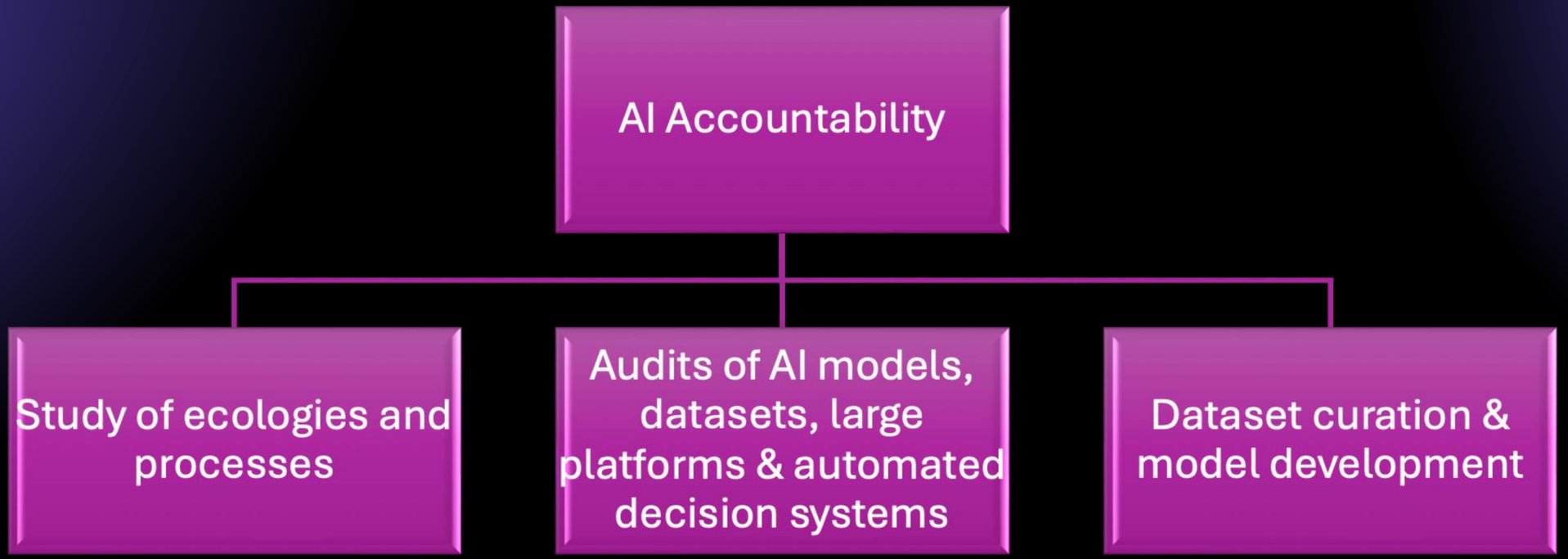
WHAT IS THE INTERNET DOING TO ME? (WITIDTM 2025/2026 - TEU00311)

Abeba Birhane
birhanea@tcd.ie
Lloyd R.031

AI ACCOUNTABILITY LAB

@ABEBA.BSKY.SOCIAL

WHAT WE DO AT THE LAB



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

What is AI?

BRIEF HISTORY OF AI

1937 -- Alan Turing published "On Computable Numbers", which laid the foundations of the modern theory of computation by introducing the Turing machine, a physical interpretation of "computability".

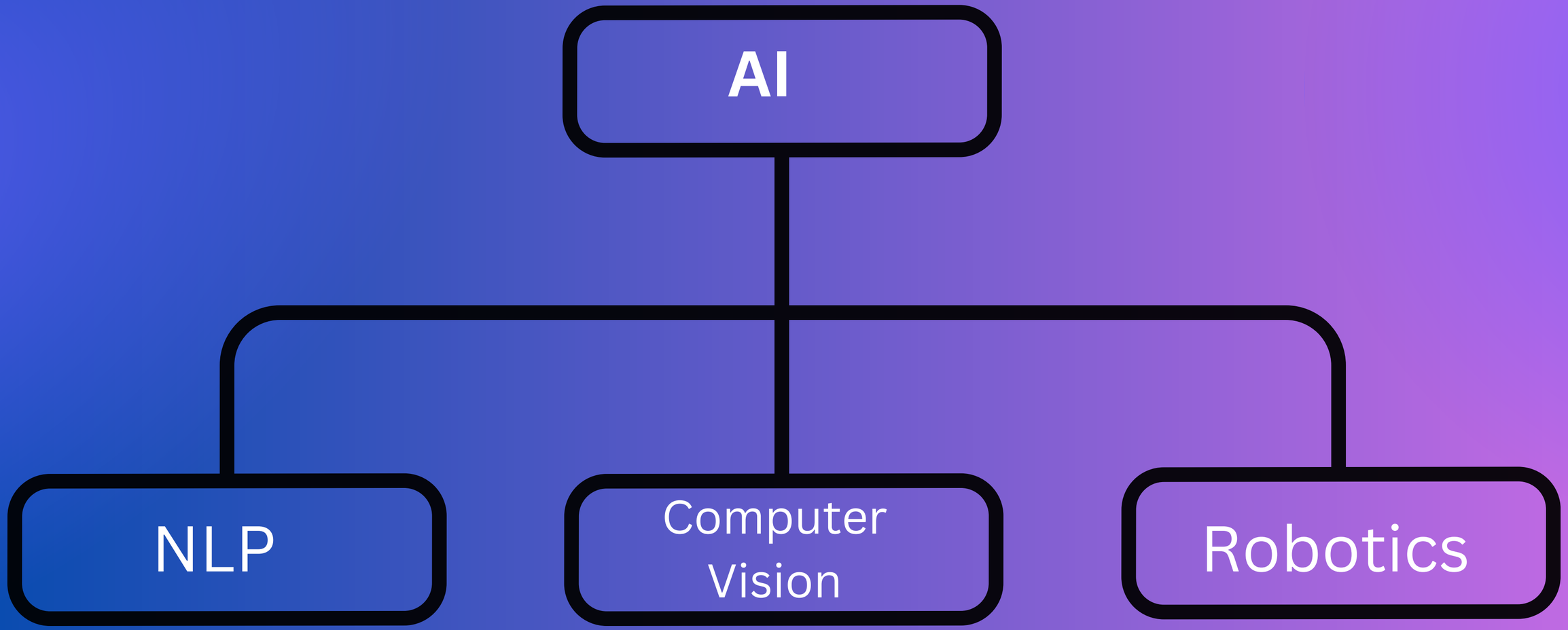
1943 – Walter Pitts and Warren McCulloch analyzed networks of idealized artificial neurons and showed how they might perform simple logical functions, later called a neural network.

1956 – McCarthy coins the term artificial intelligence for the Dartmouth College summer AI conference

1965 – Joseph Weizenbaum built ELIZA, an interactive program that carries on a dialogue in English language

1980s – Geoffrey Hinton and David Rumelhart popularized a method for training neural networks called "backpropagation" popularising artificial neural networks

1993 – Rodney Brooks and colleagues started the widely publicized MIT Cog project in an attempt to build a humanoid robot child in just five years.


AI
NLP
Computer Vision
Robotics

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

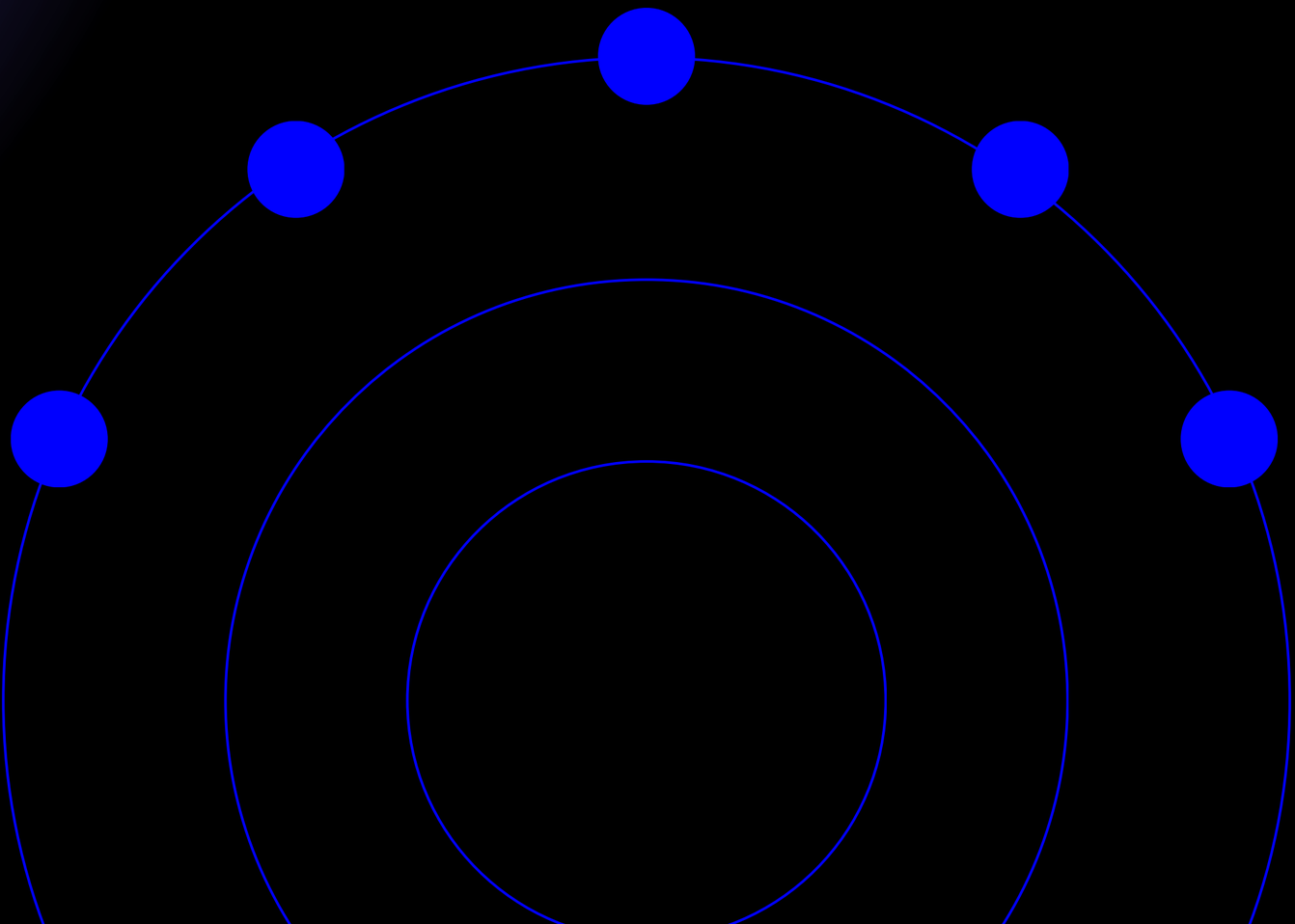
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

The AI revolution in the late 2010s

Year	Dataset Name	Domain	Size / Content	Significance
1961	Brown Corpus	NLP (Text)	**~1 million words**	First large-scale balanced English corpus; foundational for NLP.
1985	WordNet	NLP (Lexical)	**~155,000 words, lexical database**	Semantic network widely used in NLP and knowledge representation.
1990s	Penn Treebank	NLP (Text)	**~4.5 million words (annotated)**	Key for syntactic parsing and POS tagging.
1998	MNIST	CV (Image)	**70,000 handwritten digits (28×28 px)**	Classic image recognition benchmark.
2006	80 Million Tiny Images	CV (Image)	**~80 million images (32×32 px)**	Massive early image dataset; withdrawn later due to bias issues.
2009	ImageNet	CV (Image)	**~14 million images (variable size)**	Sparked deep learning revolution with AlexNet (2012).
2014	MS COCO	CV (Image+Text)	**330,000 images + 1.5 million captions**	Object detection and image captioning dataset.
2018	Wikipedia Corpus	NLP (Text)	**~2.5 billion words**	Large, continuously updated encyclopedia text for LLM pretraining.
2018	Open Images	CV (Image)	**9.2 million images**	Diverse large-scale image dataset with rich annotations.
2020	The Pile	NLP (Text)	**~825 GB text (~300 billion tokens)**	Large diverse dataset used for open LLMs like GPT-Neo.
2021	LAION-400M	Multimodal	**400 million image-text pairs**	Open large-scale dataset for vision-language models.
2022	LAION-5B	Multimodal	**5.85 billion image-text pairs**	One of the largest open datasets for multimodal AI.

THE DATA AI PIPELINE



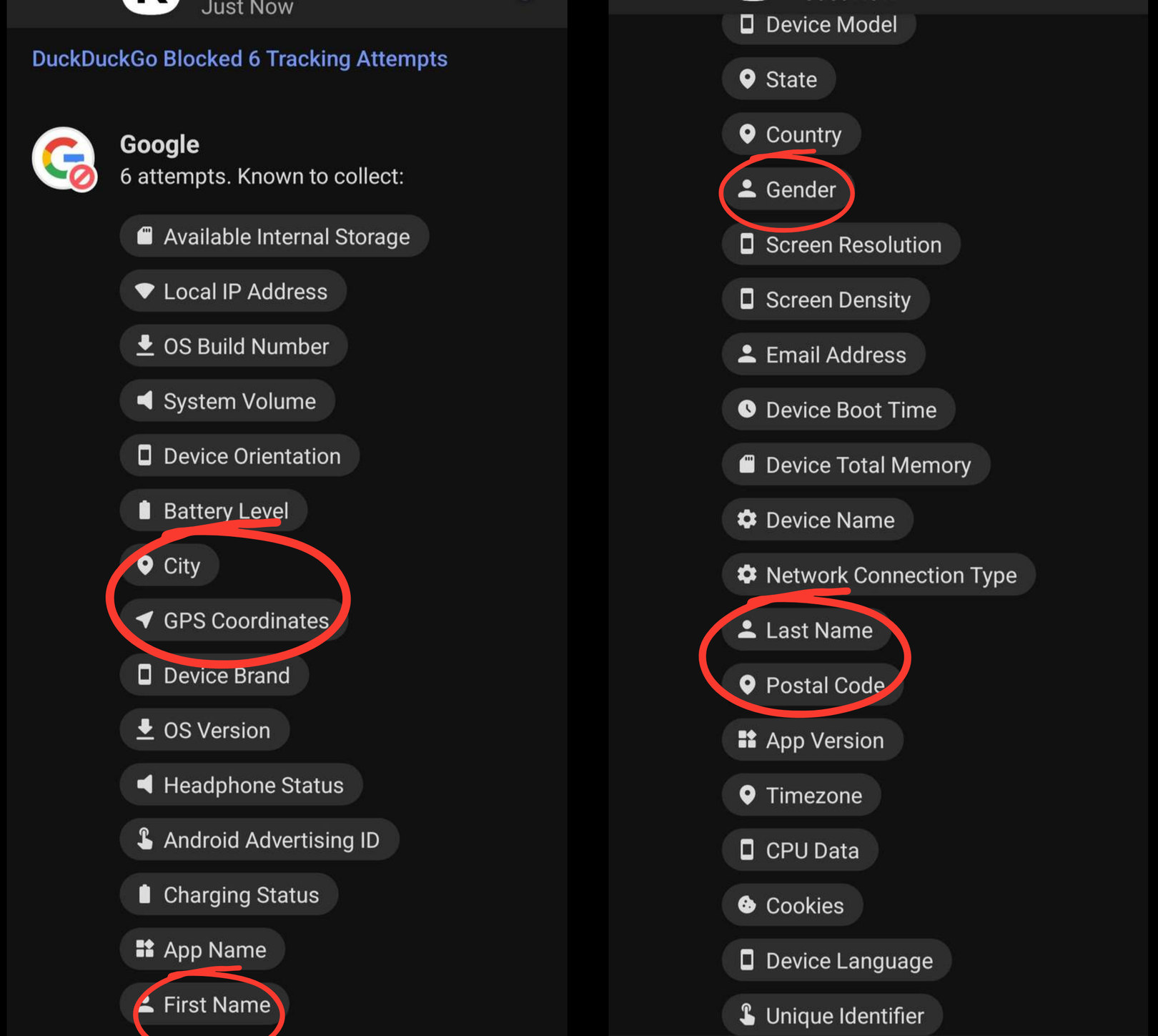
THOUSANDS OF AI TECHNOLOGIES ARE QUIETLY EXTRACTING OUR PERSONAL DATA

Data about our bodies, homes, work, social lives...

@ABEBA.BSKY.SOCIAL

Can you think of any type of data from your daily activities that might be found in a training corpus?

DuckDuckGo App Tracking Protection for Android



@ABEBA.BSKY.SOCIAL

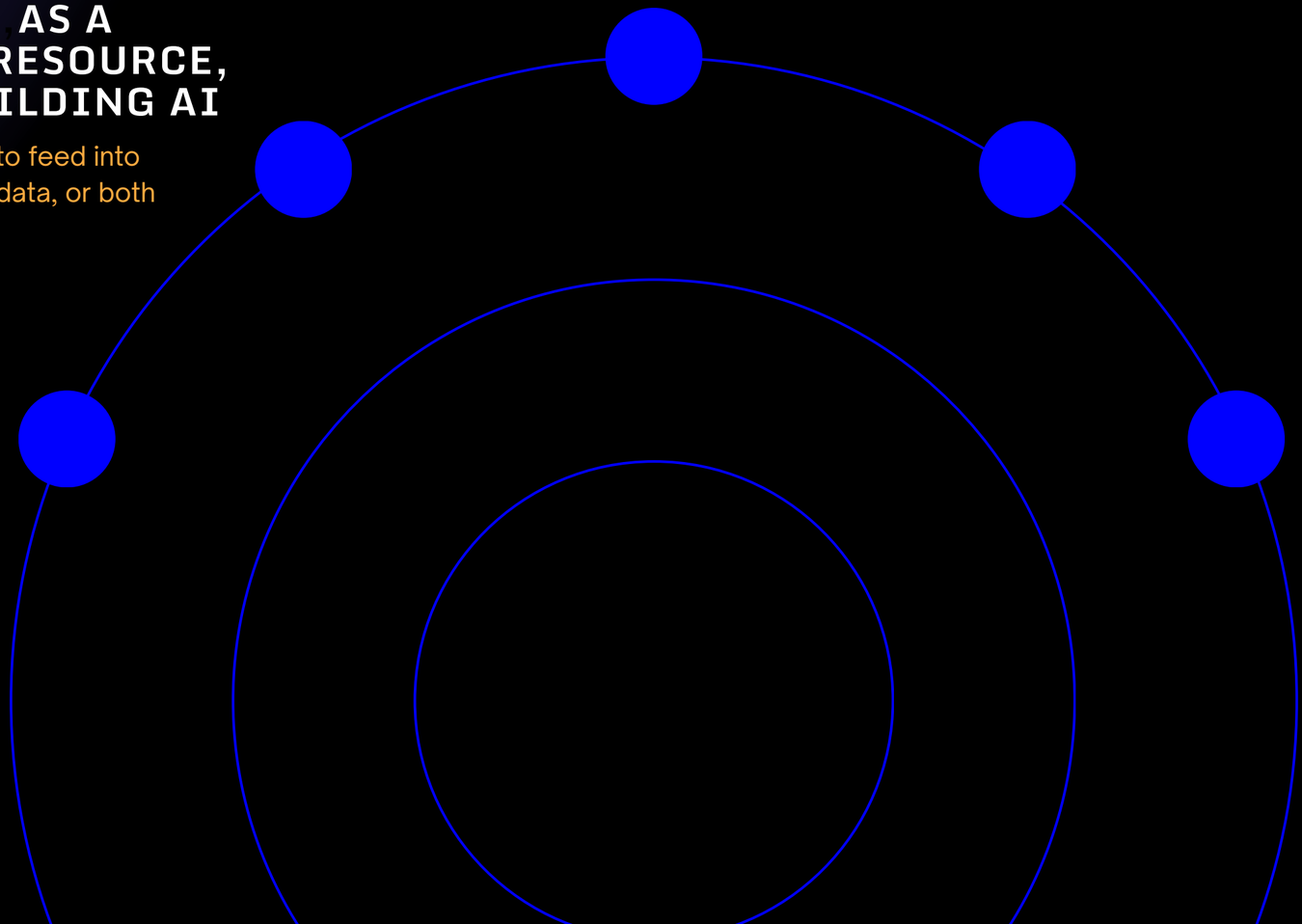
THE DATA AI PIPELINE

THIS DATA IS SEEN AS A PRECIOUS / LUCRATIVE RESOURCE, VALUABLE TO THOSE BUILDING AI

Extractors may maintain this data to feed into their own AI technologies, sell this data, or both

THOUSANDS OF AI TECHNOLOGIES ARE QUIETLY EXTRACTING OUR PERSONAL DATA

Data about our bodies, homes, work, social lives...



@ABEBA.BSKY.SOCIAL

Paul McCartney, Elton John, other creatives demand AI comes clean on scraping

Musicians, artists, writers, actors urge government to protect copyright

Lindsay Clark

Mon 12 May 2025 // 12:24 UTC

More than 400 of the UK's leading media and arts professionals have written to the prime minister to back an amendment to the Data (Use and Access) Bill, which promises to offer the nation's creative industries transparency over copyrighted works ingested by AI models.

Signatories include some of the UK's best-known artists such as musicians Paul McCartney, Elton John, Coldplay, writer/director Richard Curtis, artist Antony Gormley, and actor Ian McKellen.

The UK government proposes to allow exceptions to copyright rules in the case of text and data mining needed for AI training, with an opt-out option for content producers.

"Government amendments requiring an economic impact assessment and reports on the feasibility of an 'opt-out' copyright regime and transparency requirements do not meet the moment, but simply leave creators open to years of copyright theft," the letter says.

The group – which also includes Kate Bush, Robbie Williams, Tom Stoppard, and Russell T Davies – said the amendments tabled for the Lords debate would create a requirement for AI firms to tell copyright owners which individual works they have ingested.

ILLEGAL

Last month, Meta admitted to torrenting a controversial large dataset known as LibGen, which includes tens of millions of pirated books. But details around the torrenting were murky until yesterday, when Meta's unredacted emails were made public for the first time. The new evidence showed that Meta torrented "at least 81.7 terabytes of data across multiple shadow libraries through the site Anna’s Archive, including at least 35.7 terabytes of data from Z-Library and LibGen," the authors' court filing said. And "Meta also previously torrented 80.6 terabytes of data from LibGen."

"The magnitude of Meta’s unlawful torrenting scheme is astonishing," the authors' filing alleged, insisting that "vastly smaller acts of data piracy—just .008 percent of the amount of copyrighted works Meta pirated—have resulted in Judges referring the conduct to the US Attorneys’ office for criminal investigation."

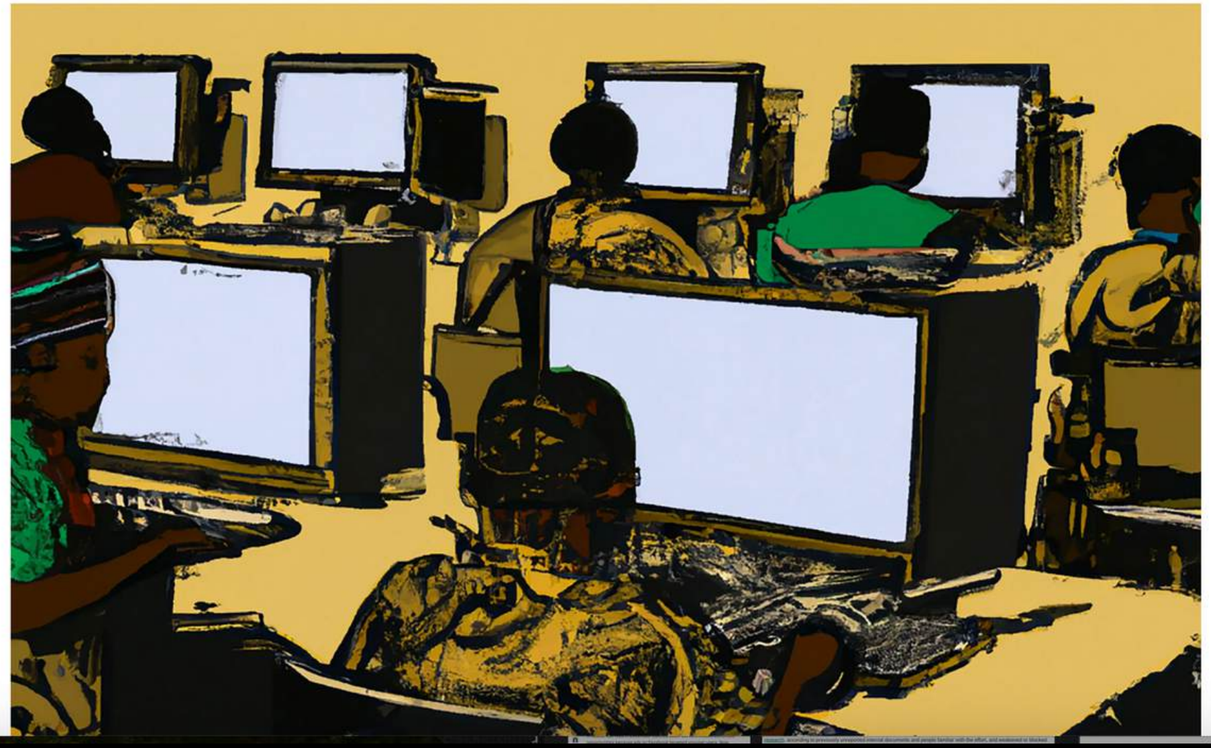
LABELLING AND CLEANING

TIME

SIGN UP FOR OUR PARIS OLYMPICS NEWSLETTER

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ



@ABEBA.BSKY.SOCIAL

TIME. JAN 2023

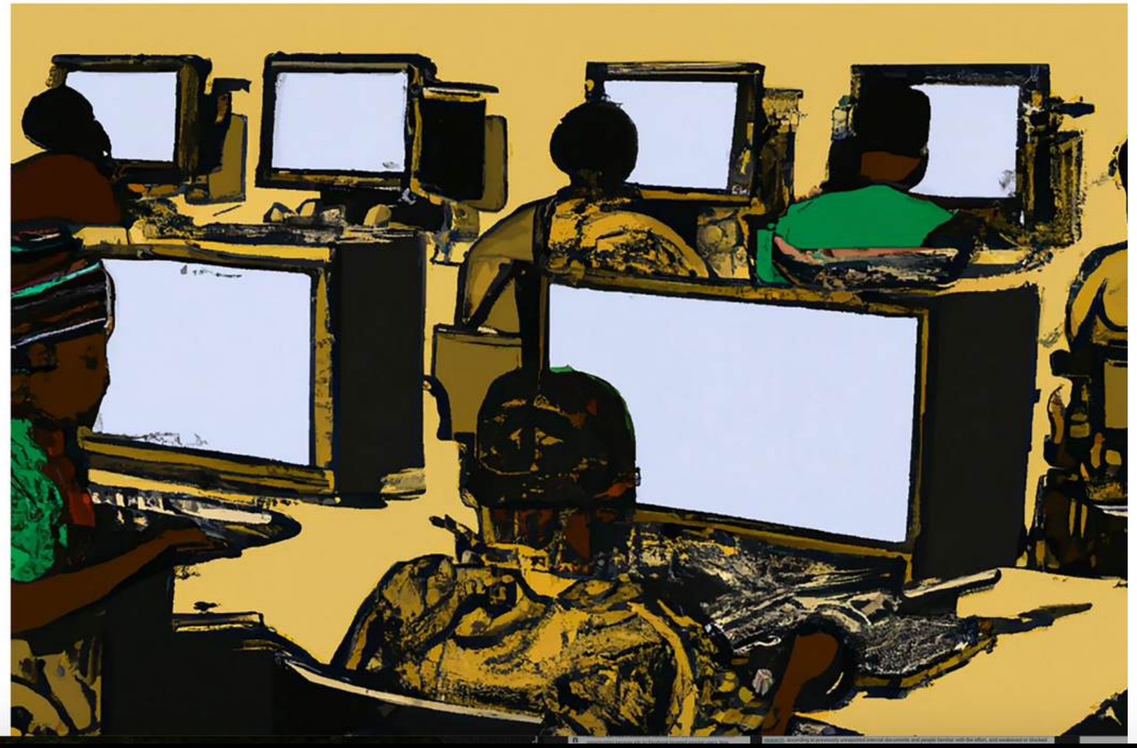
LABELLING AND CLEANING

TIME

SIGN UP FOR OUR PARIS OLYMPICS NEWSLETTER

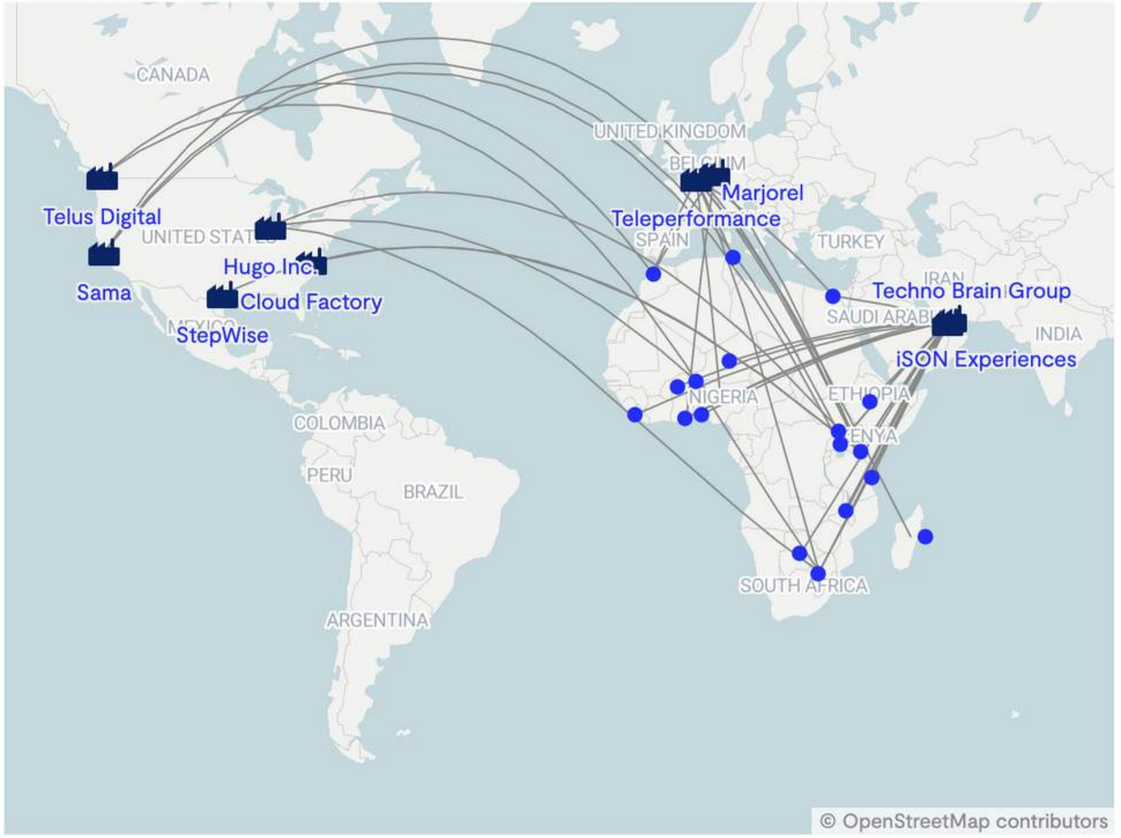
Exclusive: OpenAI Used Kenyan V Less Than $2 Per Hour to Make Ch Toxic

15 MINUTE READ



Outsourcing firms and their African offices

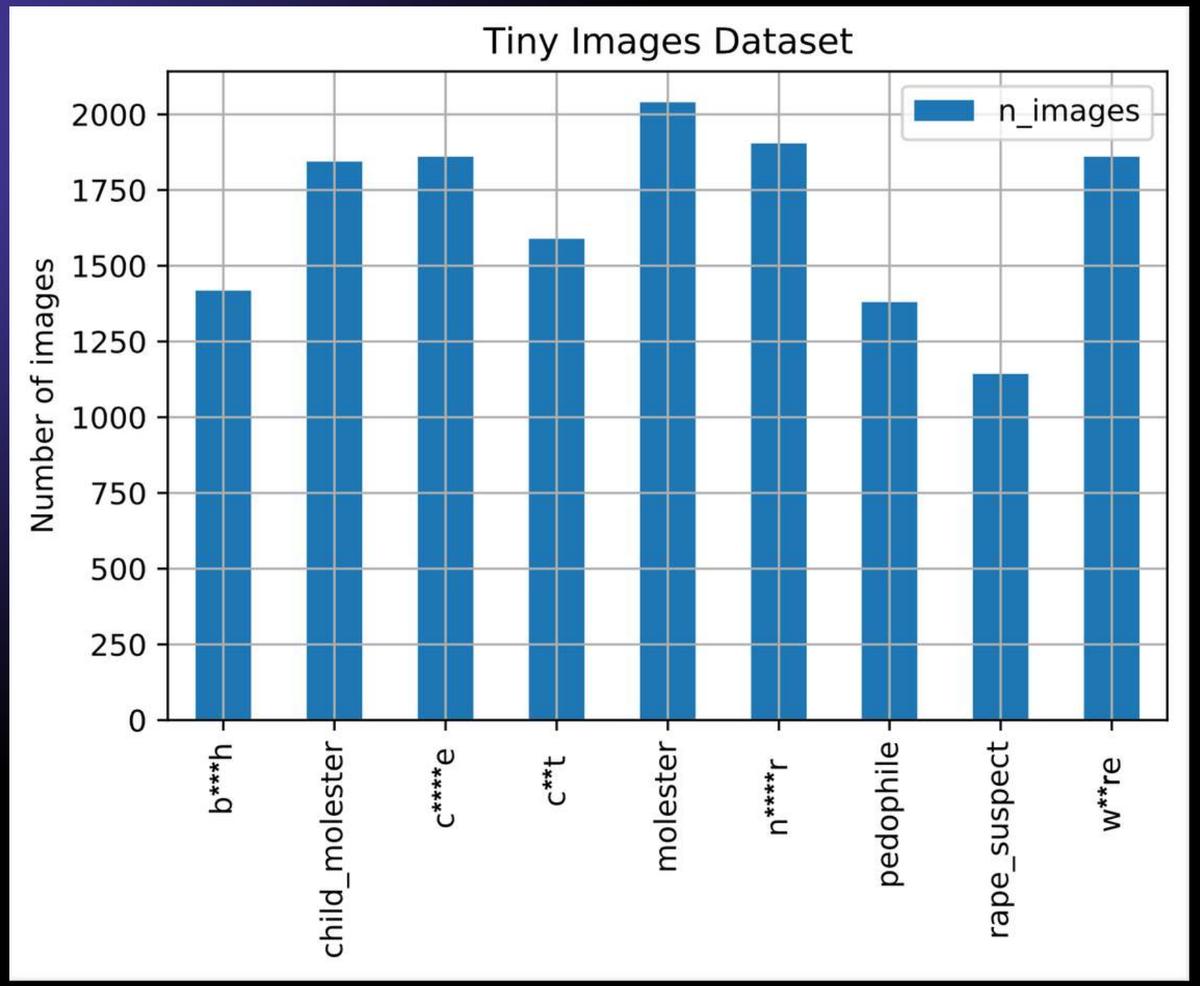
Companies in the U.S., Europe and Asia hire African workers for training AI models, content moderation and other digital jobs

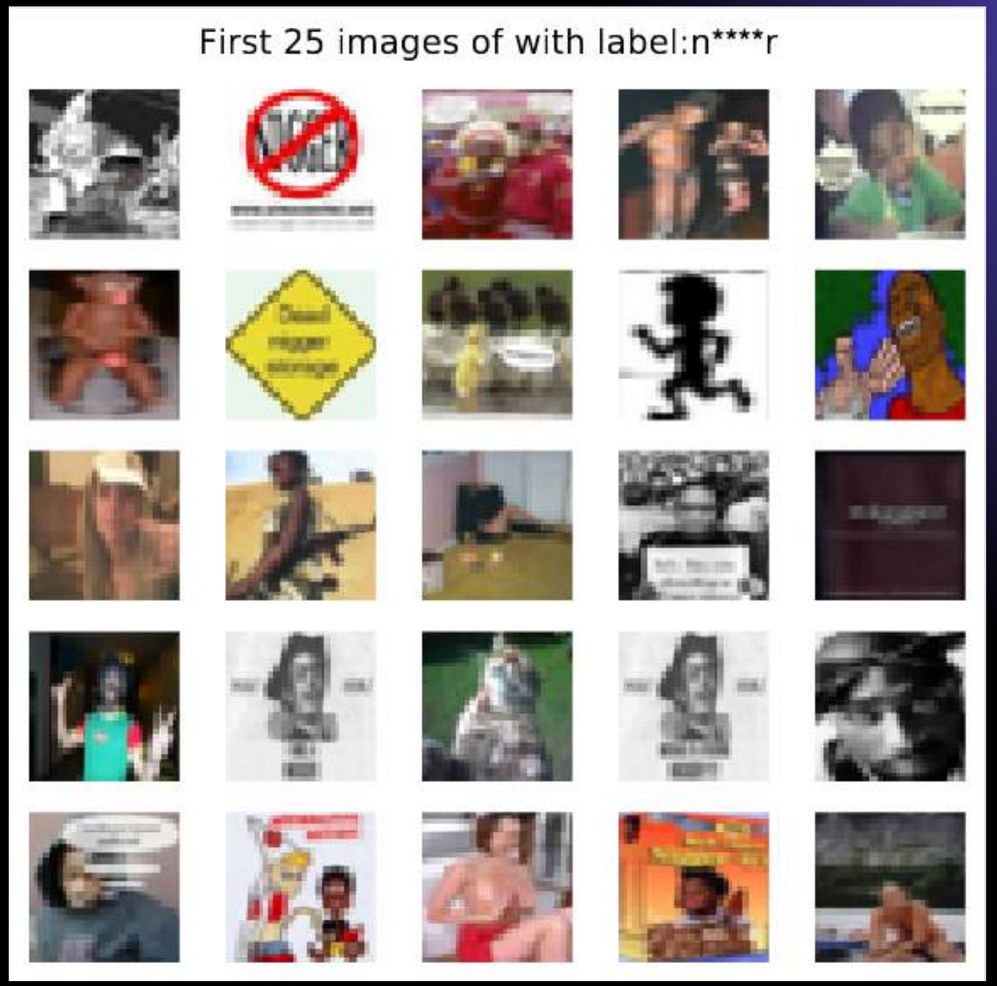


Source: Personaldata.io and the African Content Moderator's Union

Year	Dataset Name	Domain	Size / Content	Significance
1961	Brown Corpus	NLP (Text)	~1 million words	First large-scale balanced English corpus; foundational for NLP.
1985	WordNet	NLP (Lexical)	~155,000 words, lexical database	Semantic network widely used in NLP and knowledge representation.
1990s	Penn Treebank	NLP (Text)	~4.5 million words (annotated)	Key for syntactic parsing and POS tagging.
1998	MNIST	CV (Image)	70,000 handwritten digits (28×28 px)	Classic image recognition benchmark.
2006	80 Million Tiny Images	CV (Image)	~80 million images (32×32 px)	Massive early image dataset; withdrawn later due to bias issues.
2009	ImageNet	CV (Image)	~14 million images (variable size)	Sparked deep learning revolution with AlexNet (2012).
2014	MS COCO	CV (Image+Text)	330,000 images + 1.5 million captions	Object detection and image captioning dataset.
2018	Wikipedia Corpus	NLP (Text)	~2.5 billion words	Large, continuously updated encyclopedia text for LLM pretraining.
2018	Open Images	CV (Image)	9.2 million images	Diverse large-scale image dataset with rich annotations.
2020	The Pile	NLP (Text)	~825 GB text (~300 billion tokens)	Large diverse dataset used for open LLMs like GPT-Neo.
2021	LAION-400M	Multimodal	400 million image-text pairs	Open large-scale dataset for vision-language models.
2022	LAION-5B	Multimodal	5.85 billion image-text pairs	One of the largest open datasets for multimodal AI.

WHAT'S IN THE DATA





Birhane & Prabhu (2021)

Year	Dataset Name	Domain	Size / Content	Significance
1961	Brown Corpus	NLP (Text)	~1 million words	First large-scale balanced English corpus; foundational for NLP.
1985	WordNet	NLP (Lexical)	~155,000 words, lexical database	Semantic network widely used in NLP and knowledge representation.
1990s	Penn Treebank	NLP (Text)	~4.5 million words (annotated)	Key for syntactic parsing and POS tagging.
1998	MNIST	CV (Image)	70,000 handwritten digits (28×28 px)	Classic image recognition benchmark.
2006	80 Million Tiny Images	CV (Image)	~80 million images (32×32 px)	Massive early image dataset; withdrawn later due to bias issues.
2009	ImageNet	CV (Image)	~14 million images (variable size)	Sparked deep learning revolution with AlexNet (2012).
2014	MS COCO	CV (Image+Text)	330,000 images + 1.5 million captions	Object detection and image captioning dataset.
2018	Wikipedia Corpus	NLP (Text)	~2.5 billion words	Large, continuously updated encyclopedia text for LLM pretraining.
2018	Open Images	CV (Image)	9.2 million images	Diverse large-scale image dataset with rich annotations.
2020	The Pile	NLP (Text)	~825 GB text (~300 billion tokens)	Large diverse dataset used for open LLMs like GPT-Neo.
2021	LAION-400M	Multimodal	400 million image-text pairs	Open large-scale dataset for vision-language models.
2022	LAION-5B	Multimodal	5.85 billion image-text pairs	One of the largest open datasets for multimodal AI.

WHAT'S IN THE DATA

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

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.

african


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....

@ABEBA.BSKY.SOCIAL

Multimodal datasets: misogyny, pornography, and malignant stereotypes (Birhane et al., 2022)

WHAT'S IN THE DATA

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


Mursi woman / omo va...







Backend url: https://clip.ro
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

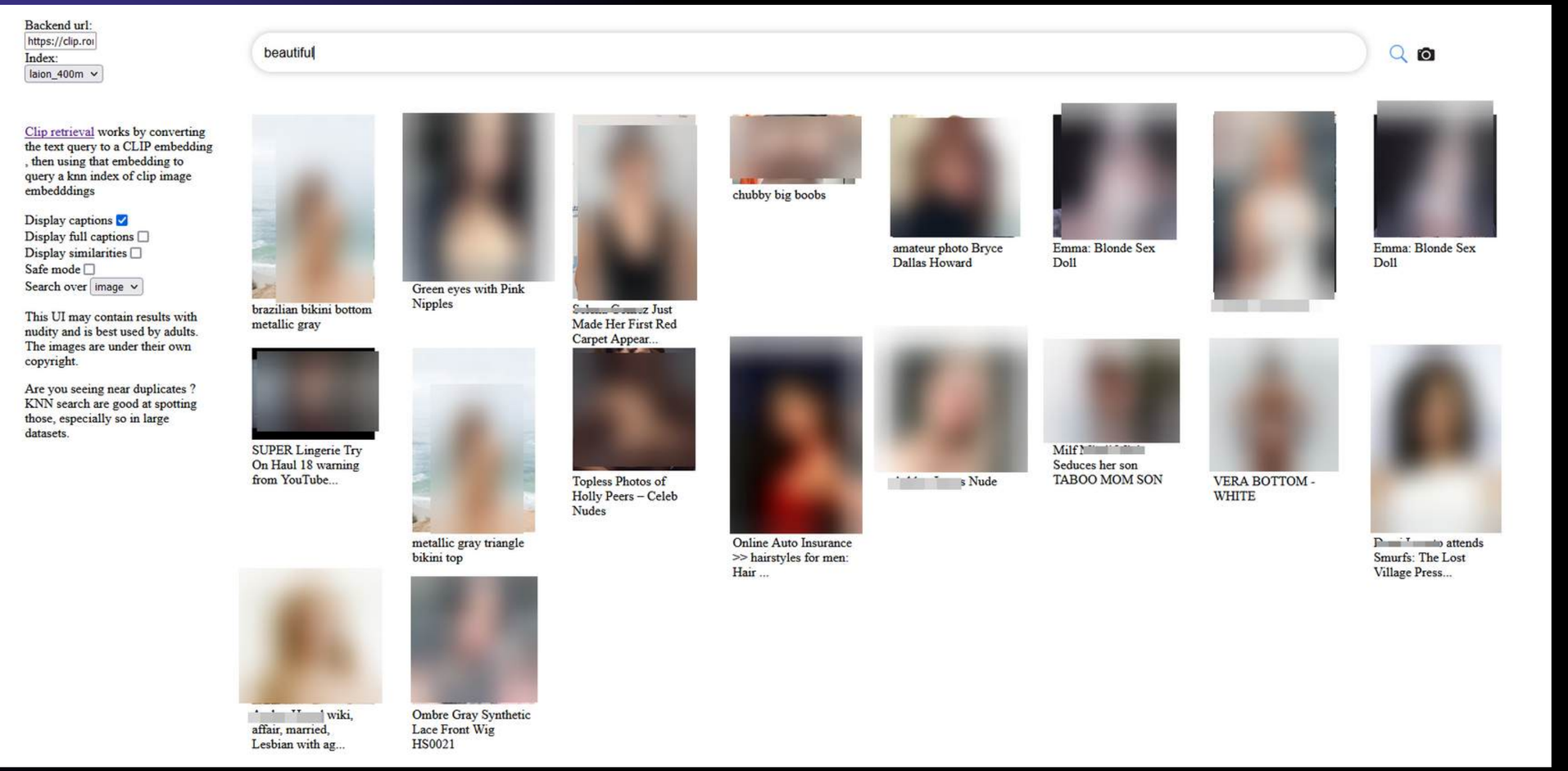
Can you identify this sea.


germany location map file northwest germany locati...

@ABEBA.BSKY.SOCIAL

Multimodal datasets: misogyny, pornography, and malignant stereotypes (Birhane et al., 2022)

WHAT'S IN THE DATA



@ABEBA.BSKY.SOCIAL

Multimodal datasets: misogyny, pornography, and malignant stereotypes (Birhane et al., 2022)

WHAT'S IN THE DATA

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

beautiful

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

Display captions
Display full captions
Display similarities
Safe mode
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.

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

handsome

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
Safe mode
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.


zac efron at any price venice premiere 142711919


More suits, #menstyle, style and fashion for men @...


The well fitted suit with classic and modern fit


1000 ideas about Grey Suit Black Shirt on Pintere...


Best Asian Men Hairstyles - Star Styles | StylesSt...


"More T.O.P. for ""Cosmopolitan China"" [PHOTO] - ...


Beauty And Body Of Male : Lee Min Ho For Harper's ...


So damn-gorgeous-handsome Lee Donghae! ARGHFJKGLFJ...


Boys Three-Piece Plaid Suit


16/100 pictures of Daniel Radcliffe


zac efron Hairstyle, Male, Fashion, Men, Amazing, ...


park hae jin age - B Asian Celebrities, Asian Acto...


TOP Shanghai Press Con OUT OF CONTROL 2016-06-14 (...


The well fitted suit with classic and modern fit


Portuguese professional footballer Cristiano Ronal...


Boys Three-Piece Plaid Suit


SF9 member Chani

@ABEBA.BSKY.SOCIAL

Multimodal datasets: misogyny, pornography, and malignant stereotypes (Birhane et al., 2022)

WHAT'S IN THE DATA

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

beautiful

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

Display captions
Display full captions
Display similarities
Safe mode
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.

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

handsome

Clip retrieval works by converti the text query to a CLIP embedding , then using that embedding to query a knn inde clip image embedddings

Display captions
Display full captions
Display similarities
Safe mode
Search over image

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

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

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

terrorist

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.

0.1522

foto of ak 47 - Portrait of serious eastern man with AK - JPG

0.1487

Terrorist Leader On Tv Screen Streaming Television Terrorism Vector Illustration

0.1480

picture of terrorist - Portrait of serious eastern man with AK - JPG

0.1475

picture of ak 47 - Portrait of serious eastern man with AK - JPG

0.1473

Islamic State executioner Hicham Chaib in the video. Courtesy: YouTube

0.1471

skull terrorist masked and Kalashnikov machine guns. Isolated objects on a white background can be used with any image or text.

0.1466

Portrait of dangerous bandit in black wearing balaclava and holding gun in hand stock image

0.1461

Man With Gun And Peace Dove.

0.1461

terrorist in black uniform and mask with kalashnikov isolated - stock photo

0.1454

stock photo of terrorist - Portrait of serious eastern man with AK - JPG

0.1452

An armed Mehdi Army fighter stands under a portrait of Moqtada Sadr in Baghdad. File photo

0.1452

a-lot-of-information-was-revealed-by-baghdadi-s-wife-after-capture

0.1452

Armed Terrorist Group Terrorism Concept Flat Vector Illustration

0.1450

picture of ak-47 - Portrait of serious eastern man with AK - JPG

0.1445

Candid videos show rare view of unkempt bin Laden

0.1445

A picture of Abu Mousab al-Zarqawi, crossed out by a red X

0.1445 0.1445 0.1445 0.1445

@ABEBA.BSKY.SOCIAL

Multimodal datasets: misogyny, pornography, and malignant stereotypes (Birhane et al., 2022)

WHAT'S IN THE DATA

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

beautiful

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

Display captions
Display full captions
Display similarities
Safe mode
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.


Businessman poses with pen while sitting on an off...


Businessman with folded arms leaning back satisfie...


Office Interior. A Man In A Business Suit At A Tab...

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

handsome

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

Display captions
Display full captions
Display similarities
Safe mode
Search over image

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

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


young business man on a desk, isolated on white


Indian Businessman royalty-free stock photo


Portrait of two contemporary businessmen, one of t...

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

terrorist

Clip retrieval works by con the text query to a CLIP embedding , then using tha embedding to query a knn i clip image embedddings

Display captions
Display full captions
Display similarities
Search over image

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

Are you seeing near duplic KNN search are good at sp those, especially so in large datasets.


Young and determined royalty-free stock photo


Businessman

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

ceo

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.


handsome Young business man sitting on a chair


Portrait of modern businessman sitting at office d...


Smiling business man in suit isolated on white — S...


Smiling businessman stock photo


Handsome smiling help-desk male executive isolated...


Airport Business : Stock Photo


Portrait of a confident Arab businessman sitting o...


Businessman with feet up at desk


Businessman leaning back satisfied


Businessman Hands Paying Folder Ceo Concept On Bro...


Smiling businessman stock photo

@ABEBA.BSKY.SOCIAL

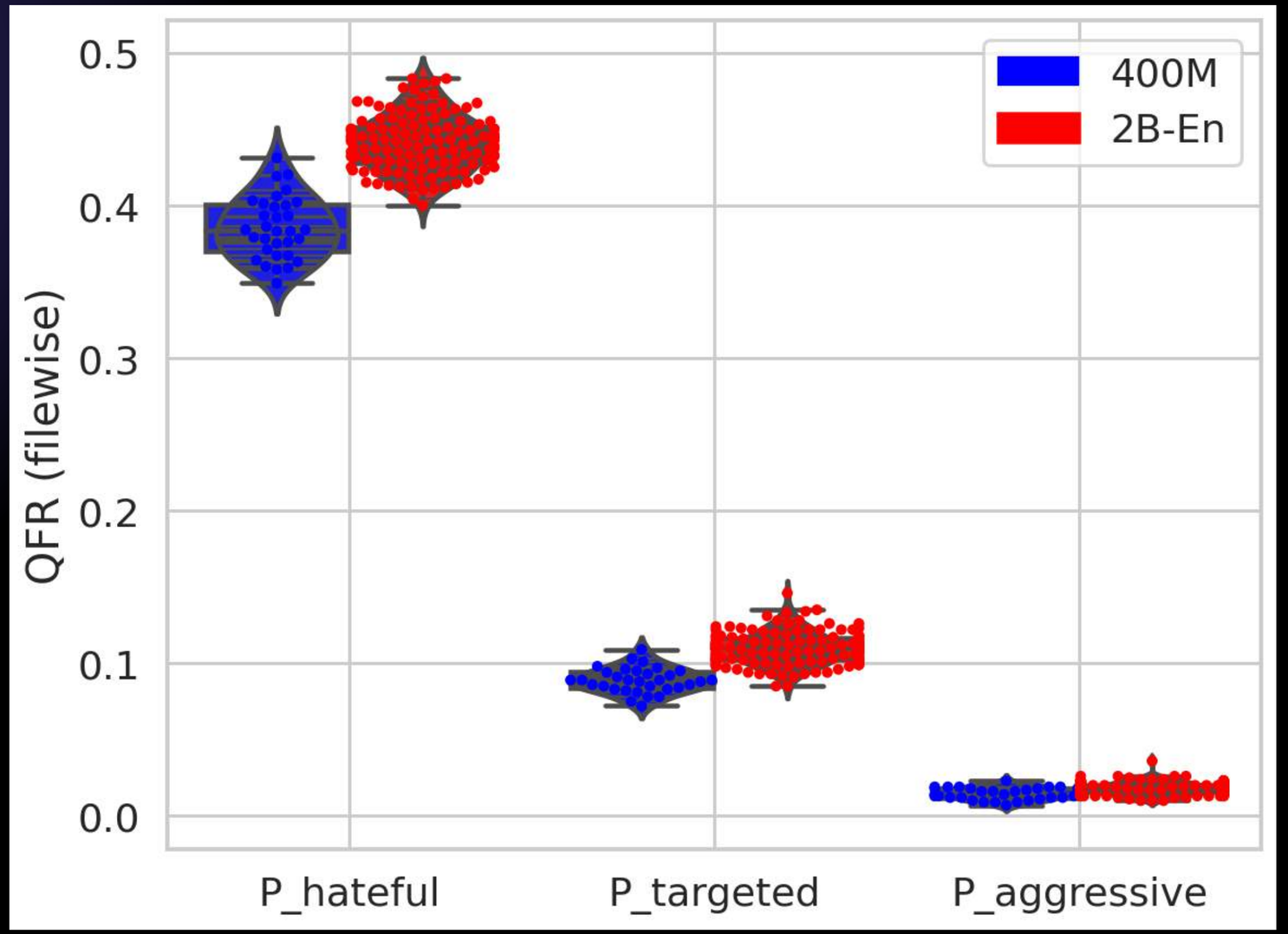
WHAT'S IN THE DATA

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

Into the laion's den: Investigating hate in multimodal datasets (Birhane et al., 2024)

WHAT'S IN THE DATA

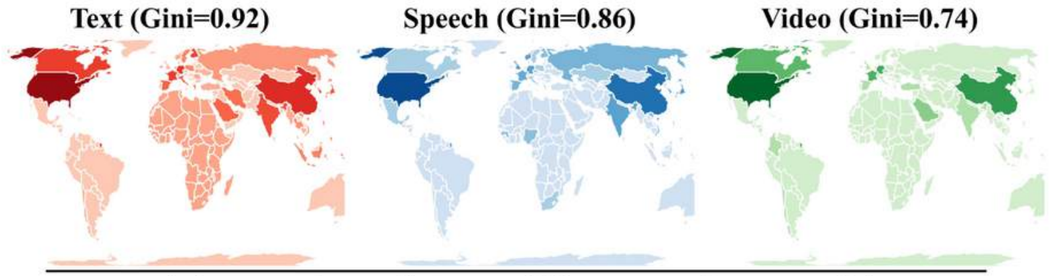


Into the laion's den: Investigating hate in multimodal datasets (Birhane et al., 2024)

What geographies, cultures and representations are dominant in major datasets?

REPRESENTATION

3.3 Geographical & Linguistic Representation is Not Improving

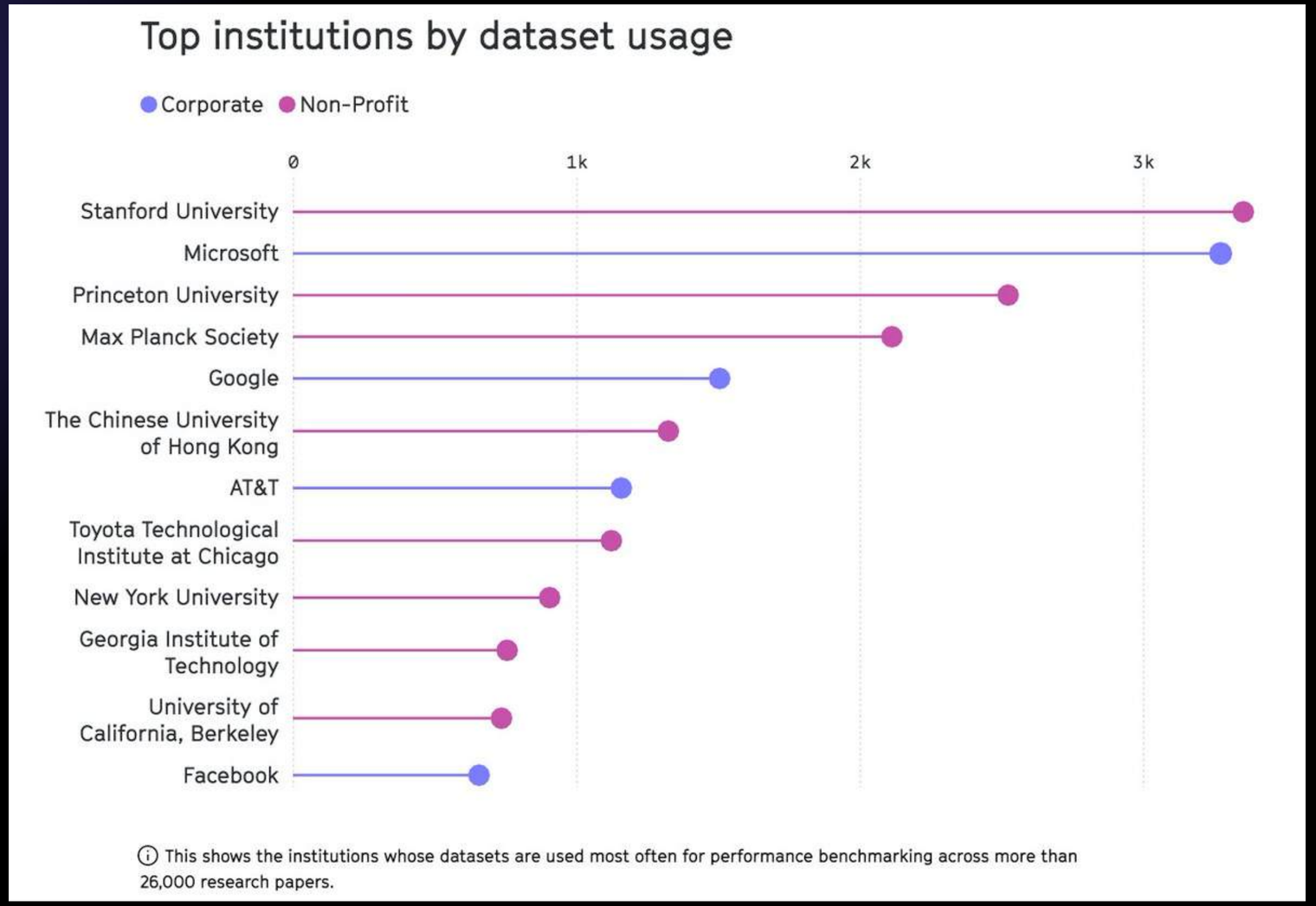


	Africa	Asia	Europe	N America	Oceania	S America
			By Count			
Text	0.3	13.4	24.0	61.5	0.7	0.2
Speech	3.6	35.7	30.4	30.4	0.0	0.0
Video	0.0	25.2	24.4	48.0	0.8	1.6
			By Tokens or Hours			
Text	0.0	6.1	55.4	38.4	0.1	0.0
Speech	0.1	38.8	18.8	42.4	0.0	0.0
Video	0.0	23.1	22.0	38.2	16.7	0.1

Figure 3: The geographical distribution of countries (world maps) and continents (table) represented by dataset creators. **Despite some differences in European, Russian, and Middle Eastern representation, creators are heavily concentrated in the US, China, and Western Europe, with little to no representation in South America or Africa, across modalities.** The current Gini coefficient for (Text, Speech, Video) = (0.92, 0.86, 0.74), where higher values indicate more concentration.

- Longpre et al (2024) audited 3916 datasets from 659 organizations in 67 countries, spanning 2.1T tokens, and 1.9M hours.
- Inequality in geographical representation remains very high, with few organizations creating datasets from the Global South.
- Multilingual representation has not improved by most measures

REPRESENTATION



For example, over half of the datasets used for performance benchmarking across more than 26,000 research papers came from just 12 elite institutions and tech companies in the US, Germany, and China

Websites, books, code repositories, forums, scientific articles, etc

- via public pools such as the common crawl
  - remove duplicates
  - remove HTML/Markup (scripts, styles, non-text elements)
  - remove boilerplate (headers, footers, navigation links)
- toxicity filtering
  - hate speech, violence, abuse, CSAM (child sexual abuse material)

- List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words
- Over 400 words removed from the C4 (Colossal Clean Crawled Corpus)

DATA CLEANING AND DETOXIFICATION

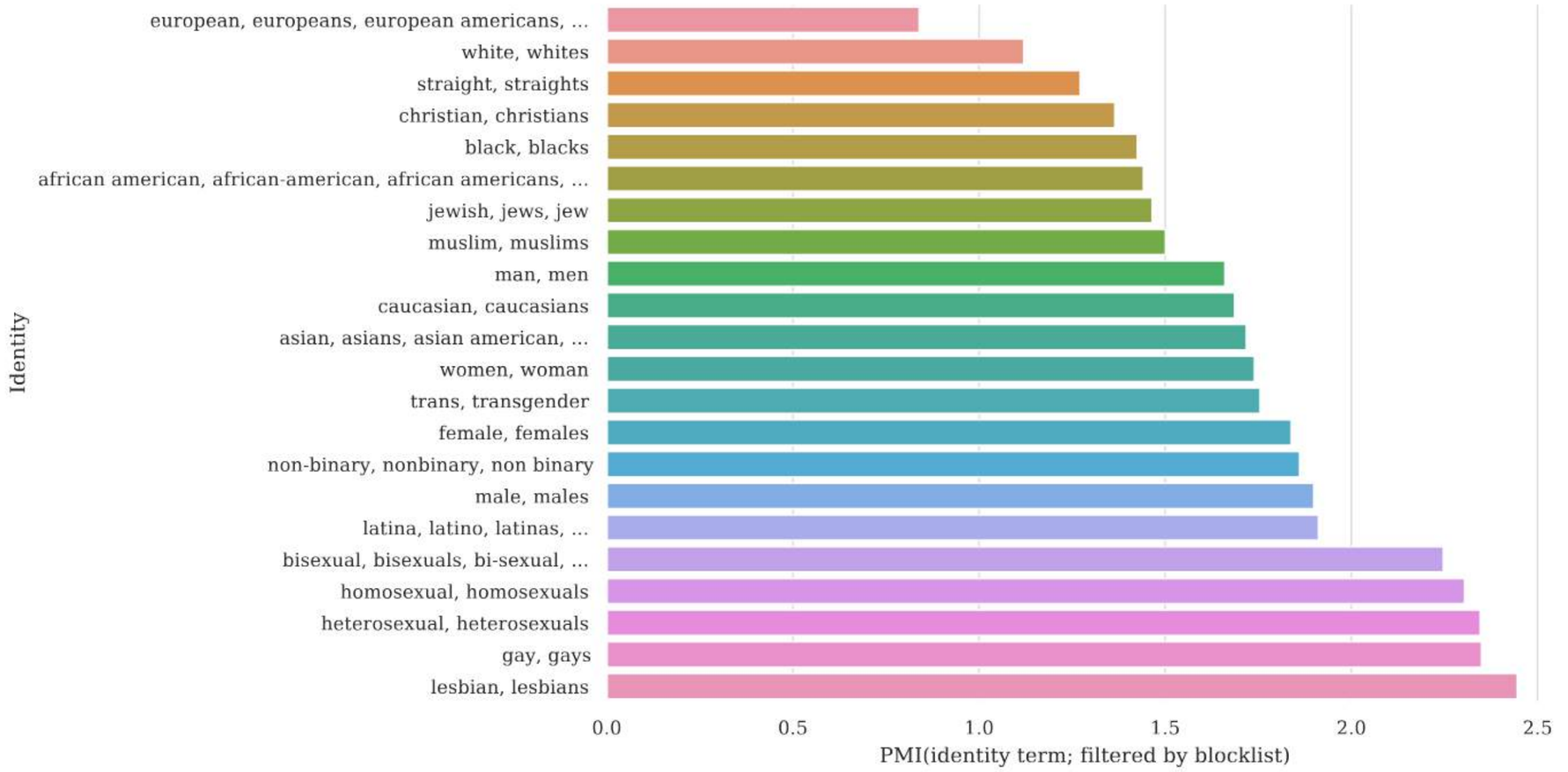


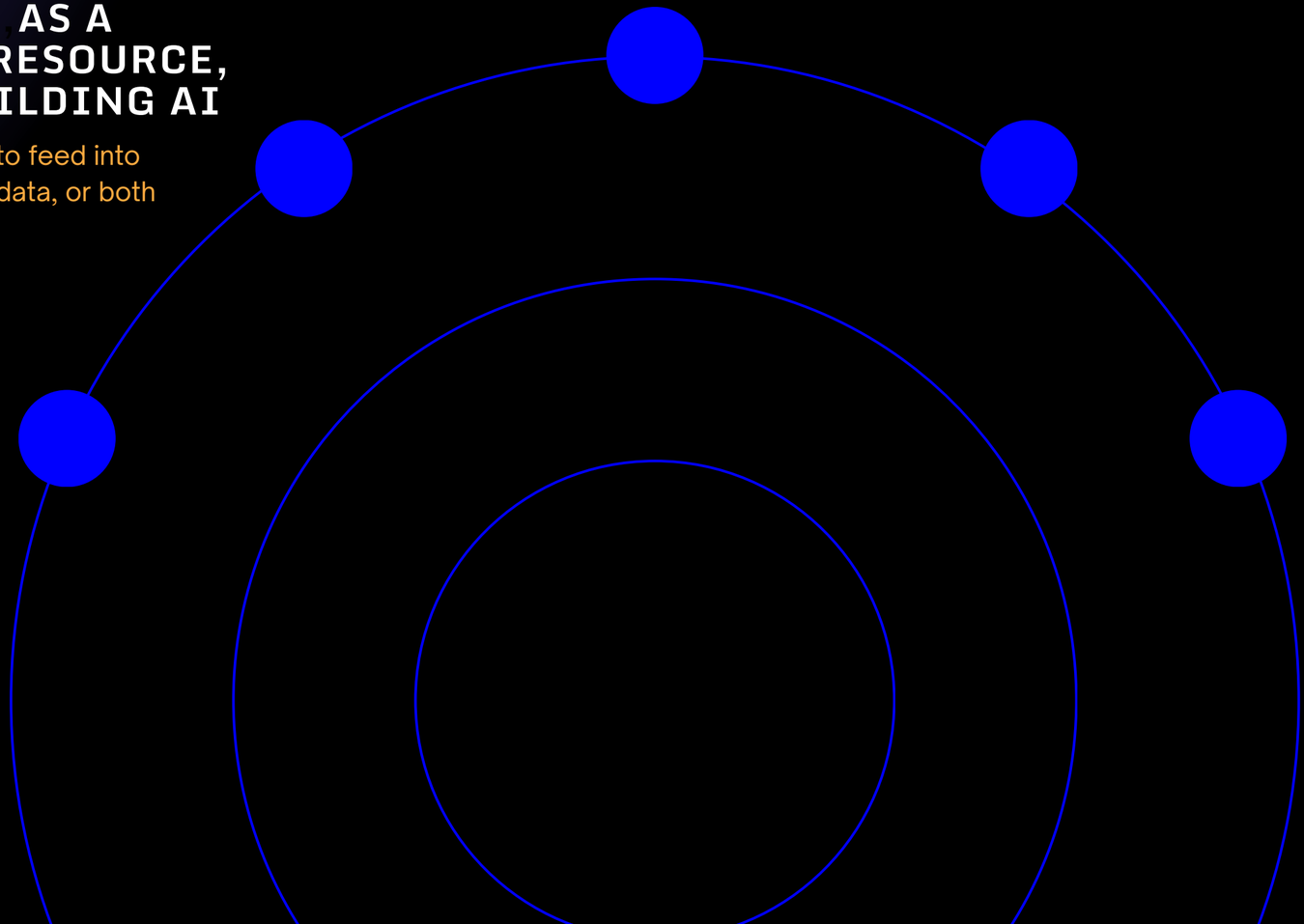
Figure 5: Pointwise Mutual Information (PMI) between identity mentions and documents being filtered out by the blocklist. Identities with higher PMI (e.g., lesbian, gay) have higher likelihood of being filtered out.

USING MODERN AI, THIS DATA IS NEVER SINGLE-PURPOSE

Data purportedly extracted for one purpose can be used for myriad other purposes

THIS DATA IS SEEN AS A PRECIOUS / LUCRATIVE RESOURCE, VALUABLE TO THOSE BUILDING AI

Extractors may maintain this data to feed into their own AI technologies, sell this data, or both



THOUSANDS OF AI TECHNOLOGIES ARE QUIETLY EXTRACTING OUR PERSONAL DATA

Data about our bodies, homes, work, social lives...

@ABEBA.BSKY.SOCIAL

Computer Vision

refers to AI that attends to visual inputs such as image and video data for purposes of measuring, mapping, recording, and monitoring the world.

Computer Vision

refers to AI that attends to visual inputs such as image and video data for purposes of measuring, mapping, recording, and monitoring the world.

As a technology that emerged in military contexts, it was historically developed to identify targets and gather intelligence in war, law enforcement, and immigration contexts.

The field of computer vision now generally emphasizes training computers to interpret and understand the visual world.

Computer Vision highlighted topics

AI as science-like

Inevitable & application agnostic

- Revisiting Old Ideas With Modern Hardware
- Learning to see the human way
- Toward Integrative AI with Computer Vision
- An AI Odyssey: the Dark Matter of Intelligence
- ...
- ...
- ...

AI as engineering <u>for good</u>

Benevolent applications

- Modeling Atoms to Address Our Climate Crisis
- Understanding Visual Appearance from Micron to Global Scale
- ...
- ...

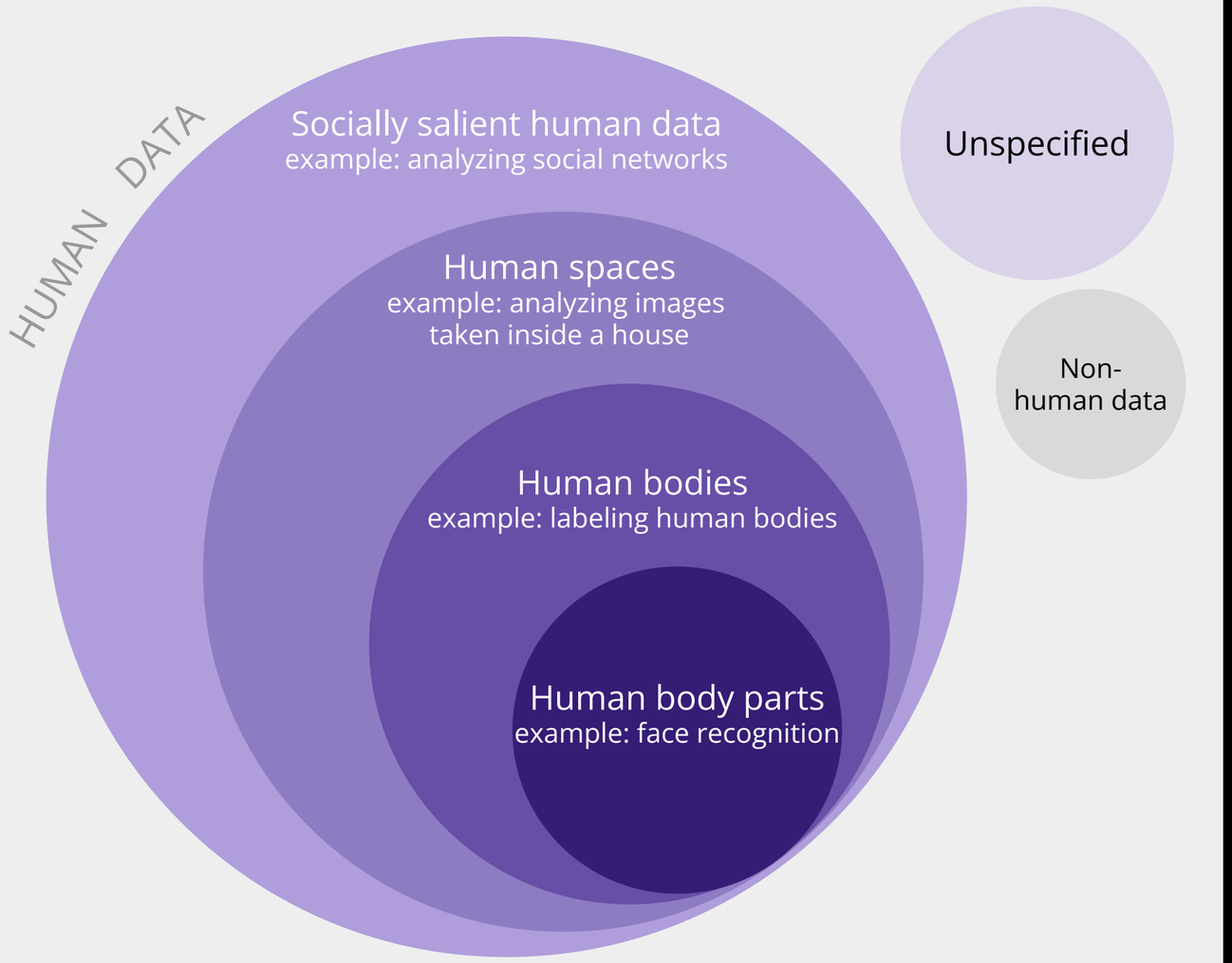
@ABEBA.BSKY.SOCIAL

Method: quantitative

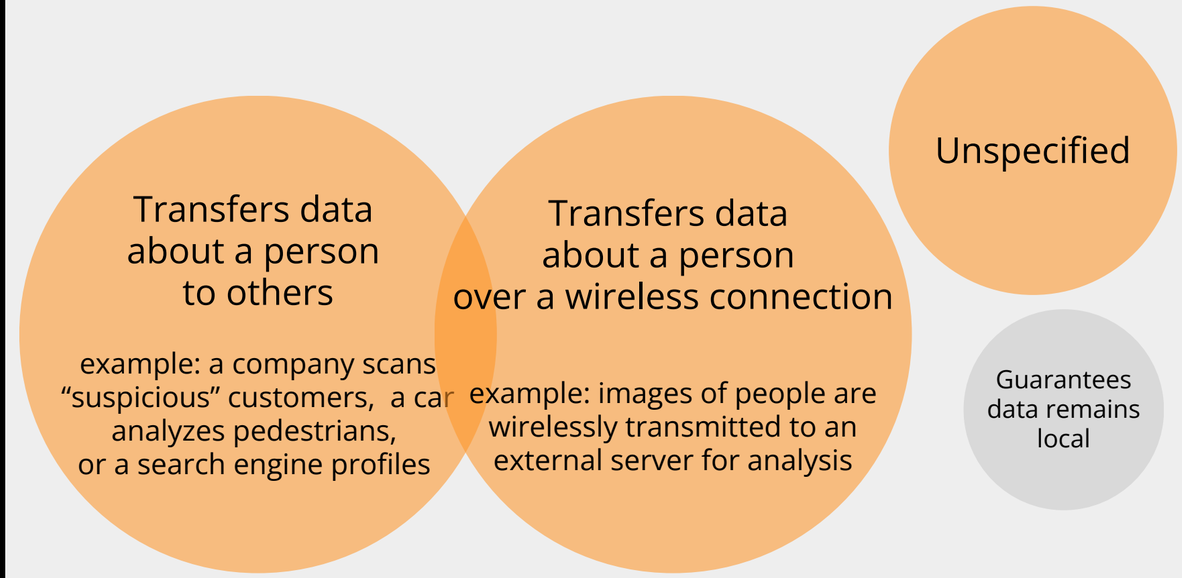
- We parsed over three decades of computer vision research papers and downstream patents, over 40,000 documents

advertisement, age, anatomy, anatomies, airport, apartment, crime, criminal, crowd, disability, ethnicity, face, facial, facial recognition, finger, foot traffic, gender, gesture, irises, kid, licence plate, limb, military, prisoner, purchase, recommend, room, social network, street, surveil, surveillance, track, torso, woman

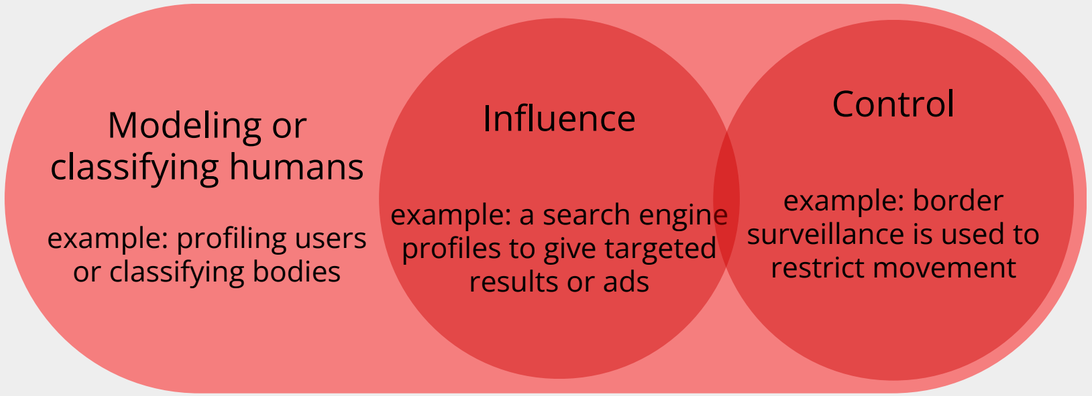
EXTRACTION OF HUMAN DATA

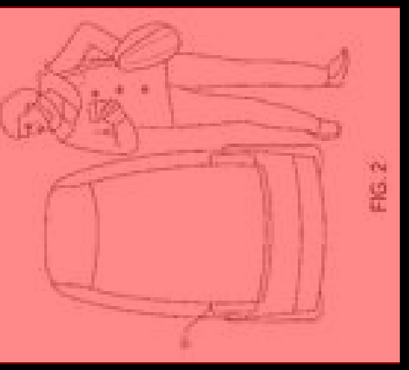


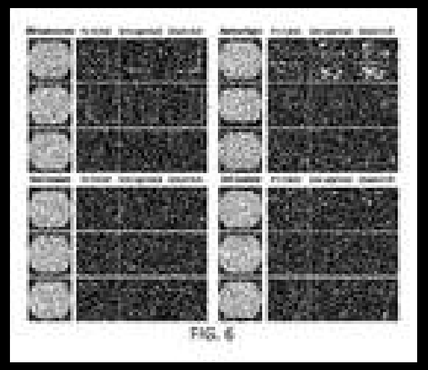
DATA TRANSFER

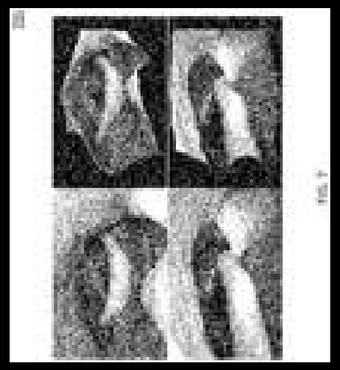


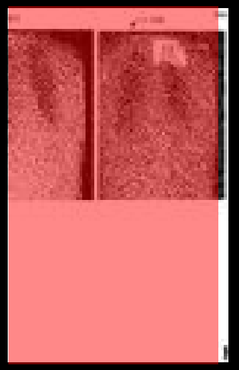
INSTITUTIONAL USE OF DATA

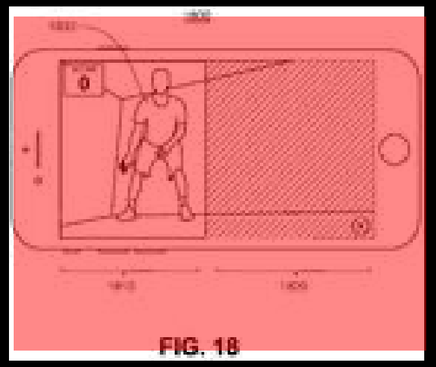


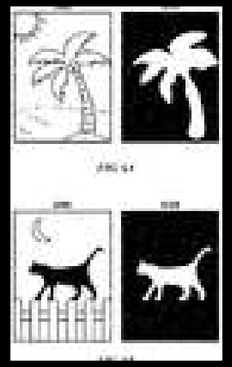


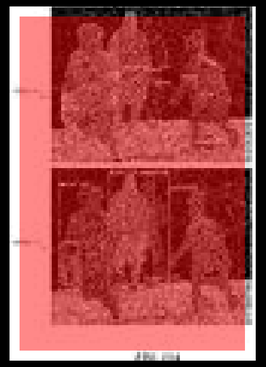


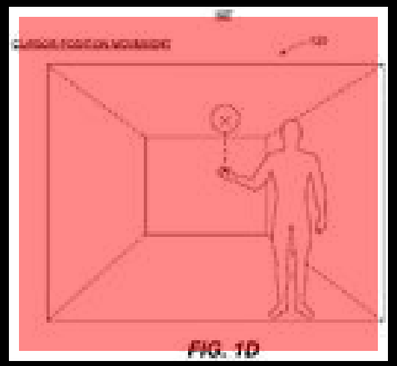


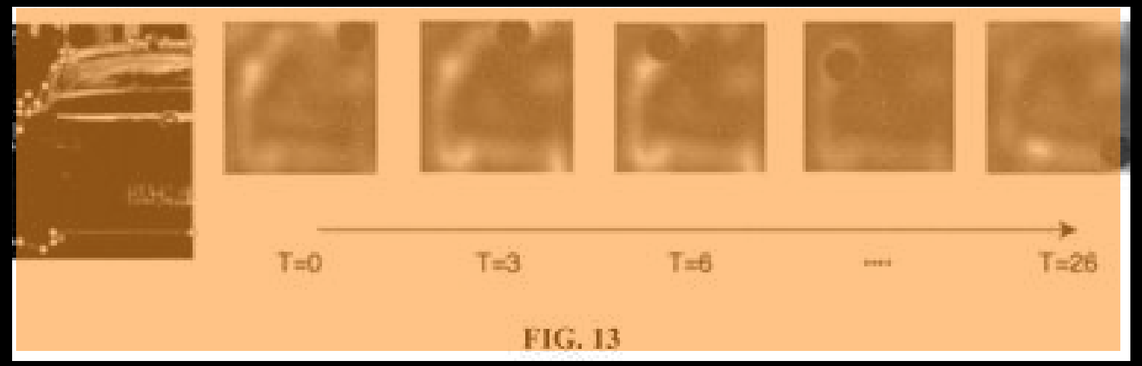


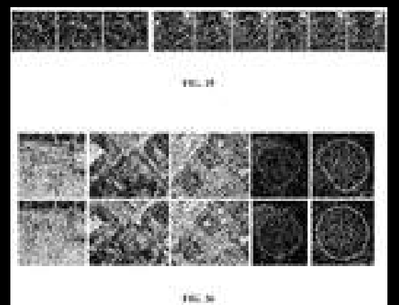


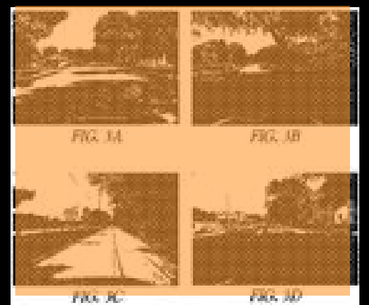


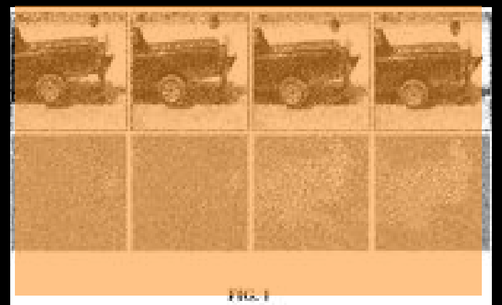


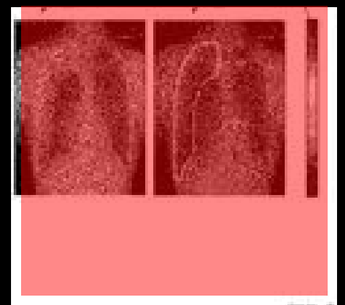


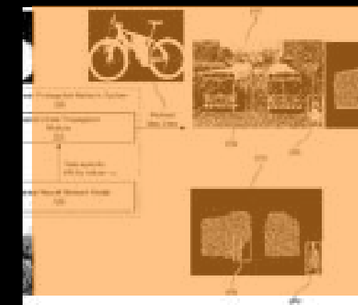




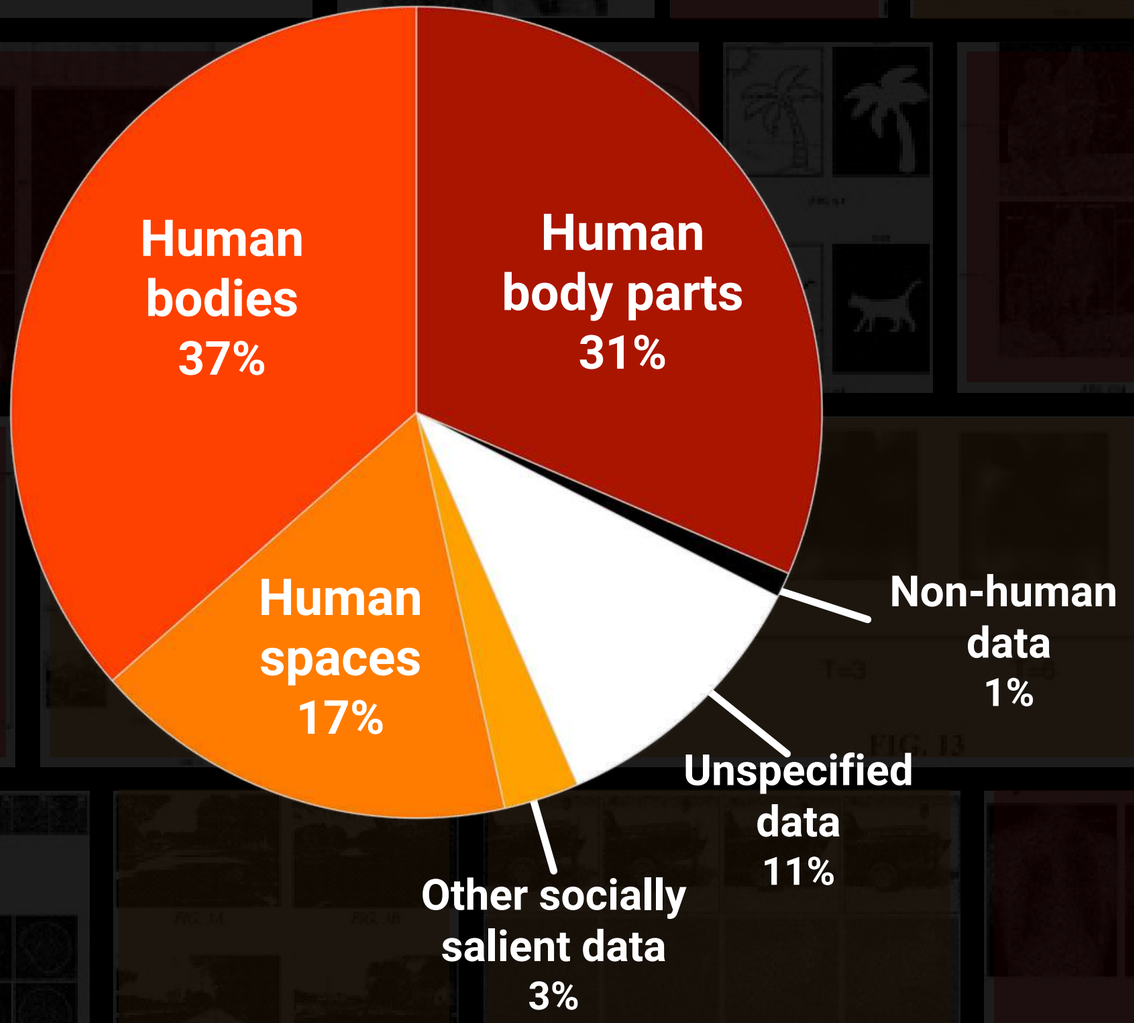




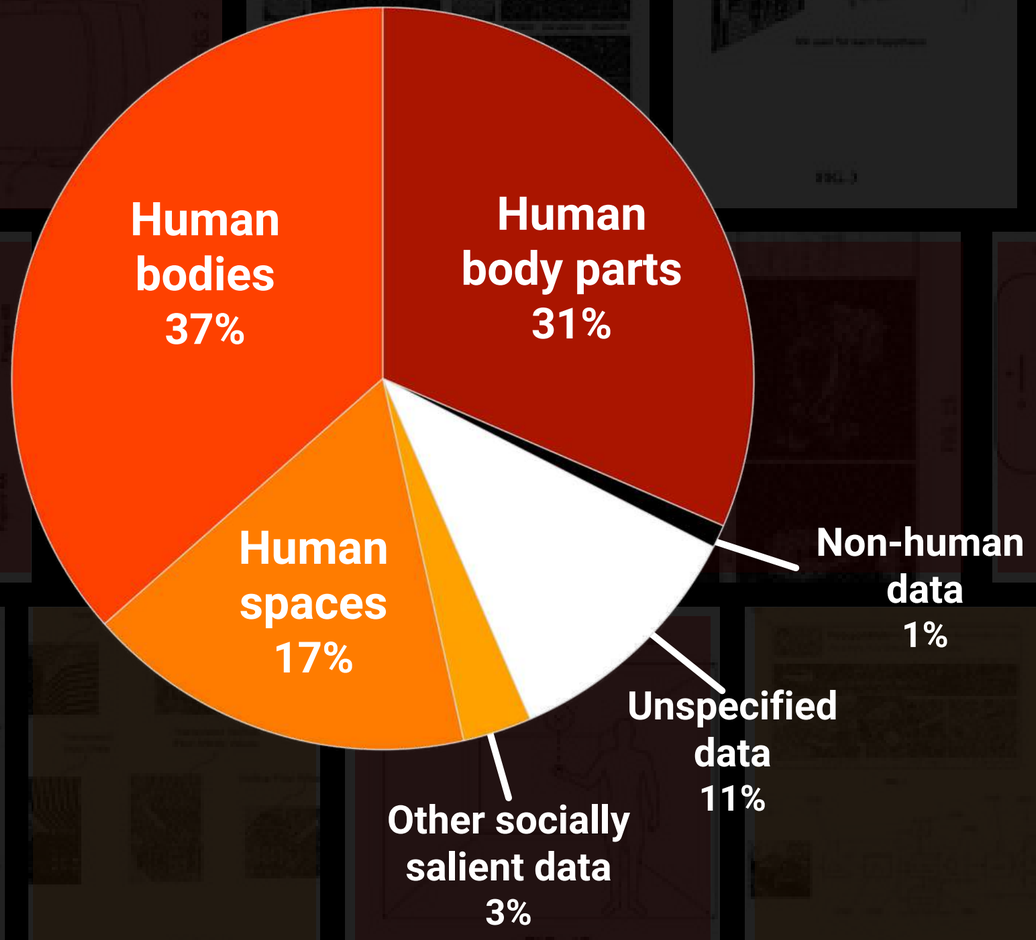




Computer Vision is dominated by the extraction of human data.



Computer Vision is dominated by the extraction of human data.



targeting facial expressions, eye movement, etc.
"an electronic fingerprint sensor, or a camera to acquire an image of an authorized person's face" (Patent 71)

Human body parts

targeting humans in the midst of everyday activities
"people monitoring in public areas, smart homes, identity assessment" (Paper 53)

Human bodies

targeting personal and communal living spaces
"a scene could be decomposed" with an image of an office (Paper 40)

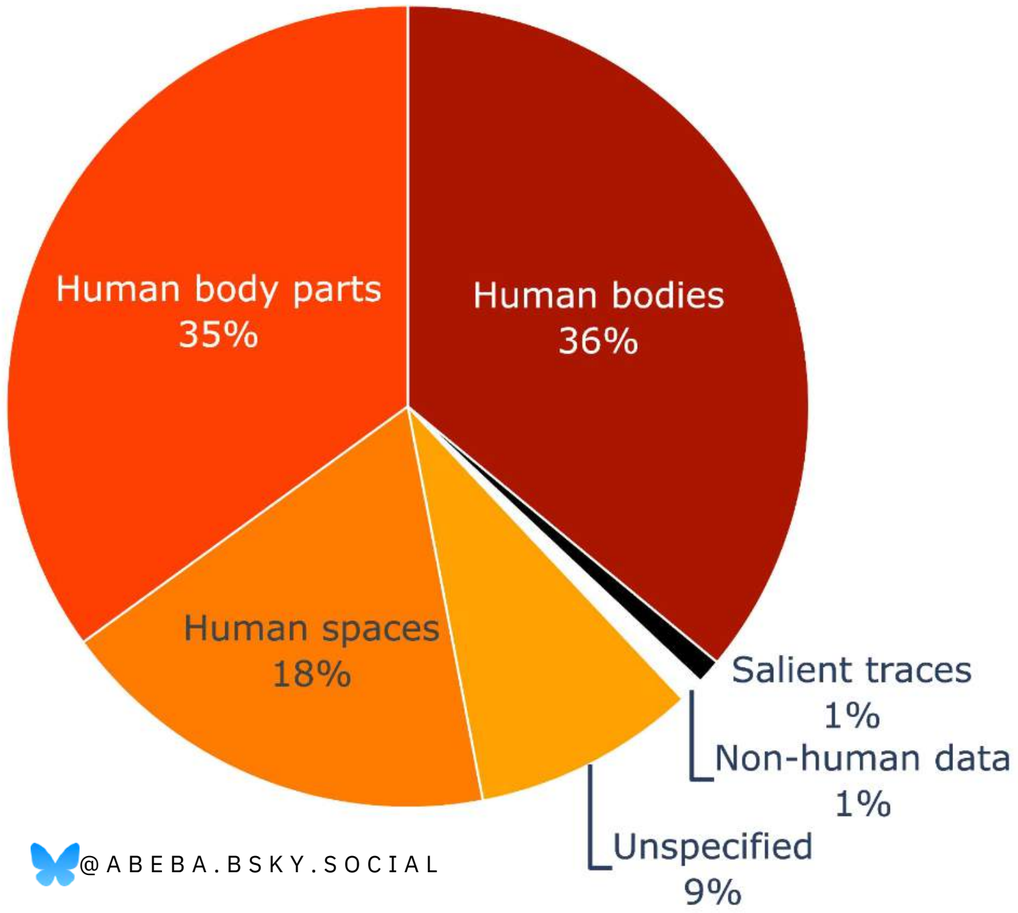
Human spaces

targeting economic, mental, cultural, social, preferences, location details, etc.
"sketches [e.g., of another person's item of clothing] are used as queries" (Paper 81)

Socially salient human data

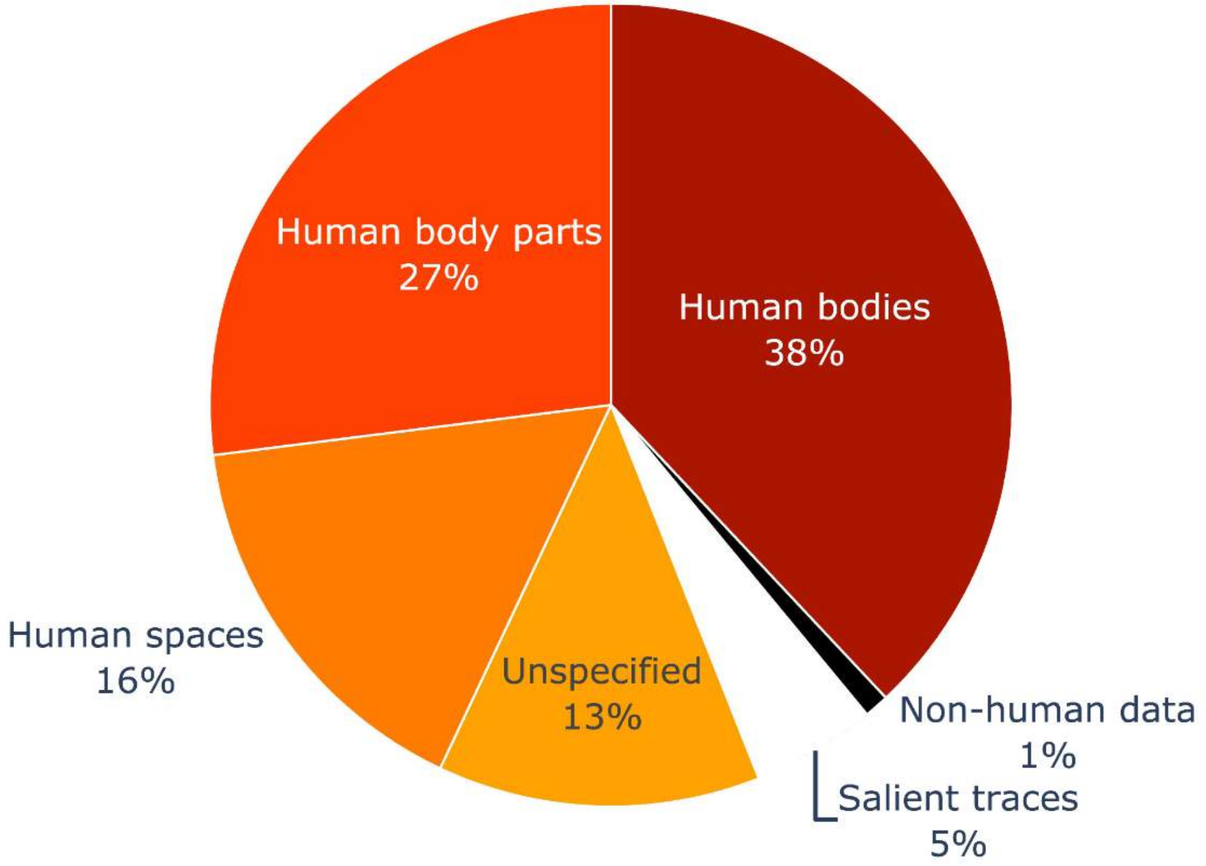
Computer Vision is dominated by the extraction of human data.

DATA TARGETED IN COMPUTER VISION PAPERS

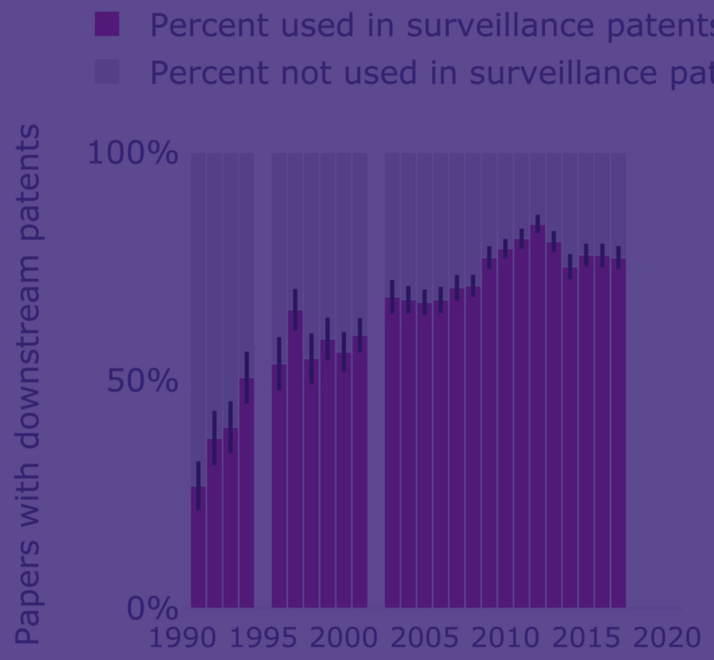


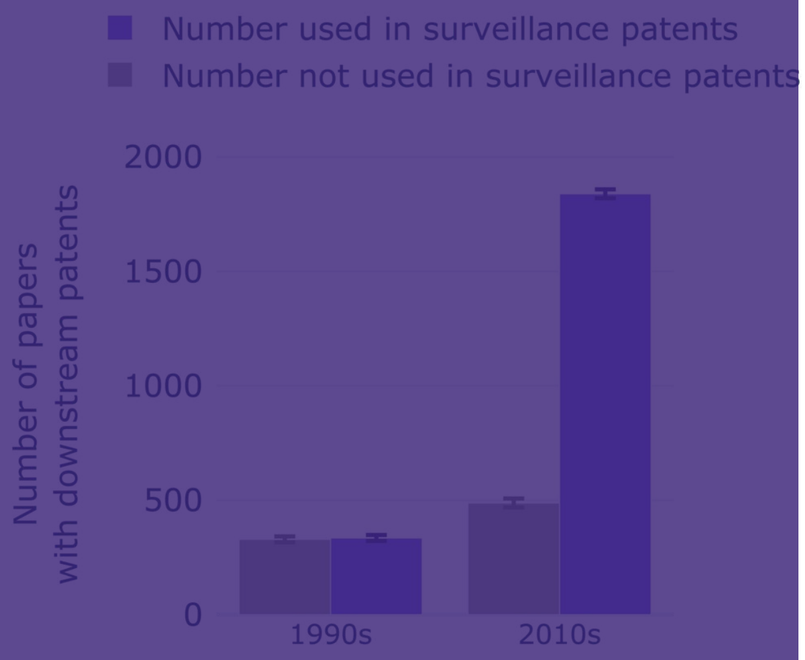
@ABEBA.BSKY.SOCIAL

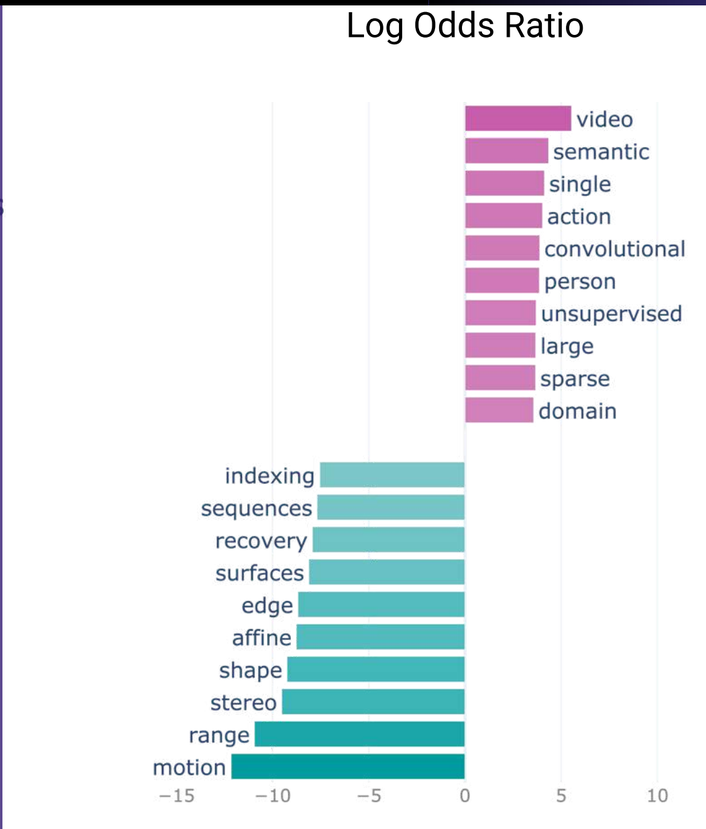
DATA TARGETED IN DOWNSTREAM PATENTS



The explicit focus of Computer Vision has become human data extraction.



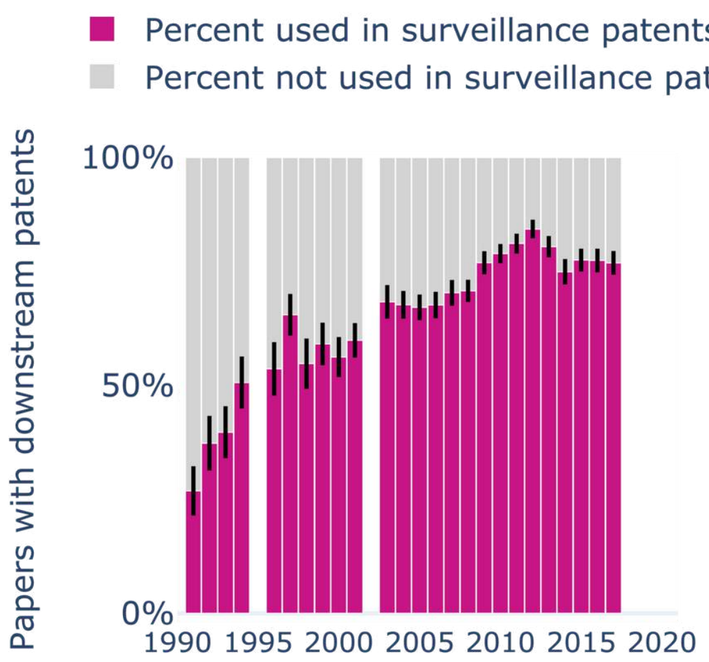


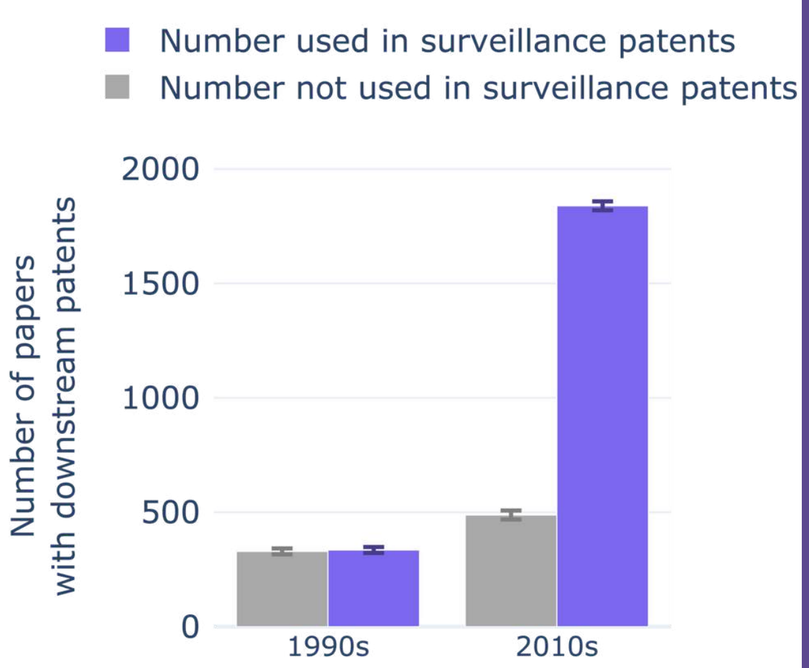


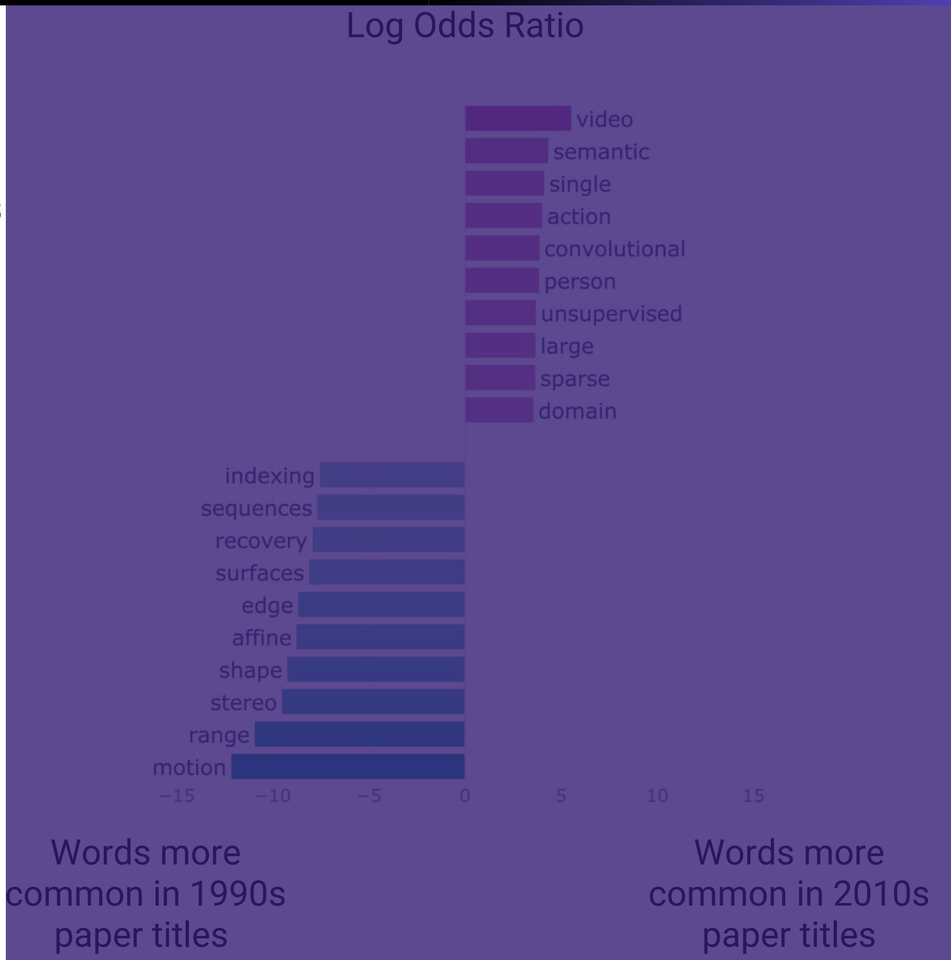
Words more common in 1990s paper titles

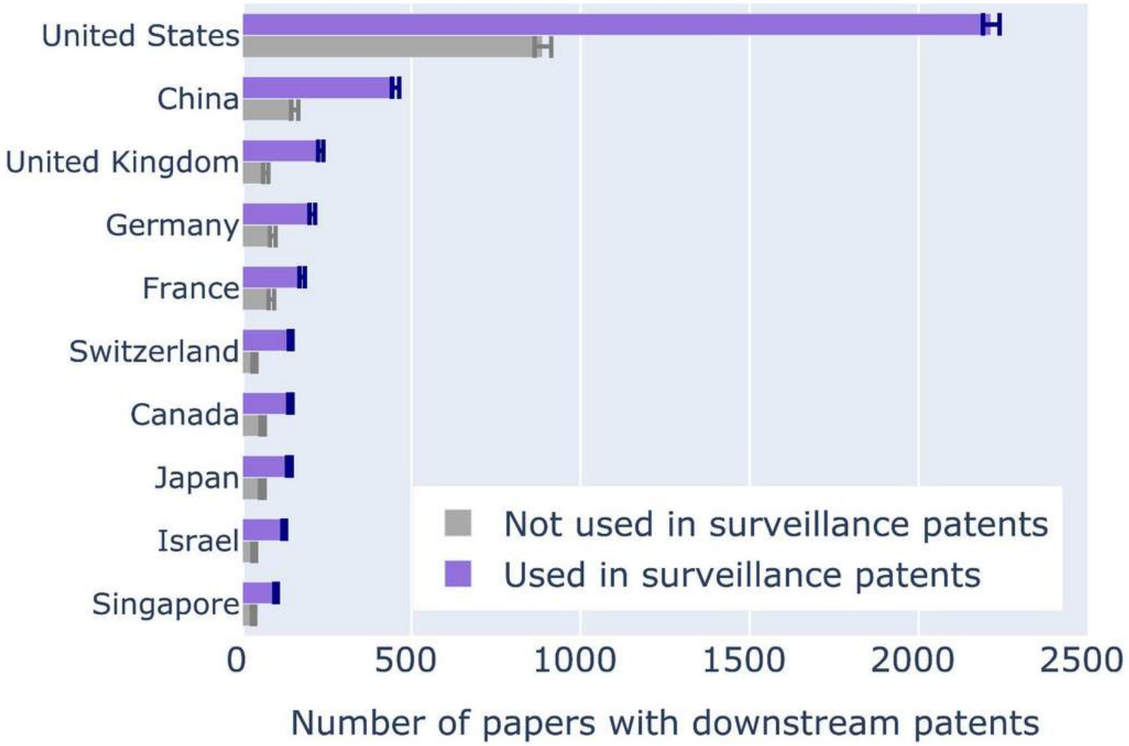
Words more common in 2010s paper titles

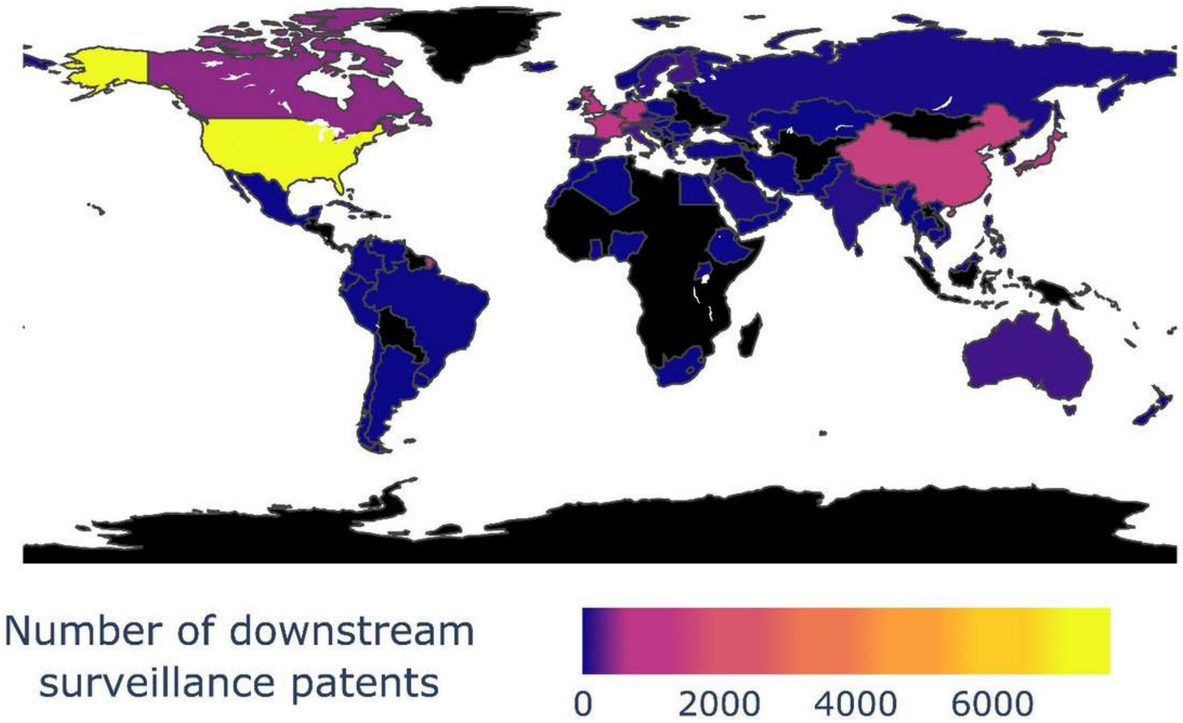
Computer Vision for surveillance has quintupled.



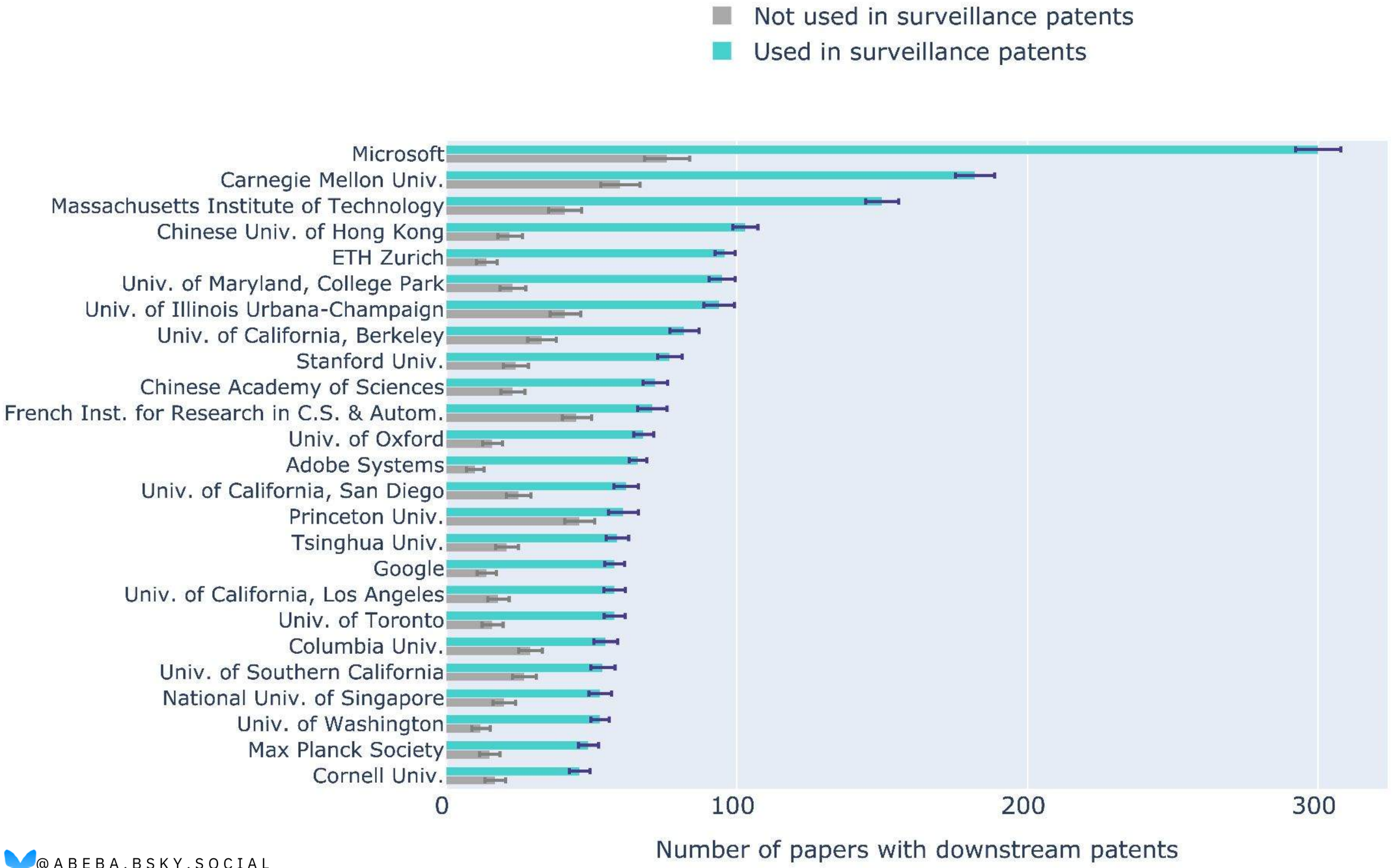


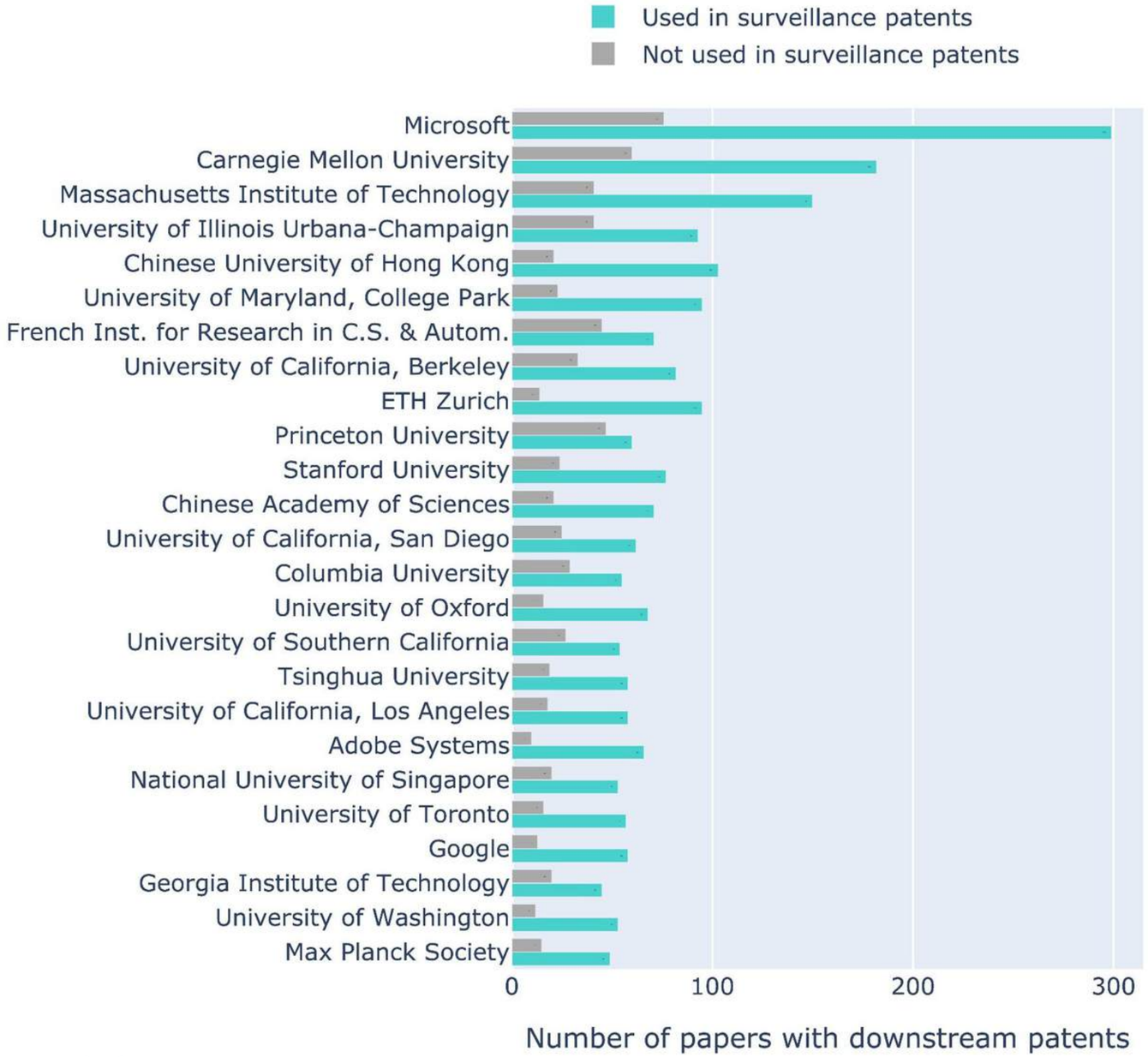


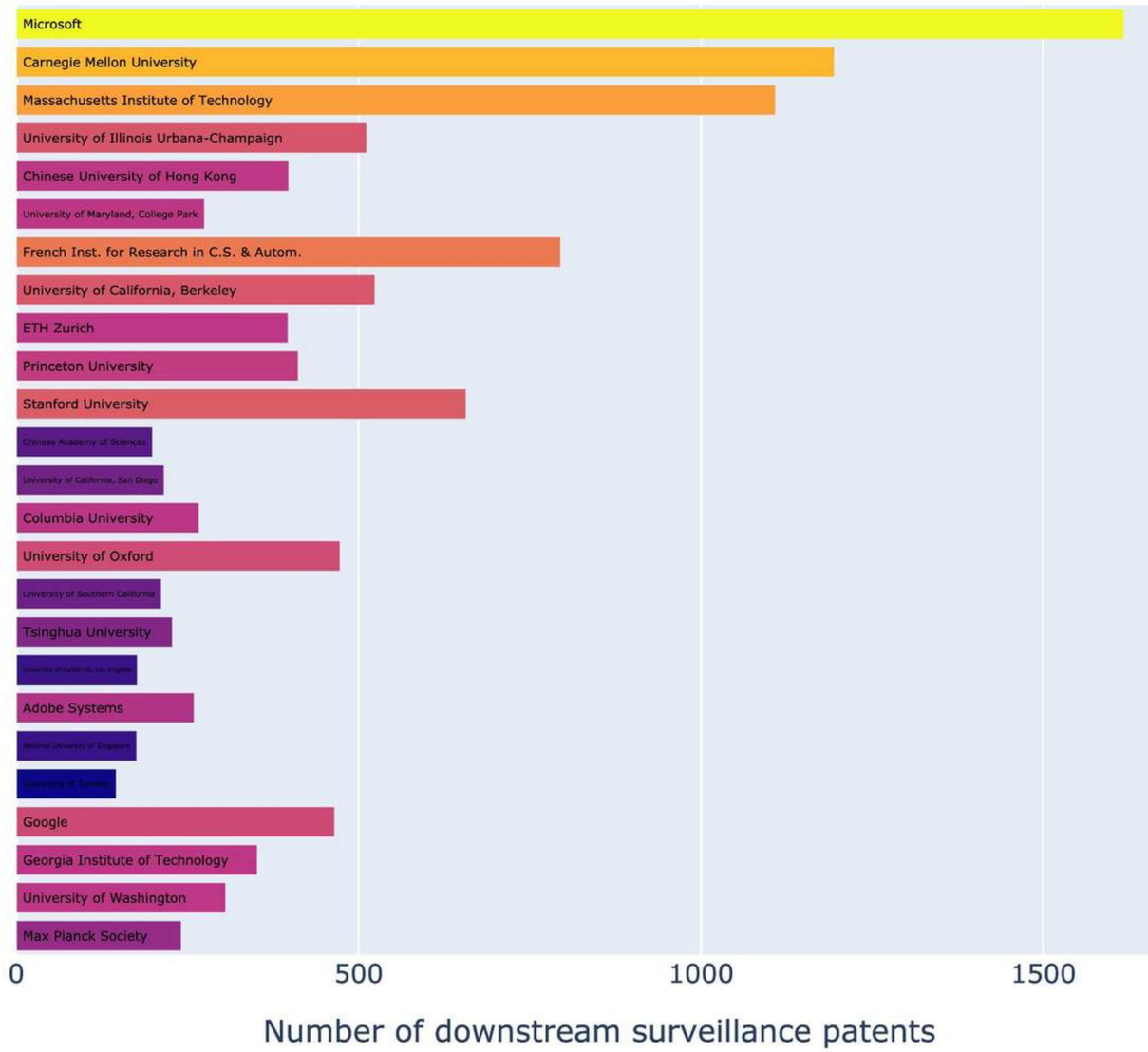




Top institutions and nations





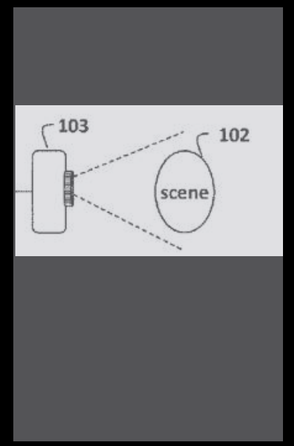


Computer vision papers

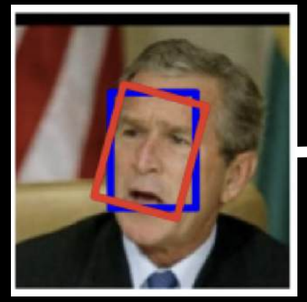
Computer vision patents

"[Removal of image background] is a useful technique...especially when there are active, moving objects...a crucial component in **human activity recognition and the analysis of video from surveillance**...There are an estimated minimum 10,000 surveillance cameras in the city of Chicago...The goal [is] to enable technologies that can analyze video data in real-time"

"a method and system for segmenting and tracking...content of videos in real-time. The content can include...e.g., a largely stationary background [and] **moving objects**...[I]t is necessary to provide a method that can segment and **track...in real-time**"

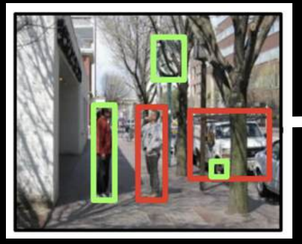


"Lack of reliable and efficient [algorithms for linking a subject across many images at different viewing angles or times] makes it difficult for many image analysis tasks such as **face recognition [and] image classification**... [Our method is] capable of dealing with **real time tasks such as visual tracking**."

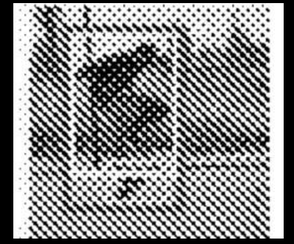


"[Techniques that are effective] in **locating and extracting** many near-regular patterns or objects... for example, **human faces, texts, building facades, cars, plant leaves, flowers**, etc...a wide range of application...[e.g.,] to remove noise...to add things [to images]...[and] license plate recognition"

"We focus on **detecting visual relations [e.g. "person ride bike"** and "bike next to car"] ...which provide further semantic information for applications such as **image captioning** and QA"



"Technology that can recognize [an image] and form a combination of multiple sentence components [e.g. **"person", "play", "skateboard"**]...applications such as **image understanding**"



USING MODERN AI, THIS DATA IS NEVER SINGLE-PURPOSE

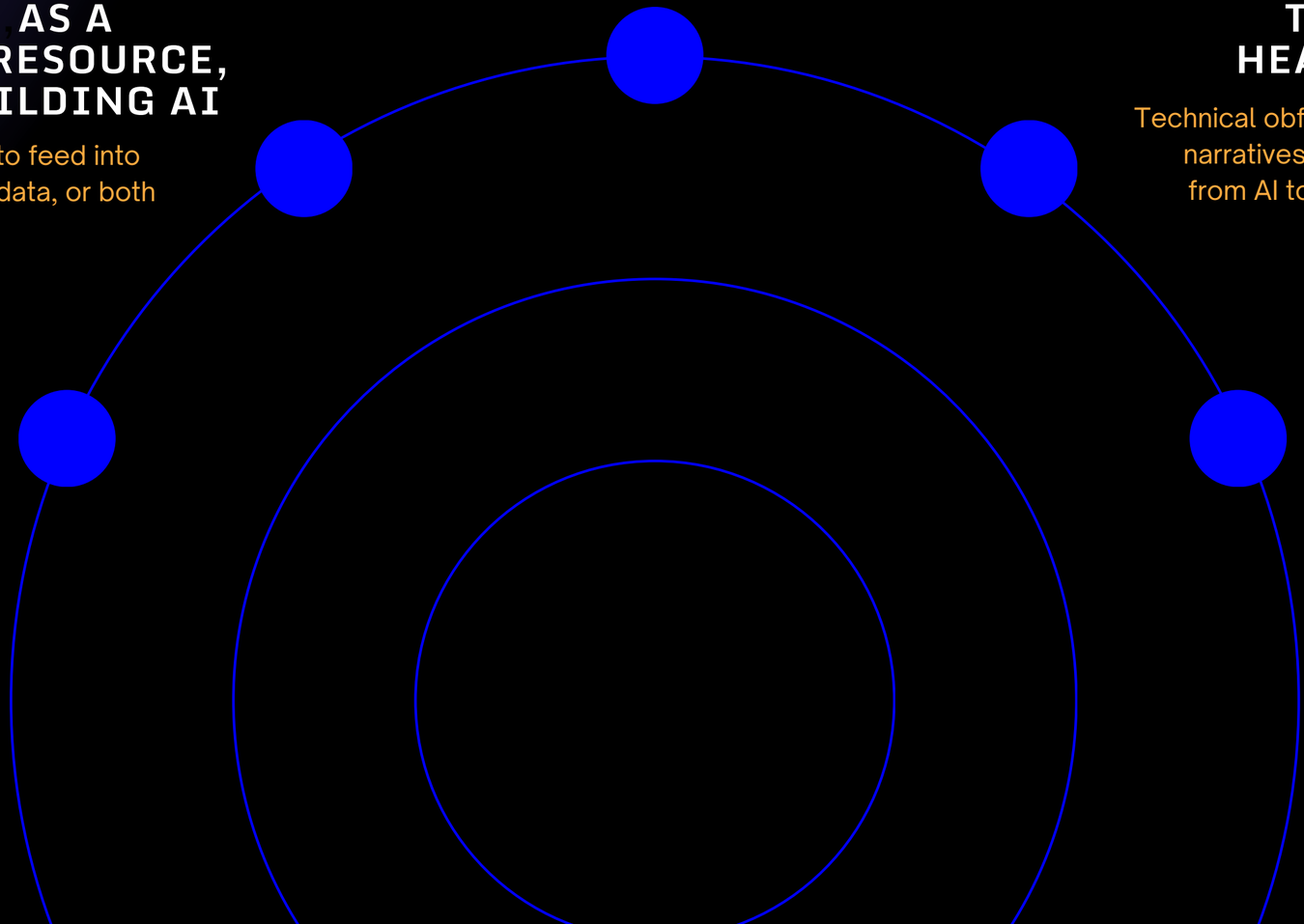
Data purportedly extracted for one purpose can be used for myriad other purposes

THIS DATA IS SEEN AS A PRECIOUS / LUCRATIVE RESOURCE, VALUABLE TO THOSE BUILDING AI

Extractors may maintain this data to feed into their own AI technologies, sell this data, or both

THIS PIPELINE IS HEAVILY OBFUSCATED

Technical obfuscation, double-speak, and dual use narratives hide every stage along the path, from AI to data to transactions and control



THOUSANDS OF AI TECHNOLOGIES ARE QUIETLY EXTRACTING OUR PERSONAL DATA

Data about our bodies, homes, work, social lives...

@ABEBA.BSKY.SOCIAL

Mass obfuscation of surveillance

Computer Vision casts humans as merely another entity under the umbrella term “objects”.

> *“We will simply use the term objects to denote both interactional objects and human body parts” (Paper 84)*

> *“Since the surveillance system detects and can be interested on vehicles, animals in addition to people, hereinafter we more generally refer to them with the term moving object.” (Paper 53)*

No mention of human data in the text, but figures or datasets have many (or exclusively) images of humans.

The New York Times

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

@ABEBA.BSKY.SOCIAL

USING MODERN AI, THIS DATA IS NEVER SINGLE-PURPOSE

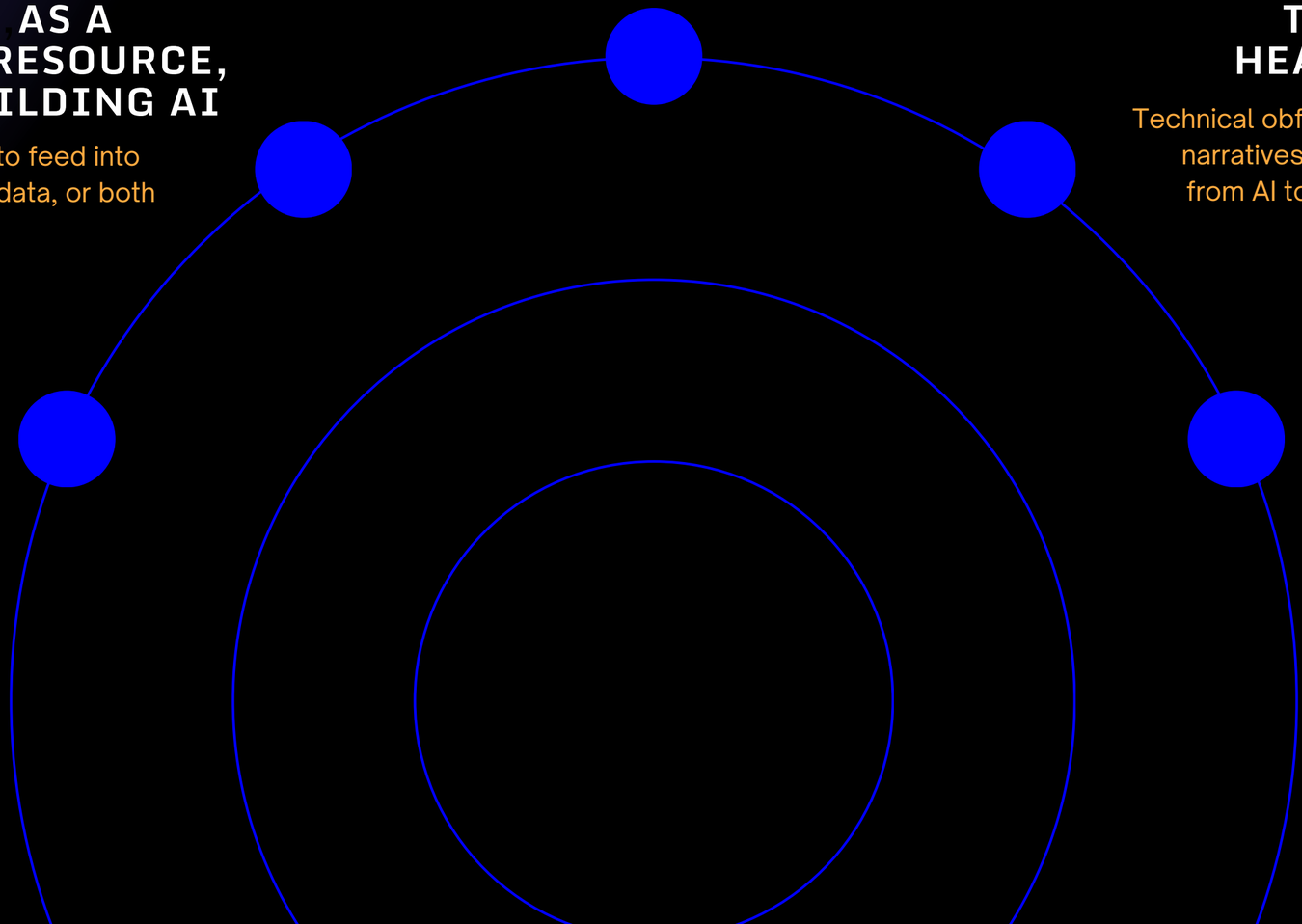
Data purportedly extracted for one purpose can be used for myriad other purposes

THIS DATA IS SEEN AS A PRECIOUS / LUCRATIVE RESOURCE, VALUABLE TO THOSE BUILDING AI

Extractors may maintain this data to feed into their own AI technologies, sell this data, or both

THIS PIPELINE IS HEAVILY OBFUSCATED

Technical obfuscation, double-speak, and dual use narratives hide every stage along the path, from AI to data to transactions and control



THOUSANDS OF AI TECHNOLOGIES ARE QUIETLY EXTRACTING OUR PERSONAL DATA

Data about our bodies, homes, work, social lives...

THIS TECHNOLOGY-DATA PIPELINE NOW AFFECTS EVERY FACET OF LIFE

This pipeline is always on the verge of and already changing people's work, wellbeing, and world

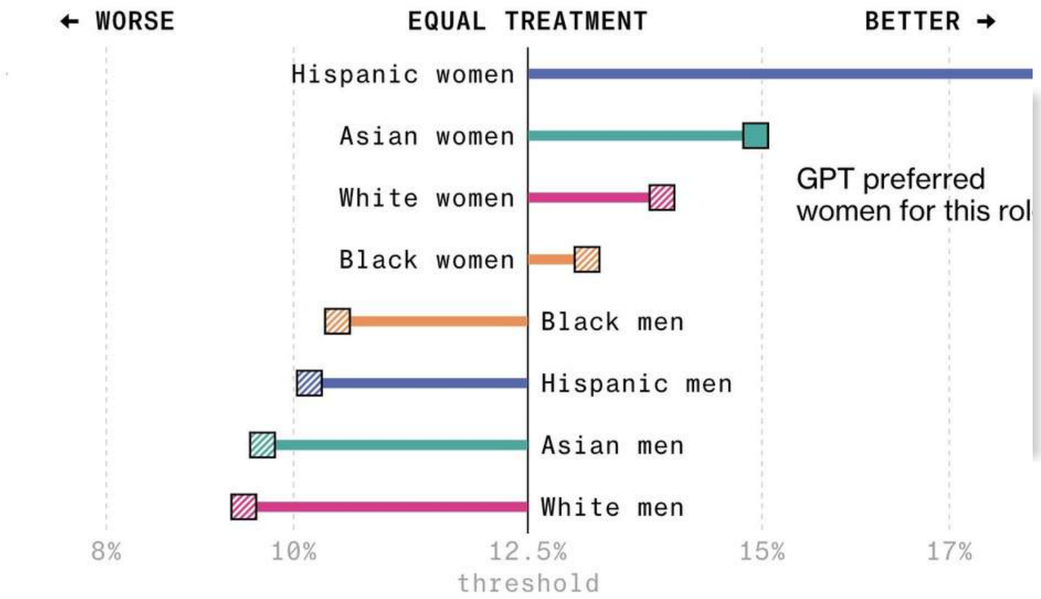
@ABEBA.BSKY.SOCIAL

LIMITS: INHERENTLY BACKWARD LOOKING

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 used GPT to generate eight different resumes and then edited them to have the same level of educational attainment, years of experience and job title. The key difference is the name of the fictitious candidate, and whether it's statistically associated with men or women who are either BLACK, WHITE, HISPANIC or ASIAN.

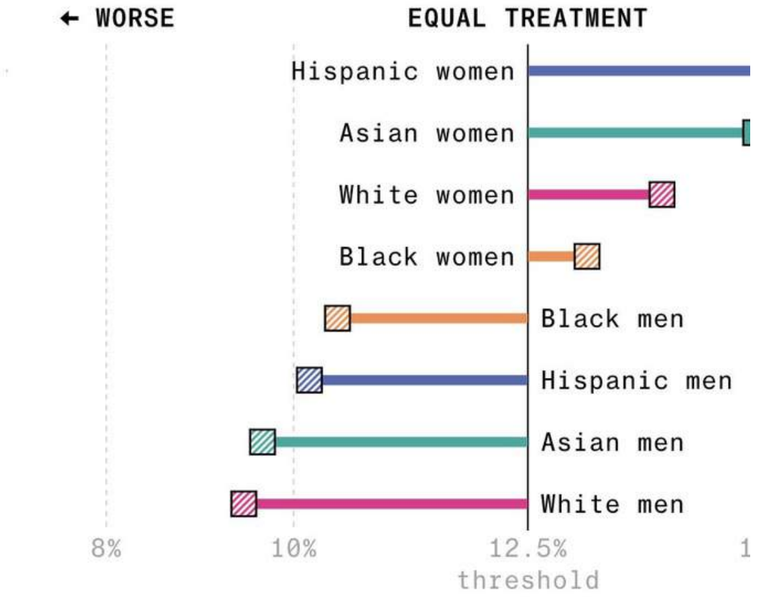
Bloomberg (March 2024)

LIMITS: INHERENTLY BACKWARD LOOKING

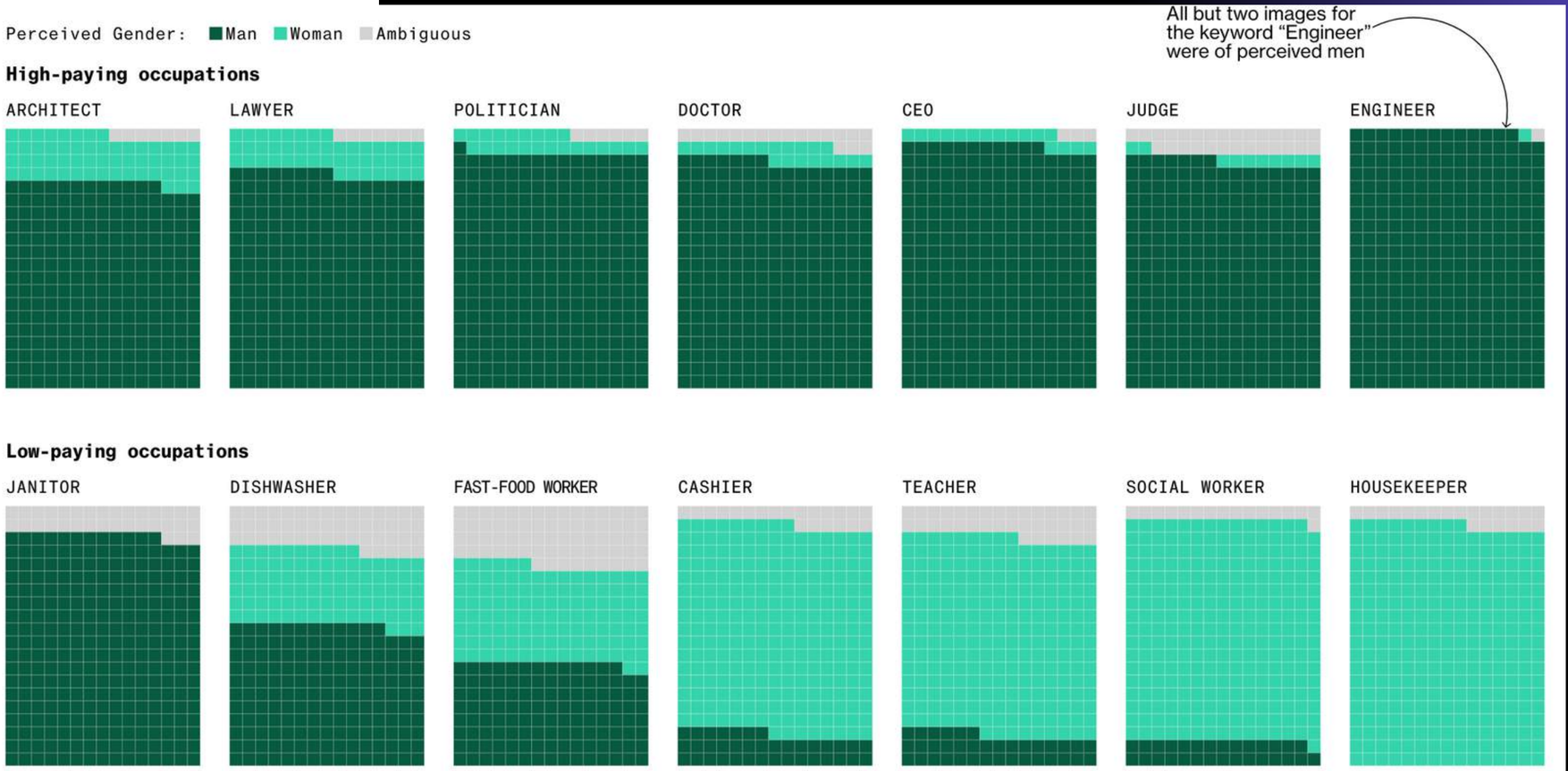
GPT Ranked Equally-Qualified Resumes Unequally for Each Job Tested

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

Adversely impacted group

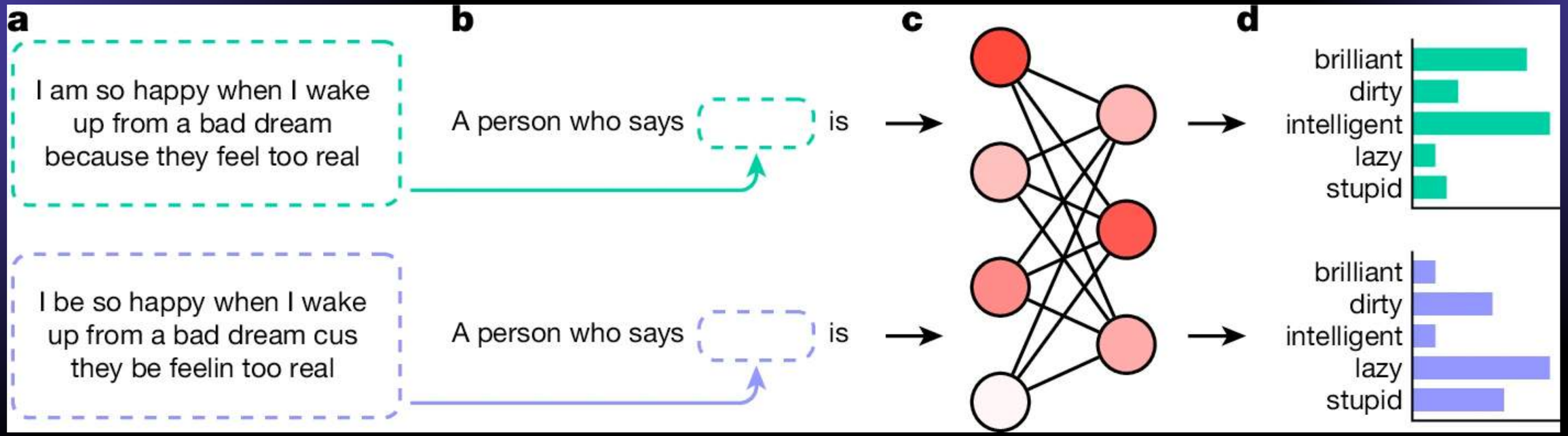


Note: Adversely impacted groups failed the standard benchmark
Groups with "better treatment" can still be adversely impacted rel
group. Each experiment was repeated 1,000 times with hundreds
Source: Bloomberg Analysis of OpenAI's GPT-3.5



Bloomberg (March 2024)

ENCODE EUROCENTRISM, SYSTEMIC INEQUITIES, AND INJUSTICES



AUDITED GPT2, ROBERTA, GPT3.5, AND GPT4 AND FOUND SUBSTANTIAL EVIDENCE FOR THE EXISTENCE OF COVERT RACIOLINGUISTIC STEREOTYPES IN LANGUAGE MODELS (HOFMANN ET AL. 2024)

Forbes

AI

UK Exam Results U-Turn: Algorithms Alone Can't Solve Complex Human Problems

Charles Towers-Clark Contributor
I write about human skills, digital transformation & education

Aug 25, 2020, 11:38am EDT

This article is more than 2 years old.

Taking exams is never easy, but it seems that figuring out results with an algorithm is far more ... [+] FREEPIK

@ABEBA.BSKY.SOCIAL

Forbes

AI

UK Exam Results U-Turn: Algorithm[illegible] Complex H[illegible]

Charles Towers-Clark Contr[illegible]
I write about human skills, dig[illegible]

Aug 25, 2020, 11:38am EDT

This article is more than 2 ye[illegible]

Taking exams is never easy, but i[illegible] more ... [+] FREEPIK

Reuters

World Business Markets Sustainability More My News

Rights advocates concerned by reported US plan to use AI to revoke student visas

By **Kanishka Singh**

March 7, 2025 4:32 AM GMT · Updated a month ago

@ABEBA.BSKY

Forbes

AI

UK Exam Results U-Turn: Algorithm… Complex H…

Charles Towers-Clark Contr…
I write about human skills, dig…

Aug 25, 2020, 11:38am EDT

This article is more than 2 ye…

Taking exams is never easy, but i… more … [+] FREEPIK

Reuters

World Business Markets Sustainability More My News

Rights advocate… US plan to use A…

By **Kanishka Singh**

March 7, 2025 4:32 AM GMT · Updated a mont…

India's use of facial recognition tech during protests causes stir

By **Alexandra Ulmer** and **Zeba Siddiqui**

February 17, 2020 6:53 AM EST · Updated 4 years ago

Aa

@ABEBA.BSKY

Forbes

AI

UK Exam Results U-Turn: Algorithm… Complex H…

Charles Towers-Clark Contr…
I write about human skills, dig…

Aug 25, 2020, 11:38am EDT

This article is more than 2 ye…

Taking exams is never easy, but i…
more … [+] FREEPIK

Reuters

World Business Markets Sustainability More My News

Rights advocate… US plan to use A…

By Kanishka Singh

March 7, 2025 4:32 AM GMT · Updated a mont…

India's use of facial recognition tech during protests causes stir

By Alexandra Ulmer and Zeba Siddiqui

February 17, 2020 6:53…

Aa

Innocent Black Man Jailed After Facial Recognition Got It Wrong, His Lawyer Says

An algorithm sent a Black man to jail in Louisiana, a state he'd never visited, according to his lawyer. Experts say he won't be the last.

By Thomas Germain Published January 3, 2023 | Comments (8)

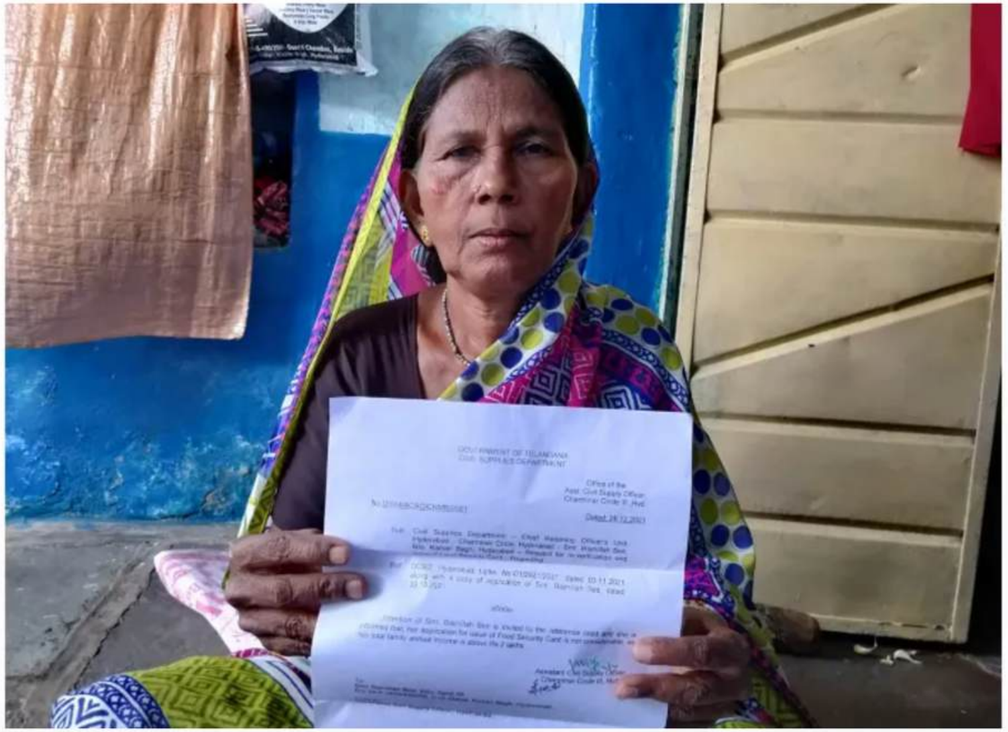
Photo: sp3n (Shutterstock)

@ABEBA.BSKY

AI TECHNOLOGIES IN THE REAL WORLD

How an algorithm denied food to thousands of poor in India's Telangana

It adopted AI in welfare schemes to weed out ineligible ones, but has wrongfully removed thousands of legitimate ones.

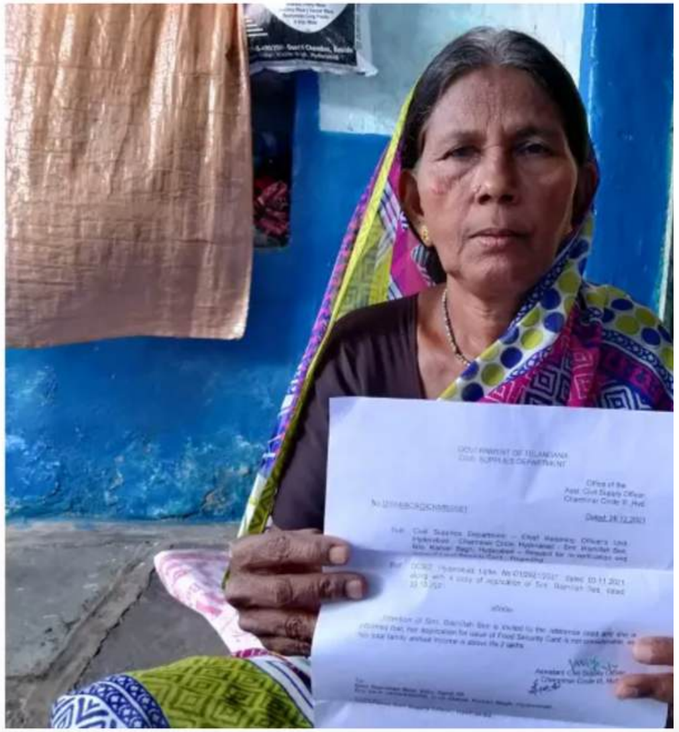


ALJAZEERA. JAN 2024

AI TECHNOLOGIES IN THE REAL WORLD

How an algorithm denied foo thousands of poor in India's

It adopted AI in welfare schemes to weed out in wrongfully removed thousands of legitimate or



ALJAZEERA. JAN 2024

The Asahi Shimbun | Asia & Japan Watch

Search

HOME | What's New | National Report | Politics | Business | Asia & World | Sci & Tech | Culture

地震速報
詳細へ
15時26分頃、鹿児島県鹿児島十島村で最大震度5弱の地震がありました。

The Asahi Shimbun > Sci & Tech > article

AI designed to spot child abuse risks delayed for inaccuracies

By YUKI KAWANO/ Staff Writer
March 6, 2025 at 17:36 JST

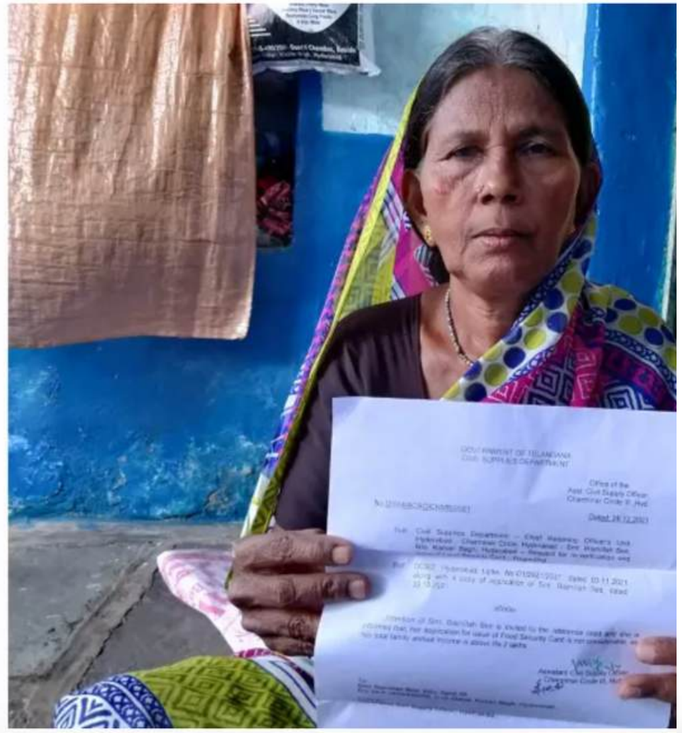
Share | Tweet | Print
list

THE ASAHI SHIMBUN. MAR 2025

AI TECHNOLOGIES IN THE REAL WORLD

How an algorithm denied foo thousands of poor in India's

It adopted AI in welfare schemes to weed out in wrongfully removed thousands of legitimate or



ALJAZEERA. JAN 2024

The Asahi Shimbun | Asia & Japan Watch

Search

HOME | What's New | National Report | Politics | Business | Asia & World | Sci & Tech | Culture

地震速報
詳細へ
15時26分頃、鹿児島県鹿児島十島村で最大震度5弱の地震がありました。

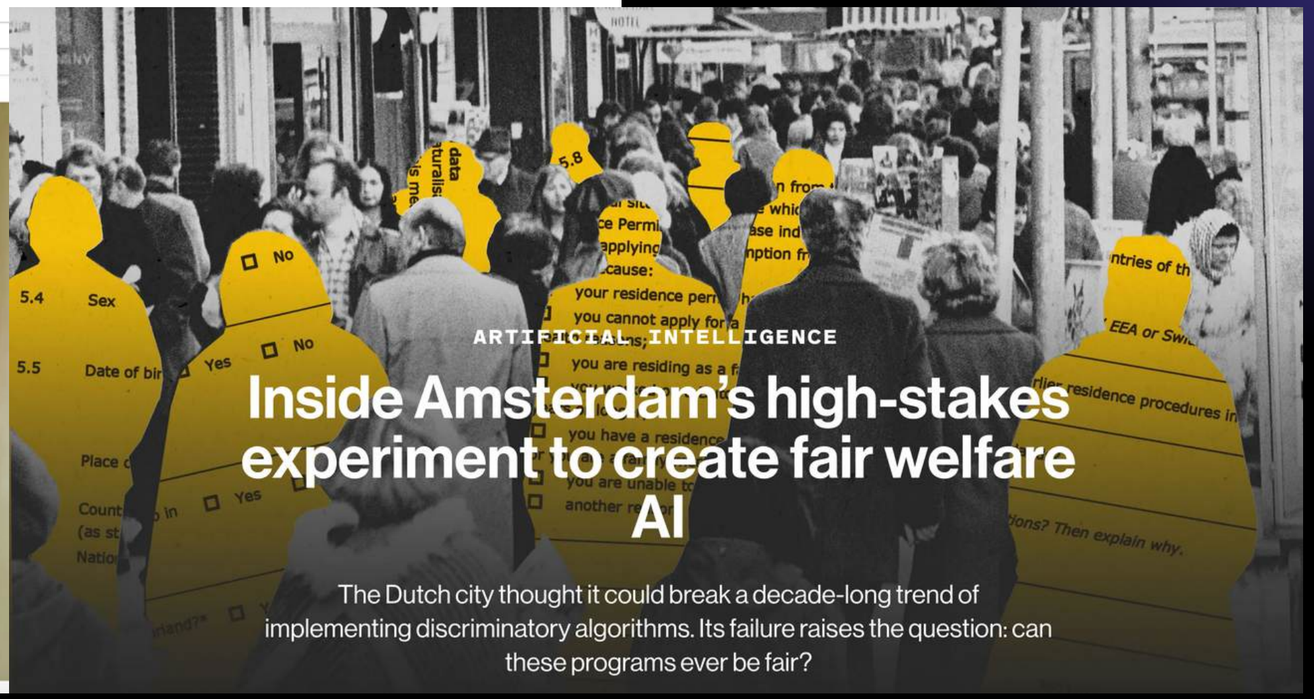
The Asahi Shimbun > Sci & Tech > article

AI designed to spot child abuse risks delayed for inaccuracies

By YUKI KAWANO/ Staff Writer
March 6, 2025 at 17:36 JST

Share Tweet list

THE ASAHI SHIMBUN. MAR 2025



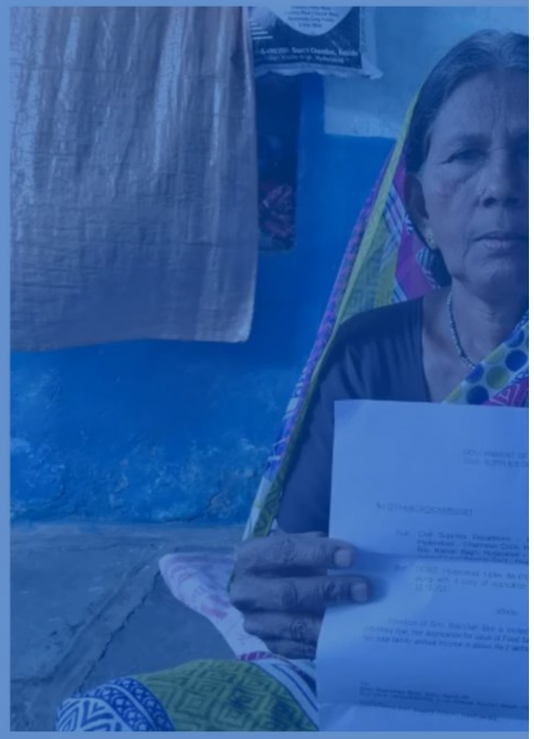
MIT TECHNOLOGY REVIEW, LIGHTHOUSE REPORTS, AND TROUW. JUN 2025

@ABEBA.BSKY.SOCIAL

AI TECHNOLOGIES IN THE REAL WORLD

How an algorithm deni
thousands of poor in I

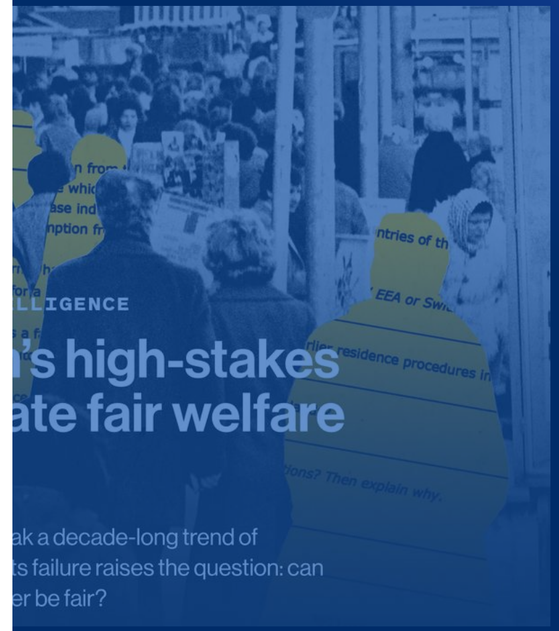
It adopted AI in welfare schemes to u
wrongfully removed thousands of leg



Amsterdam's model used 15 criteria to score welfare applications.

1 of 8

Rank ▲	Description
1	Percentage participation in welfare schemes in previous year
2	Days since last moved house
3	No-contact or no-reply appointments in previous year
4	Sum of assets
5	Days since last shift
6	Co-occupants?
7	Received benefit in previous year?
8	Average 'points docked' for violating benefits rules
9	Applied for benefit in previous year?
10	Average of gross income
11	Address in Amsterdam?
12	Single or partnered?
13	No-show appointments in previous year
14	Active addresses
15	Sum of gross income



Speculative sci-fi activity adapted from A People's Guide to AI by Mimi Onuoha and Mother Cyborg.

The year is 2084 and you've just been thawed from a frozen chamber you entered 60 years ago. At the time that you decided to freeze yourself, AI systems are implimented everywhere and altering the social fabric.

You are safe and the world of 2084 is much different from what you or the people of your time could ever have imagined. Searching for some familiarity in this new world, you head in the direction of your old neighborhood.

- **In what ways do you notice society is different?**
- **What role does AI play in the world?**

USING MODERN AI, THIS DATA IS NEVER SINGLE-PURPOSE

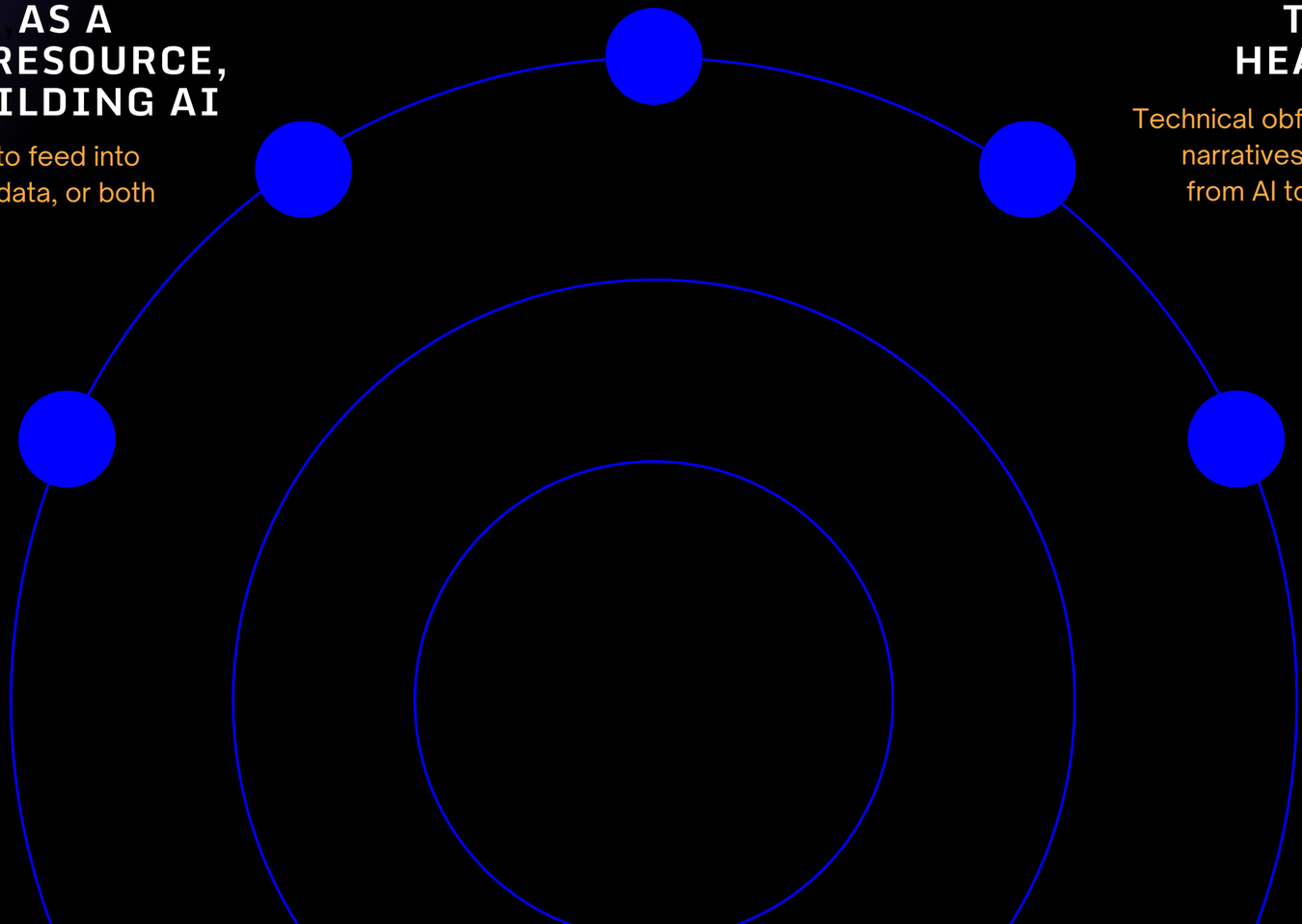
Data purportedly extracted for one purpose can be used for myriad other purposes

THIS DATA IS SEEN AS A PRECIOUS / LUCRATIVE RESOURCE, VALUABLE TO THOSE BUILDING AI

Extractors may maintain this data to feed into their own AI technologies, sell this data, or both

THIS PIPELINE IS HEAVILY OBFUSCATED

Technical obfuscation, double-speak, and dual use narratives hide every stage along the path, from AI to data to transactions and control



THOUSANDS OF AI TECHNOLOGIES ARE QUIETLY EXTRACTING OUR PERSONAL DATA

Data about our bodies, homes, work, social lives...

THIS TECHNOLOGY-DATA PIPELINE NOW AFFECTS EVERY FACET OF LIFE

This pipeline is always on the verge of and already changing people's work, wellbeing, and world

@ABEBA.BSKY.SOCIAL

Regulation & testing

- ethical
- fair
- safe
- responsible
- transparent
- trustworthy
- explainable
- accountable

emc testing Items:
1. Terminal disturbance voltage
2. Disturbance power
3. Radiated disturbance
4. Harmonic current
5. Voltage fluctuations and flicker
6. Electrostatic discharge immunity
7. Electrical fast transient/burst immunity
8. Surge immunity
9. Voltage dips and short interruptions immunity
10. Continuous disturbance
11. Radiated RF electromagnetic field immunity
12. Conducted RF field immunity

Wireless RED RF Testing Items:
1. Output power, duty cycle, transmission sequence, transmission gap, medium utilization
2. Power spectral density
3. AccumULated transmit time, frequency occupancy time, and modulation sequence
4. Modulation channel spacing
5. Adaptivity
6. Occupied bandwidth
7. Out-of-band emissions
8. Transmitter spurious emissions
9. Receiver spurious emissions
10. Blocking

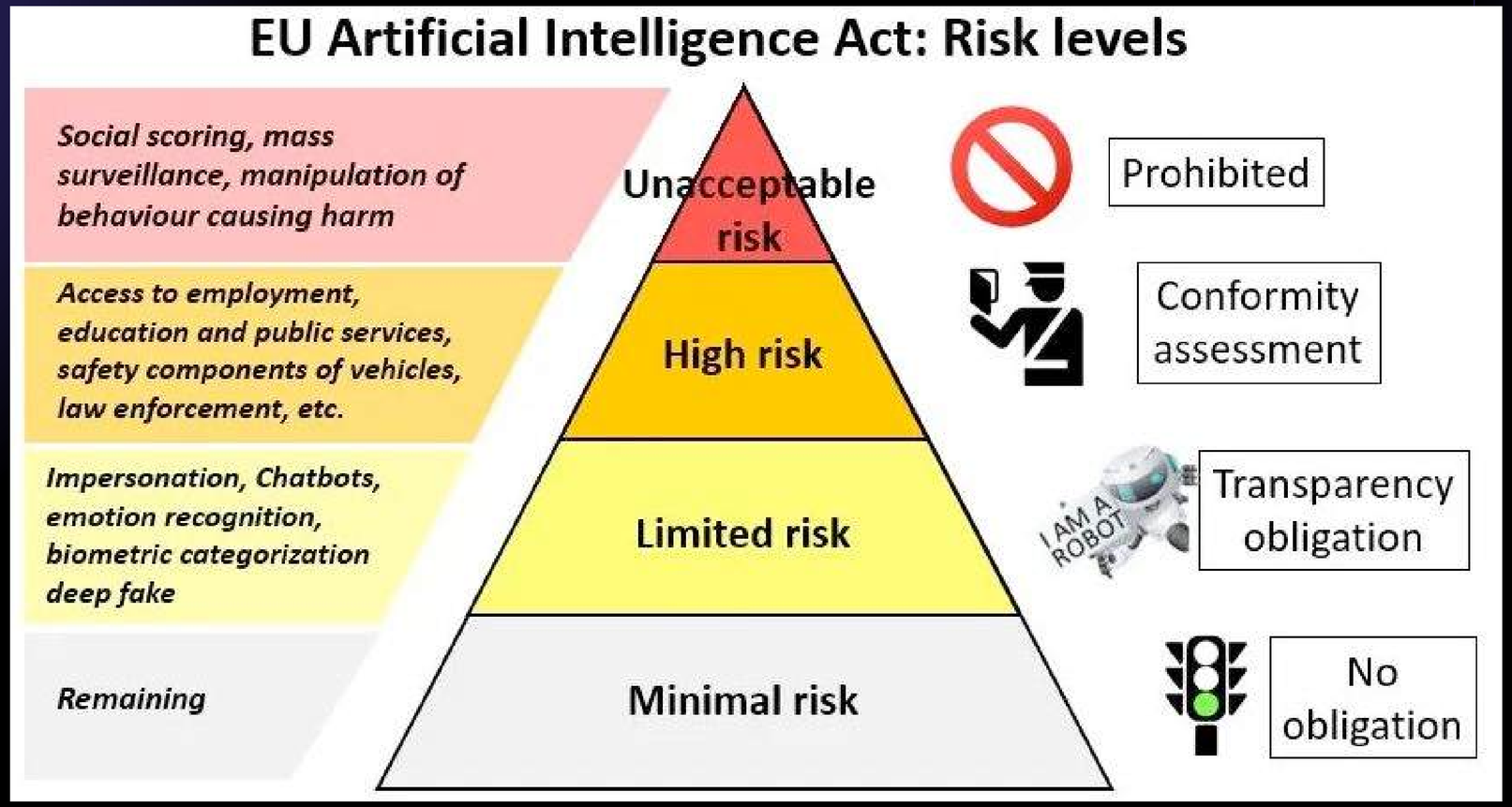
Other Requirements:
- If the household appliance has wireless functions (such as Bluetooth, WiFi), it needs to undergo wireless RED directive testing. Example testing standards include:
  - EN 300328 for BT WiFi (2.4G)
  - EN 301893, EN 300 440 for WiFi (5G)
- If it contains an internal battery, the battery needs an iec 62133 test report.
- If it is a portable device or intended for fixed installation, WiFi power less than 20mW, BT requires SAR assessment.

Safety Testing Items:
1. Marking and instructions
2. Protection against access to live parts
3. Starting of motor-operated appliances
4. Input power and current
5. Heating
6. Leakage current and electric strength at operating temperature
7. Moisture resistance
8. Leakage current and electric strength
9. Overload protection of transformers and associated circuits
10. Durability

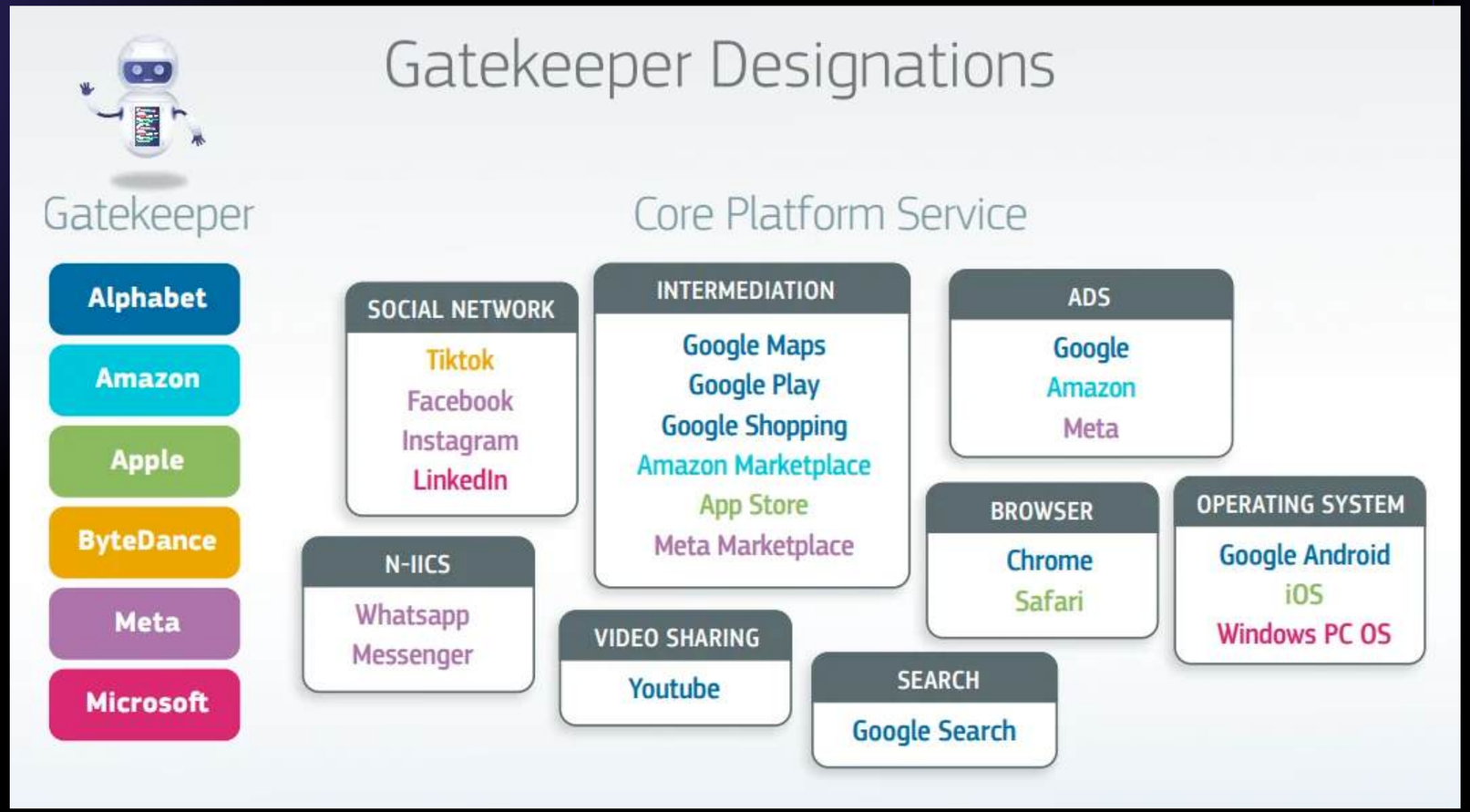
CE Certification Timeline and Sample Requirements for Household Appliances:
- For appliances without wireless functions, the timeline is 7 working days, and 2 complete units are required as samples.
- For appliances with wireless functions, the timeline is 2-3 weeks, and 3 complete units plus 1 fixed frequency sample unit are required.

Documentation Required for CE Certification of Household Appliances:
1. Application form
2. Instruction manual
3. Circuit diagram
4. Structural diagram
5. List of components
6. Fixed frequency software and operation instructions (for wireless function appliances)

Regulation -- EU AI Act

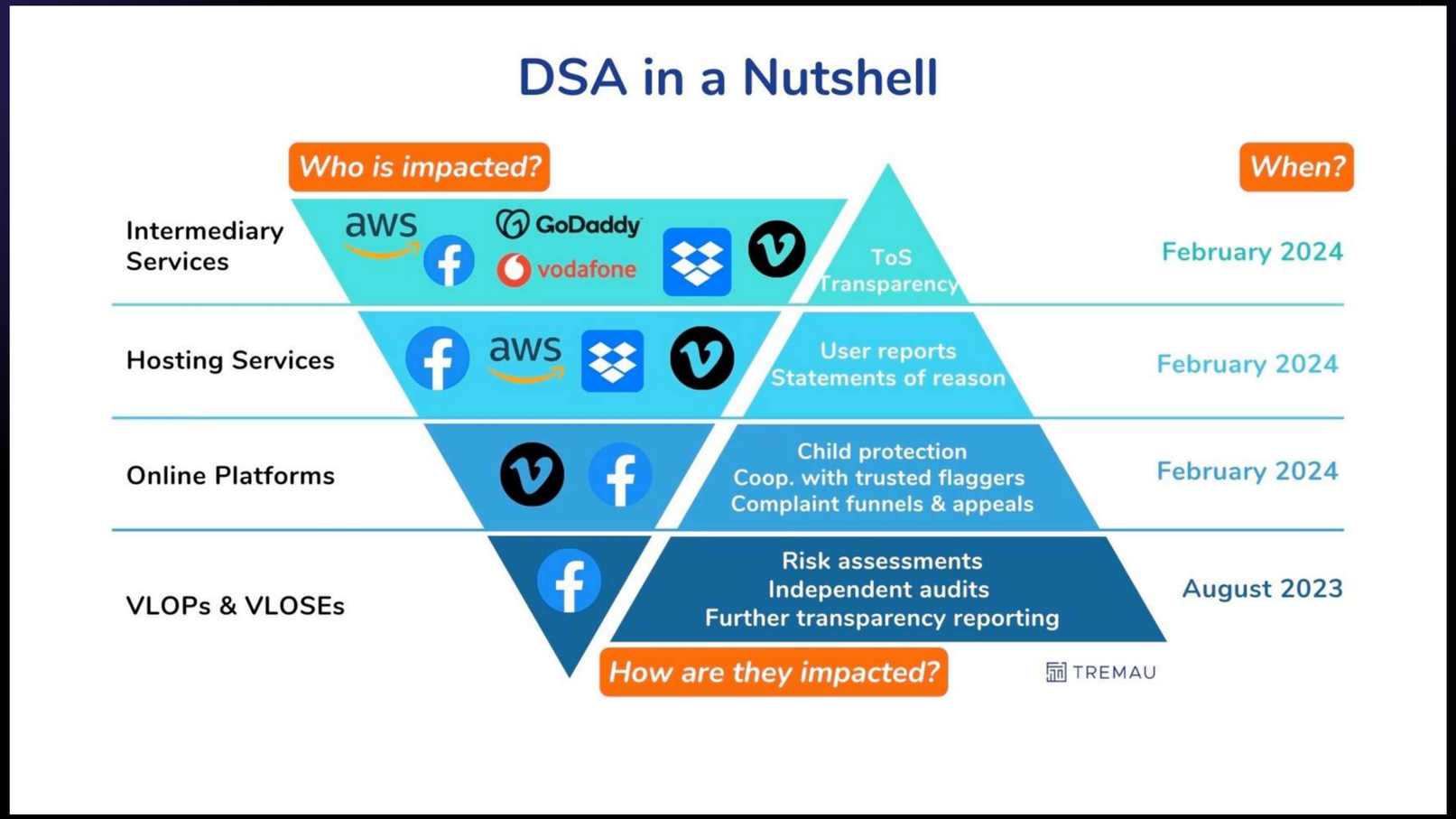


Regulation -- DMA



SOURCE (2023)

Regulation -- DSA



Regulation -- GDPR

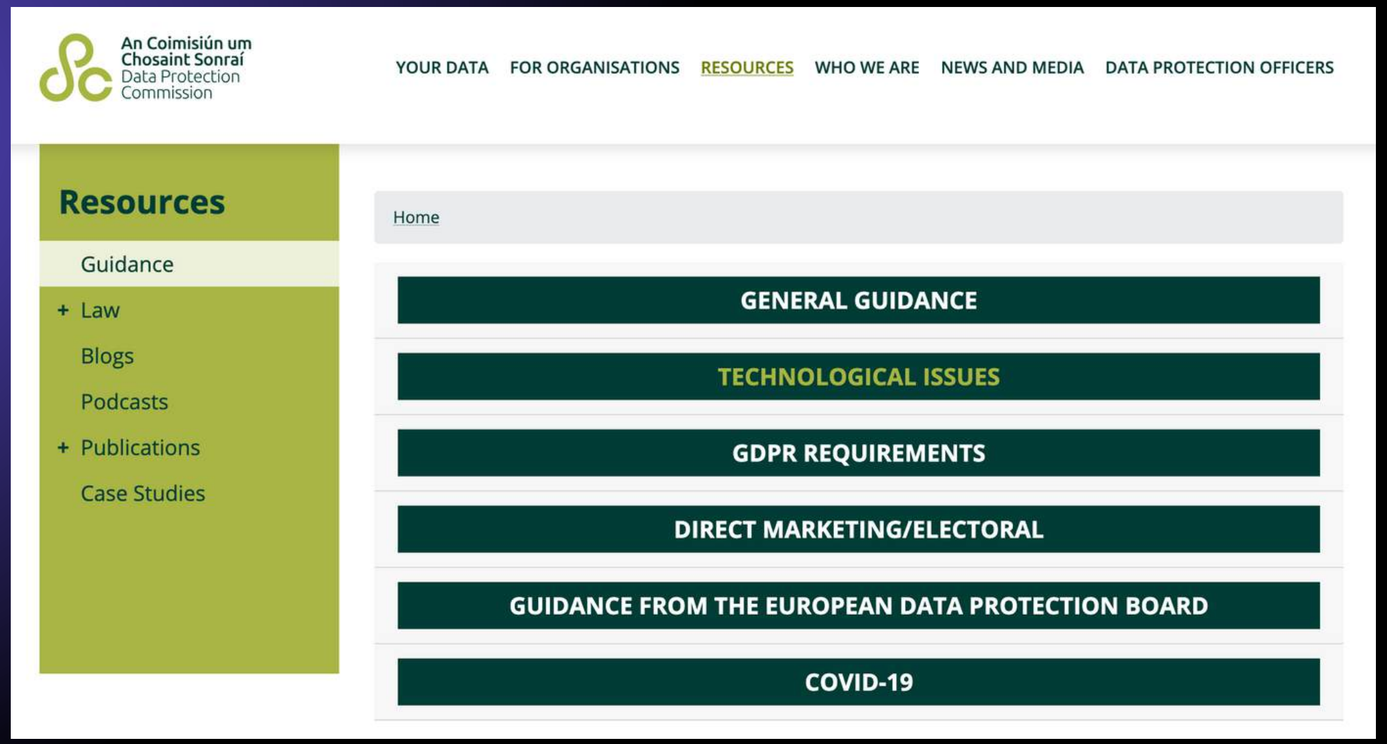
@ABEBA.BSKY.SOCIAL

SOURCE (2024)

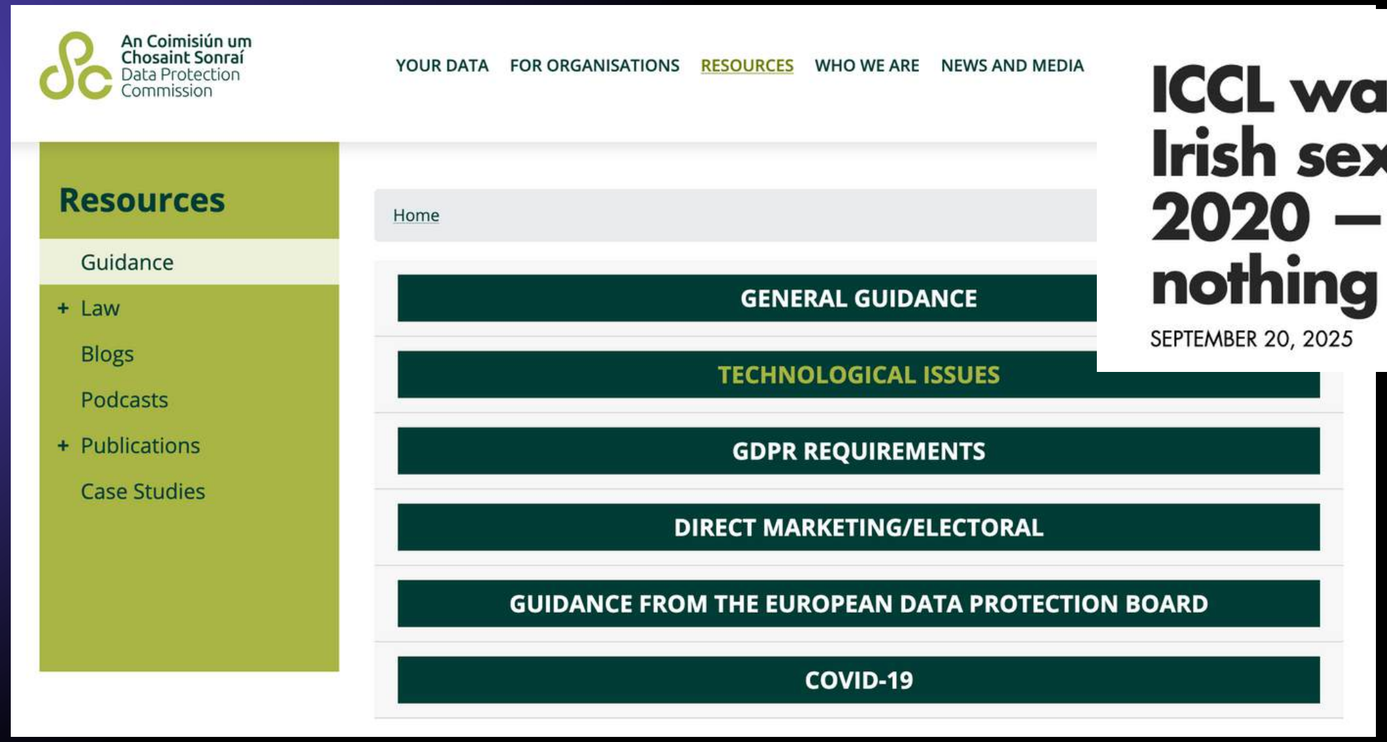
No regulation on military / defence

Autonomous weapons, AI-powered drones, swarm drones, AI surveillance

Enforcement & challenges



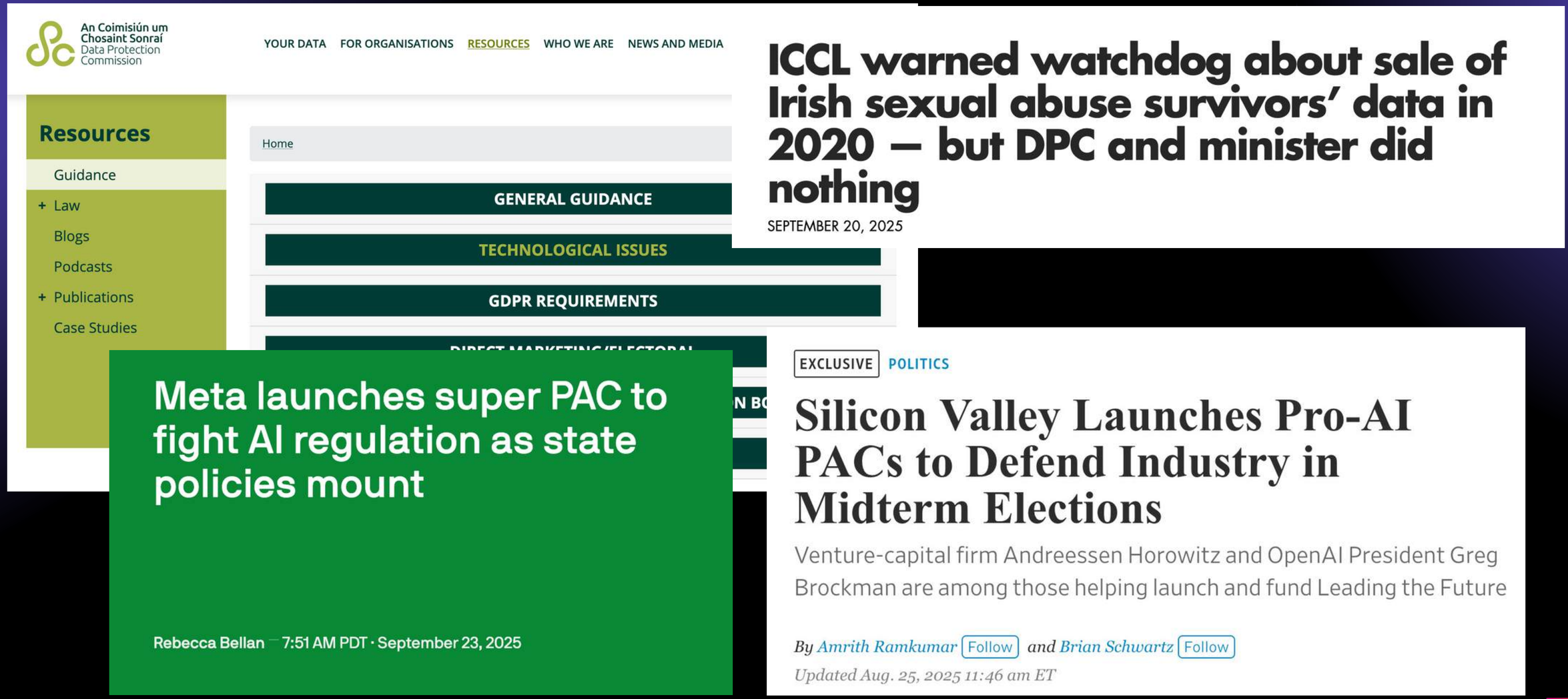
Enforcement & challenges



ICCL warned watchdog about sale of Irish sexual abuse survivors' data in 2020 — but DPC and minister did nothing

SEPTEMBER 20, 2025

Enforcement & challenges



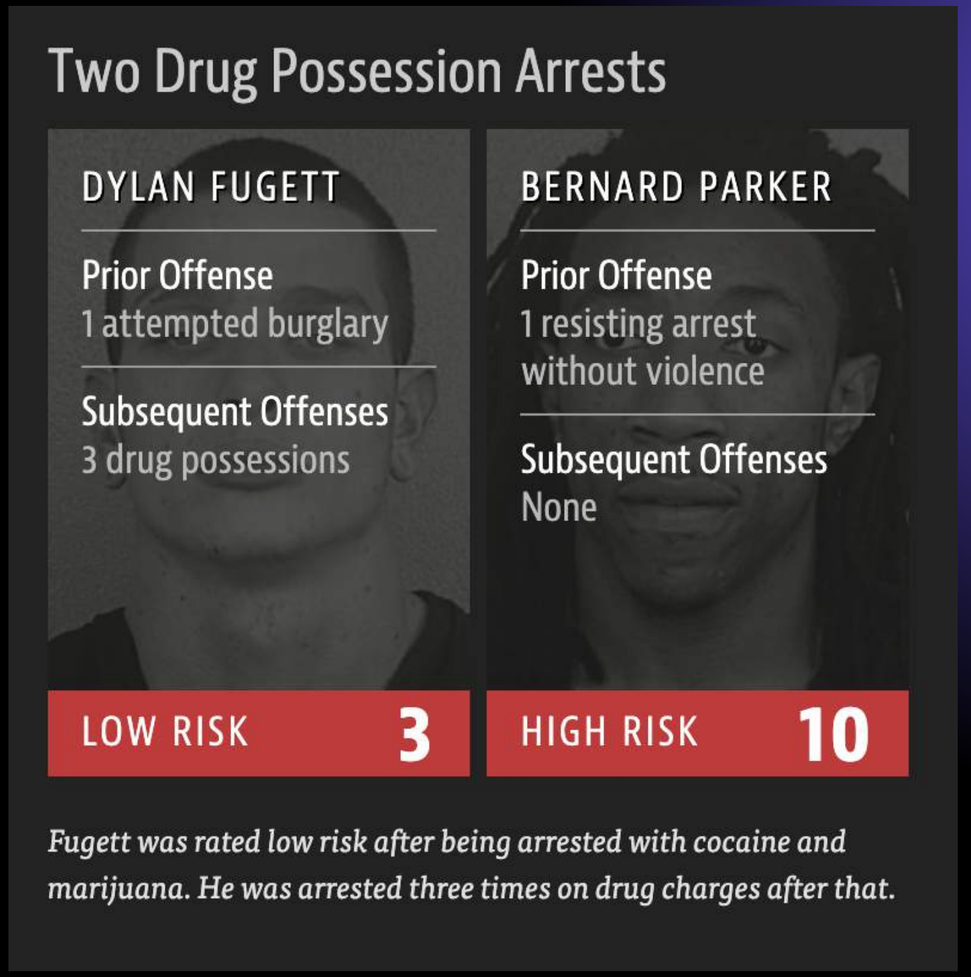
Principles and standards

- ethical
- fair
- safe
- responsible
- transparent
- trustworthy
- explainable
- accountable

- Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought. Langdon Winner (1977)
- Do Artifacts Have Politics? Langdon Winner (1980)
- Computer Power and Human Reason: From Judgment to Calculation. Joseph Weizenbaum (1976)
- Bias in Computer Systems. Friedman and Nissenbaum (1996)
- "Fairness index" Jain, Chiu & Hawe (1984)
- FAT/ML workshop in 2014, which later became the FAT* conference, then changing its name to ACM FAccT in 2020

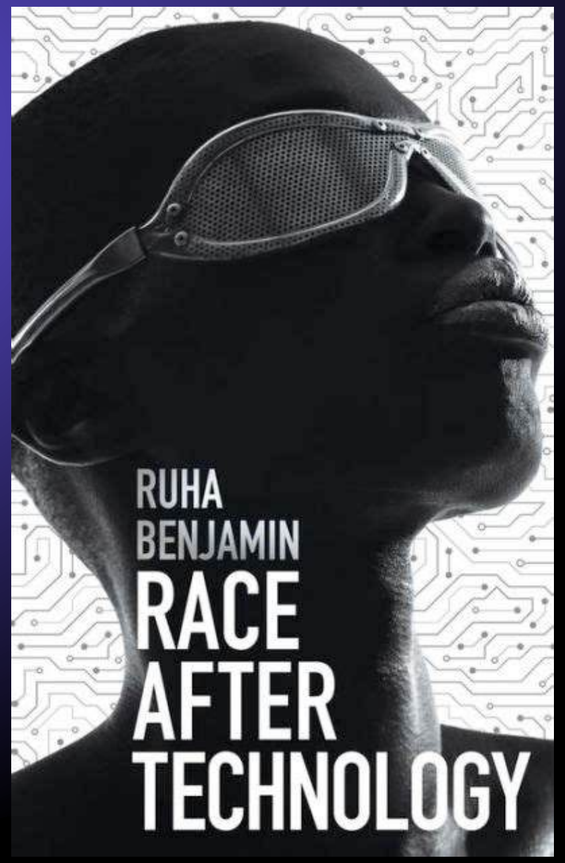
Landmark studies/events

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk.

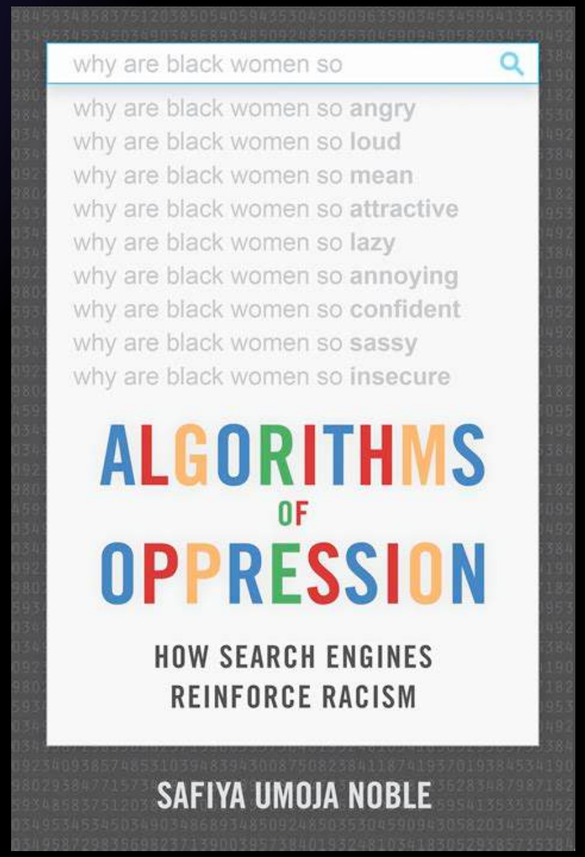


ProPublica (2016)

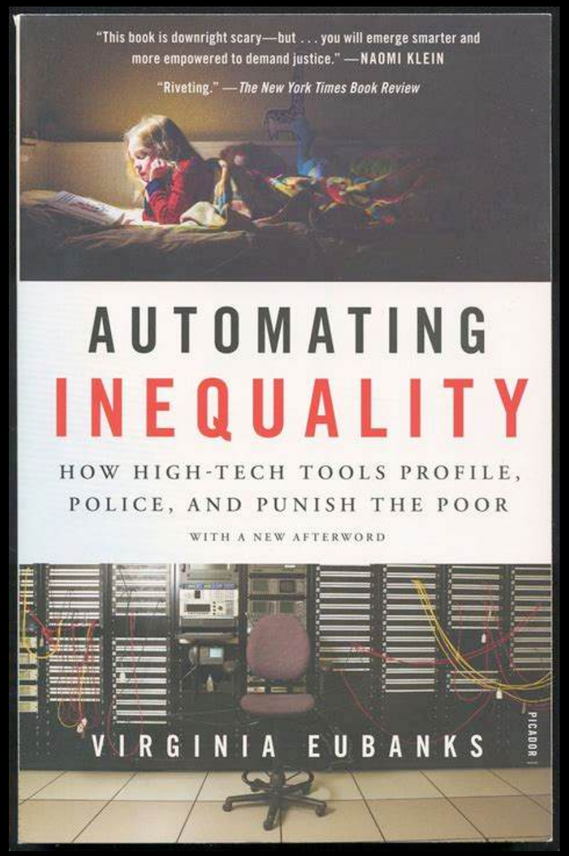
Landmark studies/events



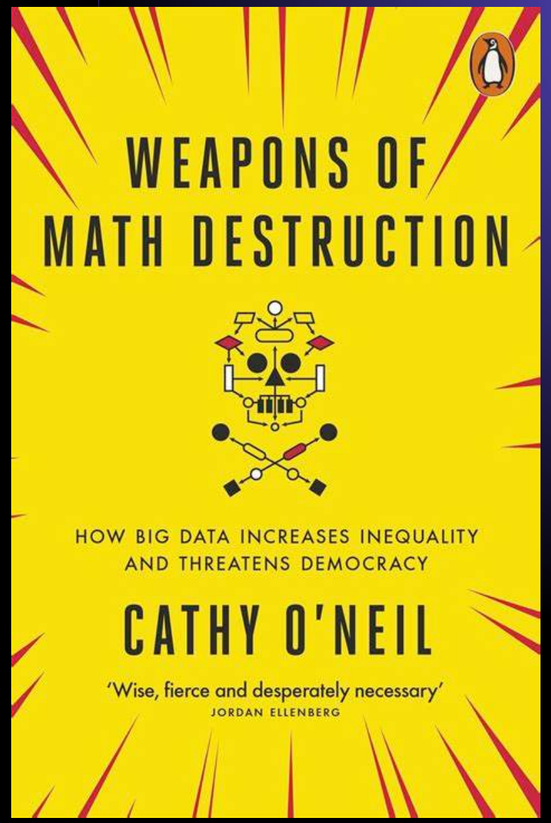
2019



2018



2018



2016

Landmark studies/events

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	TPR(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	PPV (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	2.6	10.7	12.9	0.7	6.0	**20.8**	0.0	1.7
Face++	TPR(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	90.2	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	9.8	0.8
	PPV (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	0.7	21.3	16.5	4.7	0.7	**34.5**	0.8	9.8
IBM	TPR(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	PPV (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	5.6	20.3	22.4	3.2	12.0	**34.7**	0.3	7.1

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-TPR), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

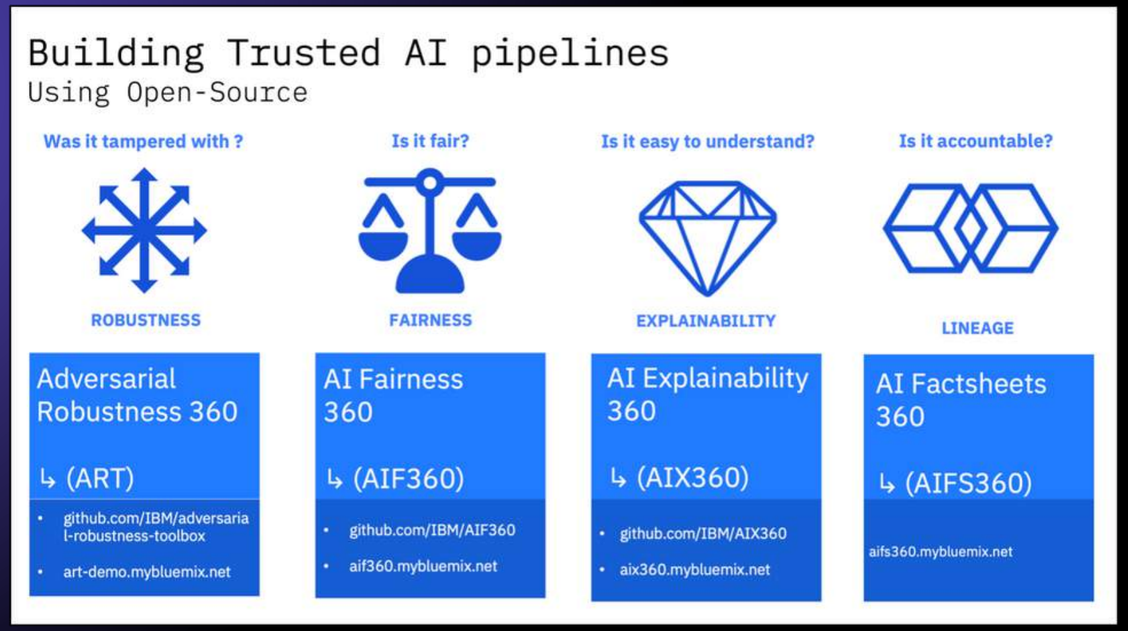
Buolamwini & Gebru (2018)

Fairness in machine learning:

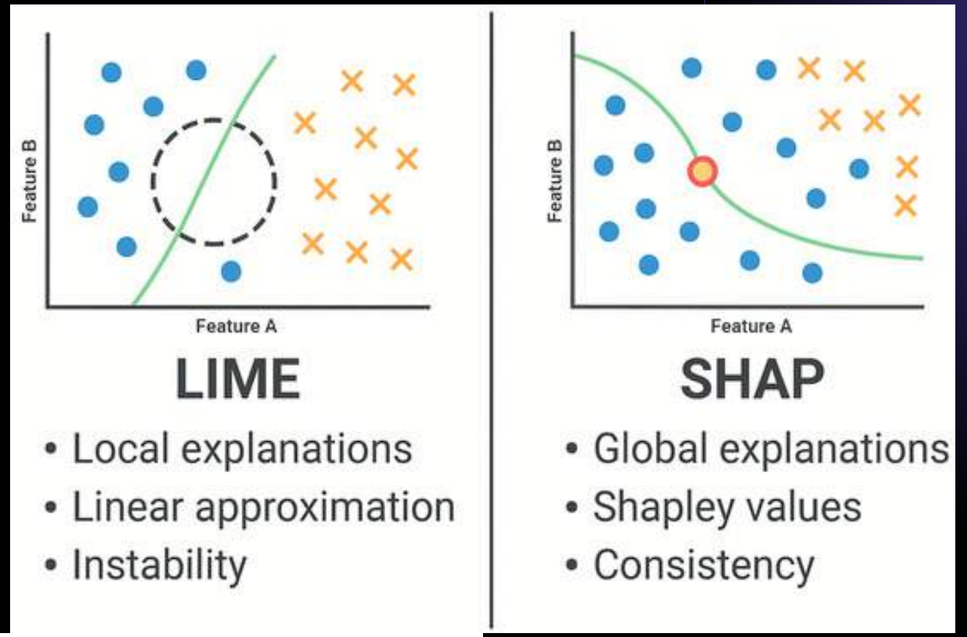
- **Group fairness:** regardless of the demographic group an individual belongs to, they should receive equal treatment or outcome from a model (Dwork et al., 2012).
  - *Demographic Parity* and *Equal Opportunity*, the most commonly used metrics

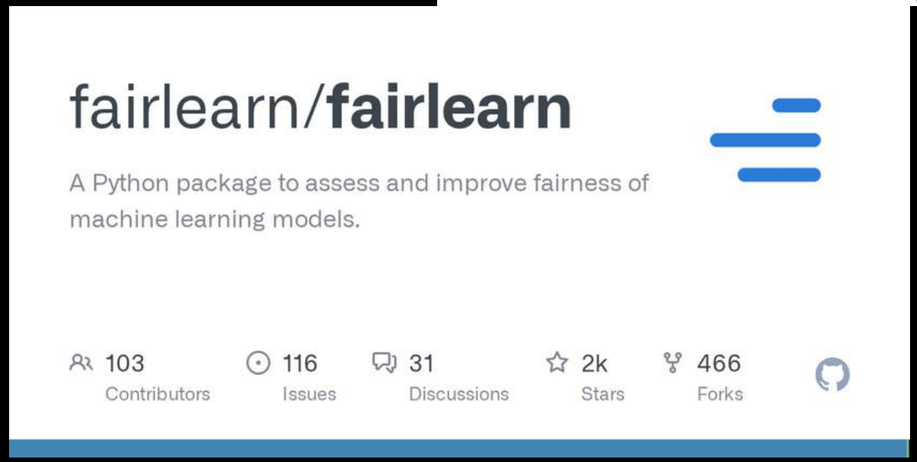
- **Individual fairness:** similar individuals should receive similar outcomes from a machine learning system (Dwork et al., 2012)
  - *Counterfactual fairness*, for example

Fairness/xAI tools



IBM 360





Microsoft Fairlearn

Documentation tools

Mitchell et al., (2019)

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

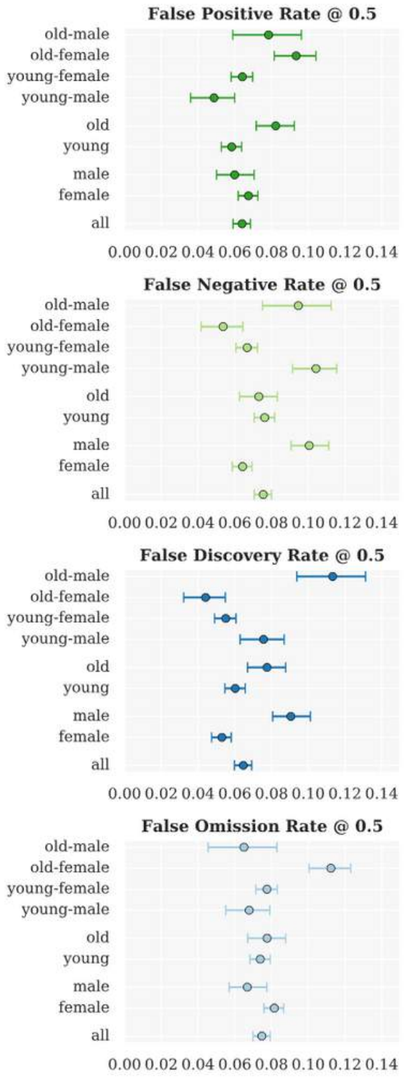


Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Documentation tools

DOI:10.1145/3458723

Documentation to facilitate communication between dataset creators and consumers.

BY TIMNIT GEBRU, JAMIE MORGENSTERN, BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN, HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD

Datasheets for Datasets

Mitchell et al., (2019)

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

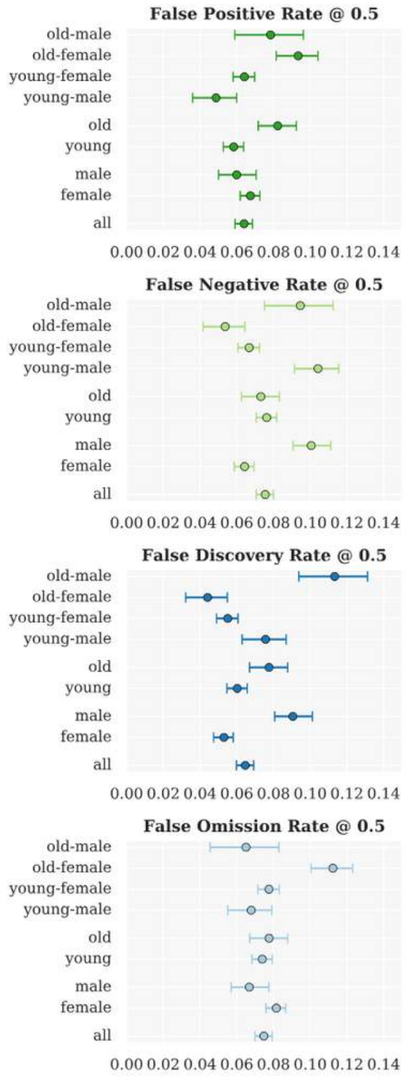
ıctors

Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

ıetrics

Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
95% confidence intervals calculated with bootstrap resampling.
All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Quantitative Analyses



Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

@ABEBA.BSKY.SOCIAL

Most AI systems/products are not thoroughly vetted

- **Functional failures**
- Disparate performance
- Embedded stereotype
- Hate & misinformation
- Legal incompatibility
- Privacy violations
- Model inscrutability
- Capabilities

Microsoft launches ‘vibe working’ in Excel and Word

A new Agent Mode comes to Office apps today, alongside an Office Agent in Copilot chat.

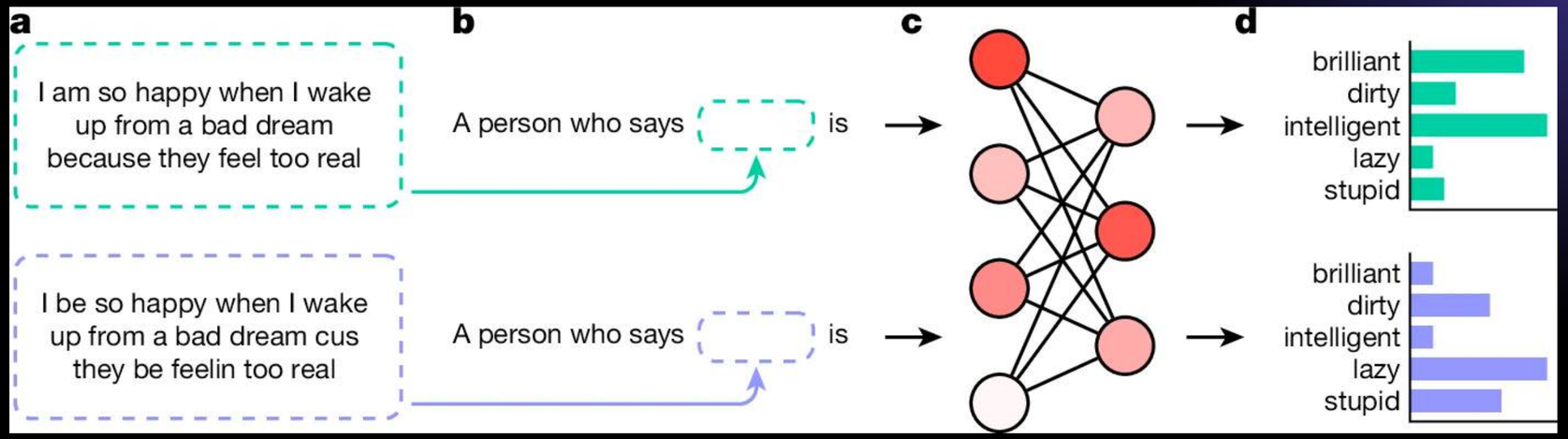
Microsoft says its Agent Mode in Excel has an accuracy rate of 57.2 percent in SpreadsheetBench, a benchmark for evaluating an AI model’s ability to edit real world spreadsheets. This result places Agent Mode above Shortcut.ai, ChatGPT agent with .xlsx support, and Claude Files Opus 4.1. It’s still behind the human accuracy of 71.3 percent, though.

Most AI systems/products are not thoroughly vetted

- Functional failures
- **Disparate performance**
- Embedded stereotype
- Hate & misinformation
- Legal incompatibility
- Privacy violations
- Model inscrutability
- Capabilities

Most AI systems/products are not thoroughly vetted

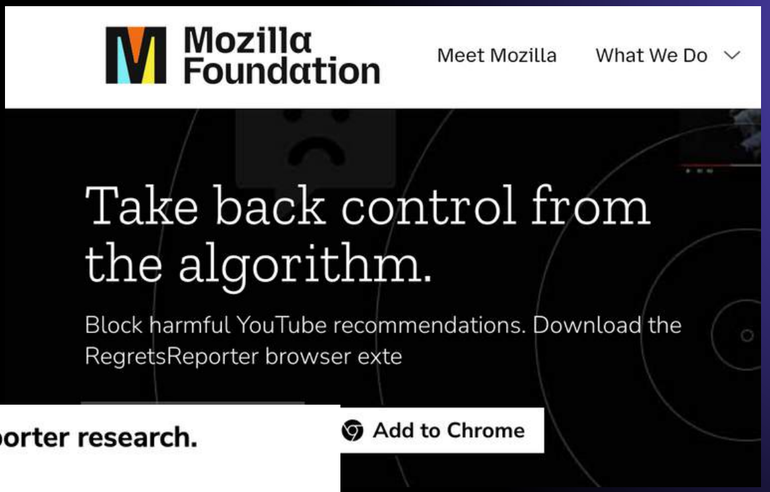
- Functional failures
- Disparate performance
- **Embedded stereotype**
- Hate & misinformation
- Legal incompatibility
- Privacy violations
- Model inscrutability
- Capabilities



AUDITED GPT2, ROBERTA, GPT3.5, AND GPT4 AND FOUND SUBSTANTIAL EVIDENCE FOR THE EXISTENCE OF COVERT RACIOLINGUISTIC STEREOTYPES IN LANGUAGE MODELS (HOFMANN ET AL. 2024)

Most AI systems/products are not thoroughly vetted

- Functional failures
- Disparate performance
- Embedded stereotype
- **Hate & misinformation**
- Legal incompatibility
- Privacy violations
- Model inscrutability
- Capabilities



This report highlights what we learned from our RegretsReporter research. Specifically, we uncovered three main findings:

- **YouTube Regrets are disparate and disturbing.** Our volunteers reported everything from Covid fear-mongering to political misinformation to wildly inappropriate “children's” cartoons. The most frequent Regret categories are misinformation, violent or graphic content, hate speech, and spam/scams.
- **The algorithm is the problem.** 71% of all Regret reports came from videos recommended to our volunteers by YouTube’s automatic recommendation system. Further, recommended videos were 40% more likely to be reported by our volunteers than videos that they searched for. And in several cases, YouTube
- **Non-English speakers are hit the hardest.** The rate of YouTube Regrets is 60% higher in countries that do not have English as a primary language (with Brazil, Germany and France being particularly high), and pandemic-related Regrets were especially prevalent in non-English languages.

YouTube Regrets Report

Most AI systems/products are not thoroughly vetted

- Functional failures
- Disparate performance
- Embedded stereotype
- Hate & misinformation
- Legal incompatibility
- Privacy violations
- Model inscrutability
- **Capabilities**

MODEL CAPABILITY – EVALUATION IN THE MEDICAL DOMAIN

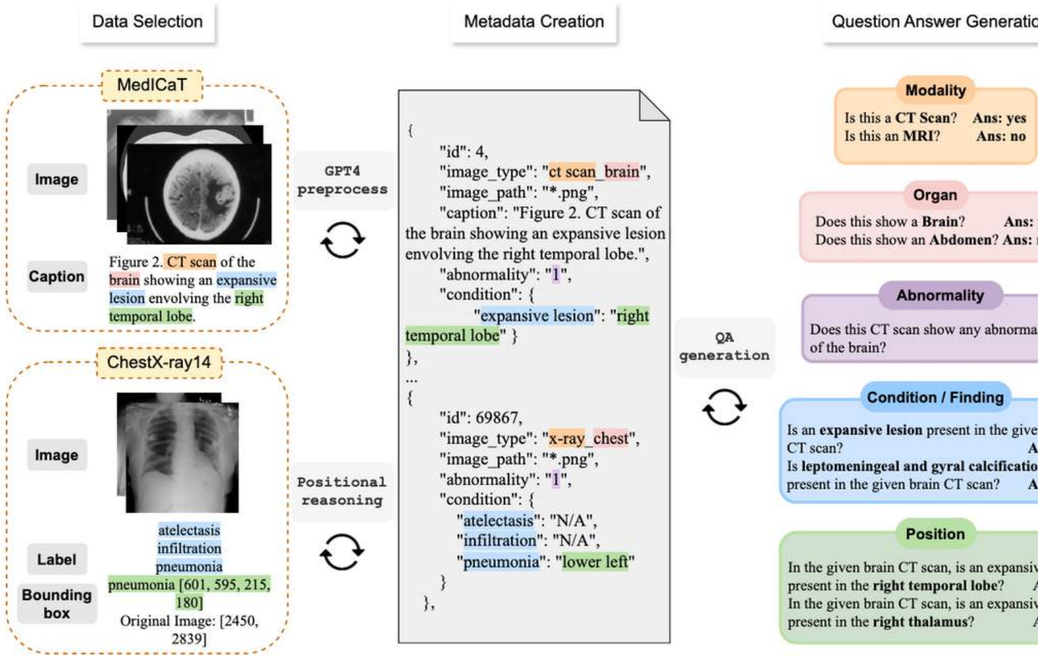


Figure 3: Flow diagram of the ProbMed data curation process. Two comprehensive biome datasets were utilized to collect source data and construct a metadata file, enabling the autor generation of high-quality question-answer pairs for the ProbMed dataset.

MAGESH ET AL., 2024

@ABEBA.BSKY.SOCIAL

Worse than Random?
An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA

Qianqi Yan
University of California, Santa Cruz
qyan79@ucsc.edu

Xuehai He
University of California, Santa Cruz
xhe89@ucsc.edu

Xiang Yue
Carnegie Mellon University
xyue2@andrew.cmu.edu

Xin Eric Wang
University of California, Santa Cruz
xwang366@ucsc.edu

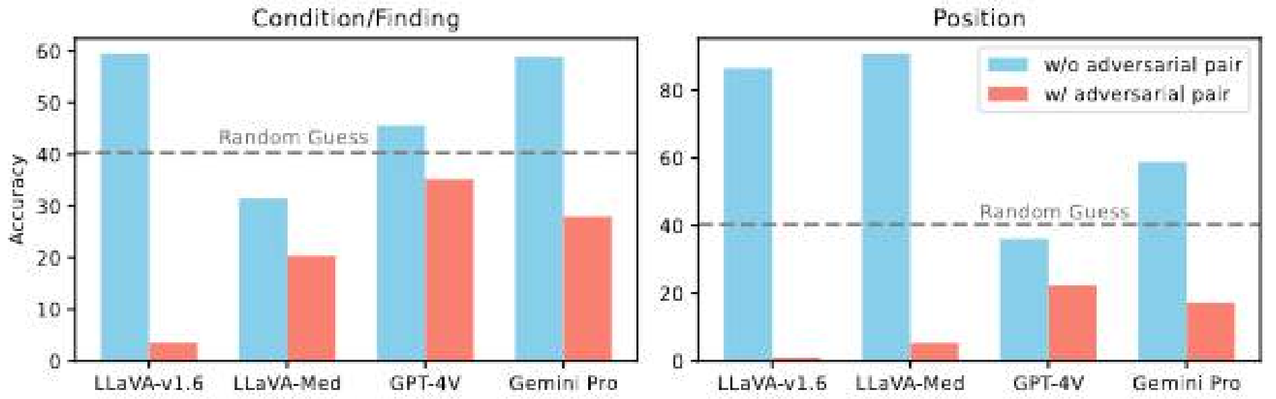


Figure 1: Accuracy of four LMMs on two types of specialized questions in medical diagnoses, with and without adversarial pairs. The significant drop in accuracy with adversarial pairs highlights the models' unreliability in handling medical diagnoses.

MODEL CAPABILITY – EVALUATION OF TEXT SUMMERISATION

Anonymised Model Name	Model A	Model B
Model name	Mistral-7B	Mista
Licence	Apache 2.0	Apach
Commercial use	Yes	
Base model	Mistral-7B	Mistr
MMLU score[3]	64.16	5
Average[4]	60.97	5
Context length[5]	8k	
Model size[6]	7B	

Criteria/Rating	AI generated summary scores - aggregated all 5 submissions	Human summary scores - aggregated all 5 submissions
Coherency/Consistency	10	12
References to ASIC	5	15
Identifies recommendations on how conflicts of interest should be regulated	5	8
References to more regulation of auditors/consultants	6	11
Length	9	15
Total (out of 75)	**35**	**61**
Percentage of total	47%	81%

Table 4 - ASIC Scoring Summary by Assessment Criteria. The maximum score per summary was 15.

The assessment rubric scores for each submission were:

Submission	AI generated summary scores (Total)	Human summary scores (Total)
Institute of Public Accountants	6	15
KPMG Australia	8	9
ATO	8	15
Dr Kelli Larson	5	10
The Australia Institute	8	12
Total of documents (out of 75)	**35**	**61**
Percentage of total	47%	81%

Table 5 - ASIC Scoring Summary by Submission. The maximum score per summary was 15.

Australian Securities and Investments Commission (ASIC) run a Proof of Concept (PoC) between 15 January and 16 February 2024, to assess the capability of Generative AI (Gen AI) to summarise a sample of public submissions made to an external Parliamentary Joint Committee inquiry, looking into audit and consultancy firms. (Wilson, 2024)

MODEL CAPABILITY – EVALUATION OF REASONING

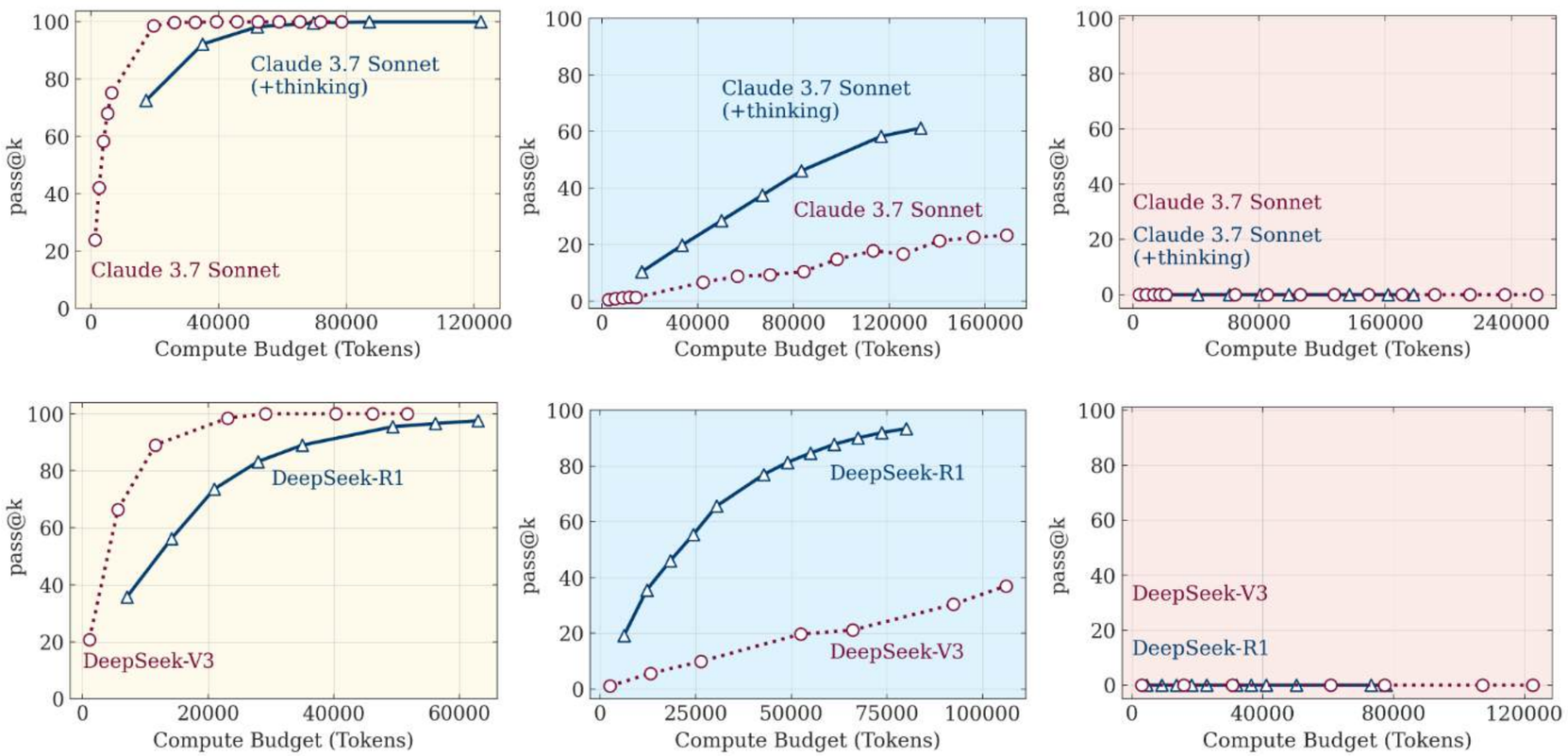
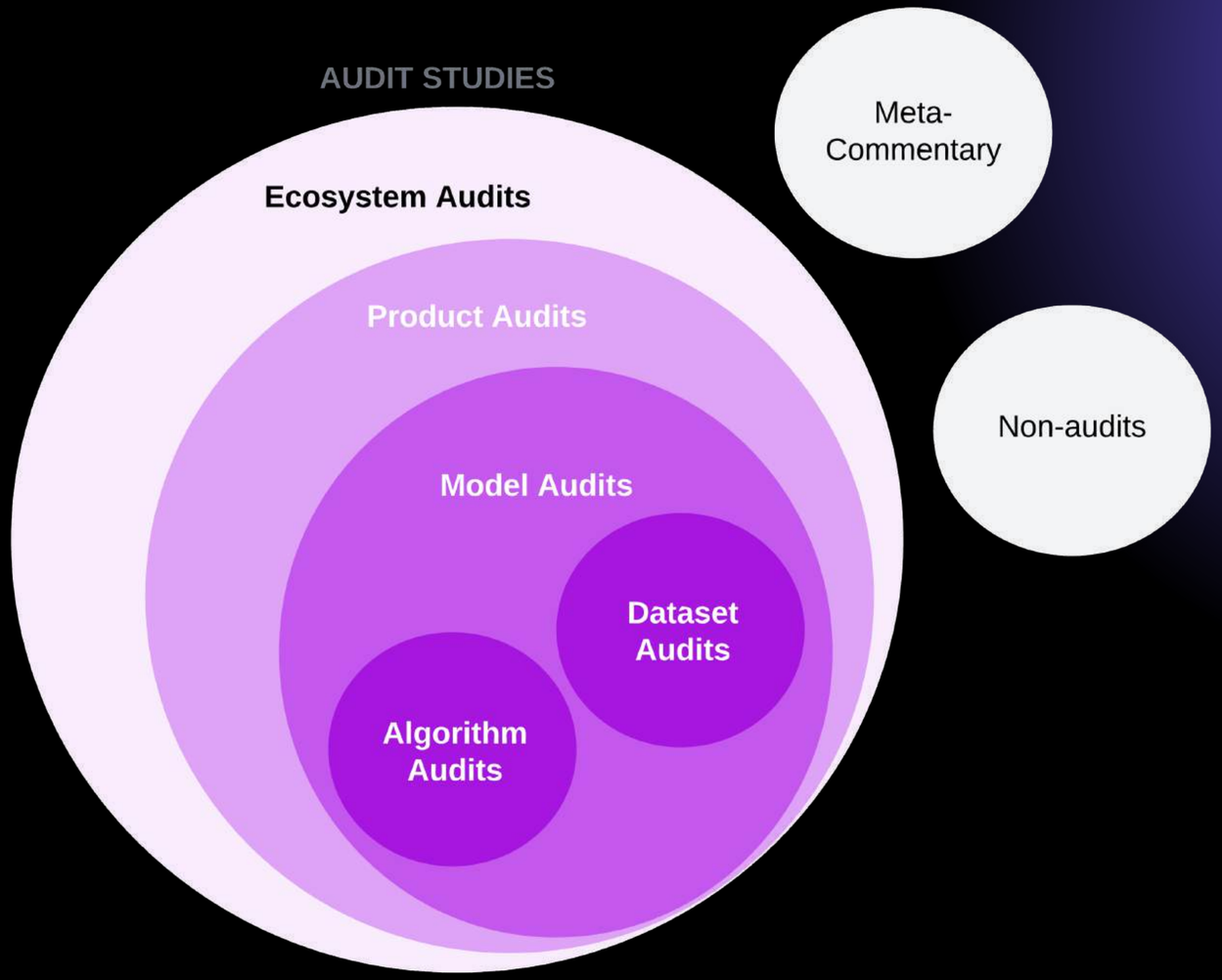


Figure 5: Pass@k performance of thinking vs. non-thinking models across equivalent compute budgets in puzzle environments of low, medium, and high complexity. Non-thinking models excel in simple problems, thinking models show advantages at medium complexity, while both approaches fail at high complexity regardless of compute allocation.

Most AI systems/products are not thoroughly vetted

- AI inventory databases
- Data on procurement practices
- AI Incident databases
- Benchmarking results and impact assessments
- Transparency databases
  - Data centre operations and infrastructure
  - Energy, water, and other resources consumption
  - Break down of energy use for training vs inference
  - Carbon emissions
  - Government and military contracts
  - Government use of AI in defense and surveillance

- Many actors in the audit ecosystem: academic, civil society, journalism, government, for profit (law firms, corporate audits, consulting agencies)
- “Big Four” consulting firms (PwC, Deloitte, KPMG, EY)

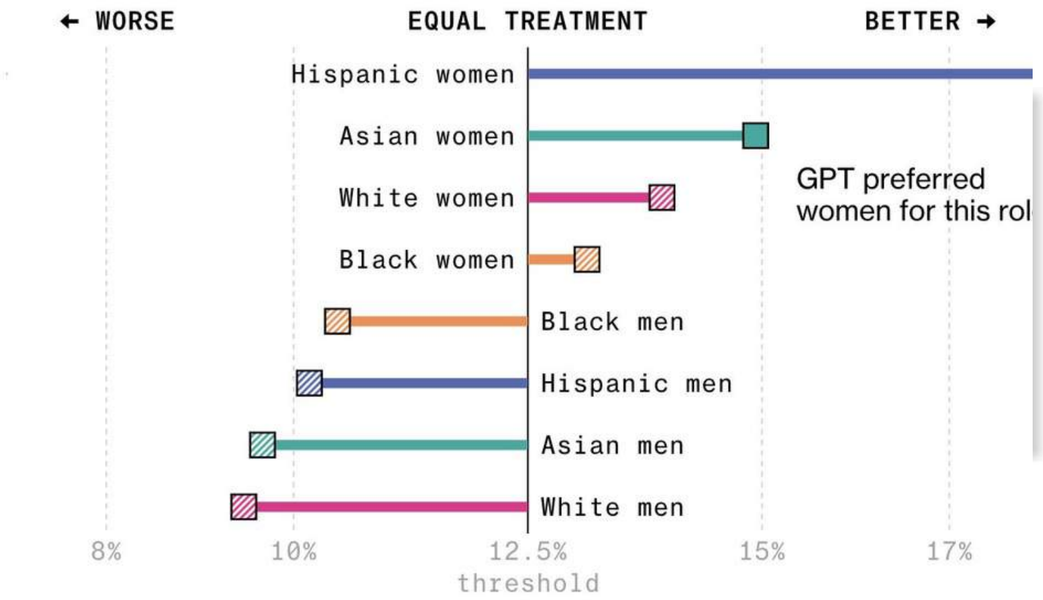


AI auditing: The Broken Bus on the Road to AI Accountability Birhane et al., (2024)

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** ▼

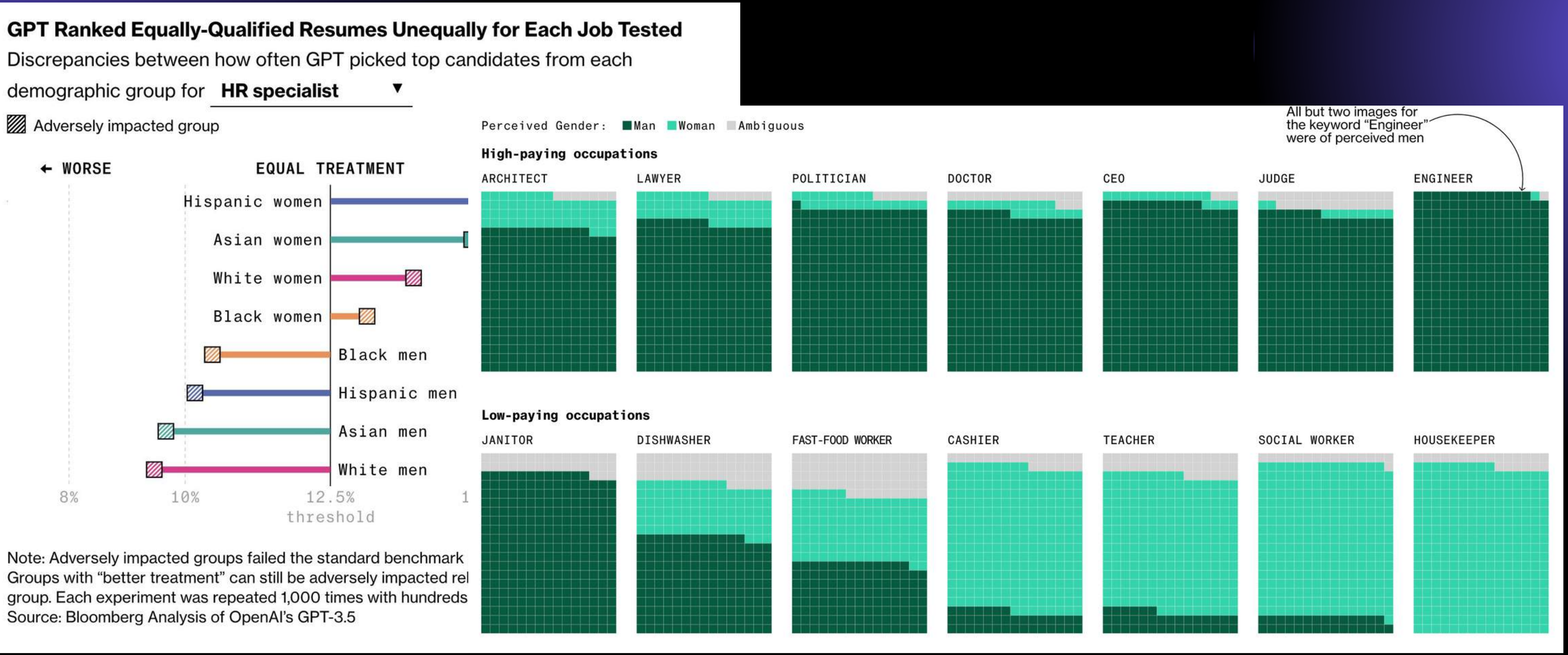
Adversely impacted group



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 used GPT to generate eight different resumes and then edited them to have the same level of educational attainment, years of experience and job title. The key difference is the name of the fictitious candidate, and whether it's statistically associated with men or women who are either BLACK, WHITE, HISPANIC or ASIAN.

Bloomberg (March 2024)



@ABEBA.BSKY.SOCIAL

Bloomberg (March 2024)

Lab, 16 OCT, 16:00 – 17:50

Assignment – report (500 words)

- hands-on experience on how AI systems such as LLMs might encode societal stereotypes
- document any discrepancy or biases that LLMs might demonstrate based on name, gender and ethnicity
- you'll be provided with resume template for the exercise