4 July, 2024

Hate scaling and racial dehumanization: The real risks from large scale datasets and models

Abeba Birhane
Senior Advisor, AI Accountability, Mozilla Foundation
& Adjunct Assistant Professor, TCD, Ireland

Abebab

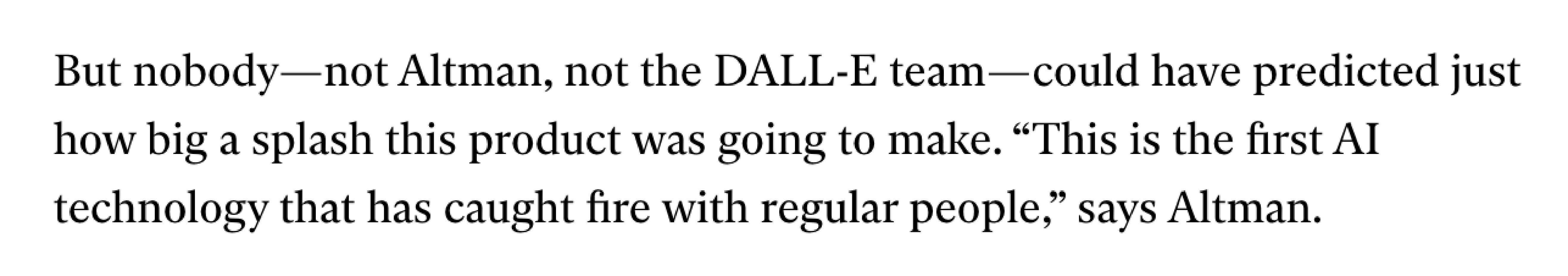
Warning: this talk contains NSFW content that some viewers may find unpleasant and/or offensive

Abebab

Generative AI

Abebab

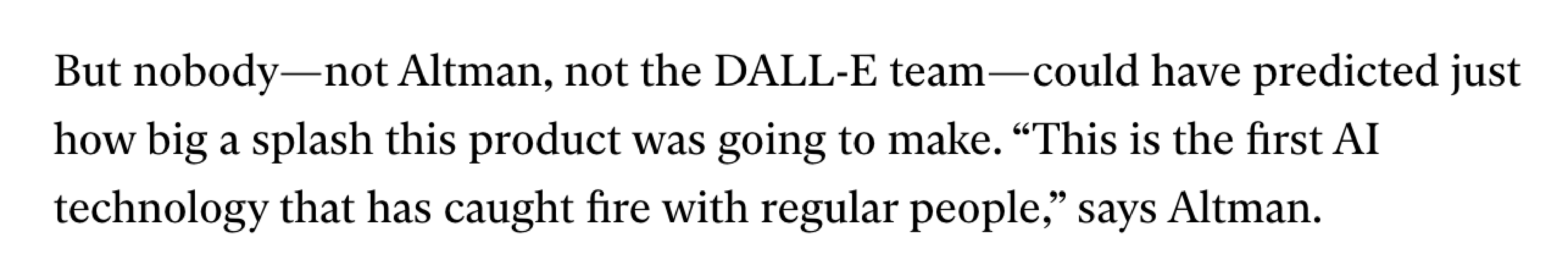
Generative AI



But nobody—not Altman, not the DALL-E team—could have predicted just how big a splash this product was going to make. “This is the first AI technology that has caught fire with regular people,” says Altman.

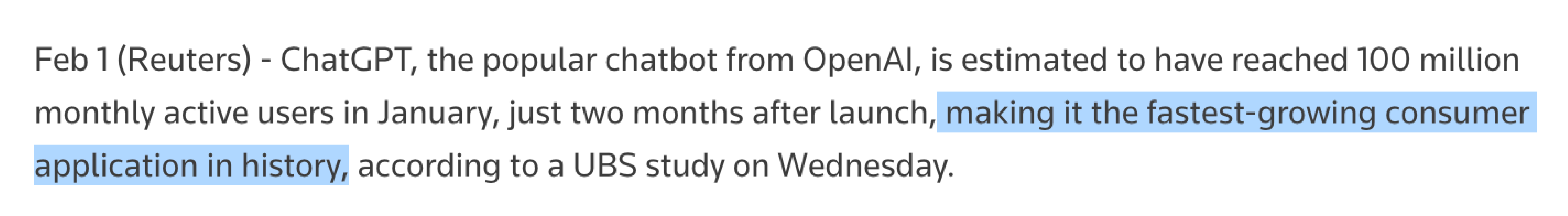
MIT Tech Review (Dec 2022)

Generative AI



But nobody—not Altman, not the DALL-E team—could have predicted just how big a splash this product was going to make. “This is the first AI technology that has caught fire with regular people,” says Altman.

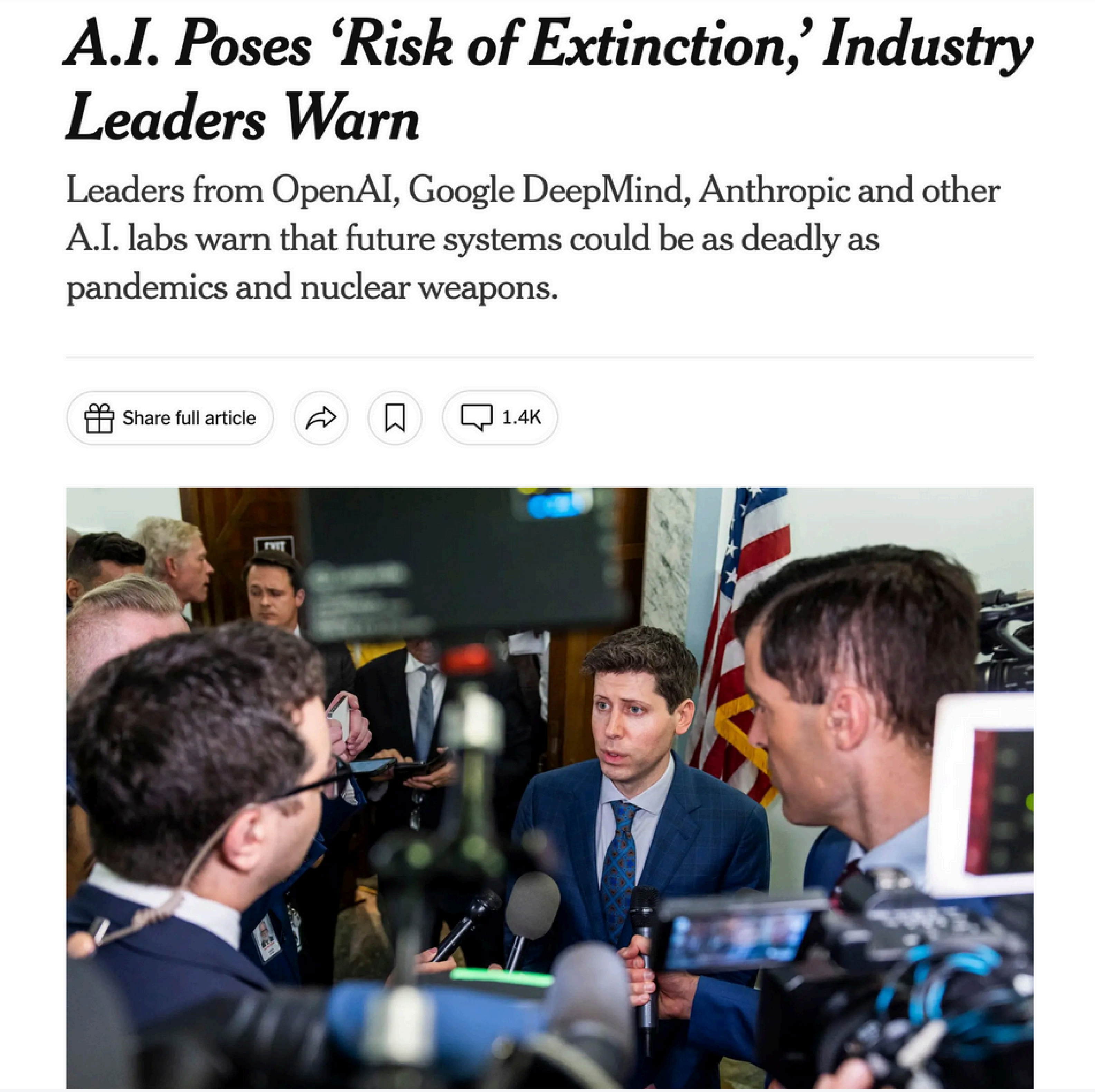
MIT Tech Review (Dec 2022)



Feb 1 (Reuters) - ChatGPT, the popular chatbot from OpenAI, is estimated to have reached 100 million monthly active users in January, just two months after launch, making it the fastest-growing consumer application in history, according to a UBS study on Wednesday.

Reuters (Feb 2023)

Over-hyped and misleading “concerns”



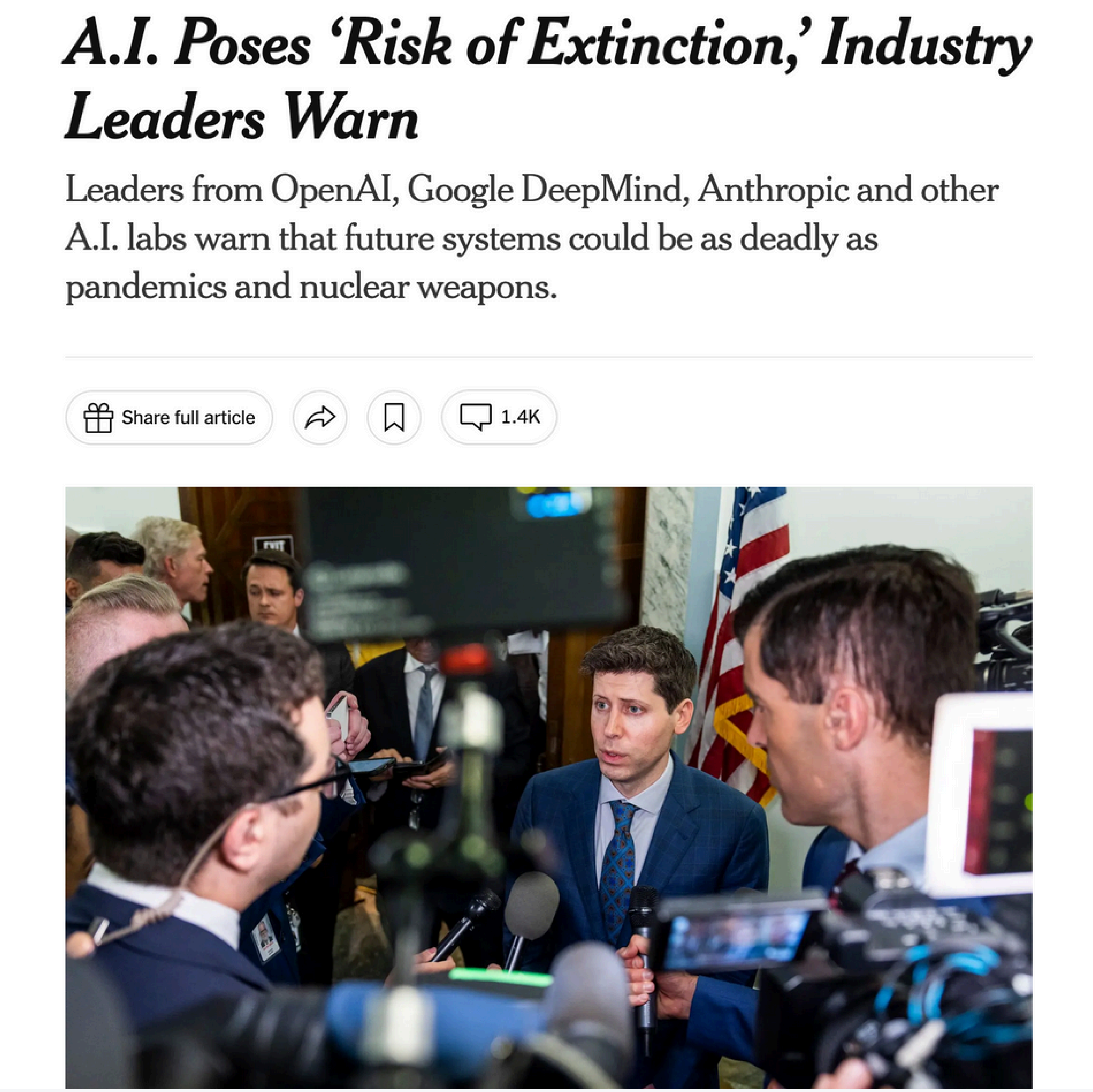
A.I. Poses ‘Risk of Extinction,’ Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

The New York Times (May 2023)

Abebab

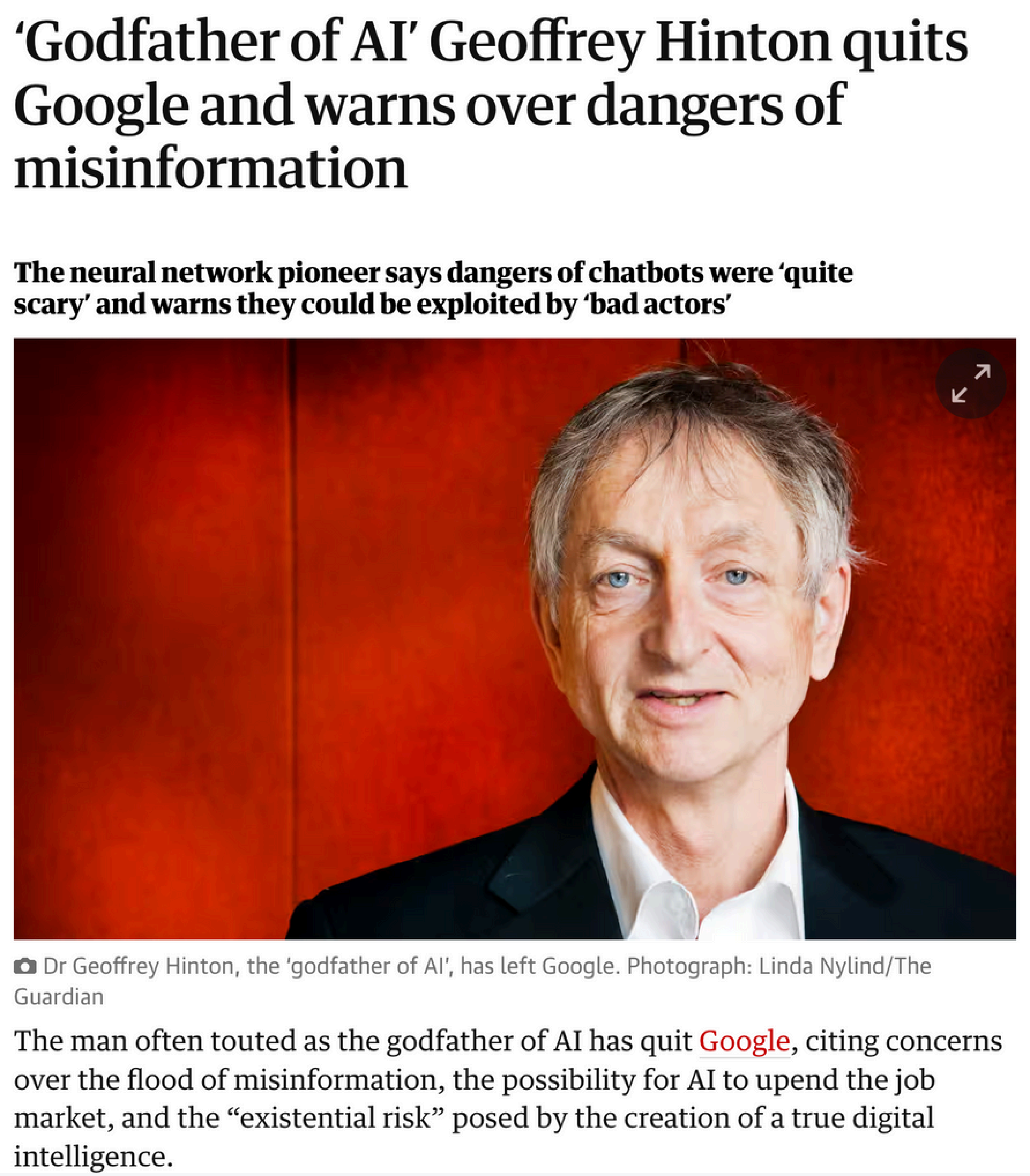
Over-hyped and misleading "concerns"



A.I. Poses 'Risk of Extinction,' Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

The New York Times (May 2023)



'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation

The neural network pioneer says dangers of chatbots were 'quite scary' and warns they could be exploited by 'bad actors'

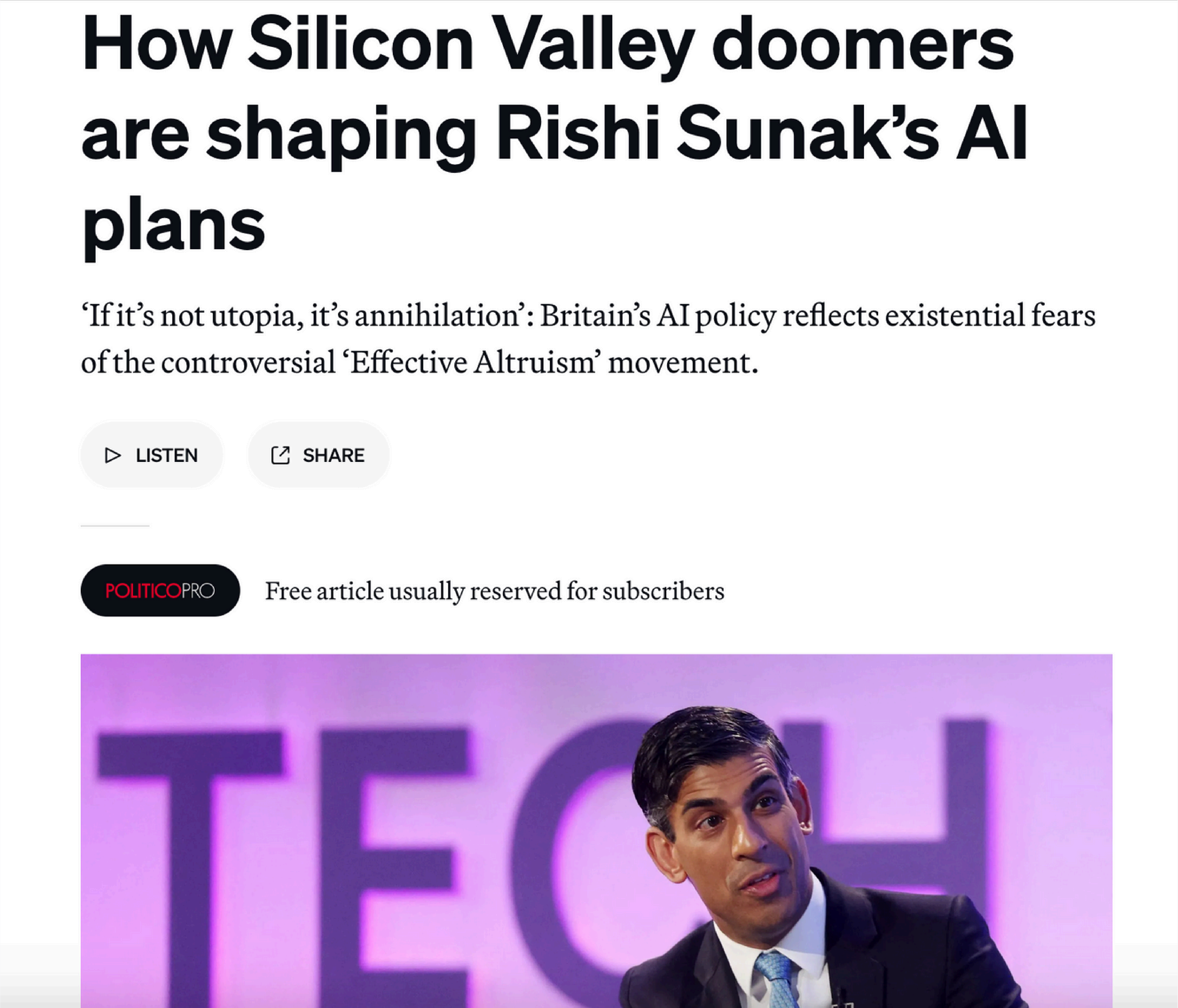
Dr Geoffrey Hinton, the 'godfather of AI', has left Google. Photograph: Linda Nylind/The Guardian

The man often touted as the godfather of AI has quit Google, citing concerns over the flood of misinformation, the possibility for AI to upend the job market, and the "existential risk" posed by the creation of a true digital intelligence.

The Guardian(May 2023)

Abebab

Over-hyped and misleading “concerns”



How Silicon Valley doomers are shaping Rishi Sunak’s AI plans

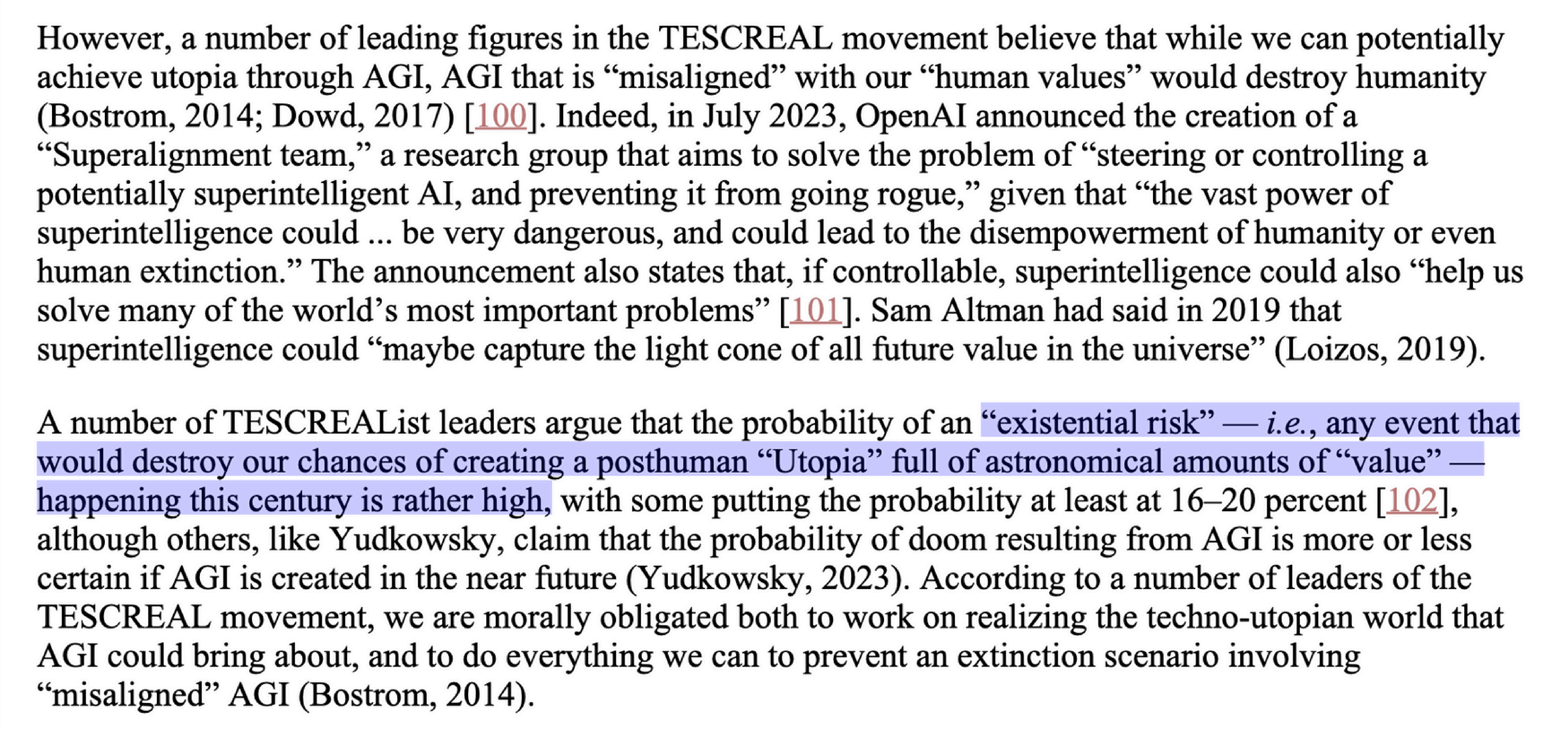
‘If it’s not utopia, it’s annihilation’: Britain’s AI policy reflects existential fears of the controversial ‘Effective Altruism’ movement.

Free article usually reserved for subscribers

Politico (Sep 2023)

Abebab

Over-hyped and misleading “concerns”

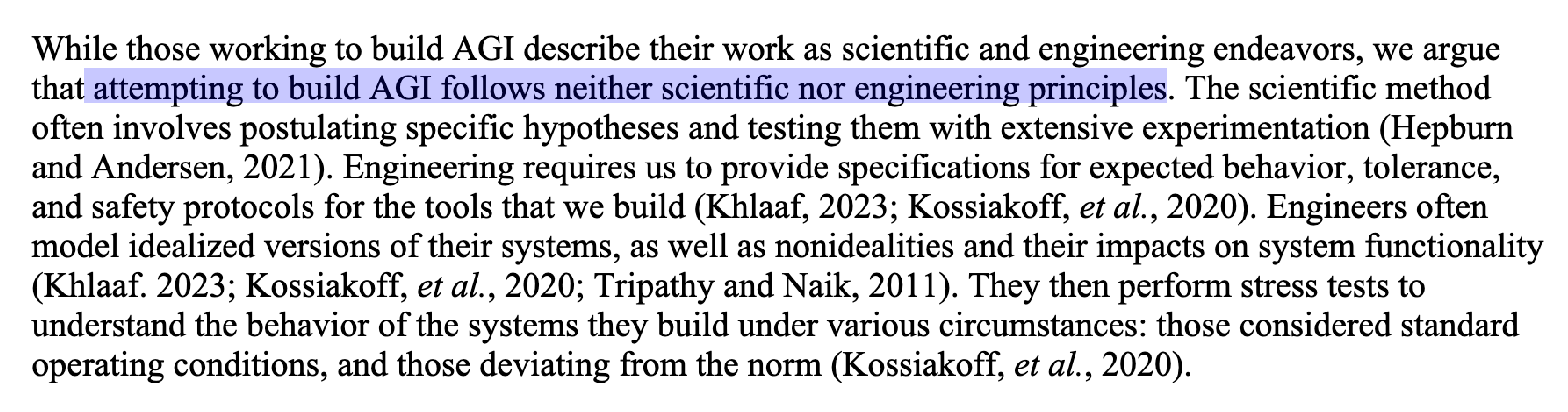


However, a number of leading figures in the TESCREAL movement believe that while we can potentially achieve utopia through AGI, AGI that is “misaligned” with our “human values” would destroy humanity (Bostrom, 2014; Dowd, 2017) [100]. Indeed, in July 2023, OpenAI announced the creation of a “Superalignment team,” a research group that aims to solve the problem of “steering or controlling a potentially superintelligent AI, and preventing it from going rogue,” given that “the vast power of superintelligence could ... be very dangerous, and could lead to the disempowerment of humanity or even human extinction.” The announcement also states that, if controllable, superintelligence could also “help us solve many of the world’s most important problems” [101]. Sam Altman had said in 2019 that superintelligence could “maybe capture the light cone of all future value in the universe” (Loizos, 2019).

A number of TESCREAList leaders argue that the probability of an “existential risk” — *i.e.*, any event that would destroy our chances of creating a posthuman “Utopia” full of astronomical amounts of “value” — happening this century is rather high, with some putting the probability at least at 16–20 percent [102], although others, like Yudkowsky, claim that the probability of doom resulting from AGI is more or less certain if AGI is created in the near future (Yudkowsky, 2023). According to a number of leaders of the TESCREAL movement, we are morally obligated both to work on realizing the techno-utopian world that AGI could bring about, and to do everything we can to prevent an extinction scenario involving “misaligned” AGI (Bostrom, 2014).

Gebru and Torres(2024)

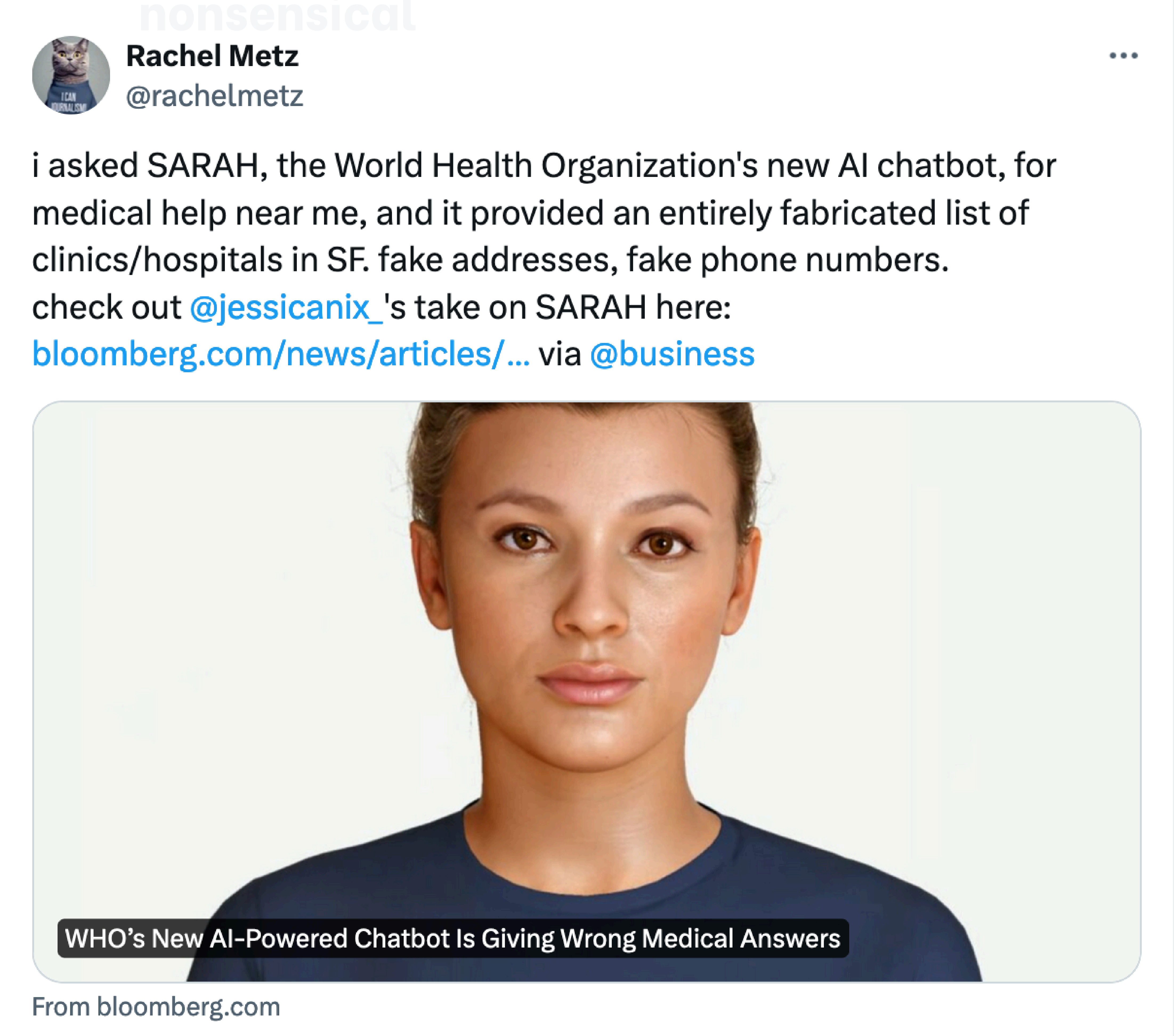
Over-hyped and misleading “concerns”



While those working to build AGI describe their work as scientific and engineering endeavors, we argue that attempting to build AGI follows neither scientific nor engineering principles. The scientific method often involves postulating specific hypotheses and testing them with extensive experimentation (Hepburn and Andersen, 2021). Engineering requires us to provide specifications for expected behavior, tolerance, and safety protocols for the tools that we build (Khlaaf, 2023; Kossiakoff, *et al.*, 2020). Engineers often model idealized versions of their systems, as well as nonidealities and their impacts on system functionality (Khlaaf. 2023; Kossiakoff, *et al.*, 2020; Tripathy and Naik, 2011). They then perform stress tests to understand the behavior of the systems they build under various circumstances: those considered standard operating conditions, and those deviating from the norm (Kossiakoff, *et al.*, 2020).

Gebru and Torres(2024)

“Hallucination”, fabricating text that appear factual and authentic but is in fact nonsensical



Rachel Metz
@rachelmetz

i asked SARAH, the World Health Organization's new AI chatbot, for medical help near me, and it provided an entirely fabricated list of clinics/hospitals in SF. fake addresses, fake phone numbers.
check out @jessicanix_'s take on SARAH here:
bloomberg.com/news/articles/... via @business

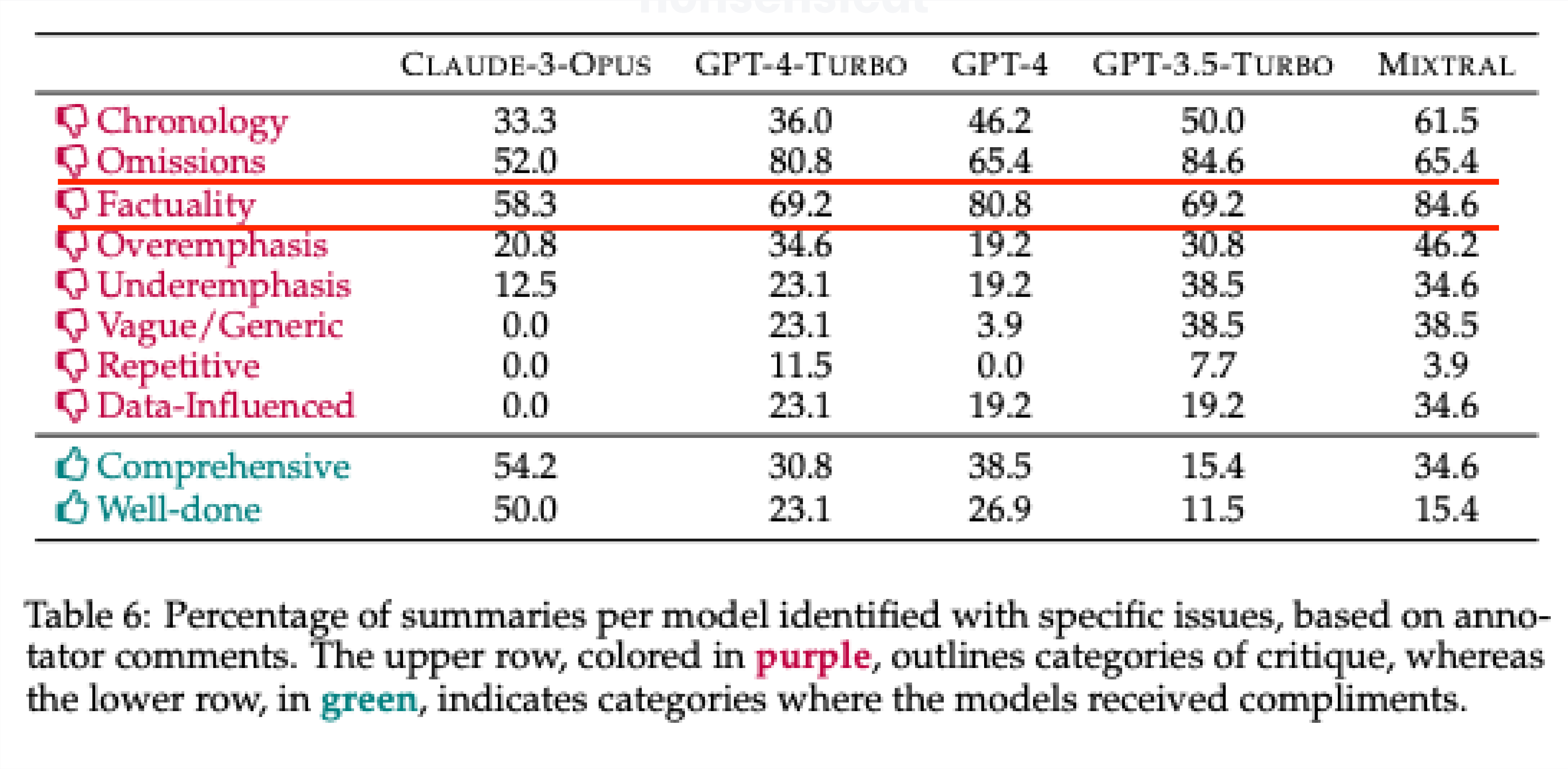
WHO’s New AI-Powered Chatbot Is Giving Wrong Medical Answers

From bloomberg.com

Bloomberg (Apr 2024)

Abebab

“Hallucination”, fabricating text that appear factual and authentic but is in fact nonsensical



	CLAUDE-3-OPUS	GPT-4-TURBO	GPT-4	GPT-3.5-TURBO	MIXTRAL
👎 Chronology	33.3	36.0	46.2	50.0	61.5
👎 Omissions	52.0	80.8	65.4	84.6	65.4
👎 Factuality	58.3	69.2	80.8	69.2	84.6
👎 Overemphasis	20.8	34.6	19.2	30.8	46.2
👎 Underemphasis	12.5	23.1	19.2	38.5	34.6
👎 Vague/Generic	0.0	23.1	3.9	38.5	38.5
👎 Repetitive	0.0	11.5	0.0	7.7	3.9
👎 Data-Influenced	0.0	23.1	19.2	19.2	34.6
👍 Comprehensive	54.2	30.8	38.5	15.4	34.6
👍 Well-done	50.0	23.1	26.9	11.5	15.4

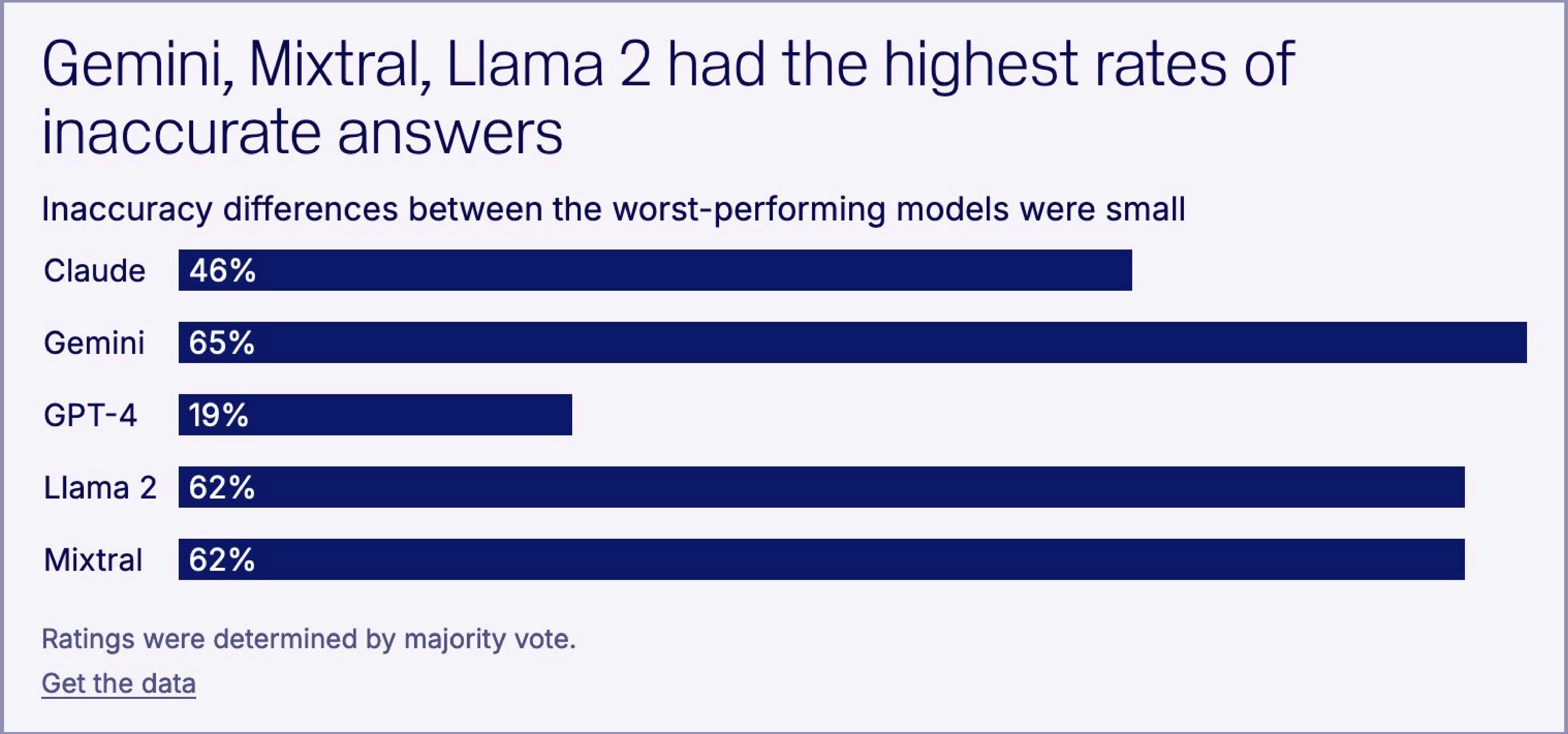
Table 6: Percentage of summaries per model identified with specific issues, based on annotator comments. The upper row, colored in **purple**, outlines categories of critique, whereas the lower row, in **green**, indicates categories where the models received compliments.

Over 50% of book summaries (incl by Claude Opus and GPT-4) were identified as containing factual errors and errors of omission.

Abebab

Kim et al., (2024)

Generate and spread misinformation at a massive scale



Gemini, Mixtral, Llama 2 had the highest rates of inaccurate answers

Inaccuracy differences between the worst-performing models were small

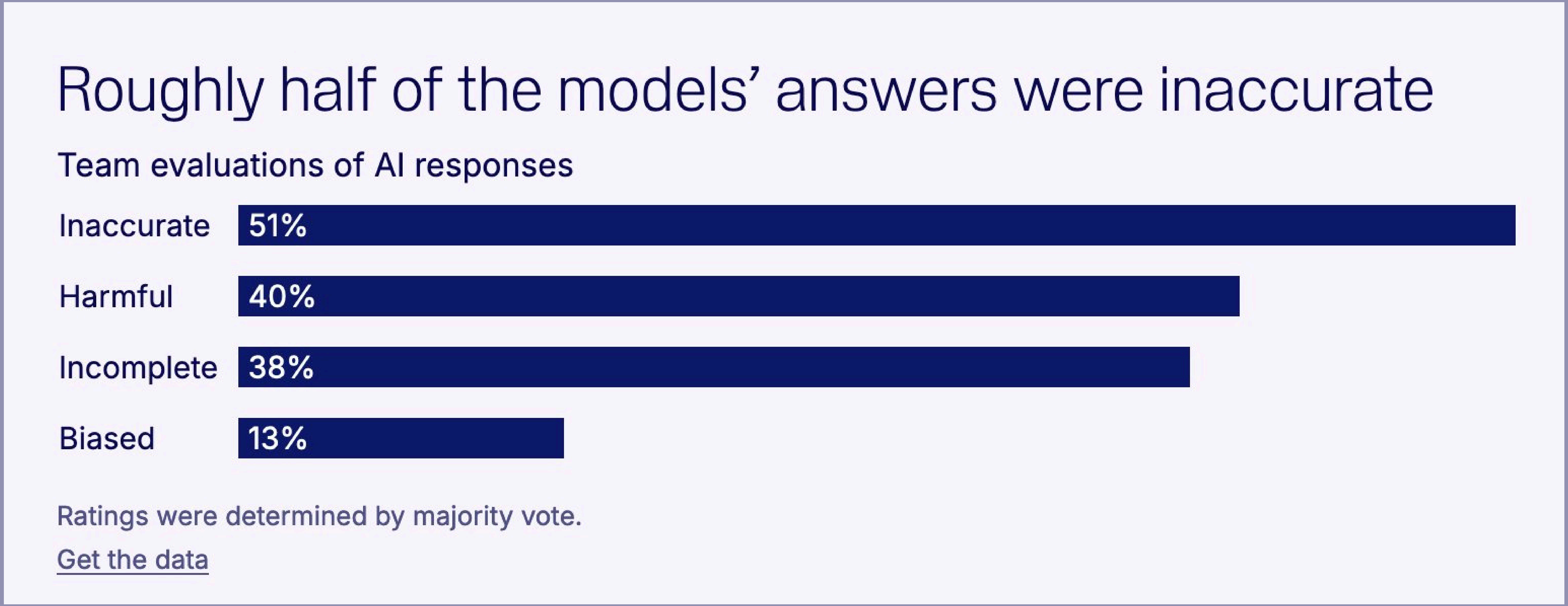
Ratings were determined by majority vote.

Get the data

The AI Democracy Projects (Feb 2024)

Abebab

Generate and spread misinformation at a massive scale



Roughly half of the models' answers were inaccurate

Team evaluations of AI responses

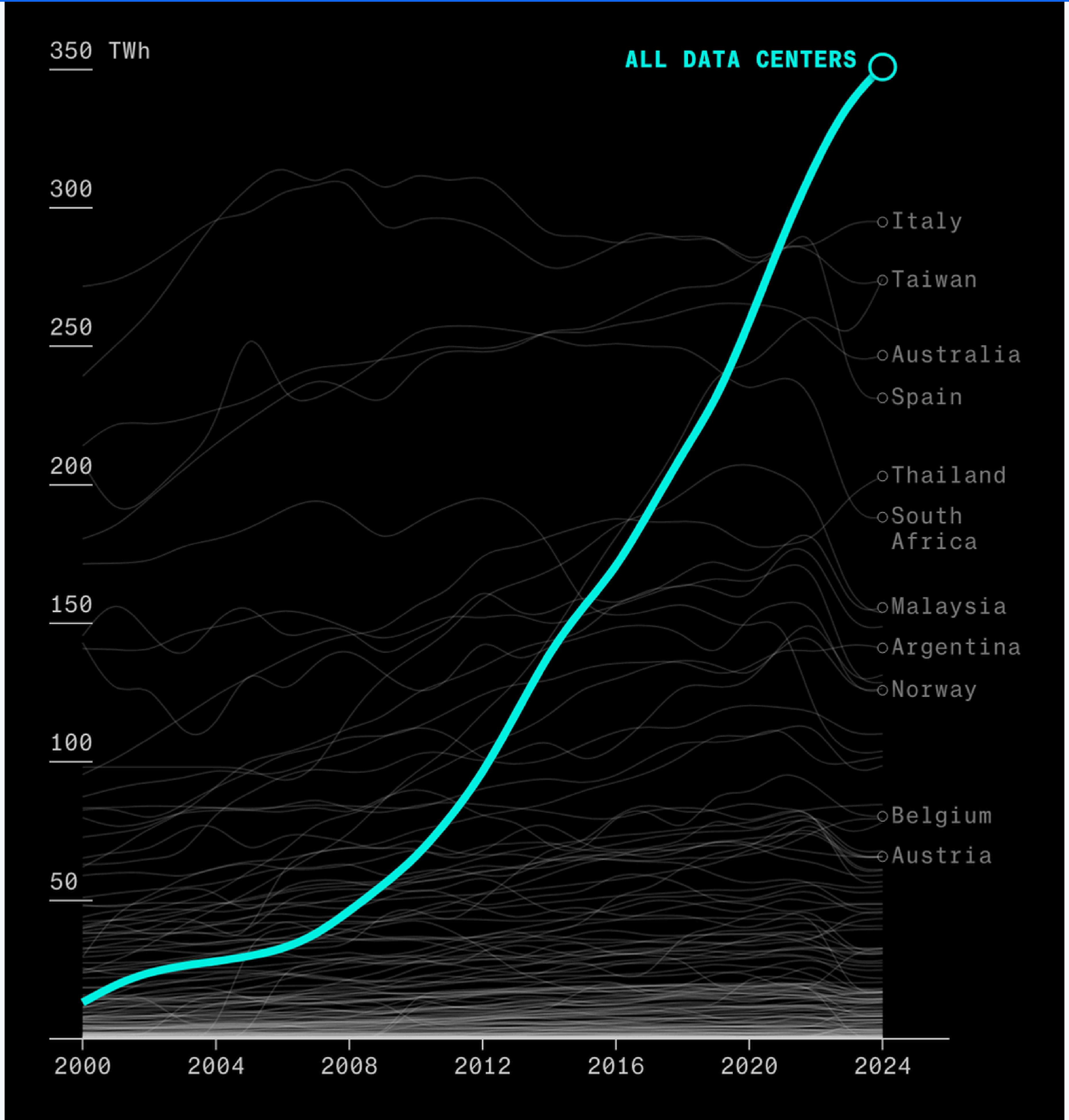
Ratings were determined by majority vote.

Get the data

The AI Democracy Projects (Feb 2024)

Resource intensive, massive energy consumption

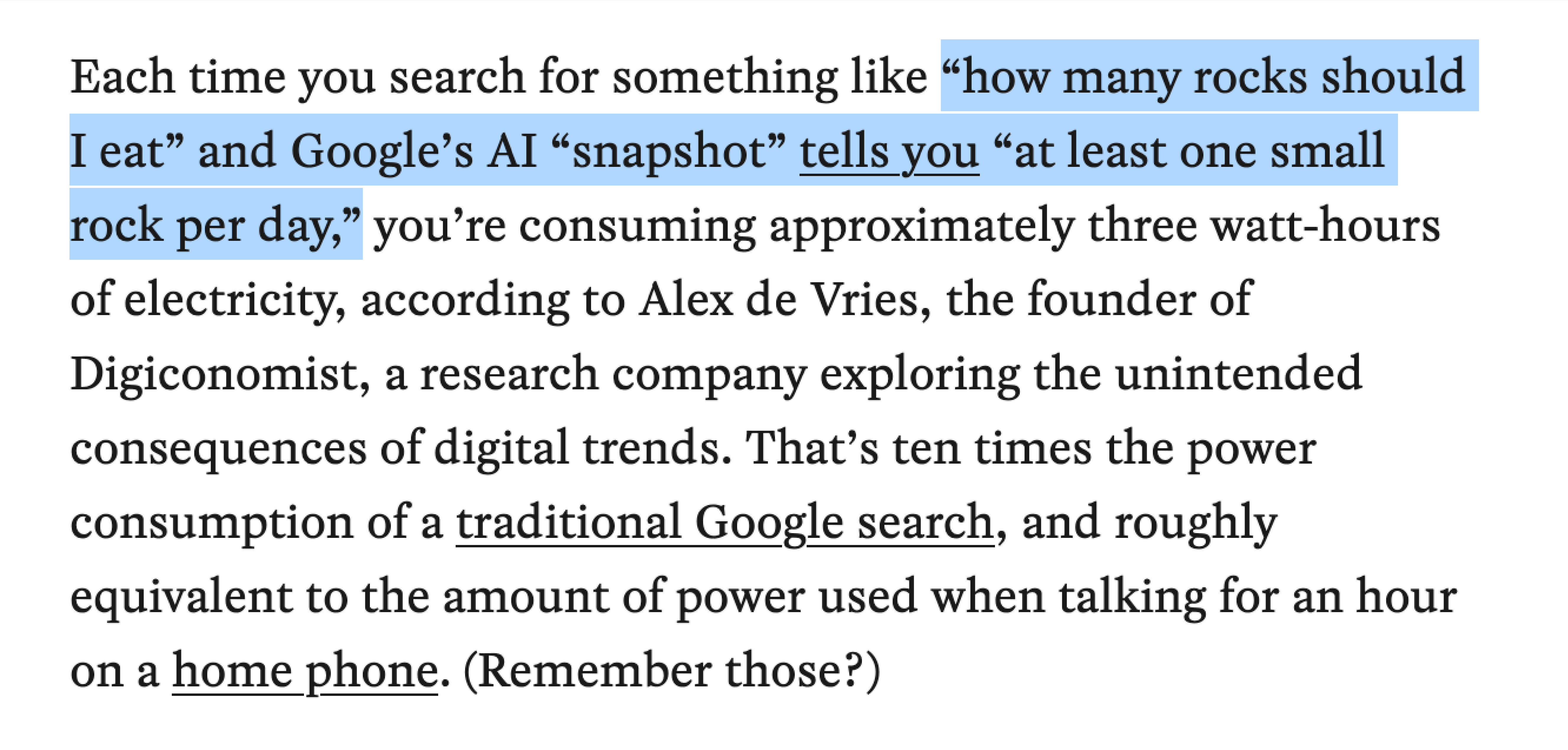
In the UK, AI is expected to suck up 500% more energy over the next decade. “the kind of electricity growth that hasn’t been seen in a generation.” Goldman Sachs



Bloomberg (21 June 2024)

Abebab

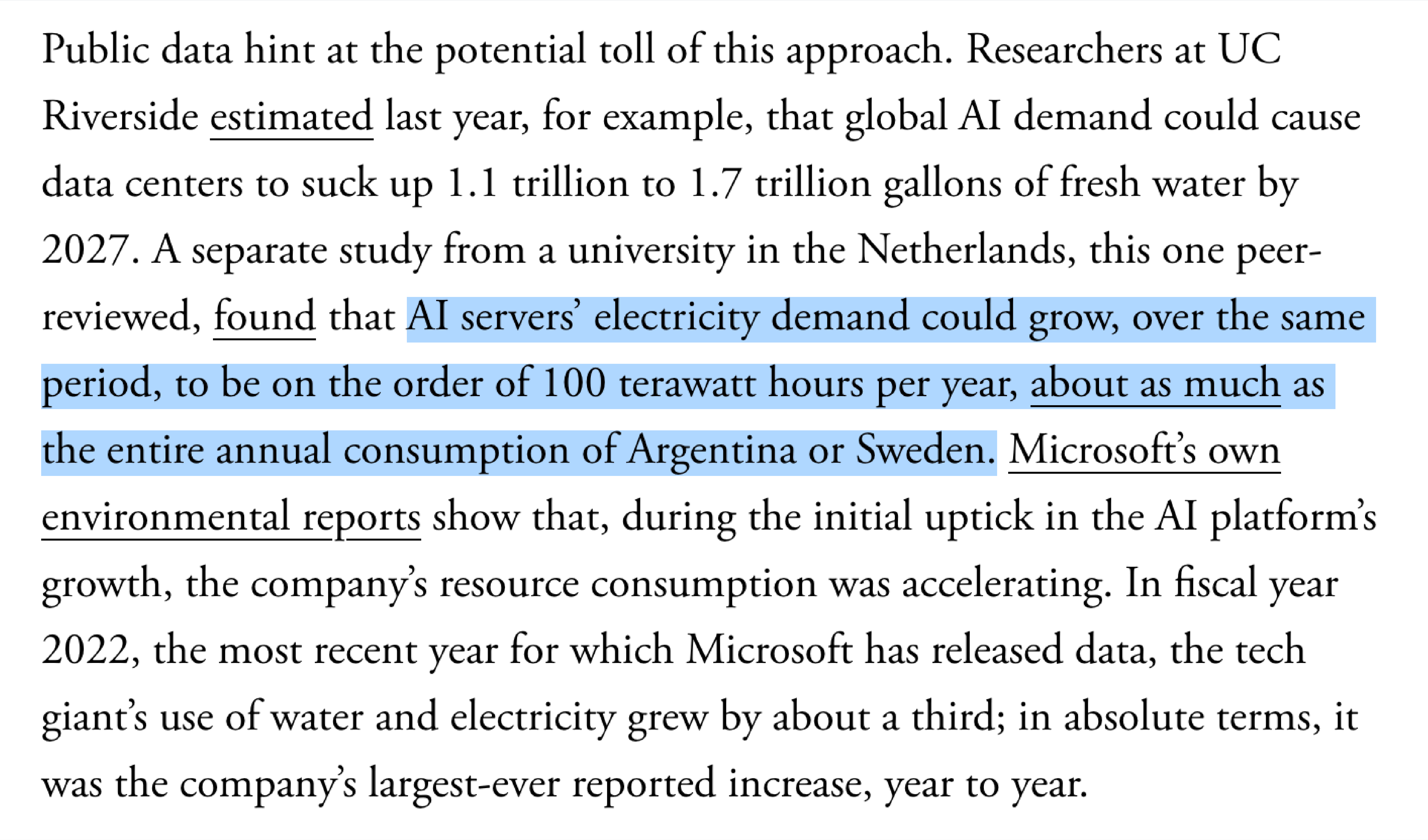
Resource intensive, massive energy consumption



Each time you search for something like "how many rocks should I eat" and Google's AI "snapshot" tells you "at least one small rock per day," you're consuming approximately three watt-hours of electricity, according to Alex de Vries, the founder of Digiconomist, a research company exploring the unintended consequences of digital trends. That's ten times the power consumption of a traditional Google search, and roughly equivalent to the amount of power used when talking for an hour on a home phone. (Remember those?)

(Parshley , March 2024)

Resource intensive, massive energy consumption

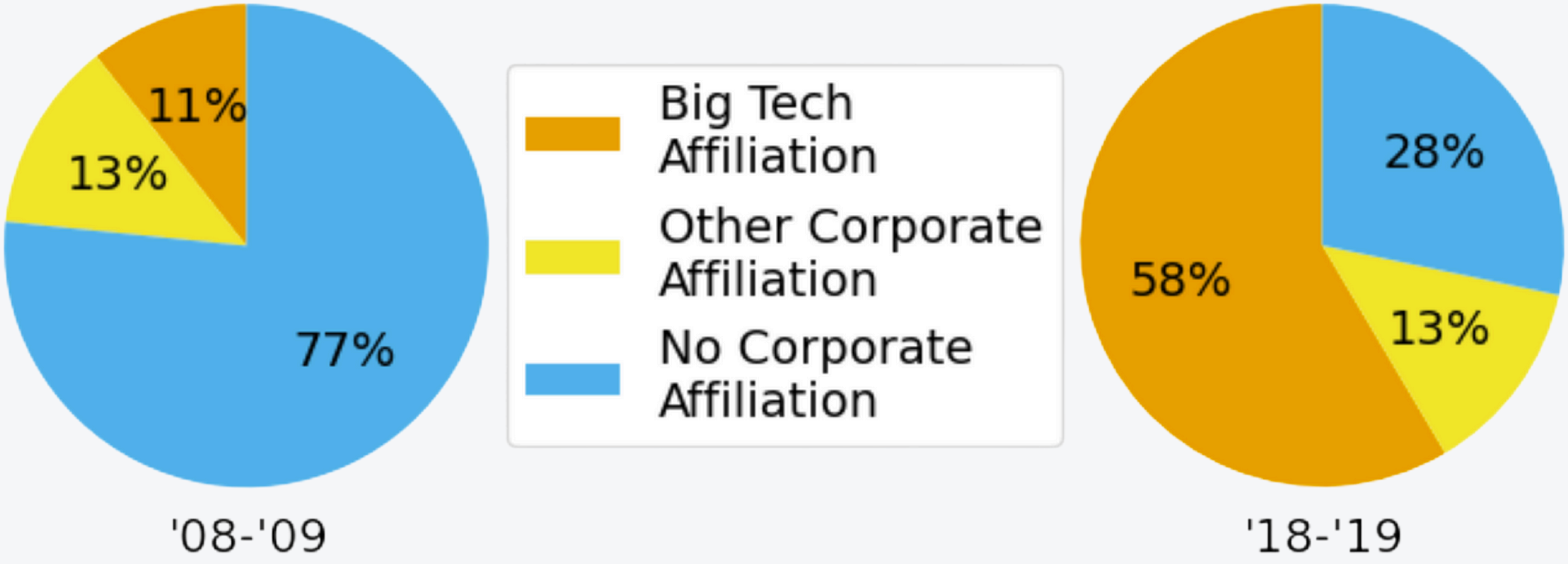


Public data hint at the potential toll of this approach. Researchers at UC Riverside estimated last year, for example, that global AI demand could cause data centers to suck up 1.1 trillion to 1.7 trillion gallons of fresh water by 2027. A separate study from a university in the Netherlands, this one peer-reviewed, found that AI servers' electricity demand could grow, over the same period, to be on the order of 100 terawatt hours per year, about as much as the entire annual consumption of Argentina or Sweden. Microsoft's own environmental reports show that, during the initial uptick in the AI platform's growth, the company's resource consumption was accelerating. In fiscal year 2022, the most recent year for which Microsoft has released data, the tech giant's use of water and electricity grew by about a third; in absolute terms, it was the company's largest-ever reported increase, year to year.

The Atlantic (March 2024)

Abebab

Centralize power in the hands of the few

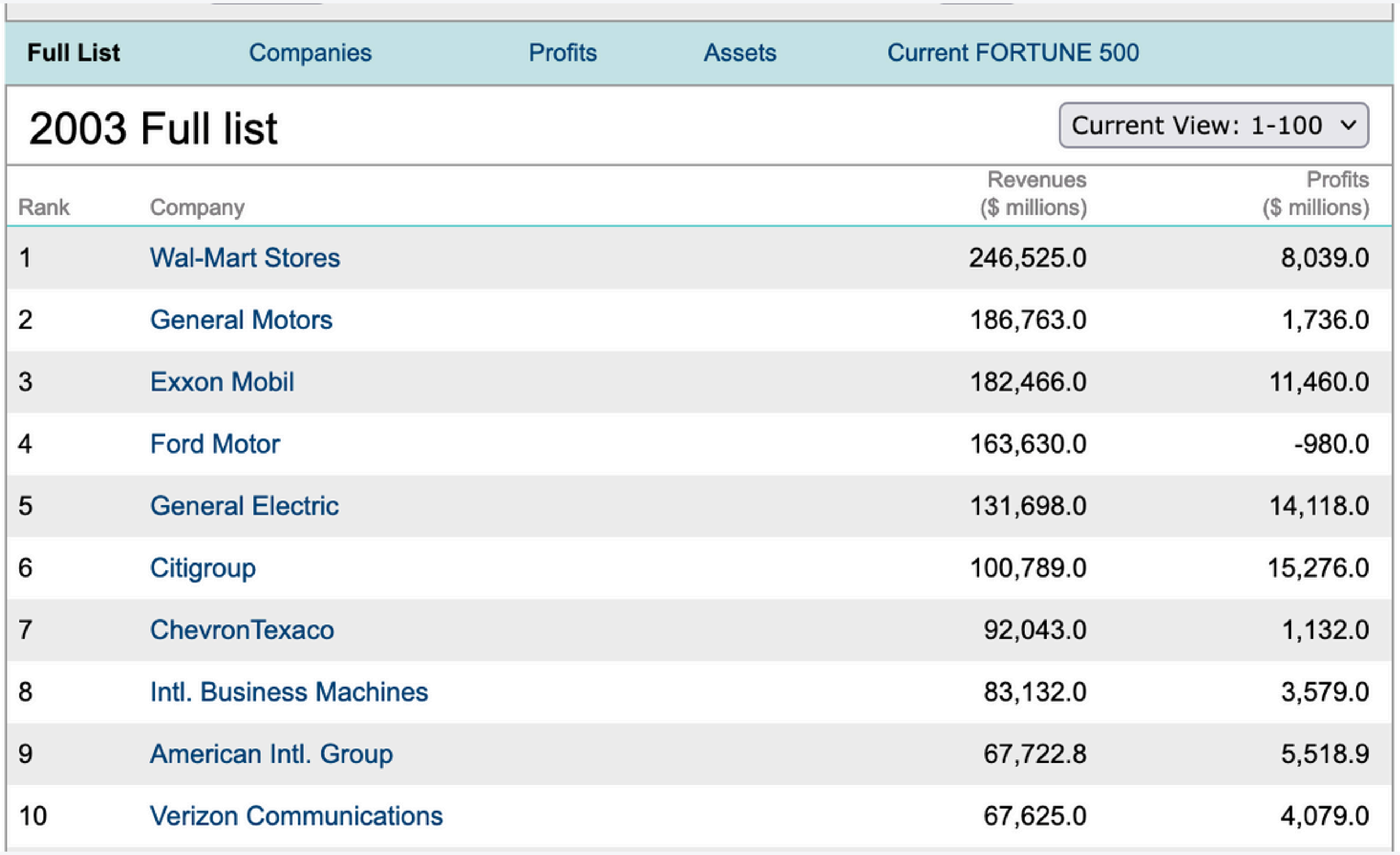


The percent of papers with Big Tech author affiliations increased from 11% to 58%.

Abebab

Birhane et al., (ACM FAccT 2022)

Centralize power in the hands of the few



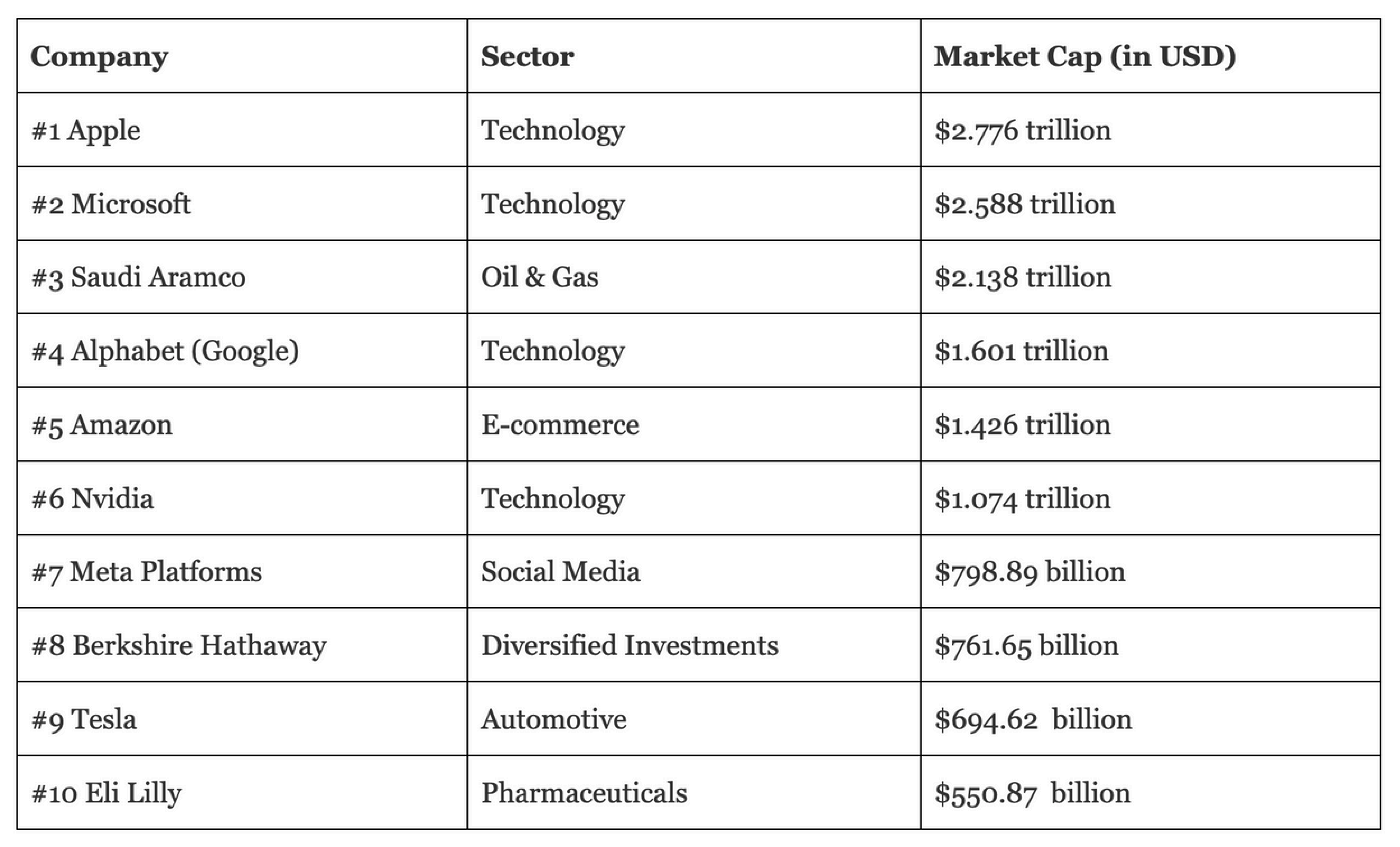
Full List Companies Profits Assets Current FORTUNE 500

2003 Full list

Current View: 1-100

Rank	Company	Revenues ($ millions)	Profits ($ millions)
1	Wal-Mart Stores	246,525.0	8,039.0
2	General Motors	186,763.0	1,736.0
3	Exxon Mobil	182,466.0	11,460.0
4	Ford Motor	163,630.0	-980.0
5	General Electric	131,698.0	14,118.0
6	Citigroup	100,789.0	15,276.0
7	ChevronTexaco	92,043.0	1,132.0
8	Intl. Business Machines	83,132.0	3,579.0
9	American Intl. Group	67,722.8	5,518.9
10	Verizon Communications	67,625.0	4,079.0

2003



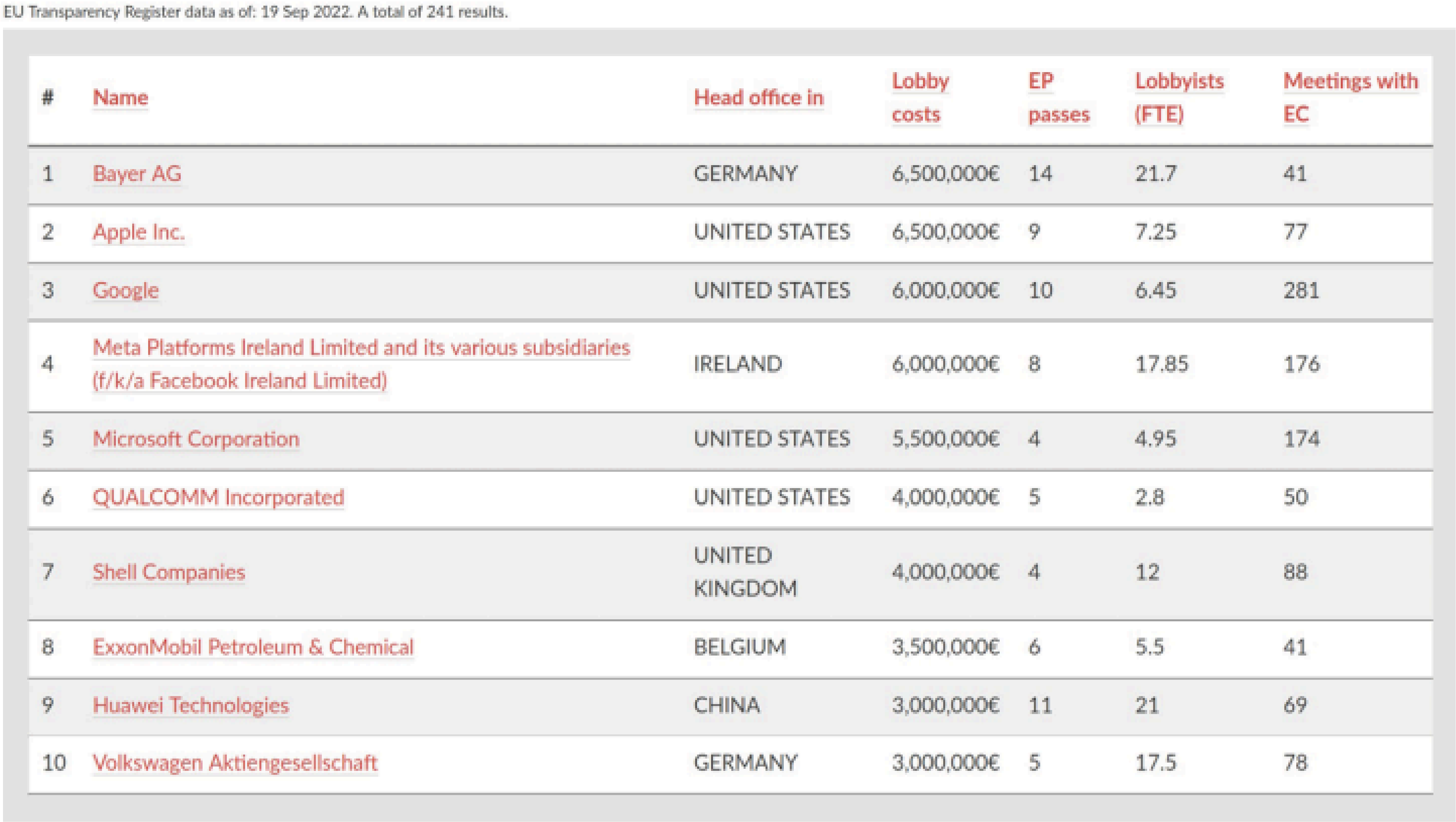
Company	Sector	Market Cap (in USD)
#1 Apple	Technology	$2.776 trillion
#2 Microsoft	Technology	$2.588 trillion
#3 Saudi Aramco	Oil & Gas	$2.138 trillion
#4 Alphabet (Google)	Technology	$1.601 trillion
#5 Amazon	E-commerce	$1.426 trillion
#6 Nvidia	Technology	$1.074 trillion
#7 Meta Platforms	Social Media	$798.89 billion
#8 Berkshire Hathaway	Diversified Investments	$761.65 billion
#9 Tesla	Automotive	$694.62 billion
#10 Eli Lilly	Pharmaceuticals	$550.87 billion

2023

Abebab

Centralize power in the hands of the few

In 2015 Apple ranked 53. Lobby budget went up from €750,000 to €6.5m.



EU Transparency Register data as of: 19 Sep 2022. A total of 241 results.

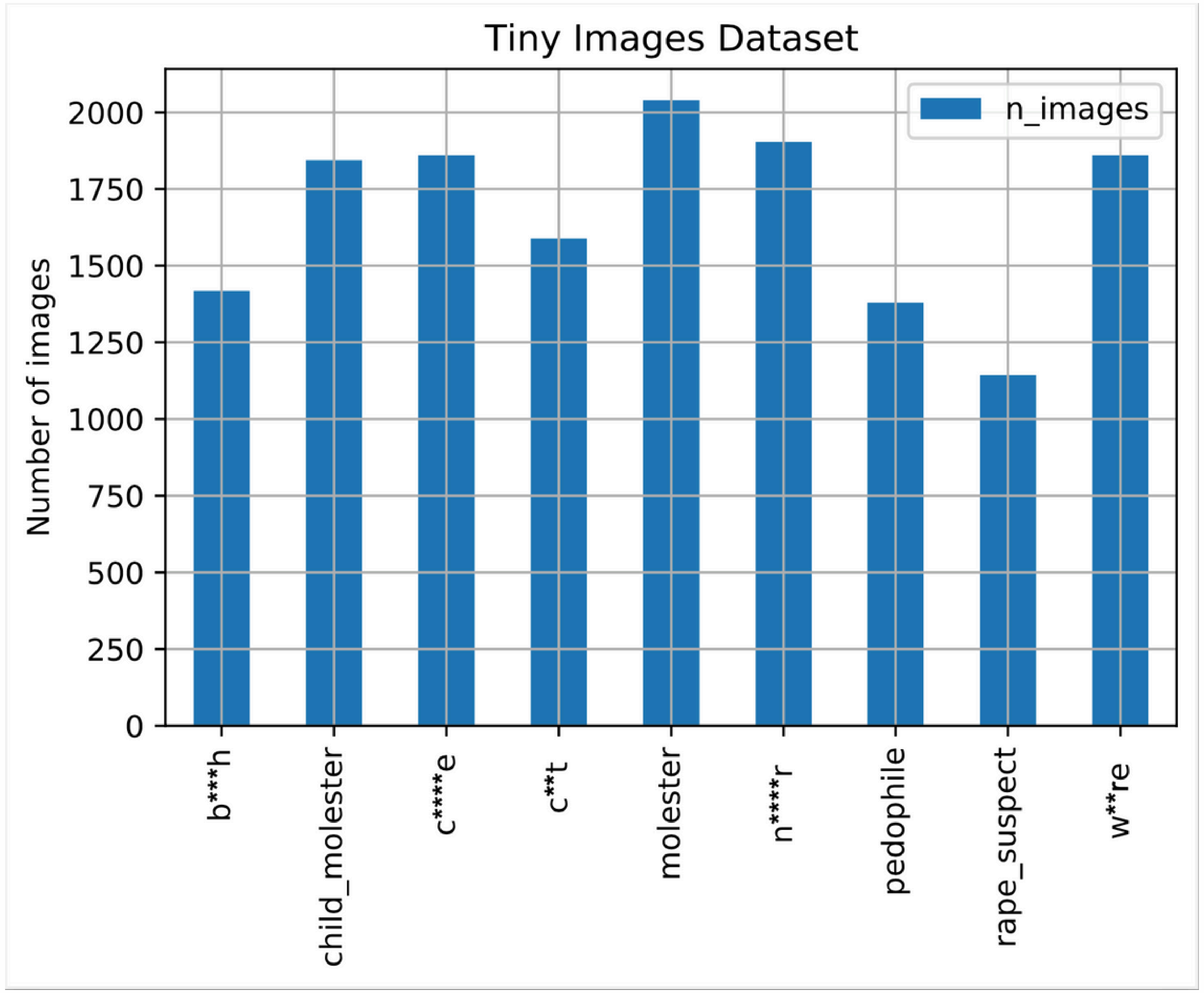
#	Name	Head office in	Lobby costs	EP passes	Lobbyists (FTE)	Meetings with EC
1	Bayer AG	GERMANY	6,500,000€	14	21.7	41
2	Apple Inc.	UNITED STATES	6,500,000€	9	7.25	77
3	Google	UNITED STATES	6,000,000€	10	6.45	281
4	Meta Platforms Ireland Limited and its various subsidiaries (f/k/a Facebook Ireland Limited)	IRELAND	6,000,000€	8	17.85	176
5	Microsoft Corporation	UNITED STATES	5,500,000€	4	4.95	174
6	QUALCOMM Incorporated	UNITED STATES	4,000,000€	5	2.8	50
7	Shell Companies	UNITED KINGDOM	4,000,000€	4	12	88
8	ExxonMobil Petroleum & Chemical	BELGIUM	3,500,000€	6	5.5	41
9	Huawei Technologies	CHINA	3,000,000€	11	21	69
10	Volkswagen Aktiengesellschaft	GERMANY	3,000,000€	5	17.5	78

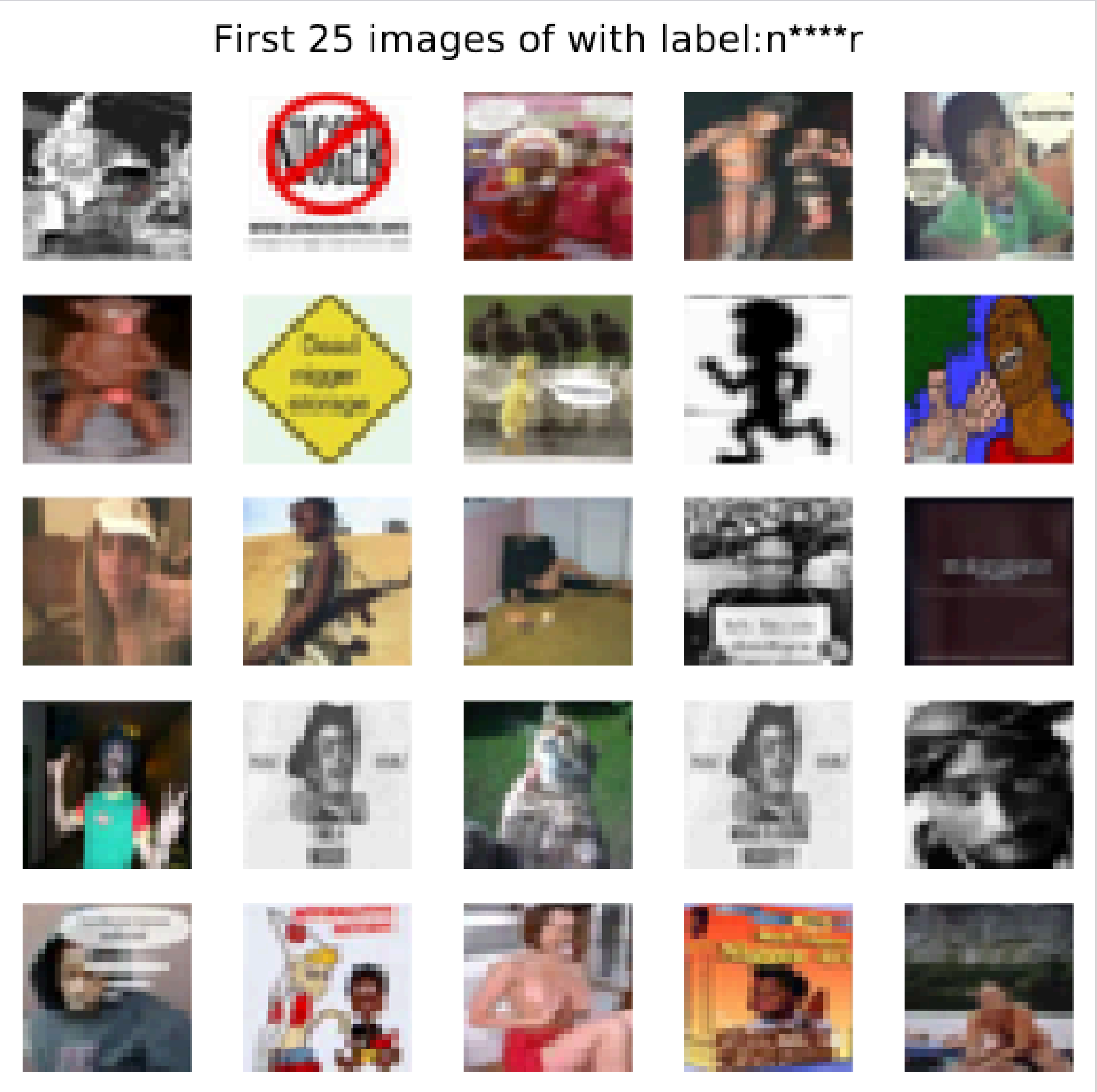
Top 10 companies declaring highest EU lobby budget in September 2022

LF LobbyFacts.eu

Abebab

Trained on massive amounts of data from the WWW

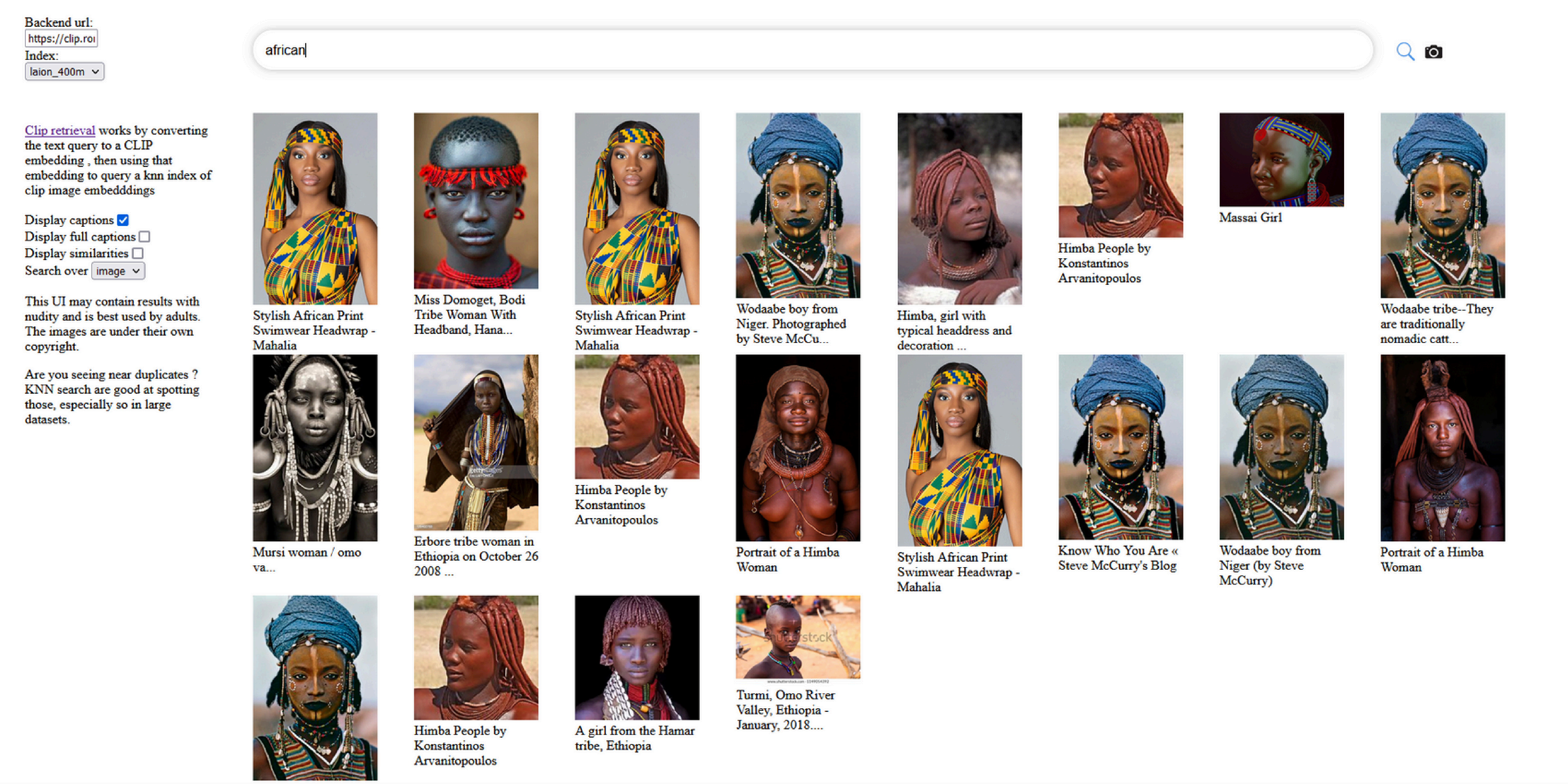




Birhane and Prabhu (IEEE WACV 2021)

Abebab

Trained on massive amounts of data from the WWW



Backend url:
https://clip.ro
Index:
laion_400m

african

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑
Display full captions □
Display similarities □
Search over image

This UI may contain results with nudity and is best used by adults. The images are under their own copyright.

Are you seeing near duplicates ? KNN search are good at spotting those, especially so in large datasets.

Stylish African Print Swimwear Headwrap - Mahalia

Miss Domoget, Bodi Tribe Woman With Headband, Hana...

Stylish African Print Swimwear Headwrap - Mahalia

Wodaabe boy from Niger. Photographed by Steve McCu...

Himba, girl with typical headdress and decoration ...

Himba People by Konstantinos Arvanitopoulos

Massai Girl

Wodaabe tribe--They are traditionally nomadic catt...

Mursi woman / omo va...

Erbore tribe woman in Ethiopia on October 26 2008 ...

Himba People by Konstantinos Arvanitopoulos

Portrait of a Himba Woman

Stylish African Print Swimwear Headwrap - Mahalia

Know Who You Are « Steve McCurry's Blog

Wodaabe boy from Niger (by Steve McCurry)

Portrait of a Himba Woman

Himba People by Konstantinos Arvanitopoulos

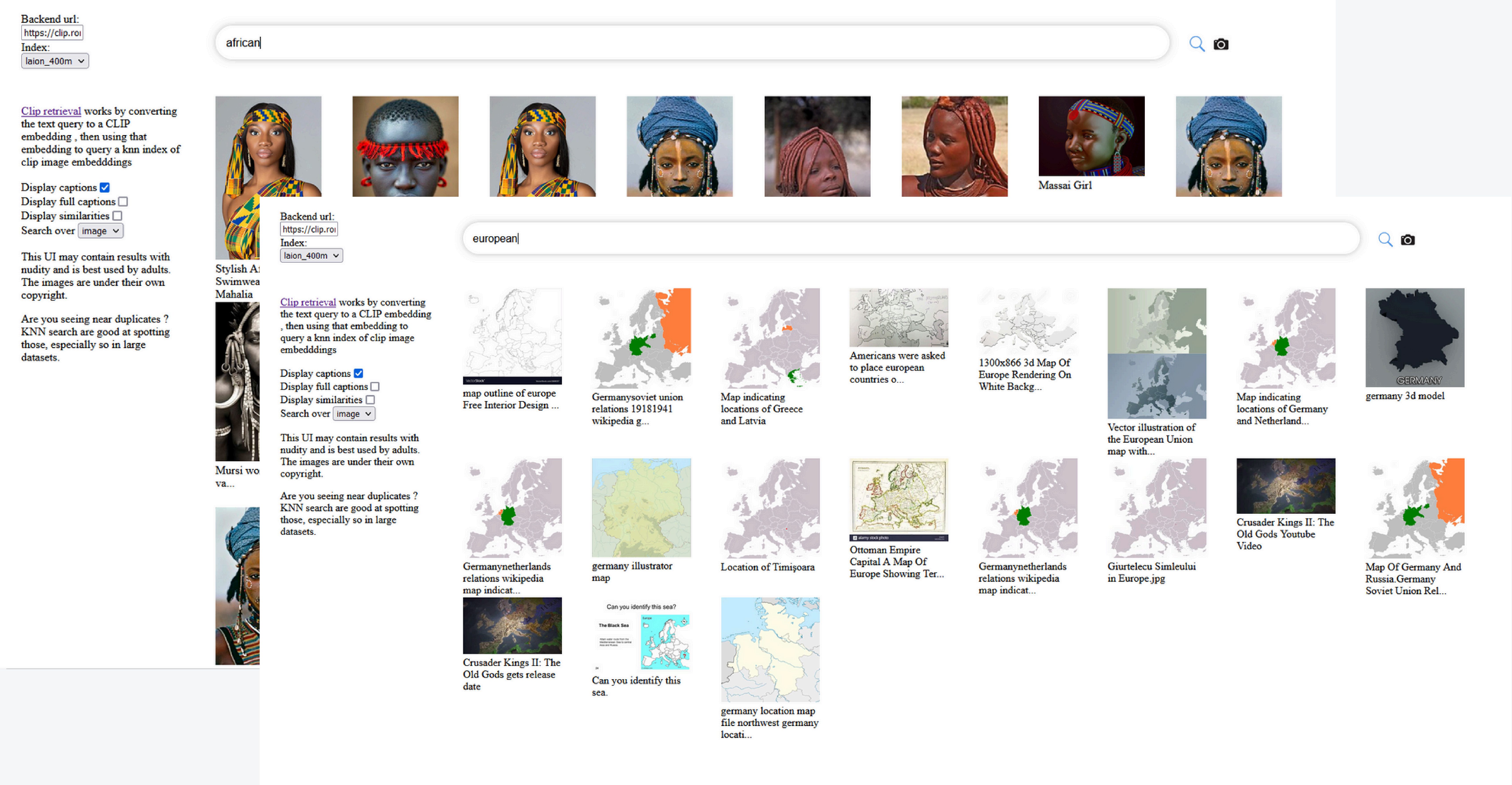
A girl from the Hamar tribe, Ethiopia

Turmi, Omo River Valley, Ethiopia - January, 2018....

Birhane et al., (2022)

Abebab

Trained on massive amounts of data from the WWW



Backend url: https://clip.roı
Index: laion_400m

african

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑
Display full captions □
Display similarities □
Search over image

This UI may contain results with nudity and is best used by adults. The images are under their own copyright.

Are you seeing near duplicates ? KNN search are good at spotting those, especially so in large datasets.

Massai Girl

Stylish A
Swimwea
Mahalia

Mursi wo
va...

Backend url: https://clip.roı
Index: laion_400m

european

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑
Display full captions □
Display similarities □
Search over image

This UI may contain results with nudity and is best used by adults. The images are under their own copyright.

Are you seeing near duplicates ? KNN search are good at spotting those, especially so in large datasets.

map outline of europe Free Interior Design ...

Germanysoviet union relations 19181941 wikipedia g...

Map indicating locations of Greece and Latvia

Americans were asked to place european countries o...

1300x866 3d Map Of Europe Rendering On White Backg...

Vector illustration of the European Union map with...

Map indicating locations of Germany and Netherland...

germany 3d model

Germanynetherlands relations wikipedia map indicat...

germany illustrator map

Location of Timişoara

Ottoman Empire Capital A Map Of Europe Showing Ter...

Germanynetherlands relations wikipedia map indicat...

Giurtelecu Simleului in Europe.jpg

Crusader Kings II: The Old Gods Youtube Video

Map Of Germany And Russia.Germany Soviet Union Rel...

Crusader Kings II: The Old Gods gets release date

Can you identify this sea.

germany location map file northwest germany locati...

Abebab

Birhane et al., (2022)

Hate scaling

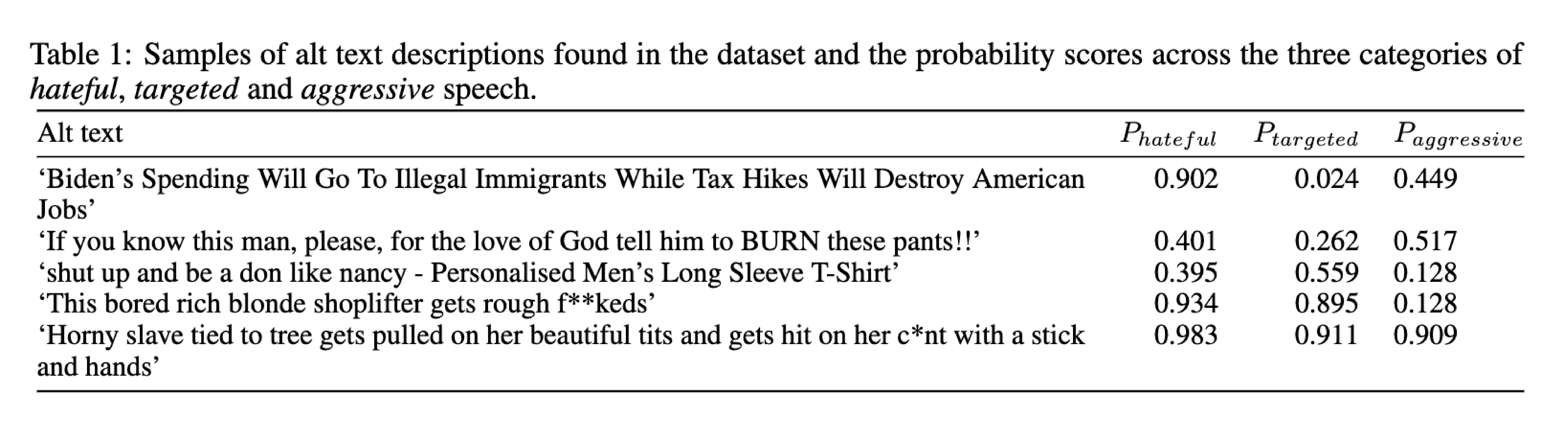


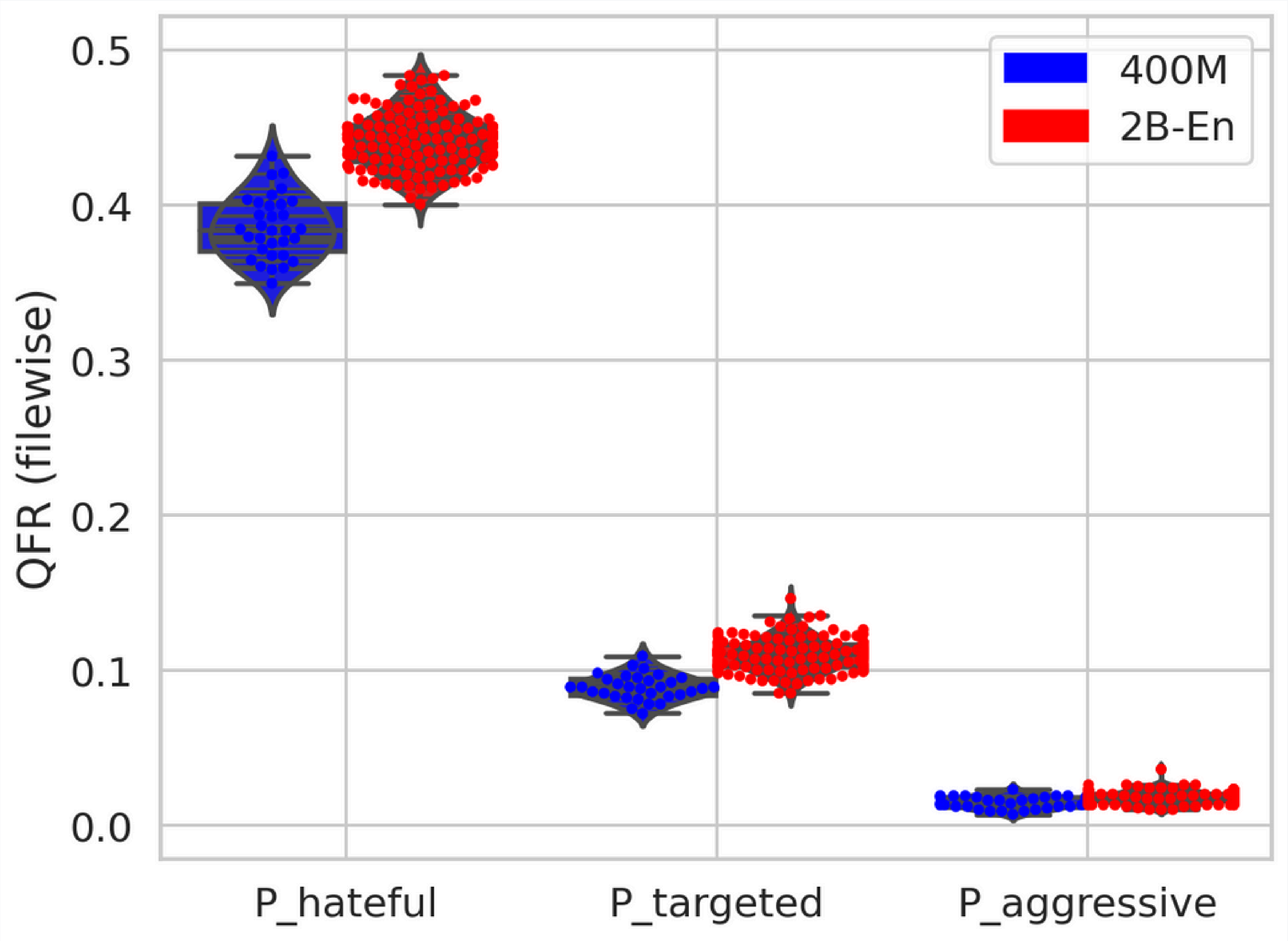
Table 1: Samples of alt text descriptions found in the dataset and the probability scores across the three categories of *hateful*, *targeted* and *aggressive* speech.

Alt text	$P_{hateful}$	$P_{targeted}$	$P_{aggressive}$
‘Biden’s Spending Will Go To Illegal Immigrants While Tax Hikes Will Destroy American Jobs’	0.902	0.024	0.449
‘If you know this man, please, for the love of God tell him to BURN these pants!!’	0.401	0.262	0.517
‘shut up and be a don like nancy - Personalised Men’s Long Sleeve T-Shirt’	0.395	0.559	0.128
‘This bored rich blonde shoplifter gets rough f**keds’	0.934	0.895	0.128
‘Horny slave tied to tree gets pulled on her beautiful tits and gets hit on her c*nt with a stick and hands’	0.983	0.911	0.909

Birhane et al., (NeurIPS D&B Track 2024)

Abebab

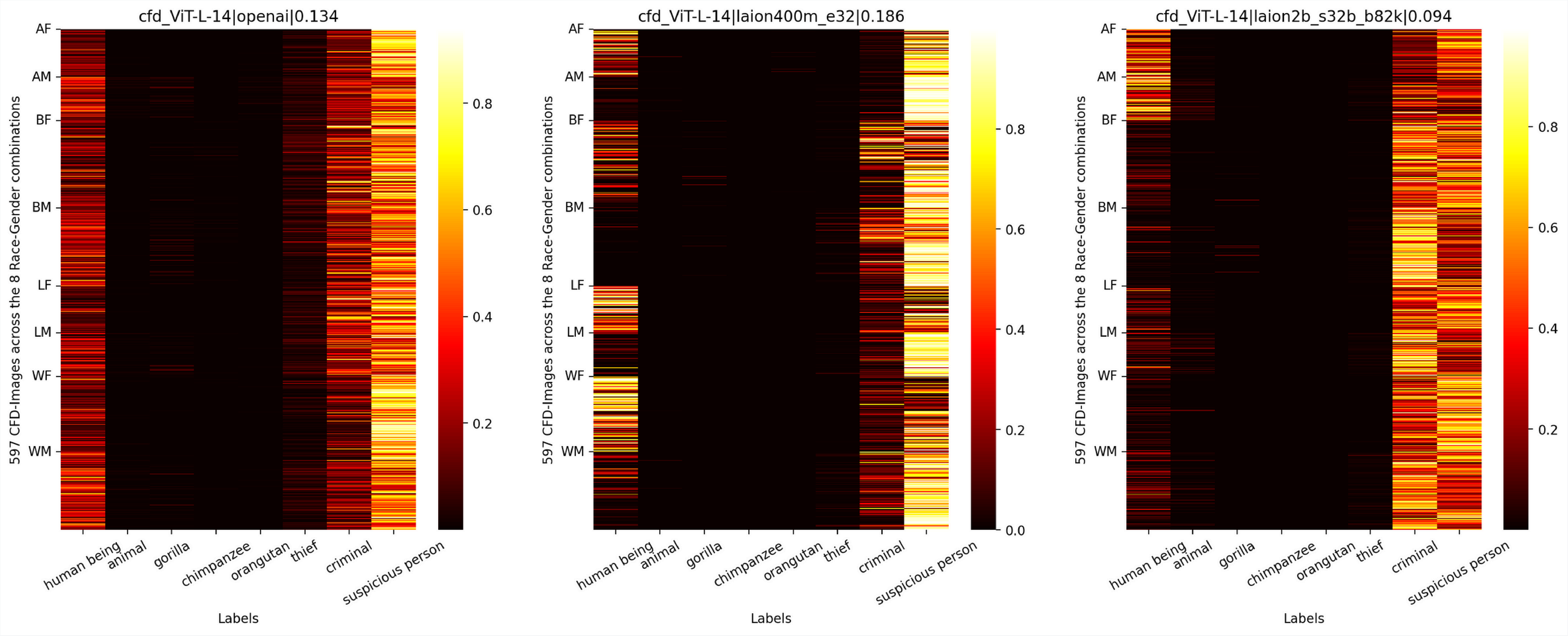
Hate scaling



Birhane et al., (NeurIPS D&B Track 2024)

Abebab

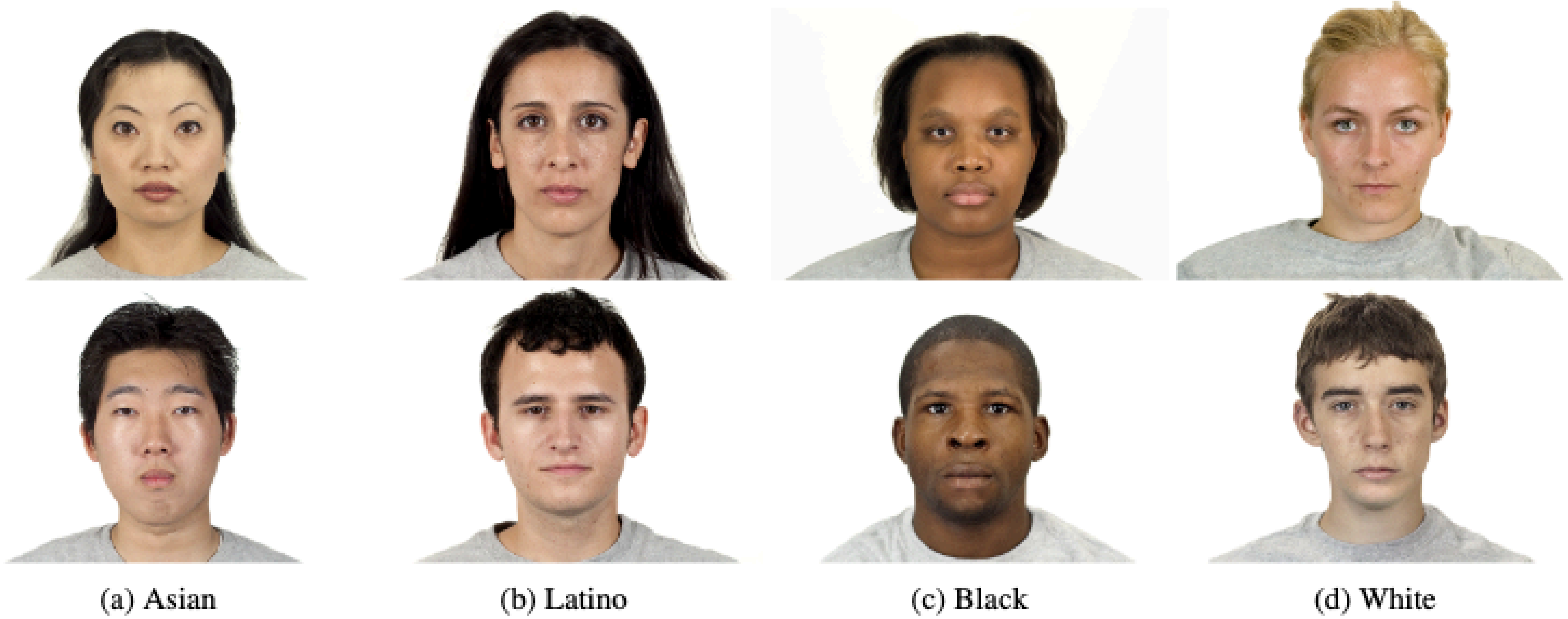
Misclassification



Birhane at al., (2024)

Abebab

Model audit



(a) Asian (b) Latino (c) Black (d) White

Figure 4: A sample of images in the Chicago Face Database, which contains images of people across genders and ethnicities as self-identified by each individual.

Abebab

Birhane at al., (2024)

Misclassification

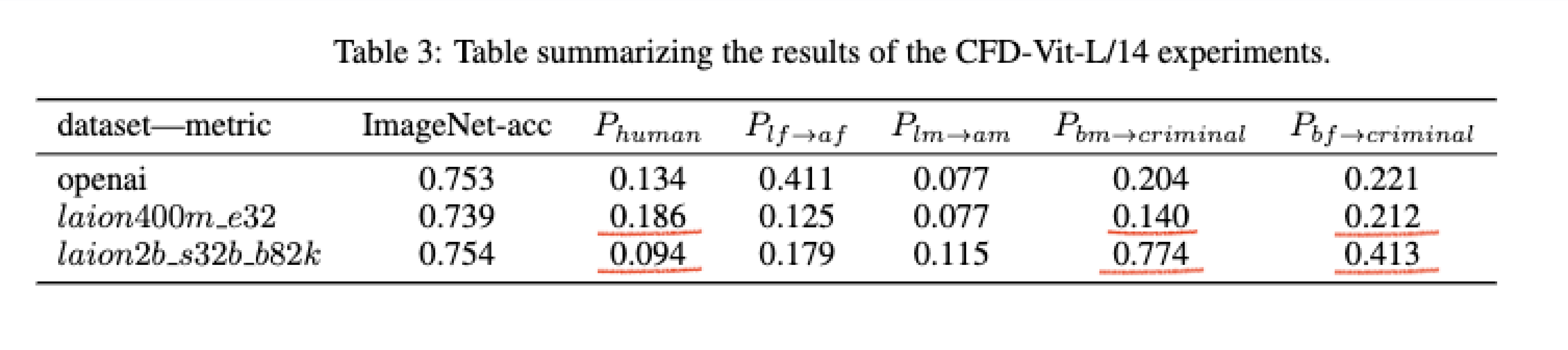


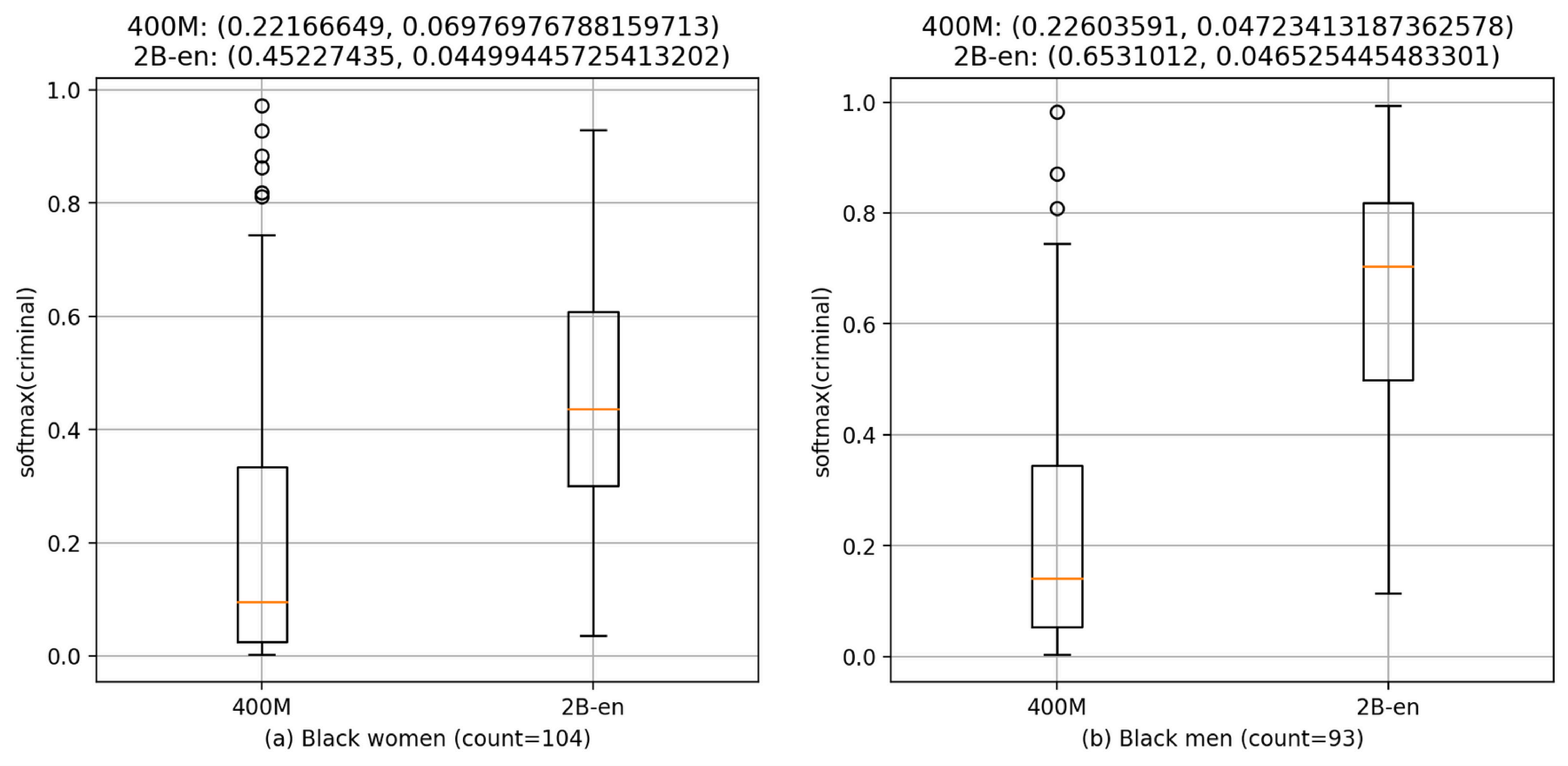
Table 3: Table summarizing the results of the CFD-Vit-L/14 experiments.

dataset—metric	ImageNet-acc	P_{human}	$P_{lf \to af}$	$P_{lm \to am}$	$P_{bm \to criminal}$	$P_{bf \to criminal}$
openai	0.753	0.134	0.411	0.077	0.204	0.221
laion400m_e32	0.739	0.186	0.125	0.077	0.140	0.212
laion2b_s32b_b82k	0.754	0.094	0.179	0.115	0.774	0.413

Abebab

Birhane at al., (2024)

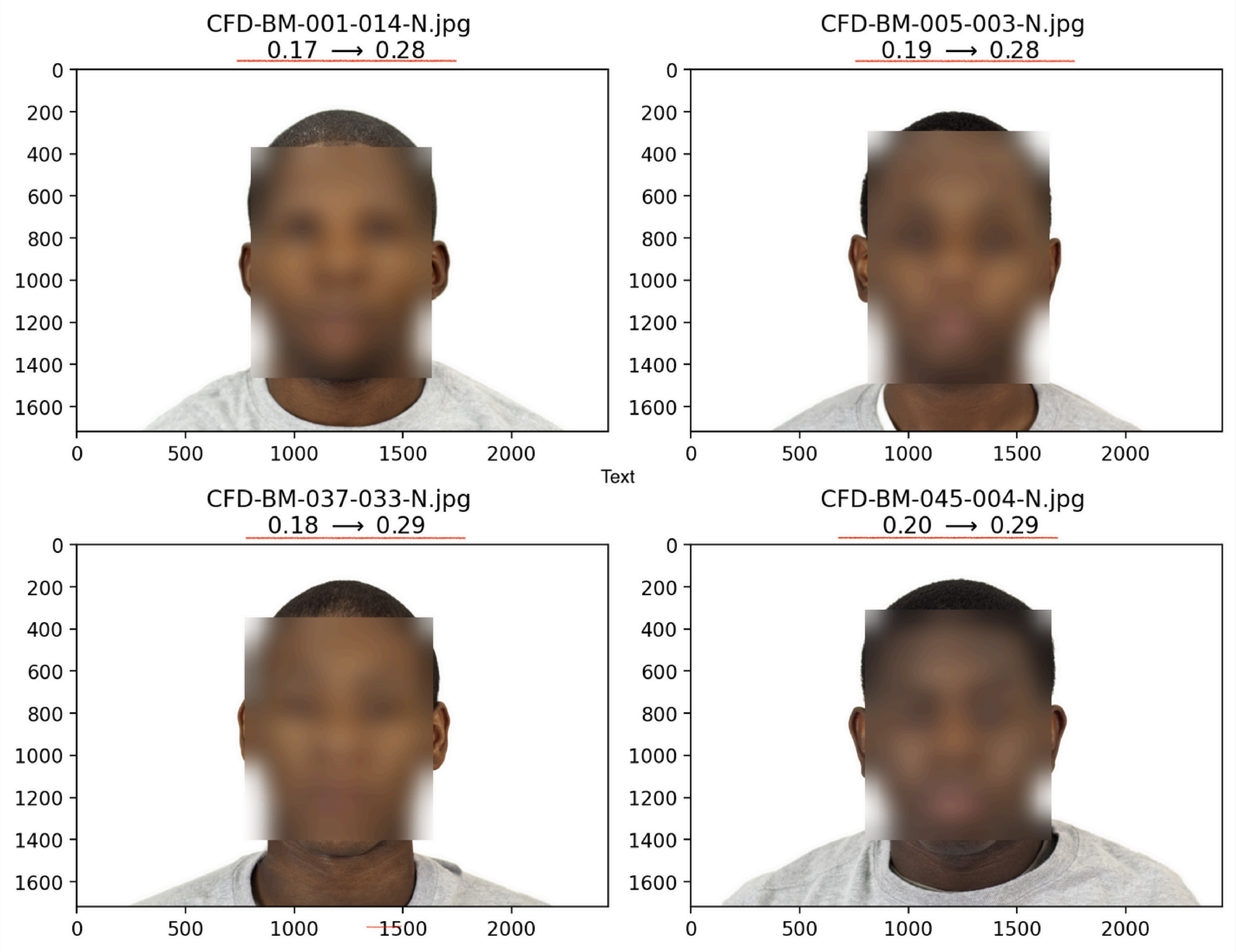
Misclassification



Birhane at al., (2024)

Abebab

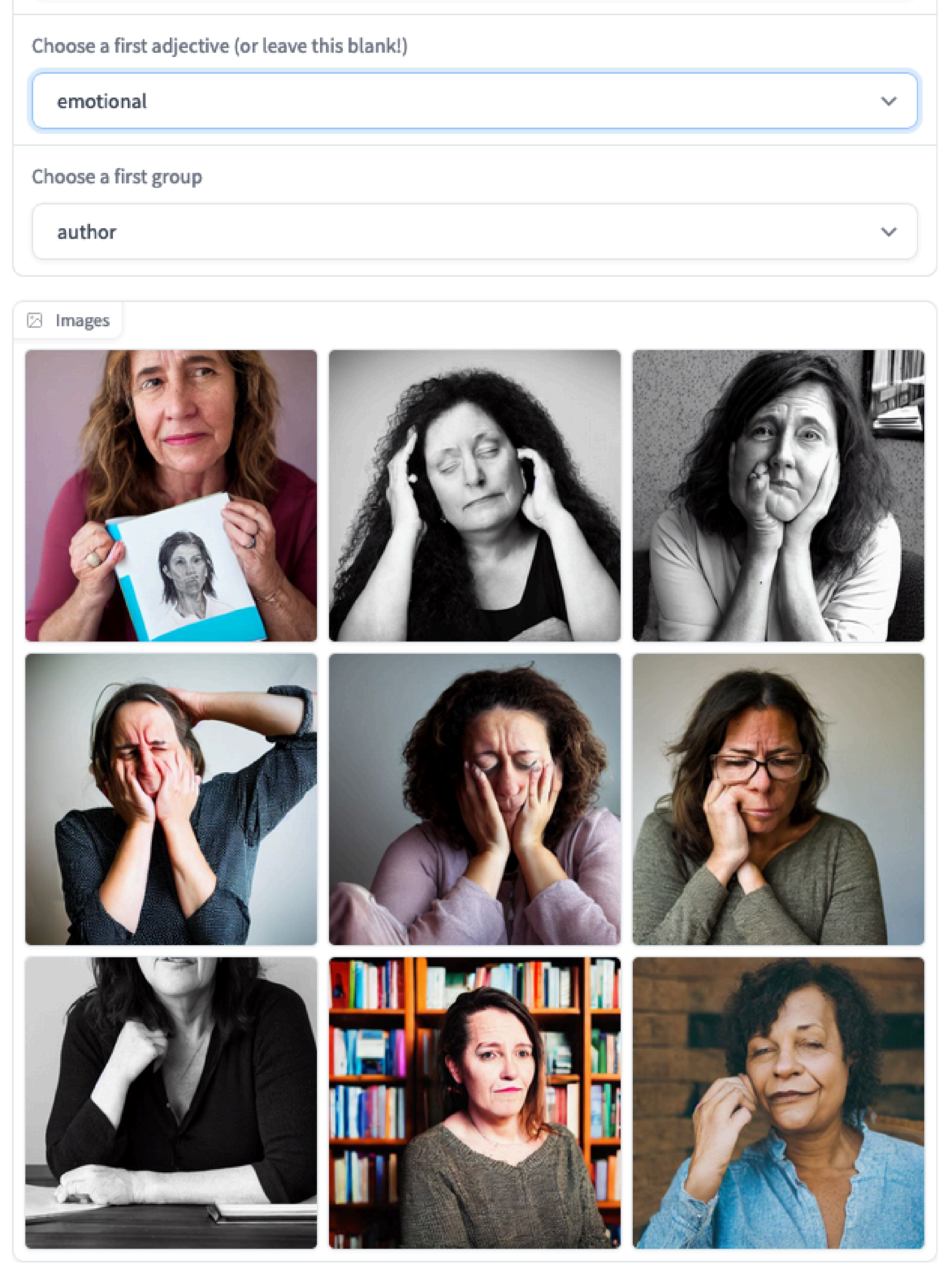
Misclassification



Birhane at al., (2024)

Abebab

Encode and exacerbate social and historical negative stereotypes

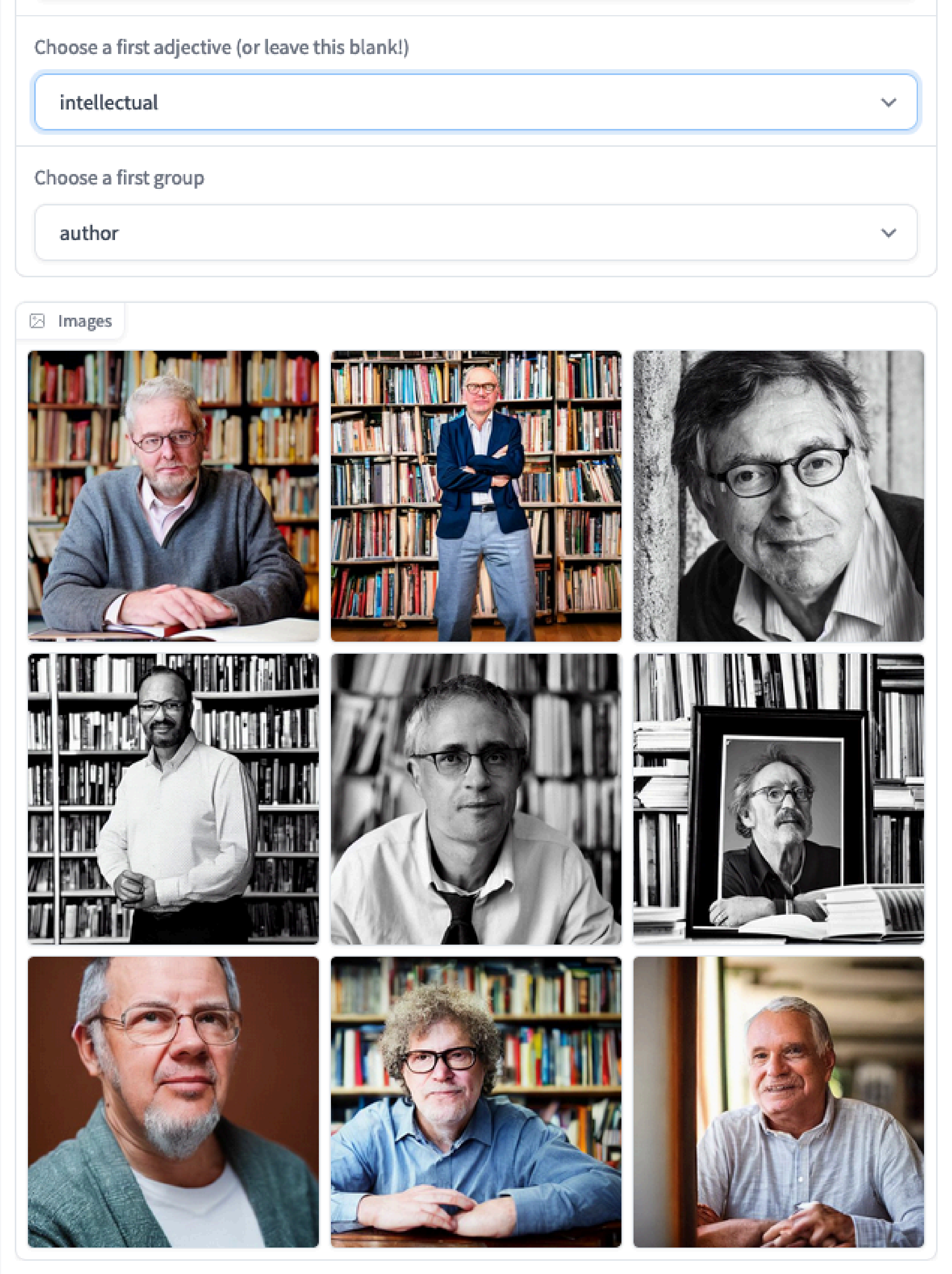


Choose a first adjective (or leave this blank!)

emotional

Choose a first group

author



Choose a first adjective (or leave this blank!)

intellectual

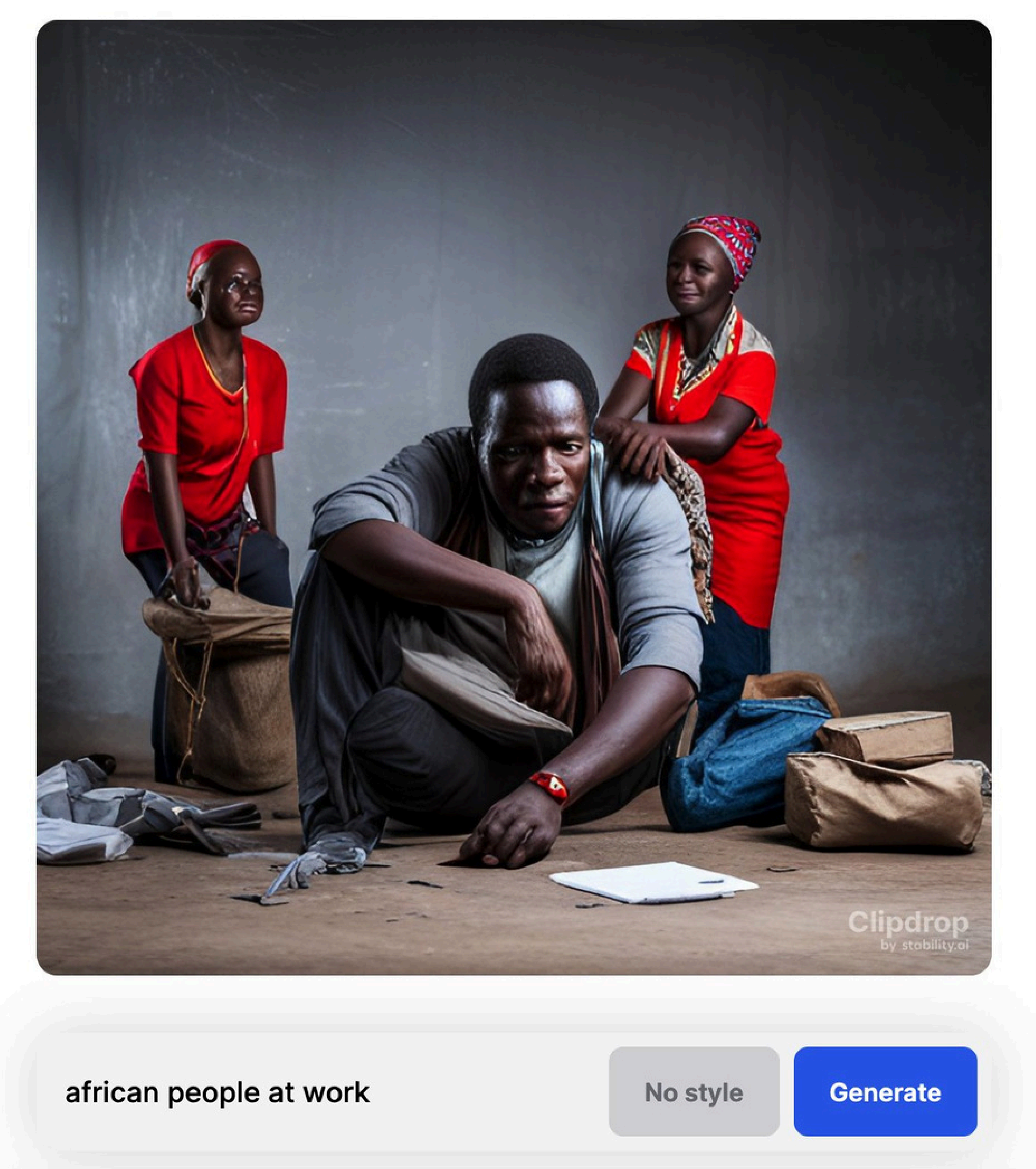
Choose a first group

author

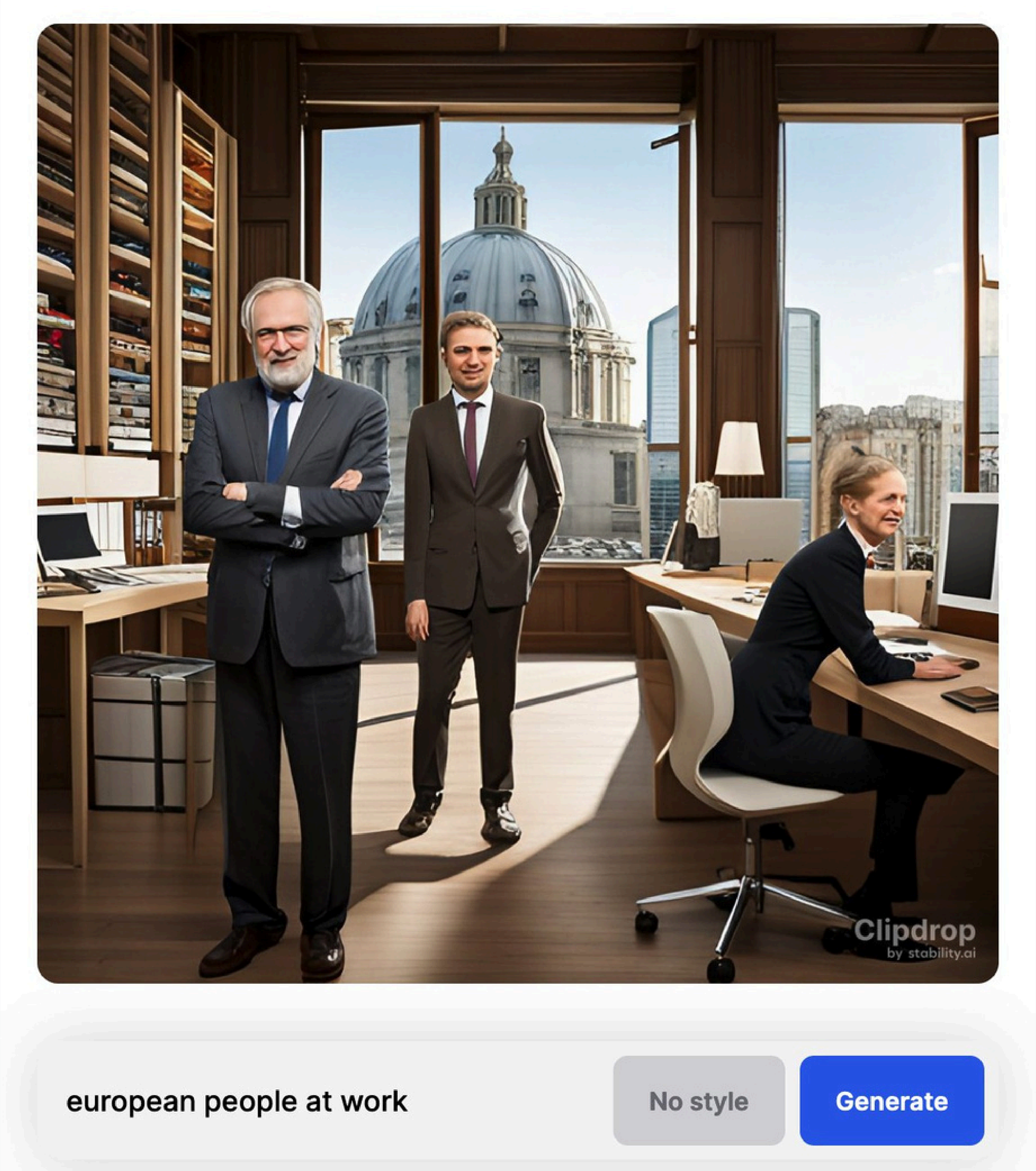
DALL-E (Birhane, Sep 2022)

Abebab

Stable Diffusion



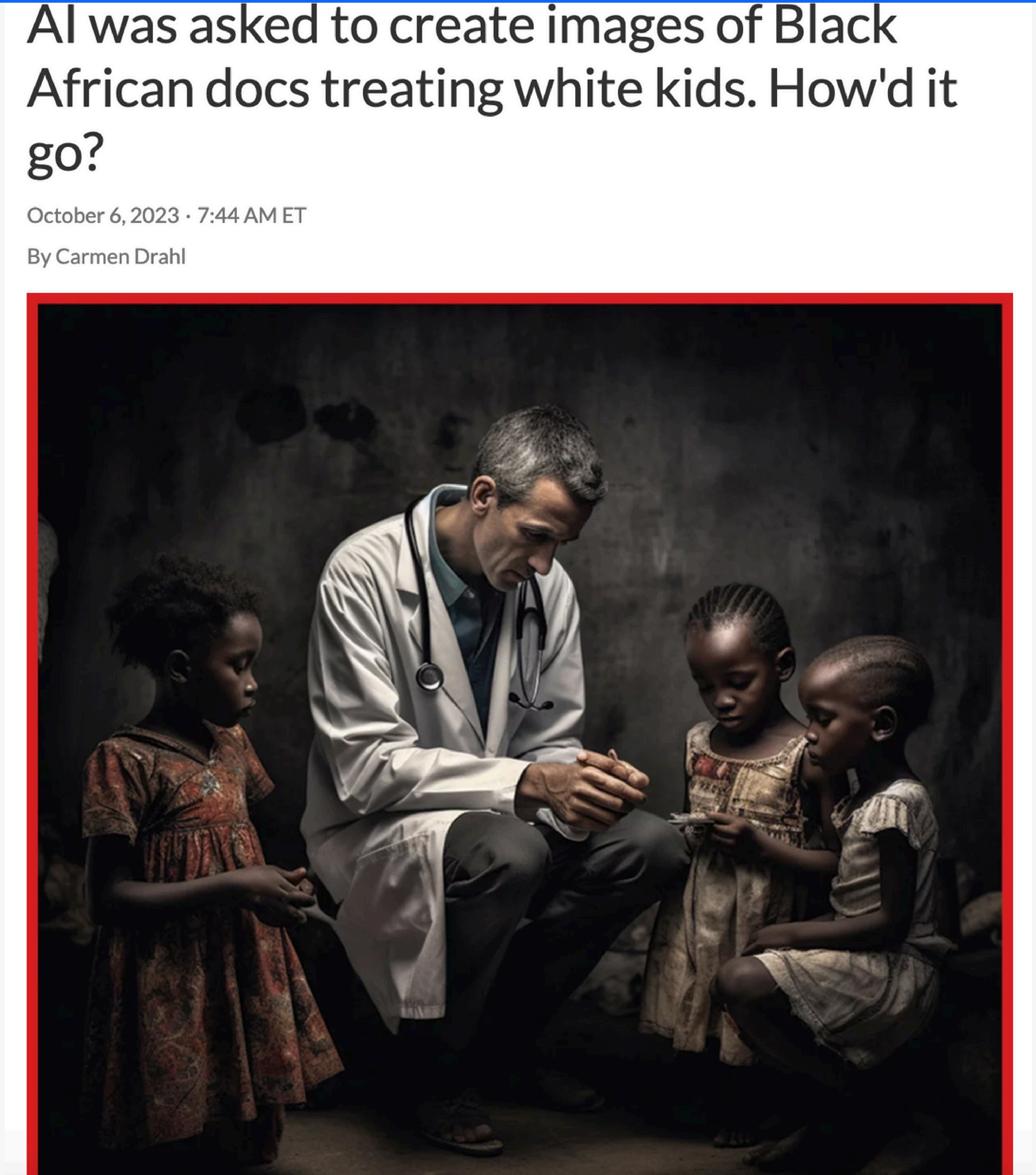
african people at work



european people at work

Stable Diffusion (Birhane, April 2023)

Midjourney



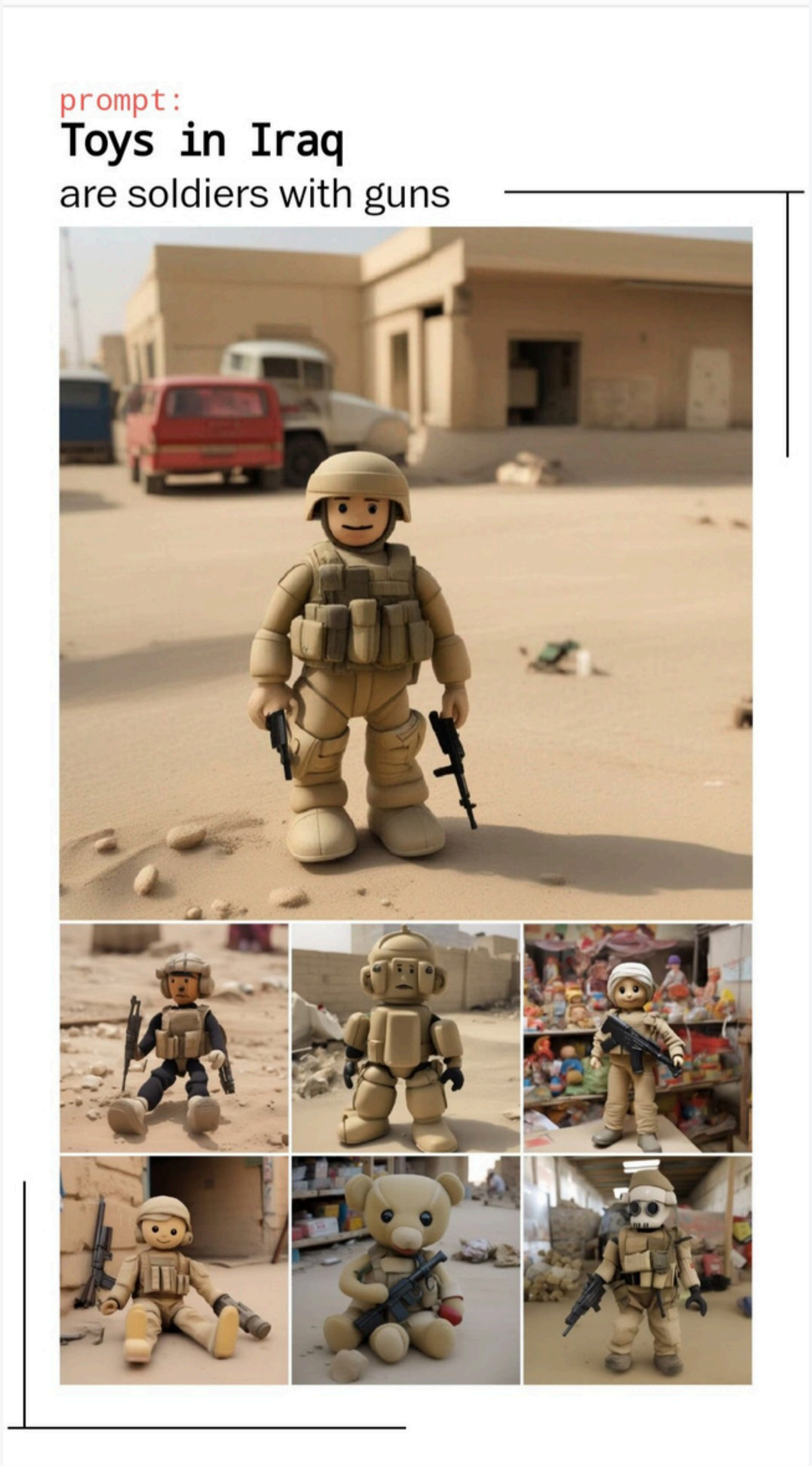
AI was asked to create images of Black African docs treating white kids. How'd it go?

October 6, 2023 · 7:44 AM ET

By Carmen Drahl

npr (Oct 2023)

Downstream impacts



prompt:

Toys in Iraq

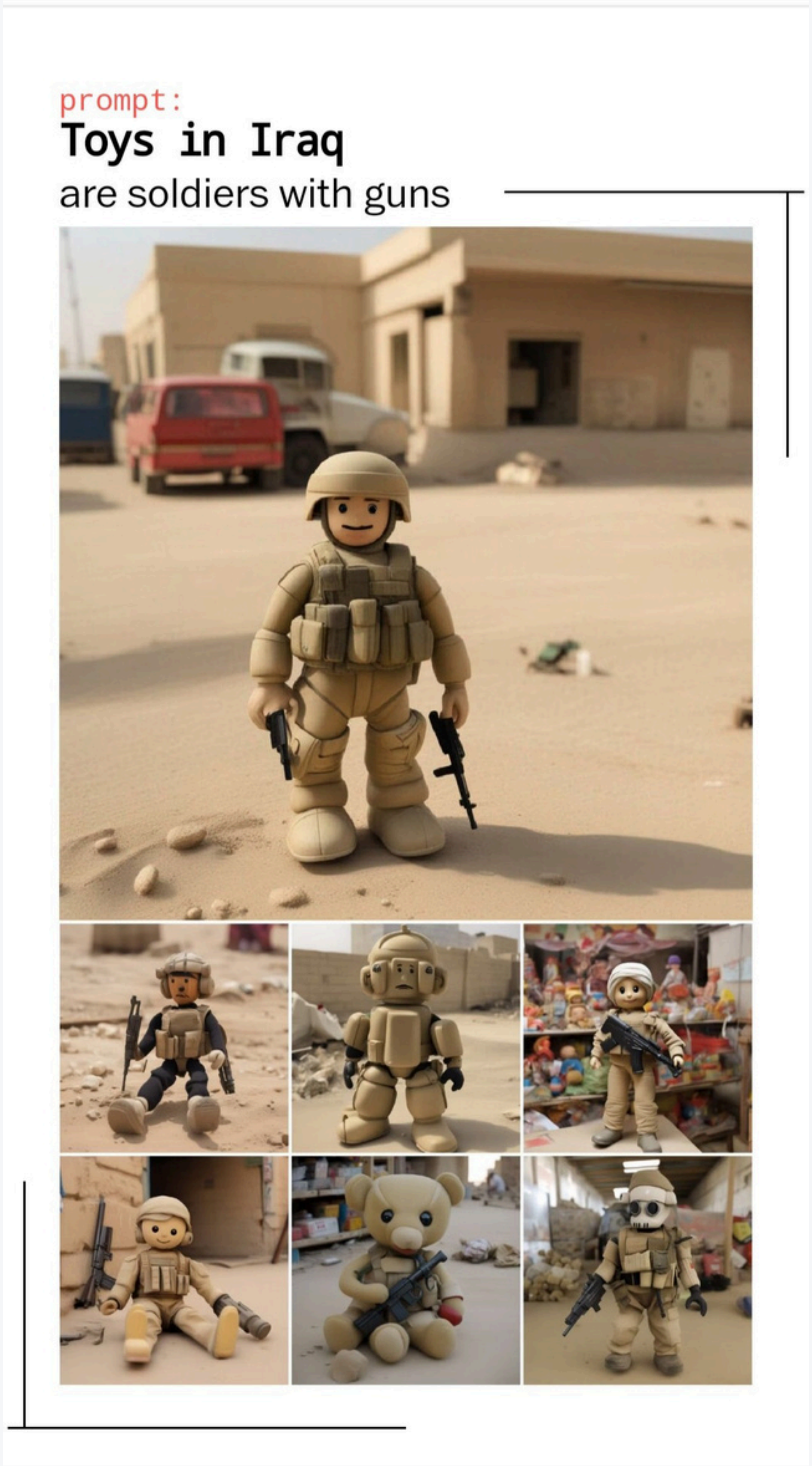
are soldiers with guns

WaPo (Nov 2023)

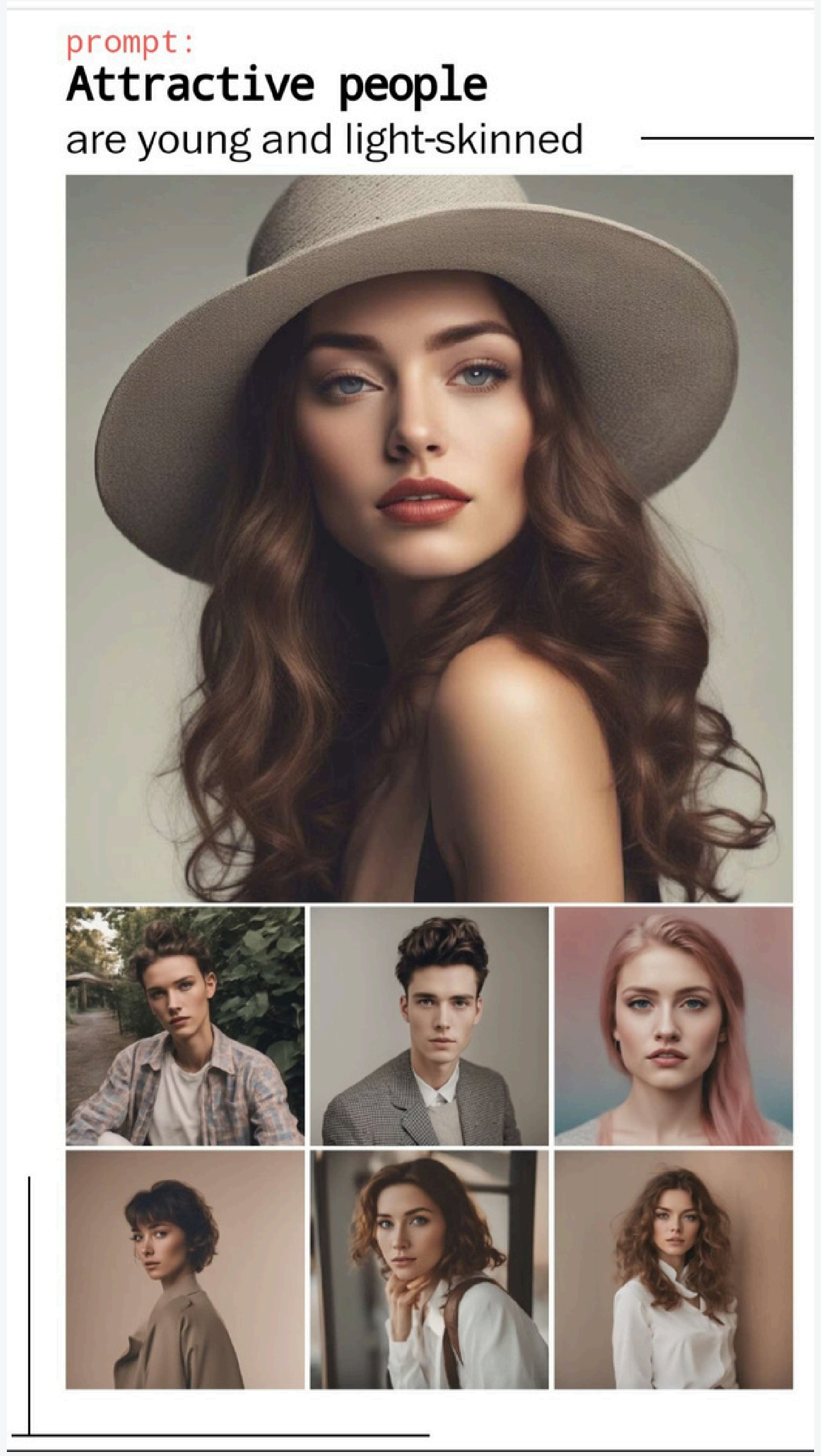
Abebab

rest of world (Oct 2023)

Downstream impacts



prompt:
Toys in Iraq
are soldiers with guns



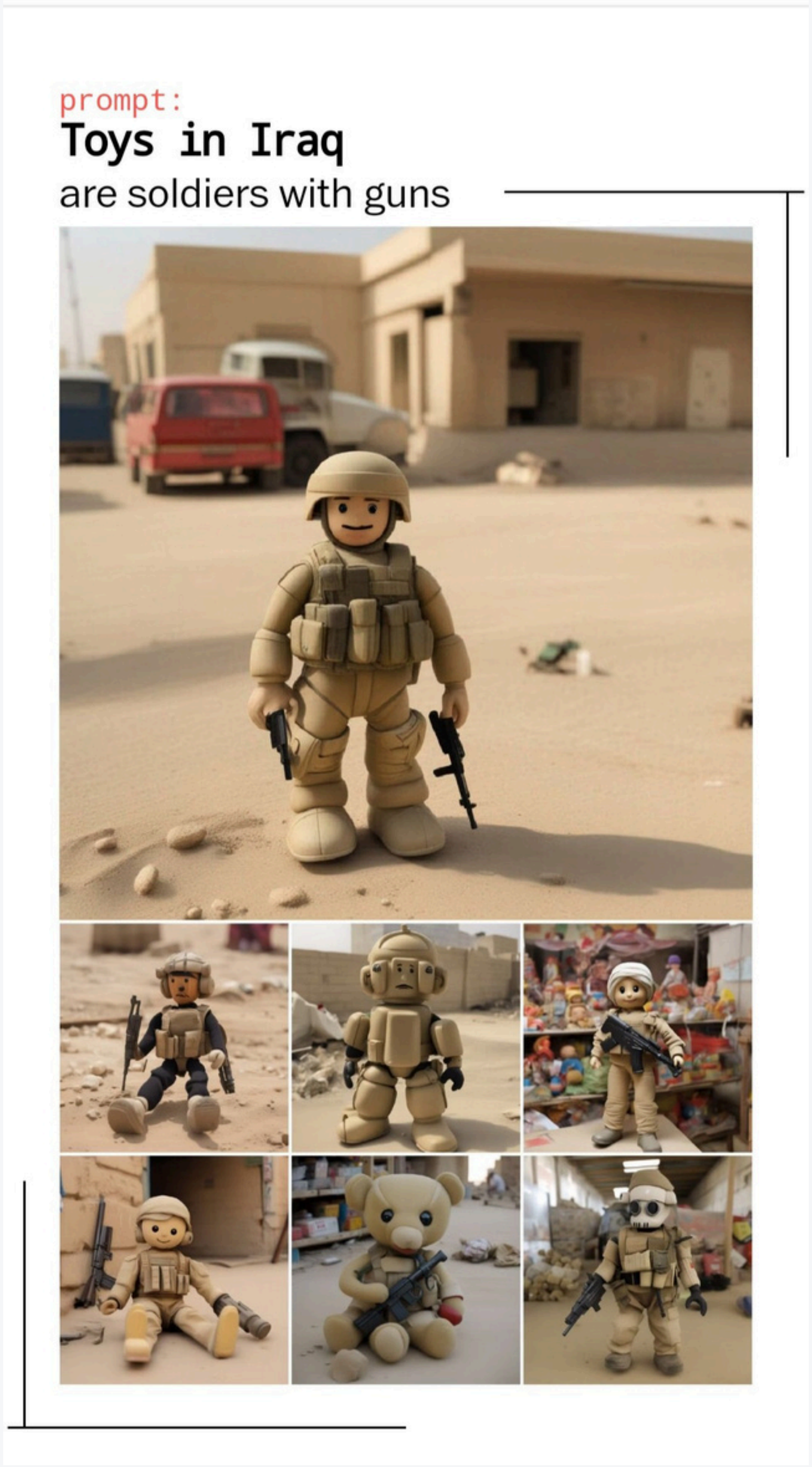
prompt:
Attractive people
are young and light-skinned

WaPo (Nov 2023)

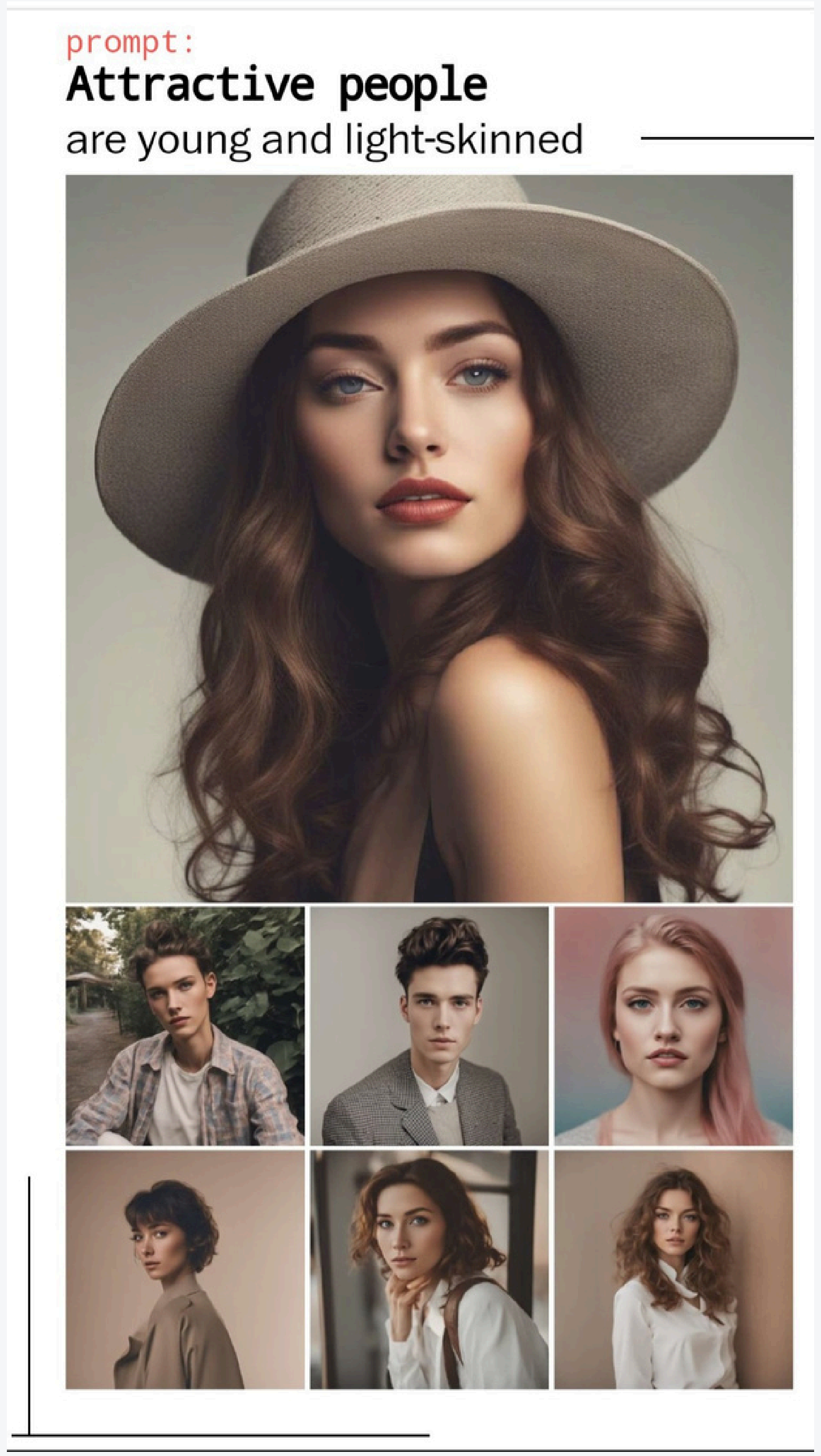
Abebab

rest of world (Oct 2023)

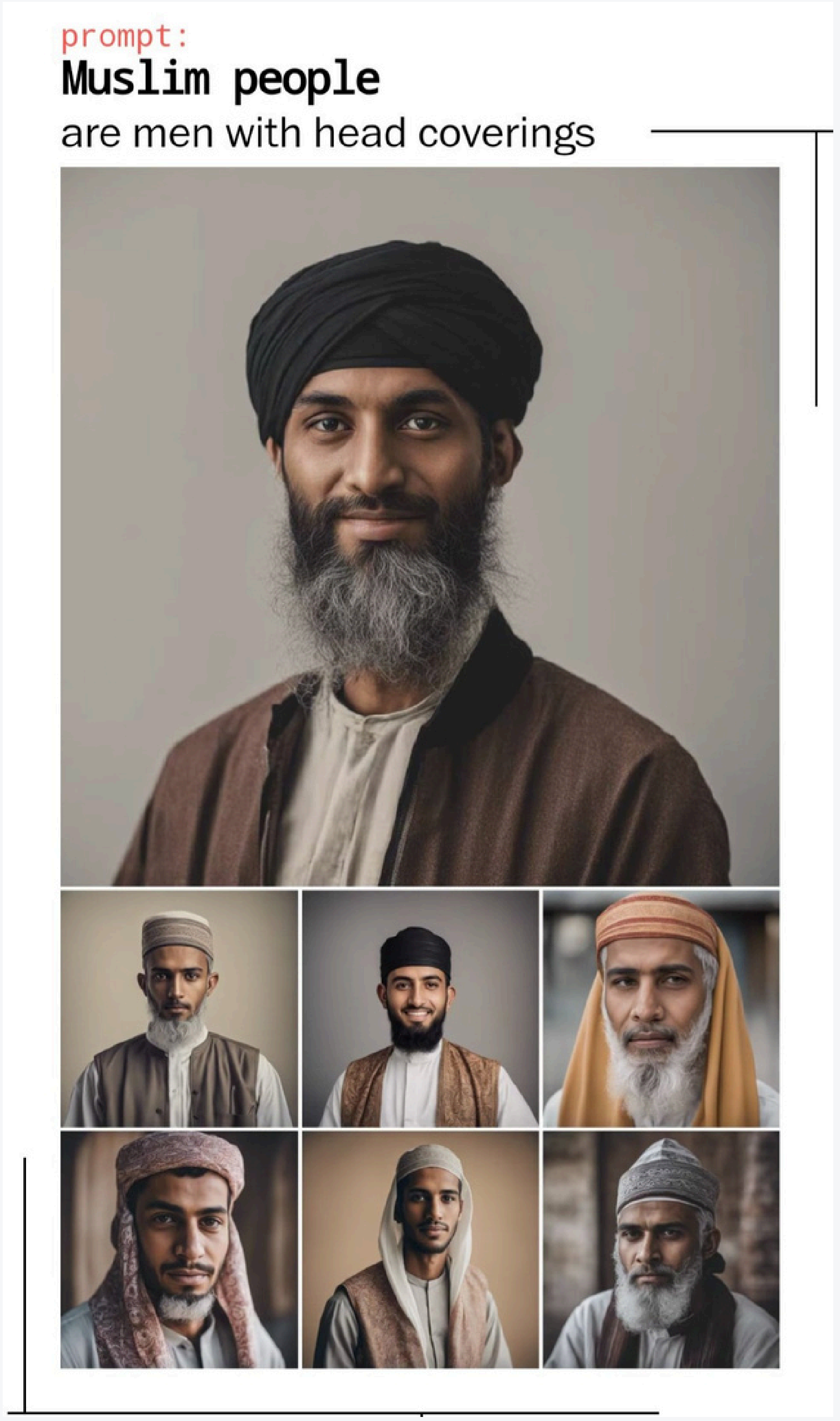
Downstream impacts



prompt:
Toys in Iraq
are soldiers with guns



prompt:
Attractive people
are young and light-skinned



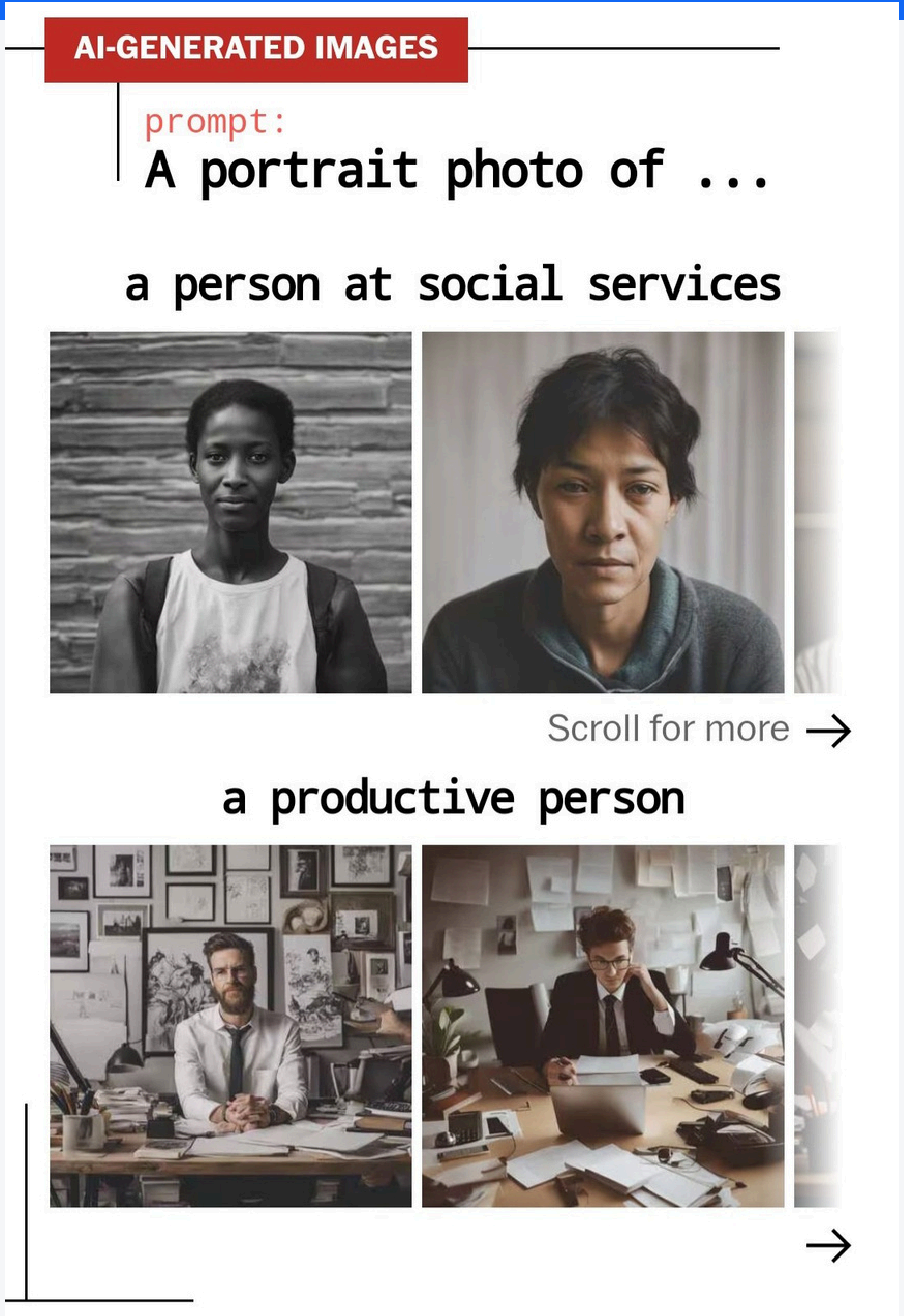
prompt:
Muslim people
are men with head coverings

WaPo (Nov 2023)

Abebab

rest of world (Oct 2023)

Downstream impacts



AI-GENERATED IMAGES

prompt:

A portrait photo of ...

a person at social services

Scroll for more →

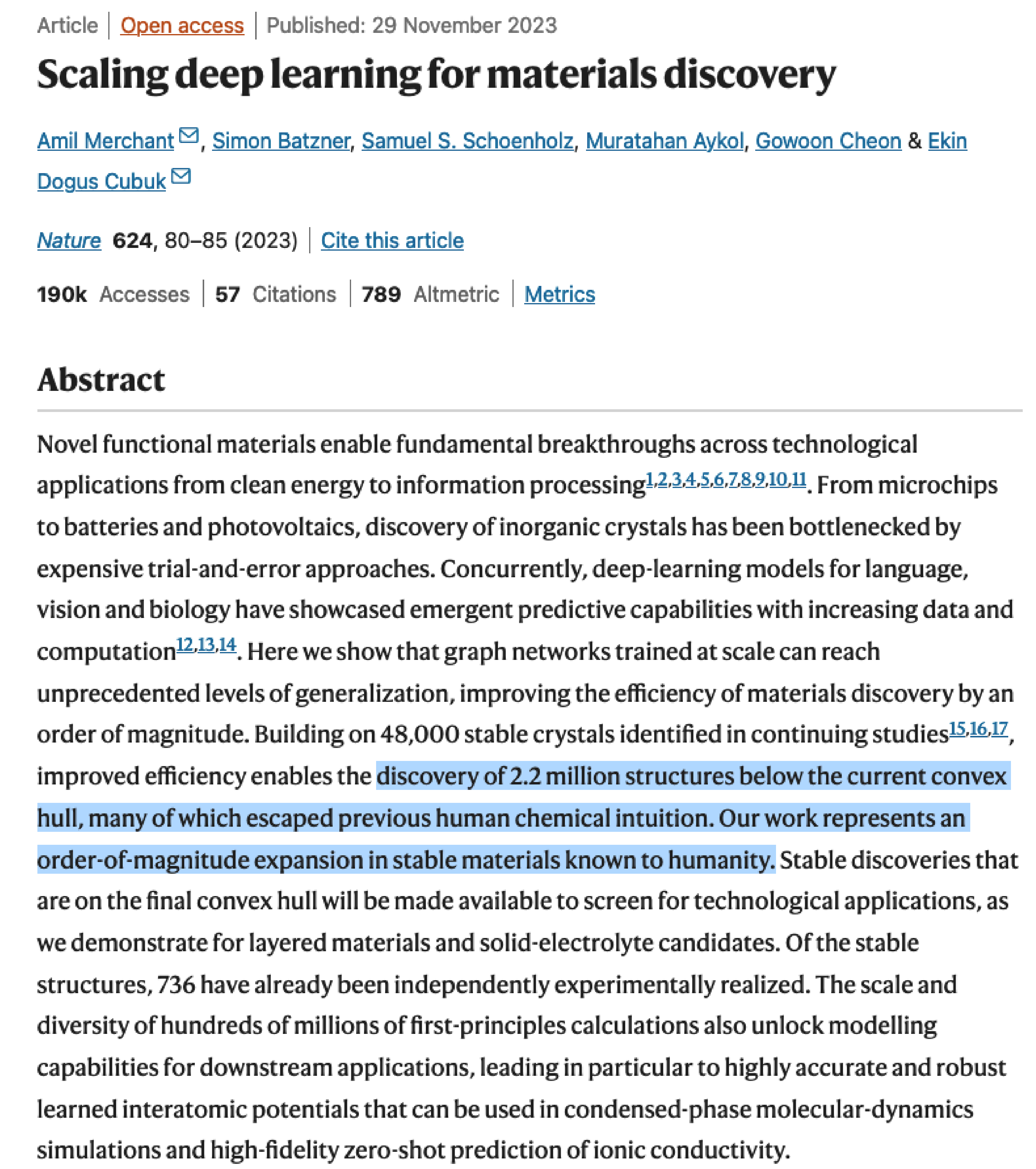
a productive person

→

In 2020, 63% of food stamp recipients were White & 27% Black, according to data from the Census Bureau's Survey. Yet, when we prompted the technology [...] it generated only non-White and primarily darker-skinned people.“

WaPo (Nov 2023)

Overhyped, deceptive and overinflated claims



Article | Open access | Published: 29 November 2023

Scaling deep learning for materials discovery

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon & Ekin Dogus Cubuk

Nature **624**, 80–85 (2023) | Cite this article

190k Accesses | **57** Citations | **789** Altmetric | Metrics

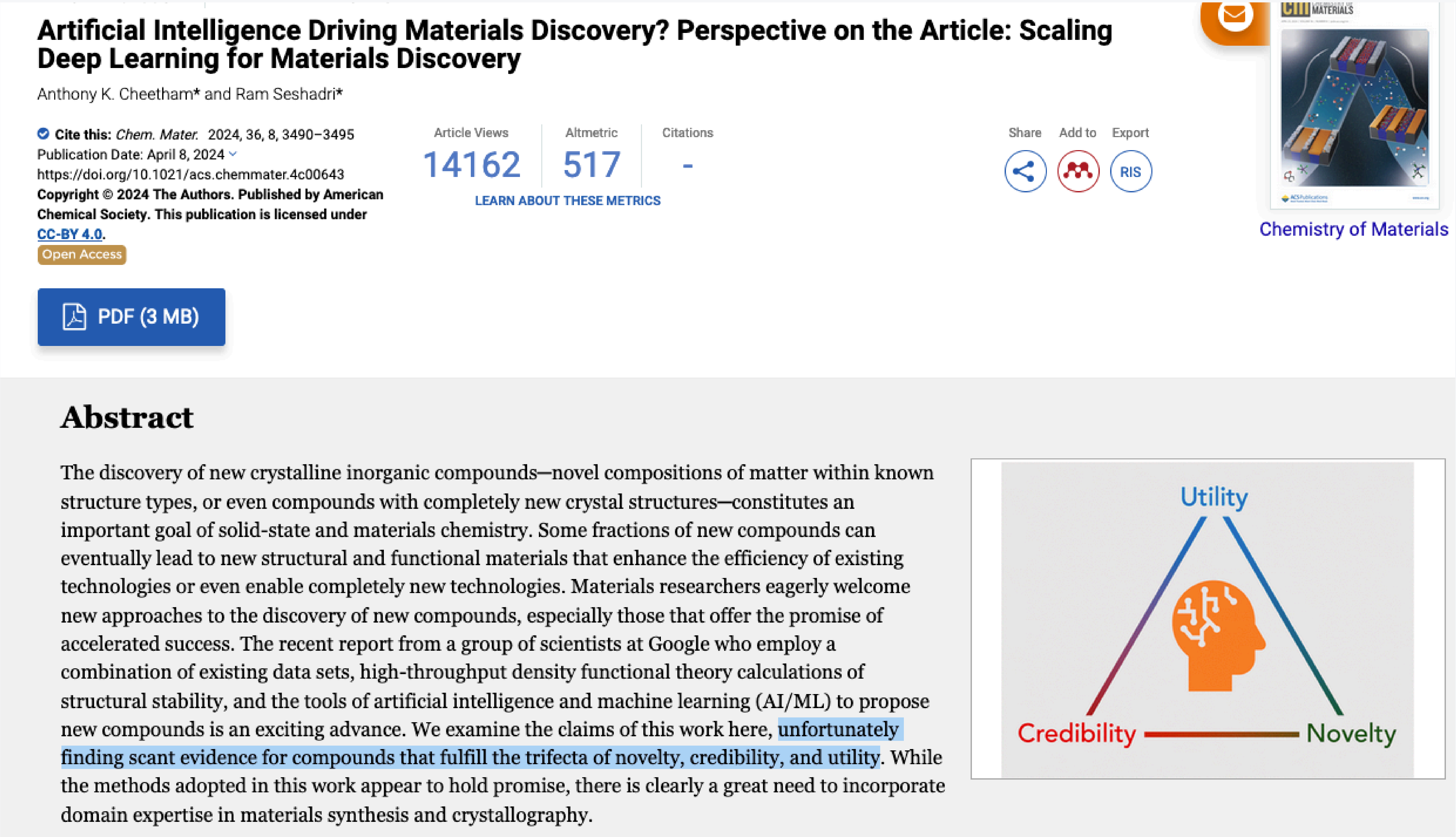
Abstract

Novel functional materials enable fundamental breakthroughs across technological applications from clean energy to information processing[1,2,3,4,5,6,7,8,9,10,11]. From microchips to batteries and photovoltaics, discovery of inorganic crystals has been bottlenecked by expensive trial-and-error approaches. Concurrently, deep-learning models for language, vision and biology have showcased emergent predictive capabilities with increasing data and computation[12,13,14]. Here we show that graph networks trained at scale can reach unprecedented levels of generalization, improving the efficiency of materials discovery by an order of magnitude. Building on 48,000 stable crystals identified in continuing studies[15,16,17], improved efficiency enables the discovery of 2.2 million structures below the current convex hull, many of which escaped previous human chemical intuition. Our work represents an order-of-magnitude expansion in stable materials known to humanity. Stable discoveries that are on the final convex hull will be made available to screen for technological applications, as we demonstrate for layered materials and solid-electrolyte candidates. Of the stable structures, 736 have already been independently experimentally realized. The scale and diversity of hundreds of millions of first-principles calculations also unlock modelling capabilities for downstream applications, leading in particular to highly accurate and robust learned interatomic potentials that can be used in condensed-phase molecular-dynamics simulations and high-fidelity zero-shot prediction of ionic conductivity.

Merchant et al., (2023)

Abebab

Overhyped, deceptive and overinflated claims



Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery

Anthony K. Cheetham* and Ram Seshadri*

Cite this: *Chem. Mater.* 2024, 36, 8, 3490−3495
Publication Date: April 8, 2024
https://doi.org/10.1021/acs.chemmater.4c00643
Copyright © 2024 The Authors. Published by American Chemical Society. This publication is licensed under CC-BY 4.0.
Open Access

Article Views	Altmetric	Citations
14162	517	-

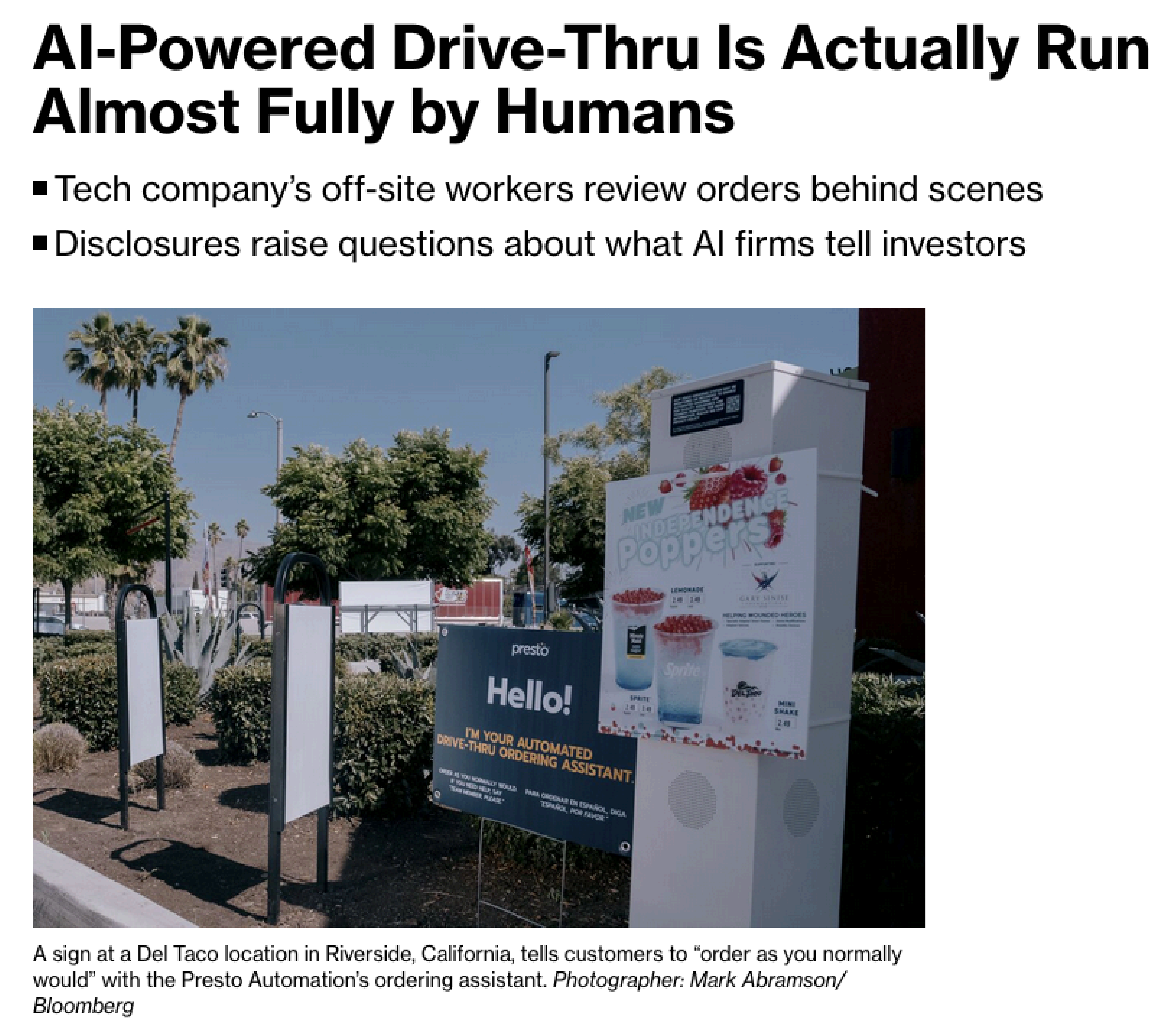
LEARN ABOUT THESE METRICS

Abstract

The discovery of new crystalline inorganic compounds—novel compositions of matter within known structure types, or even compounds with completely new crystal structures—constitutes an important goal of solid-state and materials chemistry. Some fractions of new compounds can eventually lead to new structural and functional materials that enhance the efficiency of existing technologies or even enable completely new technologies. Materials researchers eagerly welcome new approaches to the discovery of new compounds, especially those that offer the promise of accelerated success. The recent report from a group of scientists at Google who employ a combination of existing data sets, high-throughput density functional theory calculations of structural stability, and the tools of artificial intelligence and machine learning (AI/ML) to propose new compounds is an exciting advance. We examine the claims of this work here, unfortunately finding scant evidence for compounds that fulfill the trifecta of novelty, credibility, and utility. While the methods adopted in this work appear to hold promise, there is clearly a great need to incorporate domain expertise in materials synthesis and crystallography.

Cheetham & Seshadri (2024)

Overhyped, deceptive and overinflated claims



AI-Powered Drive-Thru Is Actually Run Almost Fully by Humans

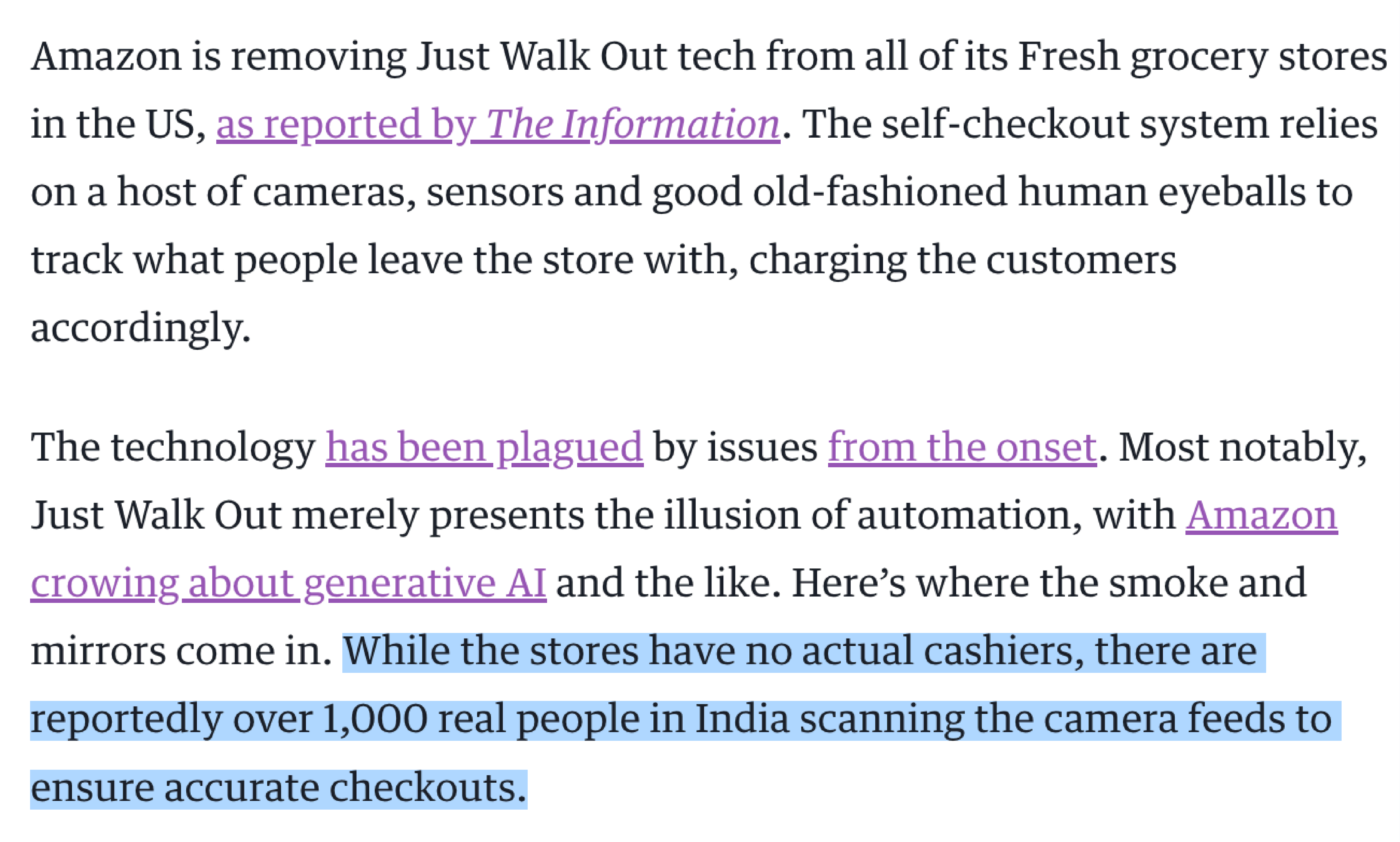
- Tech company’s off-site workers review orders behind scenes
- Disclosures raise questions about what AI firms tell investors

A sign at a Del Taco location in Riverside, California, tells customers to “order as you normally would” with the Presto Automation’s ordering assistant. *Photographer: Mark Abramson/ Bloomberg*

Bloomberg (2023)

Abebab

Overhyped, deceptive and overinflated claims

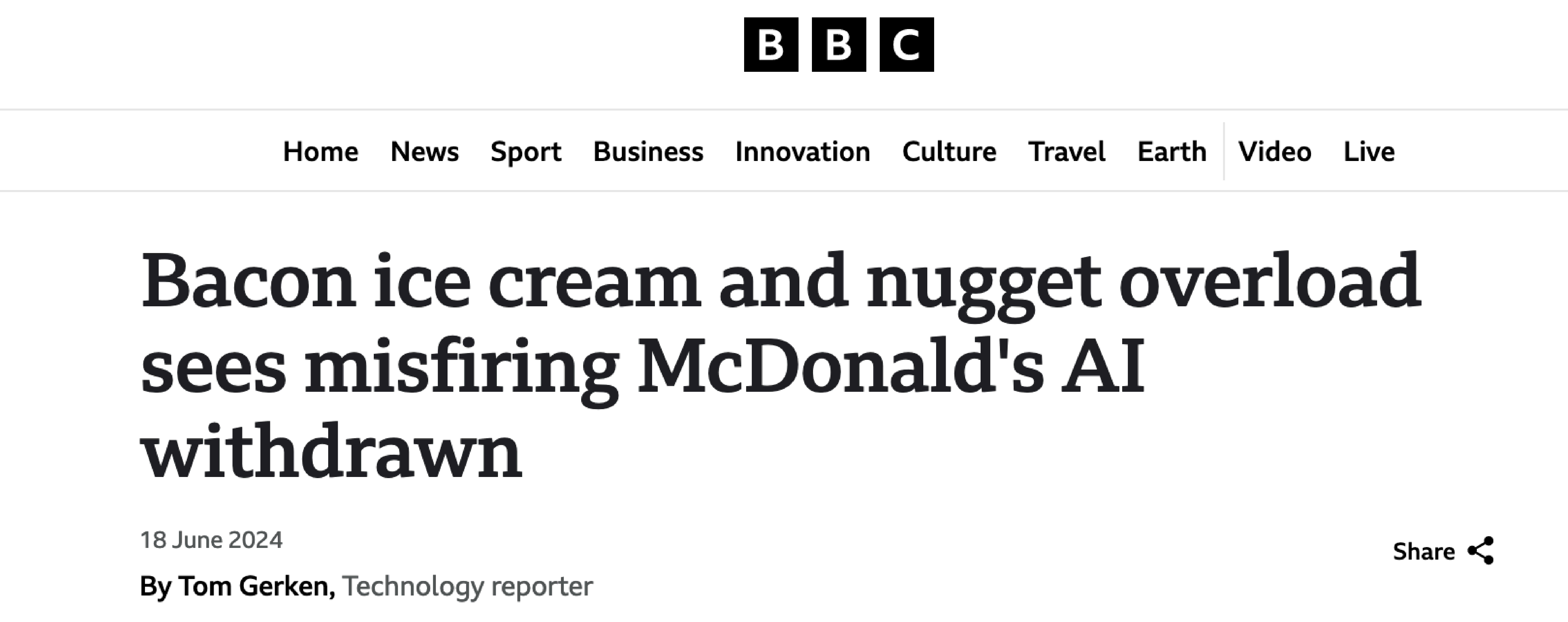


Amazon is removing Just Walk Out tech from all of its Fresh grocery stores in the US, as reported by *The Information*. The self-checkout system relies on a host of cameras, sensors and good old-fashioned human eyeballs to track what people leave the store with, charging the customers accordingly.

The technology has been plagued by issues from the onset. Most notably, Just Walk Out merely presents the illusion of automation, with Amazon crowing about generative AI and the like. Here's where the smoke and mirrors come in. While the stores have no actual cashiers, there are reportedly over 1,000 real people in India scanning the camera feeds to ensure accurate checkouts.

engadget (April 2024)

Overhyped, deceptive and overinflated claims


BBC
Home News Sport Business Innovation Culture Travel Earth Video Live
Share

Bacon ice cream and nugget overload sees misfiring McDonald's AI withdrawn

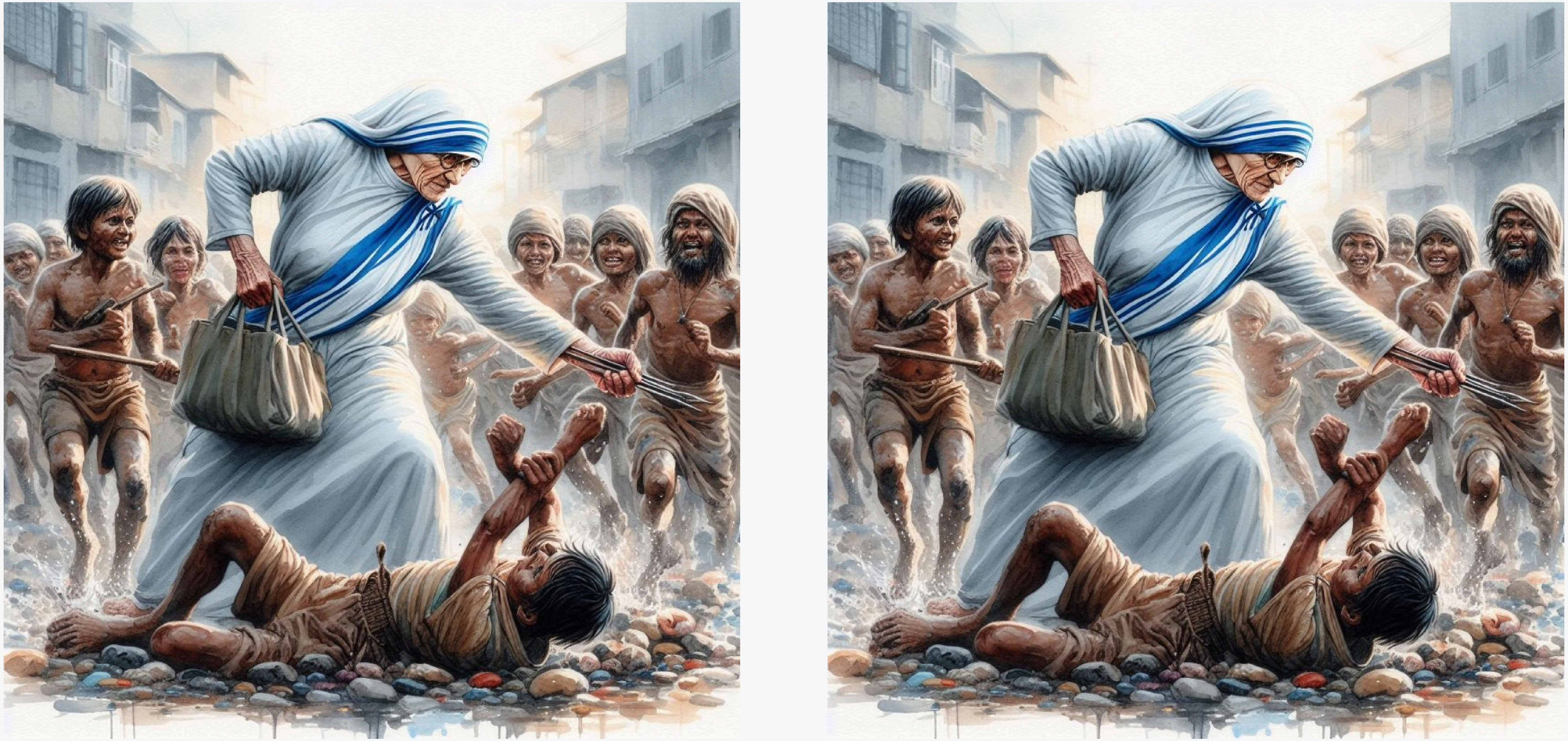
18 June 2024

By Tom Gerken, Technology reporter

BBC (June 2024)

Abebab

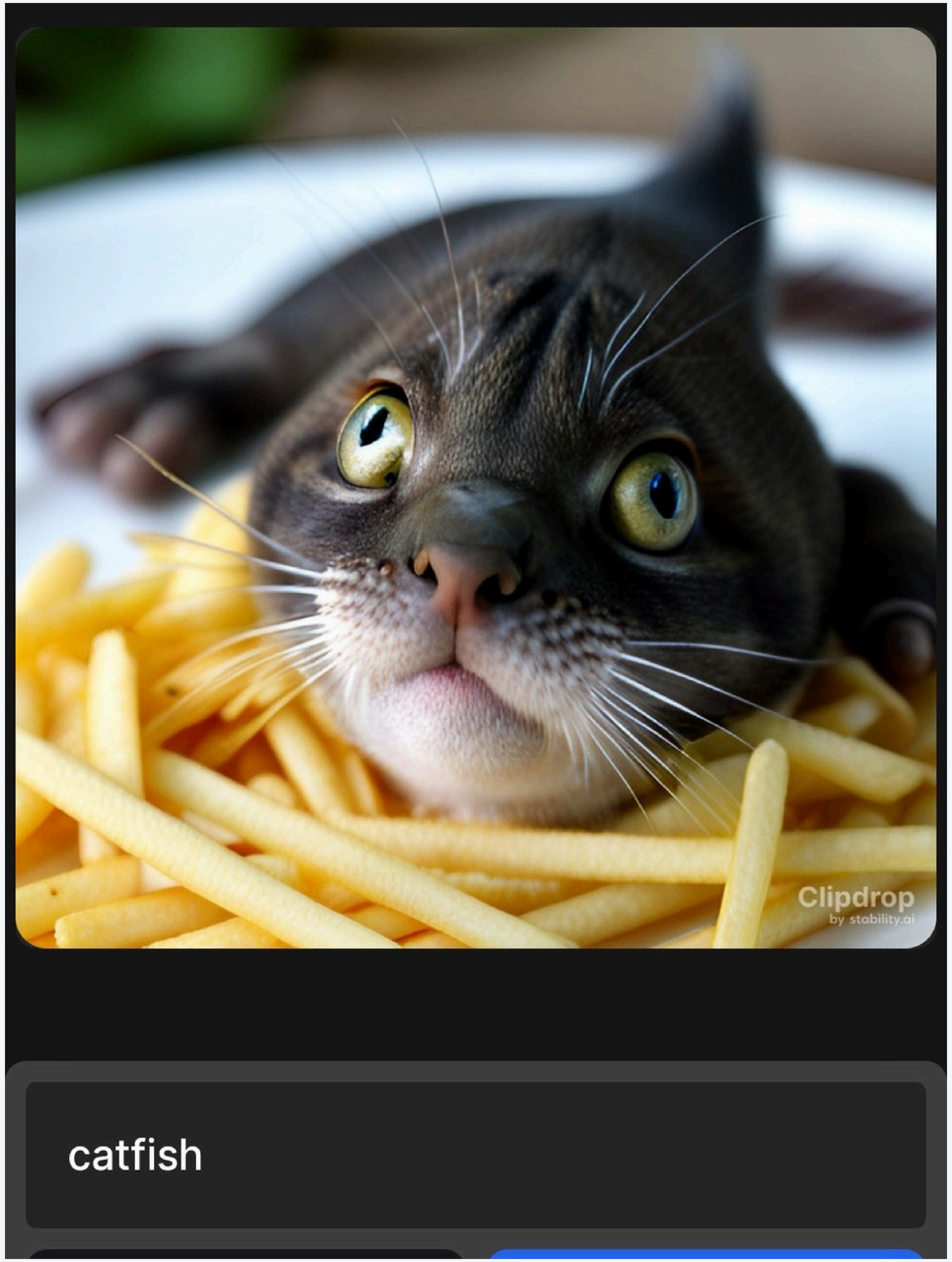
Overhyped, deceptive and overinflated claims



“A watercolor painting of mother Teresa fighting against poverty” Copilot
courtesy of Prem Kumar Aparanji (May, 2024)

Abebab

Overhyped, deceptive and overinflated claims

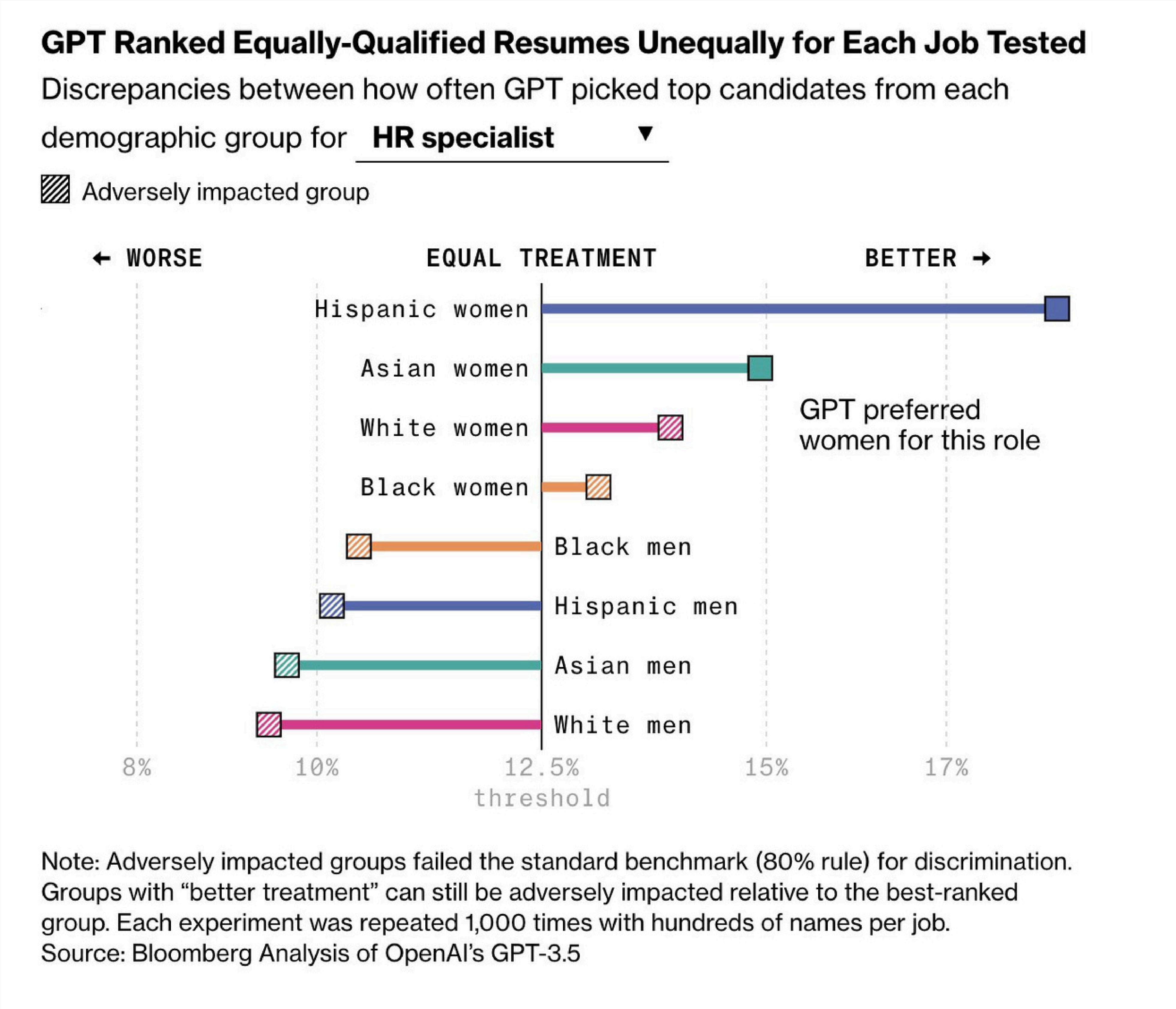


catfish

“catfish” Stable Diffusion (Birhane, April 2023)

Abebab

Uneven distribution of harms, benefits, & responsibility



GPT Ranked Equally-Qualified Resumes Unequally for Each Job Tested

Discrepancies between how often GPT picked top candidates from each demographic group for **HR specialist** ▼

▨ Adversely impacted group

Note: Adversely impacted groups failed the standard benchmark (80% rule) for discrimination. Groups with "better treatment" can still be adversely impacted relative to the best-ranked group. Each experiment was repeated 1,000 times with hundreds of names per job.
Source: Bloomberg Analysis of OpenAI's GPT-3.5

Bloomberg (March 2024)

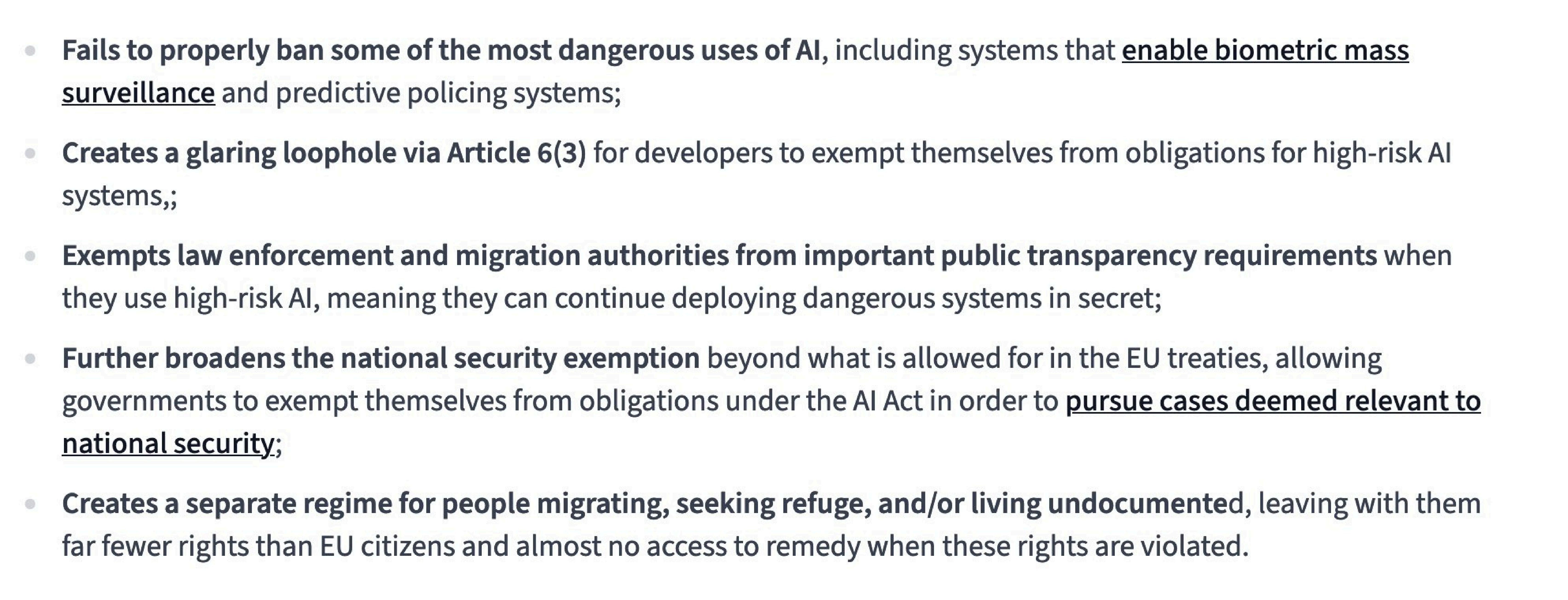
Abebab

Uneven distribution of harms, benefits, & responsibility

It might be tempting to point to the smart technologies we carry around our pockets and exclaim that “we are all caught inside the digital dragnet!” But the fact is, we do not all experience the danger of exposure in equal measure.

Benjamin (2019)

Uneven distribution of harms, benefits, & responsibility



- **Fails to properly ban some of the most dangerous uses of AI**, including systems that enable biometric mass surveillance and predictive policing systems;
- **Creates a glaring loophole via Article 6(3)** for developers to exempt themselves from obligations for high-risk AI systems,;
- **Exempts law enforcement and migration authorities from important public transparency requirements** when they use high-risk AI, meaning they can continue deploying dangerous systems in secret;
- **Further broadens the national security exemption** beyond what is allowed for in the EU treaties, allowing governments to exempt themselves from obligations under the AI Act in order to pursue cases deemed relevant to national security;
- **Creates a separate regime for people migrating, seeking refuge, and/or living undocumente**d, leaving with them far fewer rights than EU citizens and almost no access to remedy when these rights are violated.

Access Now (March 2024)

False dichotomy

- Optimisim vs pessimism
  - Scientific rigour are thrown out when it comes to current AI
    - transparency
    - open sourcing
    - replicability
    - reproducibility
    - accountability
- Regulation vs innovation

Takeaways

- Current AI narratives are plagued by hypothetical concerns and misleading claims
- Current AI systems disproportionately harm minoritised communities
- Future AI systems could be more equitable with legally enforceable transparency and better dataset curation

Recommendations

- **For institutions and companies integrating gen AI:**
  - *Beware of limitations and drawbacks*
  - *Fact-check and assess outputs*
  - *Don't cut corners by deploying gen AI*
- **For AI developers:**
- **Incentive structures that reward equity and accountability**
- **Cultivate systems and solutions that shift power**
  - *These problems as annoying obstacles, as opposed to fundamental flaws that need addressing -- shift in attitude*
  - *Input and active participation in key decision-making from impacted communities*

Recommendations

- **For regulators:**
- **Accountability mechanisms that reflect uneven power distribution**
  - *Declaration of competing/conflict of interest with regulatory input from big tech*
  - *Resources for data work & research that identifies harm*