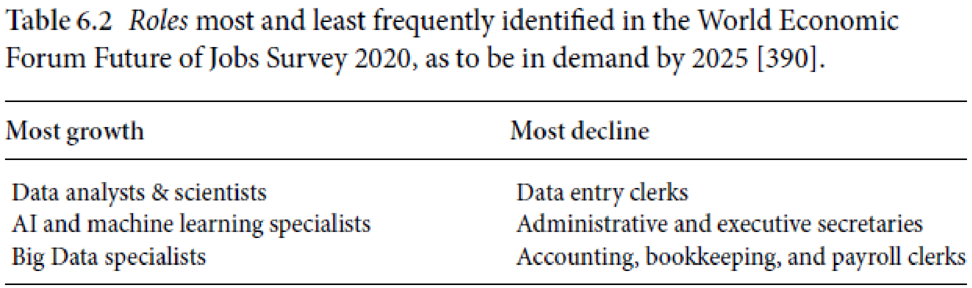
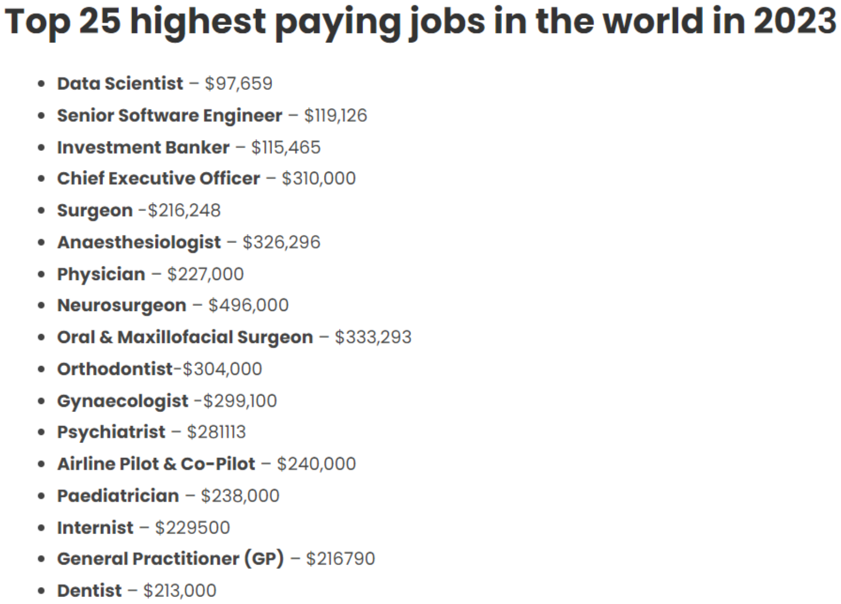


Table 6.2 *Roles* most and least frequently identified in the World Economic Forum Future of Jobs Survey 2020, as to be in demand by 2025 [390].

Most growth	Most decline
Data analysts & scientists	Data entry clerks
AI and machine learning specialists	Administrative and executive secretaries
Big Data specialists	Accounting, bookkeeping, and payroll clerks

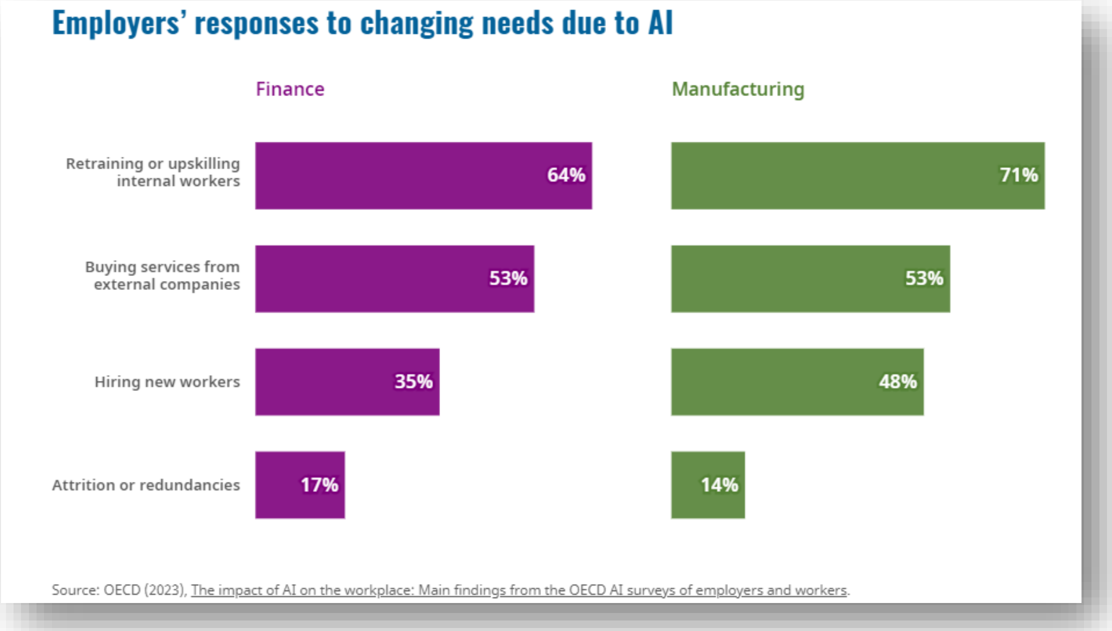


Top 25 highest paying jobs in the world in 2023

- **Data Scientist** – $97,659
- **Senior Software Engineer** – $119,126
- **Investment Banker** – $115,465
- **Chief Executive Officer** – $310,000
- **Surgeon** -$216,248
- **Anaesthesiologist** – $326,296
- **Physician** – $227,000
- **Neurosurgeon** – $496,000
- **Oral & Maxillofacial Surgeon** – $333,293
- **Orthodontist**-$304,000
- **Gynaecologist** -$299,100
- **Psychiatrist** – $281113
- **Airline Pilot & Co-Pilot** – $240,000
- **Paediatrician** – $238,000
- **Internist** – $229500
- **General Practitioner (GP)** – $216790
- **Dentist** – $213,000

Solutions

1. **Reskilling**
    - Hard to automate skills, and learn to work with machines
    - *“And as technology keeps changing, we need to focus more on continuous education throughout our lives. And yes, giving everyone the freedom to pursue purpose isn’t going to be free.* ***People like me should pay for it****, and a lot of you are going to do really well, and you should, too.” Zuckerberg (2017)*



Employers’ responses to changing needs due to AI

Source: OECD (2023), The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers.

Solutions

1. **Reskilling**
    - Hard to automate skills, and learn to work with machines
2. **Universal Basic Income**
    - Advocated by Mark Zuckerberg, Elon Musk, Jack Dorsey, Larry Page
3. **Learn to live with the freedom of not working**
    - Being a stay-at-home mom or dad, artist, author argubaly not held in as high esteem as being a succesful entrepeneur or CEO.
    - *“reexamine what we value, what we are collectively willing to pay for – whether it’s teachers, nurses, caregivers, moms or dads who stay at home, artists, all the things that are incredibly valuable to us right now but don’t rank high on the pay totem pole – that’s a* ***conversation we need to begin to have.****”* Barack Obama (2016)
    - *“[W]e have been* ***trained too long to strive and not to enjoy****”*
      Keynes (1930)

AI Risks

Immediate

Systemic

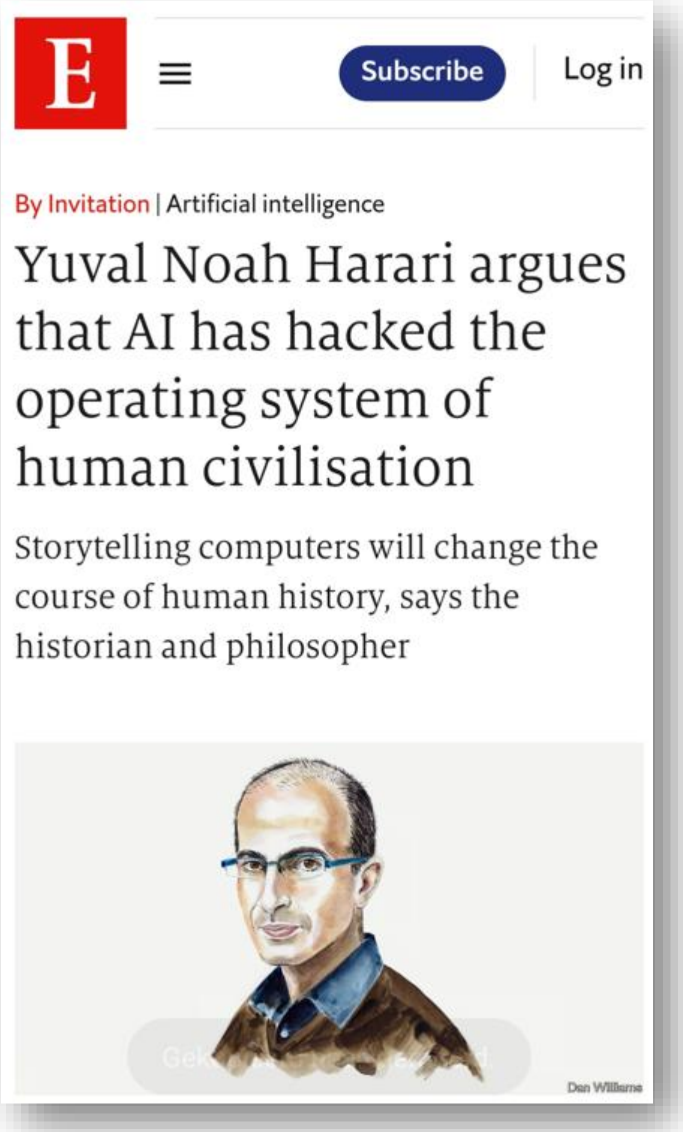
Long term

Long term risk

- The potential for an AI-dominated future:
  A world beyond human control?

- Most discussed, least relevant right now.
- Focus on this area is argued to also serve a certain big tech agenda.

Long term risk



By Invitation | Artificial intelligence

Yuval Noah Harari argues that AI has hacked the operating system of human civilisation

Storytelling computers will change the course of human history, says the historian and philosopher

Pause Giant AI Experiments: An Open Letter

“Should we risk loss of control of our civilization?”

Conclusion

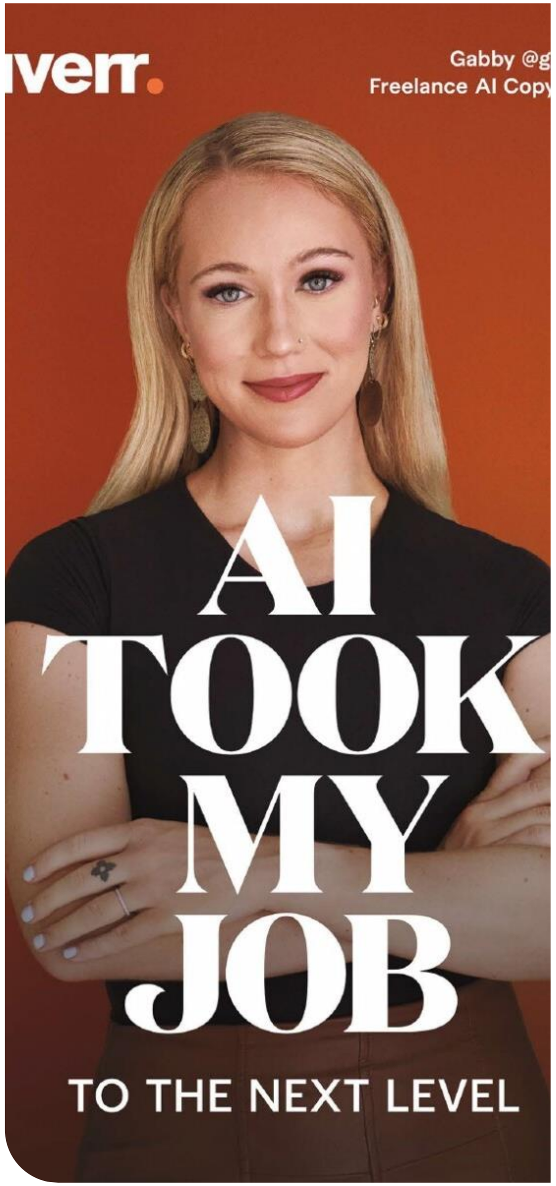
- **Embrace AI**
  - Age of AI
  - Fast moving
  - *“AI is not gonna take your job, someone who understands AI is going to take your job. Get good at it.”* Scott Galloway

- **Be aware of the risks**
  - Not just a chatGPT issue
  - Need for education
  - Not for me?

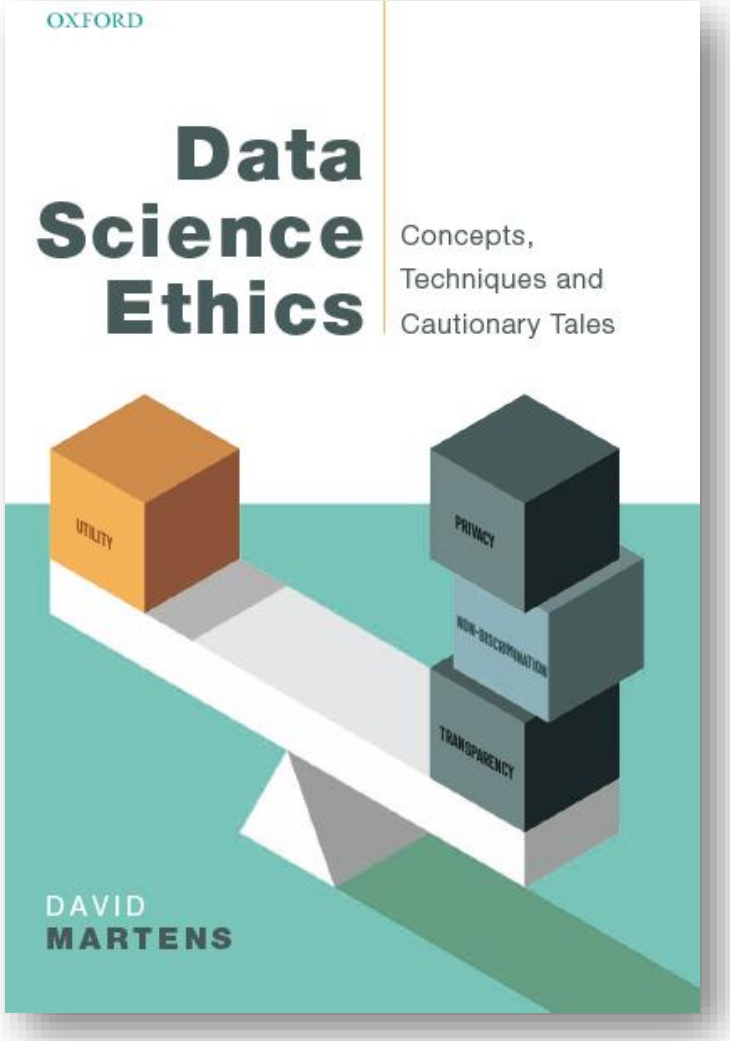
Conclusion

- **Embrace AI**
  - Age of AI
  - Fast moving
  - *"AI is not gonna take your job, someone who understands AI is going to take your job. Get good at it."* Scott Galloway

- **Be aware of the risks**
  - Not just a chatGPT issue
  - Need for education
  - Not for me?



Questions



David Martens
Data Science Ethics: Concepts, Techniques and Cautionary Tales
Oxford University Press (March 24, 2022)
272 pages
www.dsethics.com