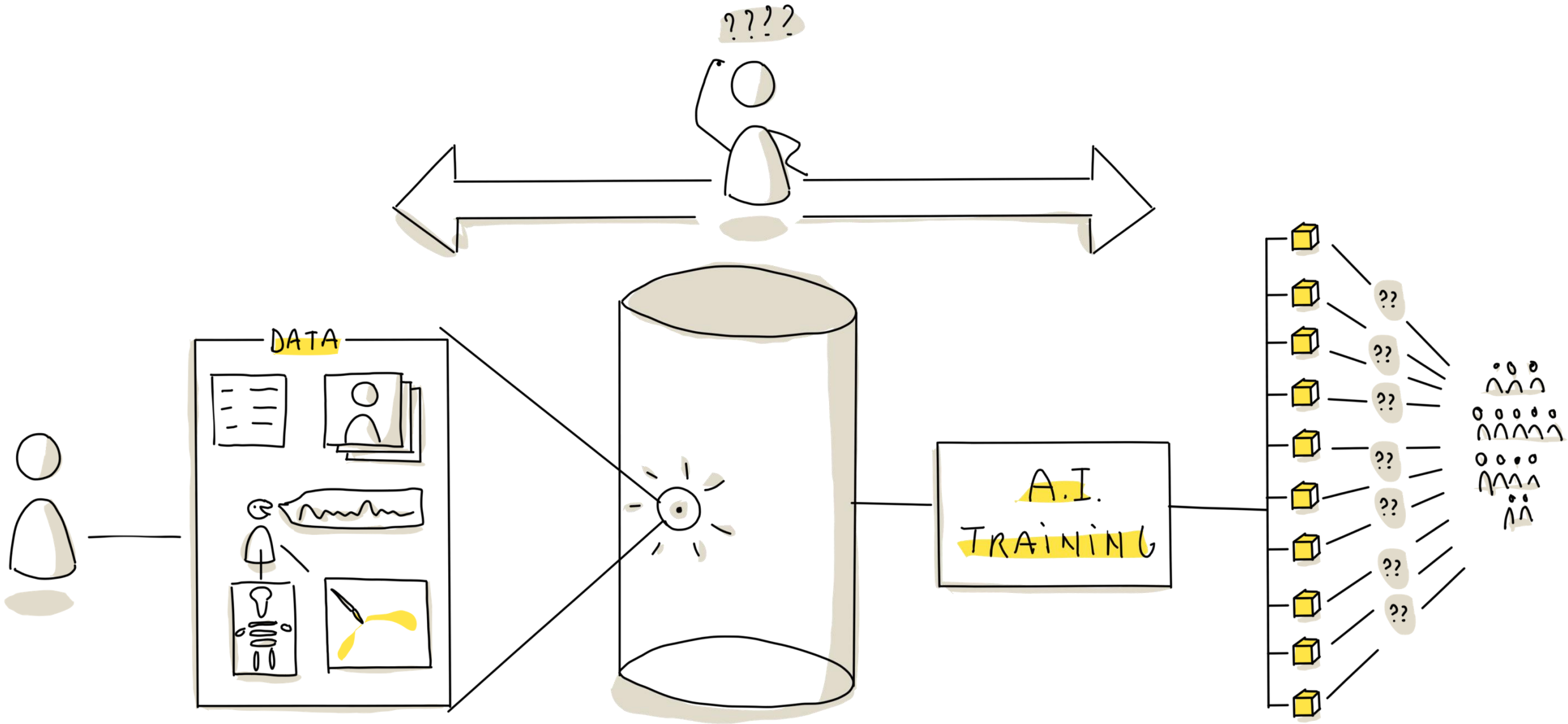
DEEL 11

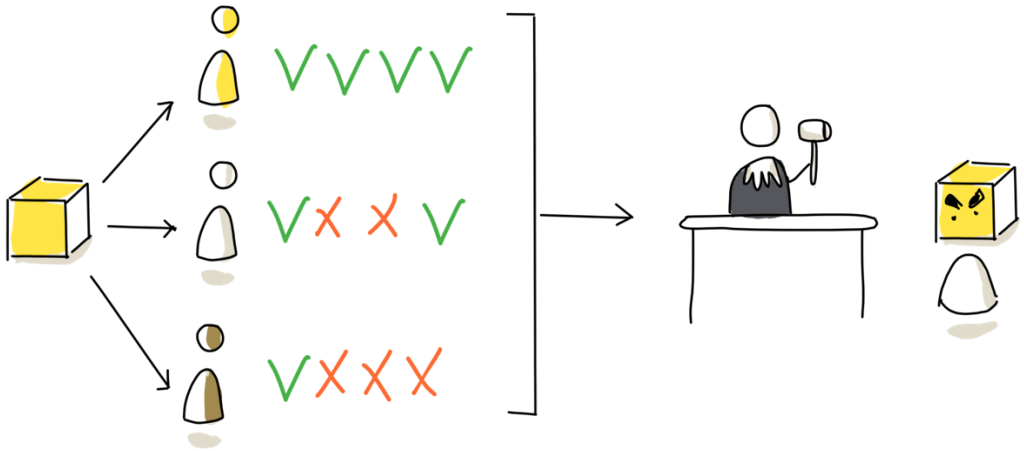
PRIVACY & ETHICS

Privacy



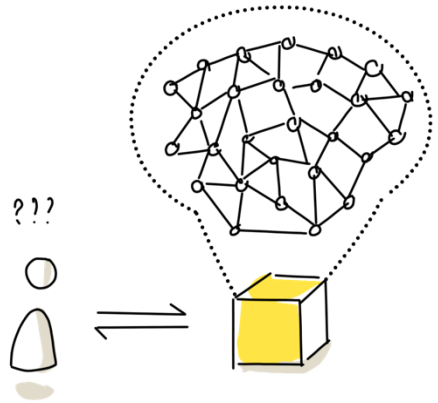
Ethics

DISCRIMINATION

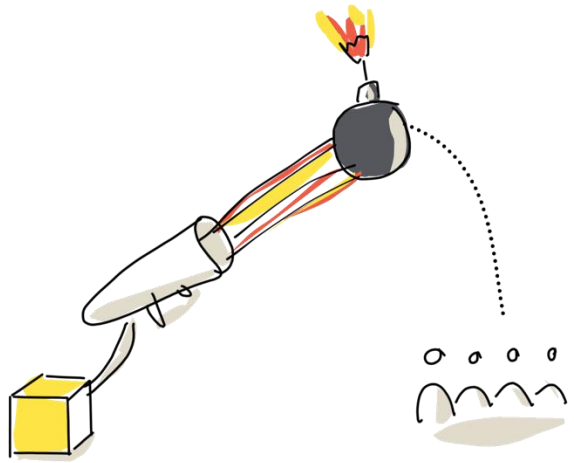


ACCOUNTABILITY

TRANSPARENCY



HUMAN THREAT

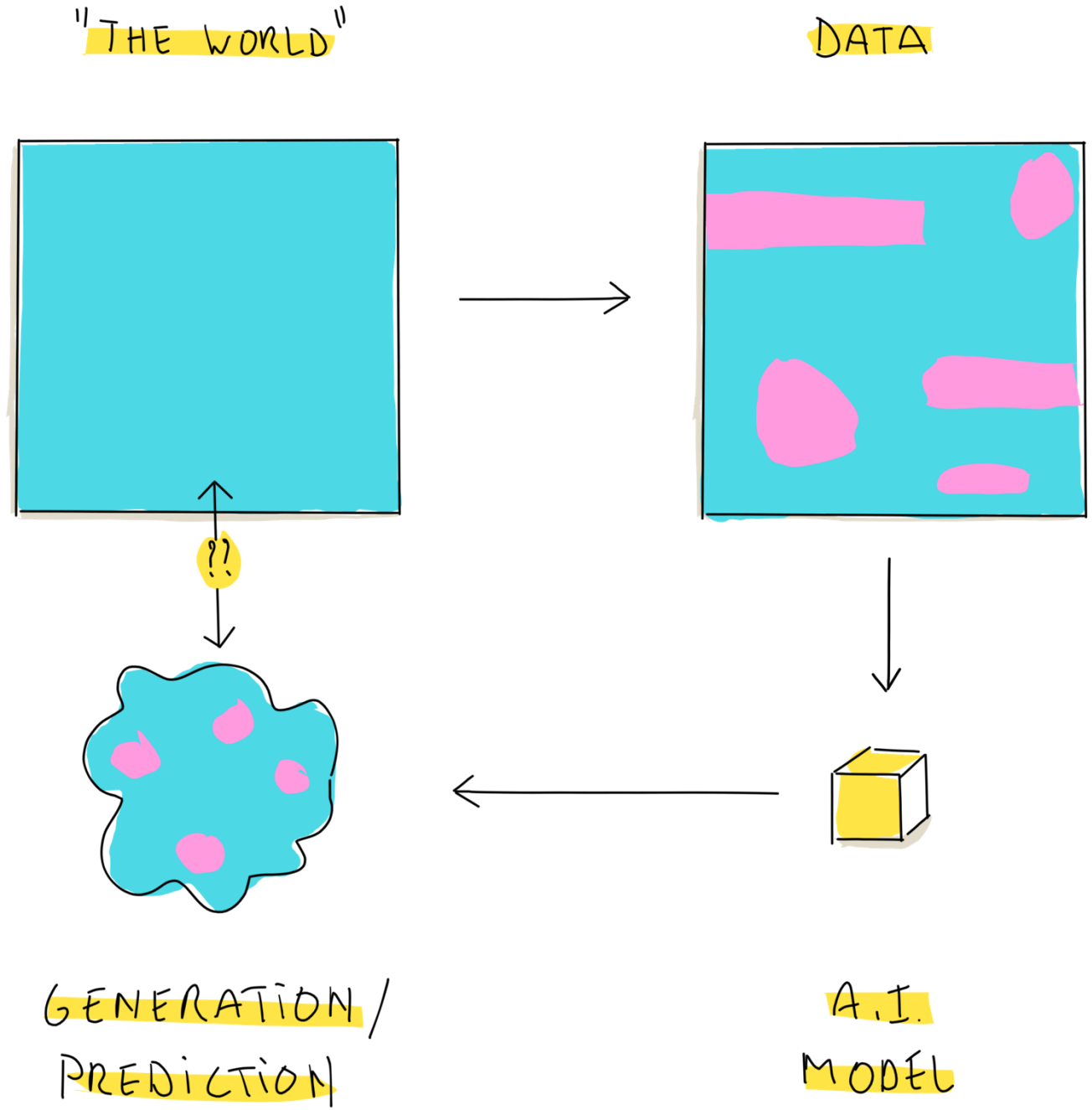


Privacy & Ethics

- Bias
- Transparency & Explainability
- Copyright
- Manipulation
- Trust and Accountability
- Rules & Regulations

Privacy & Ethics

- **Bias**
- Transparency & Explainability
- Copyright
- Manipulation
- Trust and Accountability
- Rules & Regulations


"THE WORLD"
DATA
??
GENERATION/
PREDICTION
A.I.
MODEL

Survivorship Bias

Sampling Bias

Algorithmic Bias

Confirmation Bias

Observer Bias

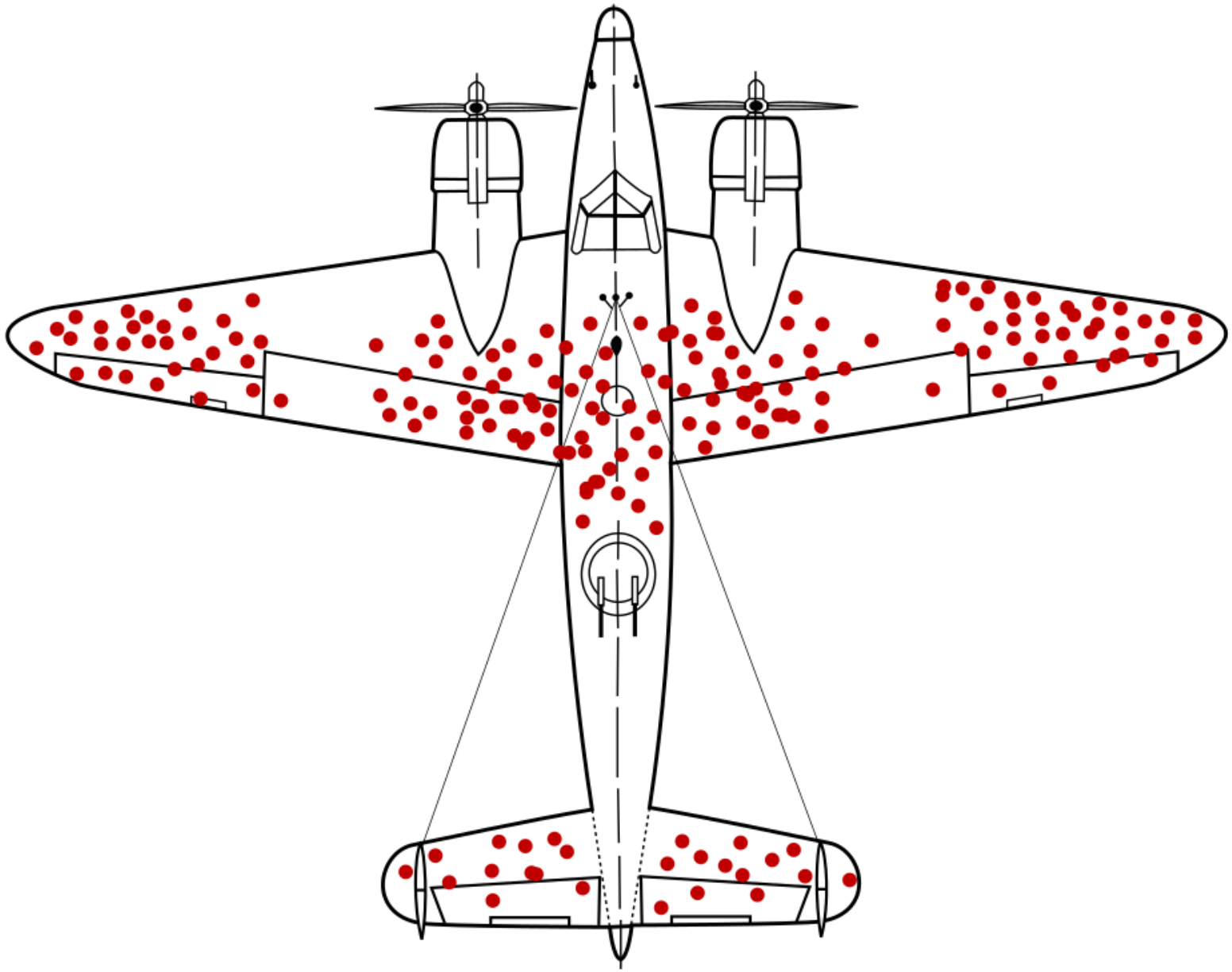
Survivorship Bias

Sampling Bias

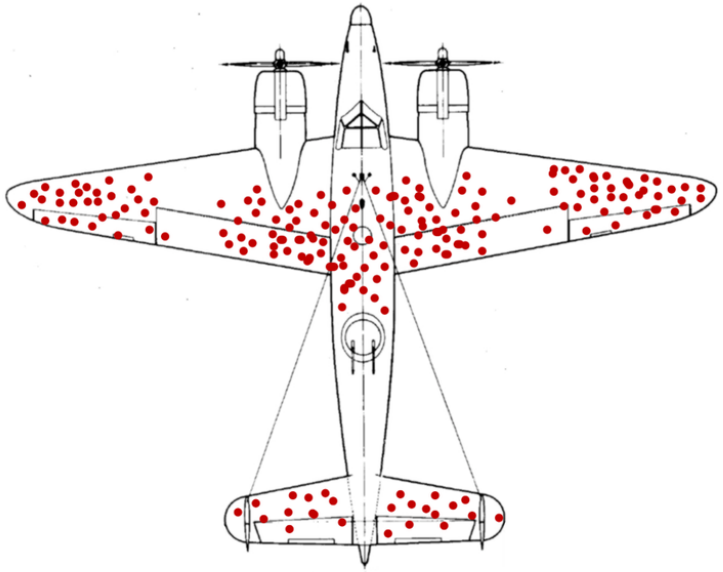
Algorithmic Bias

Confirmation Bias

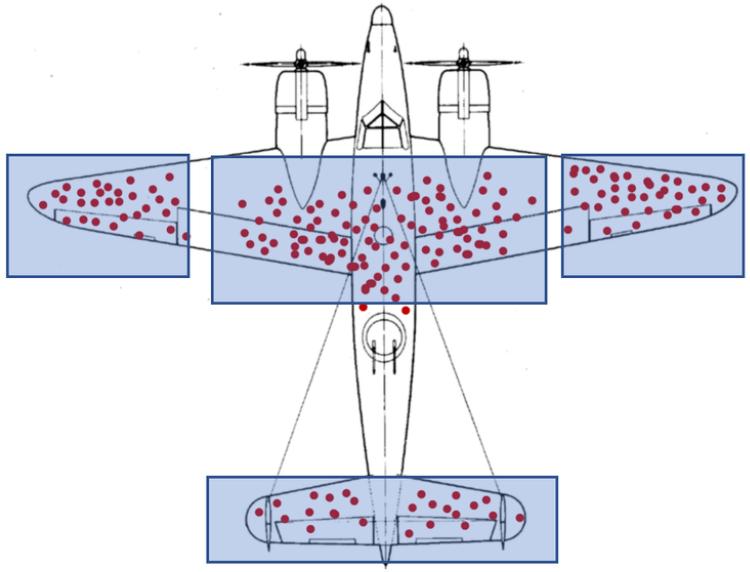
Observer Bias



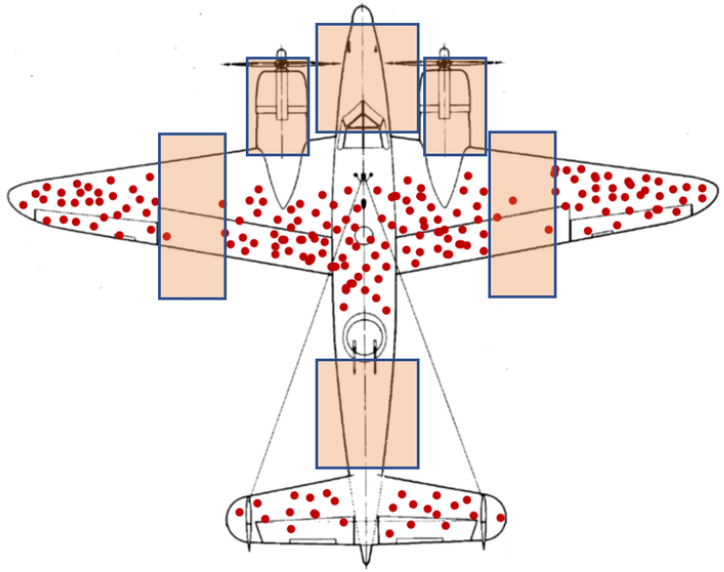
[A. Wald, World War II]



Our data if only from returning flights. Here we is a visualization of the places that bullet holes were observed.

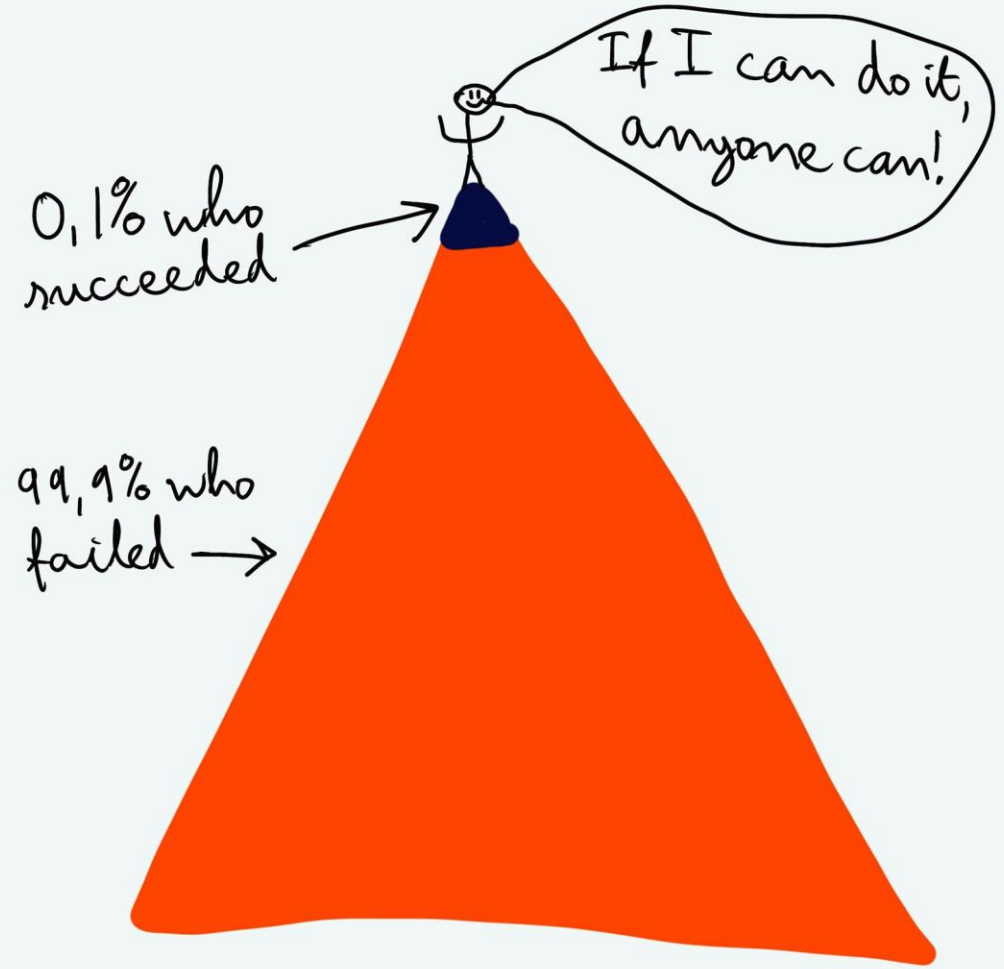


And initial guess at how to fix this might be to apply additional armor platting to the parts of the plane with the most holes…

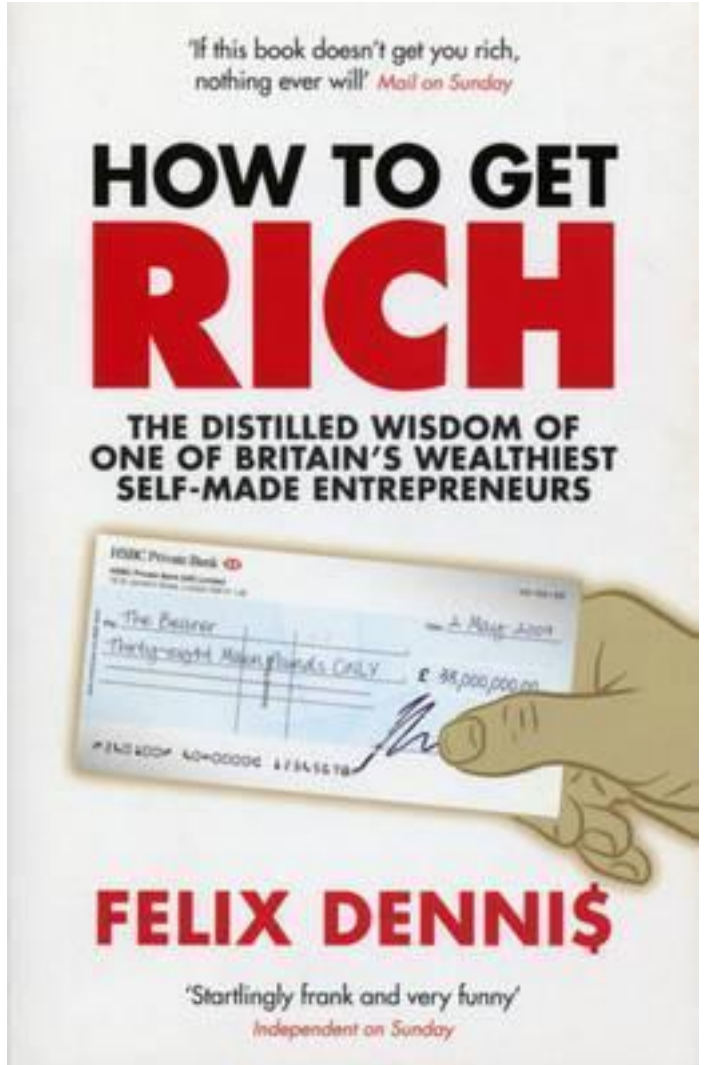


…. However this is where planes that *returned* had bullet holes. The planes we want to protect are the ones that did *not* return, so we should place armor there.

SURVIVORSHIP BIAS




SKYVUE
RISE


'If this book doesn't get you rich, nothing ever will' Mail on Sunday
HOW TO GET
RICH
THE DISTILLED WISDOM OF ONE OF BRITAIN'S WEALTHIEST SELF-MADE ENTREPRENEURS
FELIX DENNIS
'Startlingly frank and very funny'
Independent on Sunday

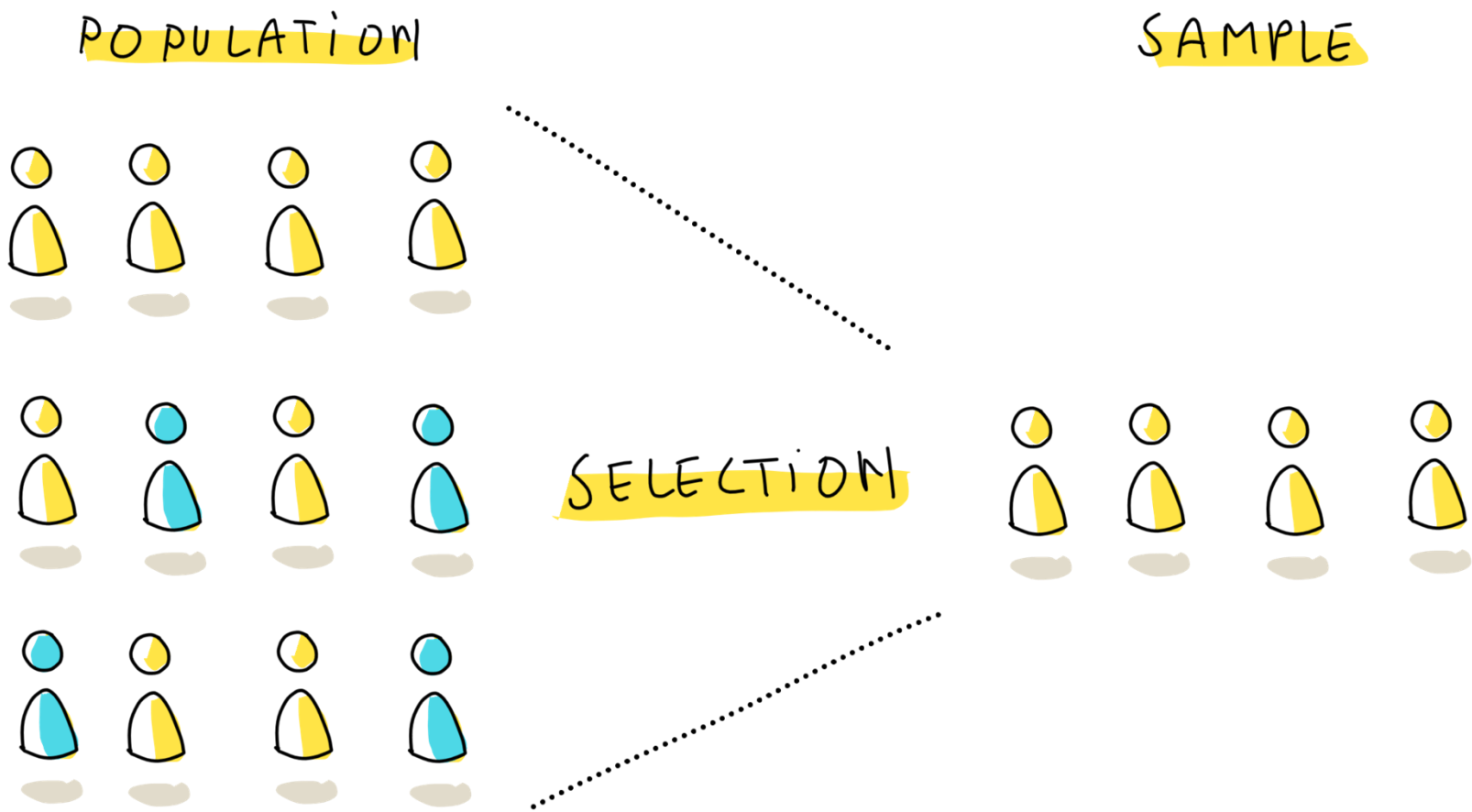
Survivorship Bias

Sampling Bias

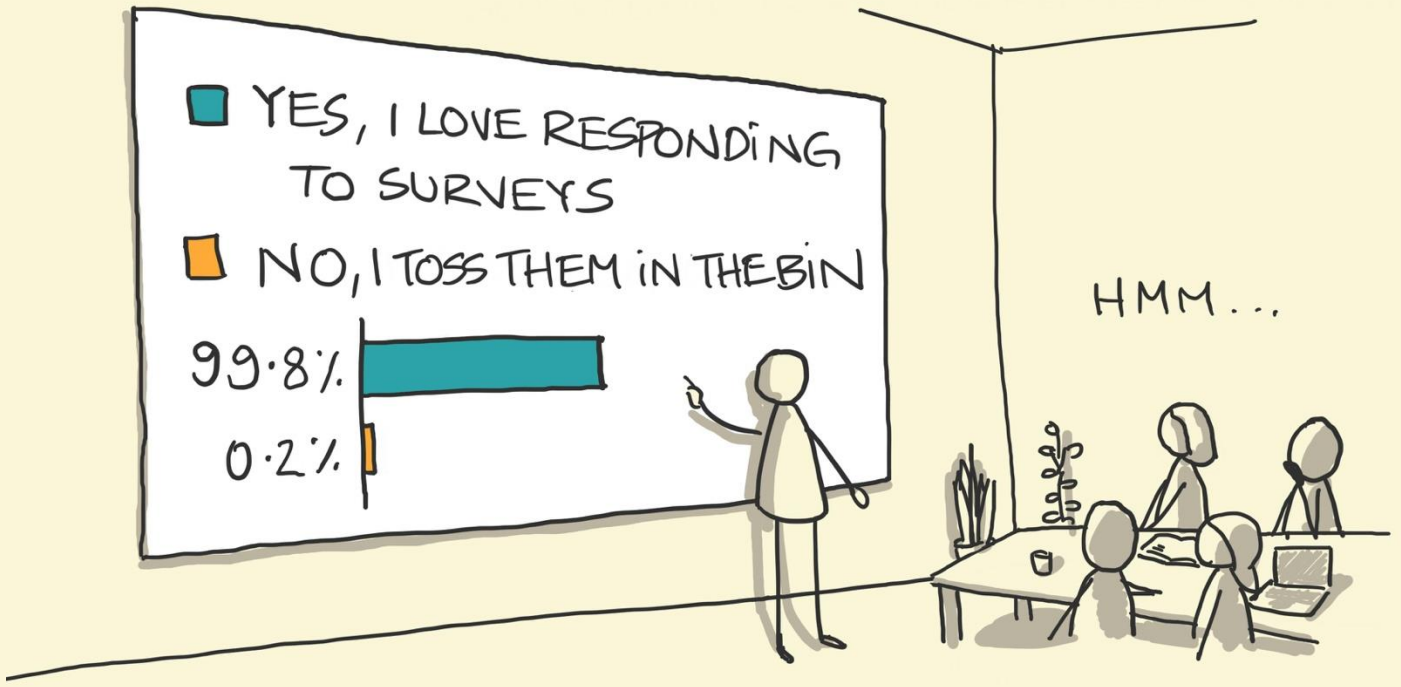
Algorithmic Bias

Confirmation Bias

Observer Bias


POPULATION
SAMPLE
SELECTION

SAMPLING BIAS



" WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS "

sketchplanations

Types of selection bias

Sampling bias

Occurs when the sample isn't representative of the target population, leading to skewed results.

Attrition bias

Occurs when participants drop out, making the remaining group no longer representative of the original sample.

Survivorship bias

Occurs when only successful observations are studied, ignoring those that did not succeed, leading to skewed conclusions.

Recall bias

Occurs when participants do not accurately remember past events or experiences, skewing the data based on current feelings or beliefs.

Self-selection bias

When individuals volunteer to participate, it often leads to the overrepresentation of more engaged or motivated users.

Nonresponse bias

When certain individuals or groups don't respond to surveys, the results reflect only the views of those who responded.

Undercoverage bias

Happens when some members of the population are inadequately represented in the sample.

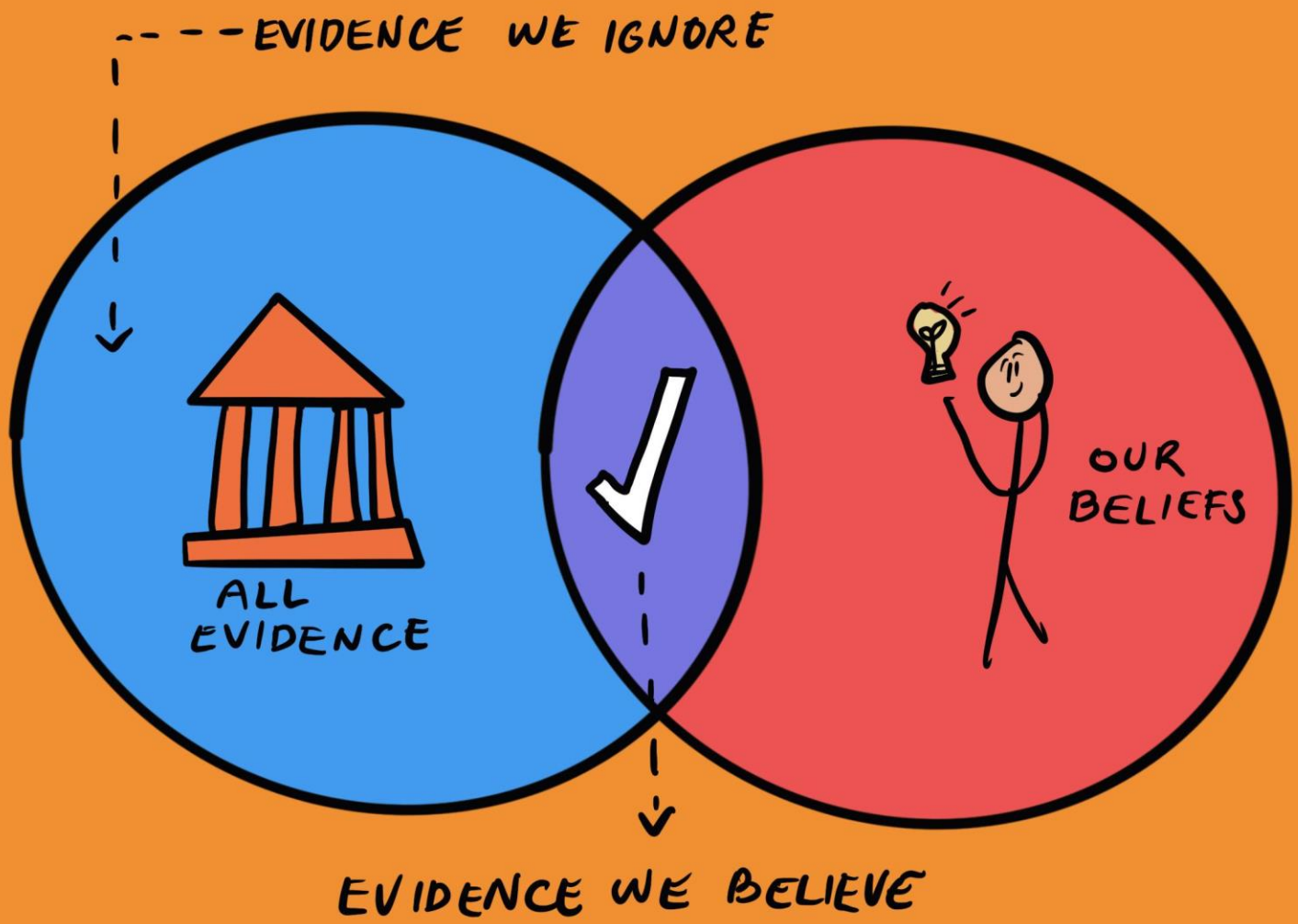
Survivorship Bias

Sampling Bias

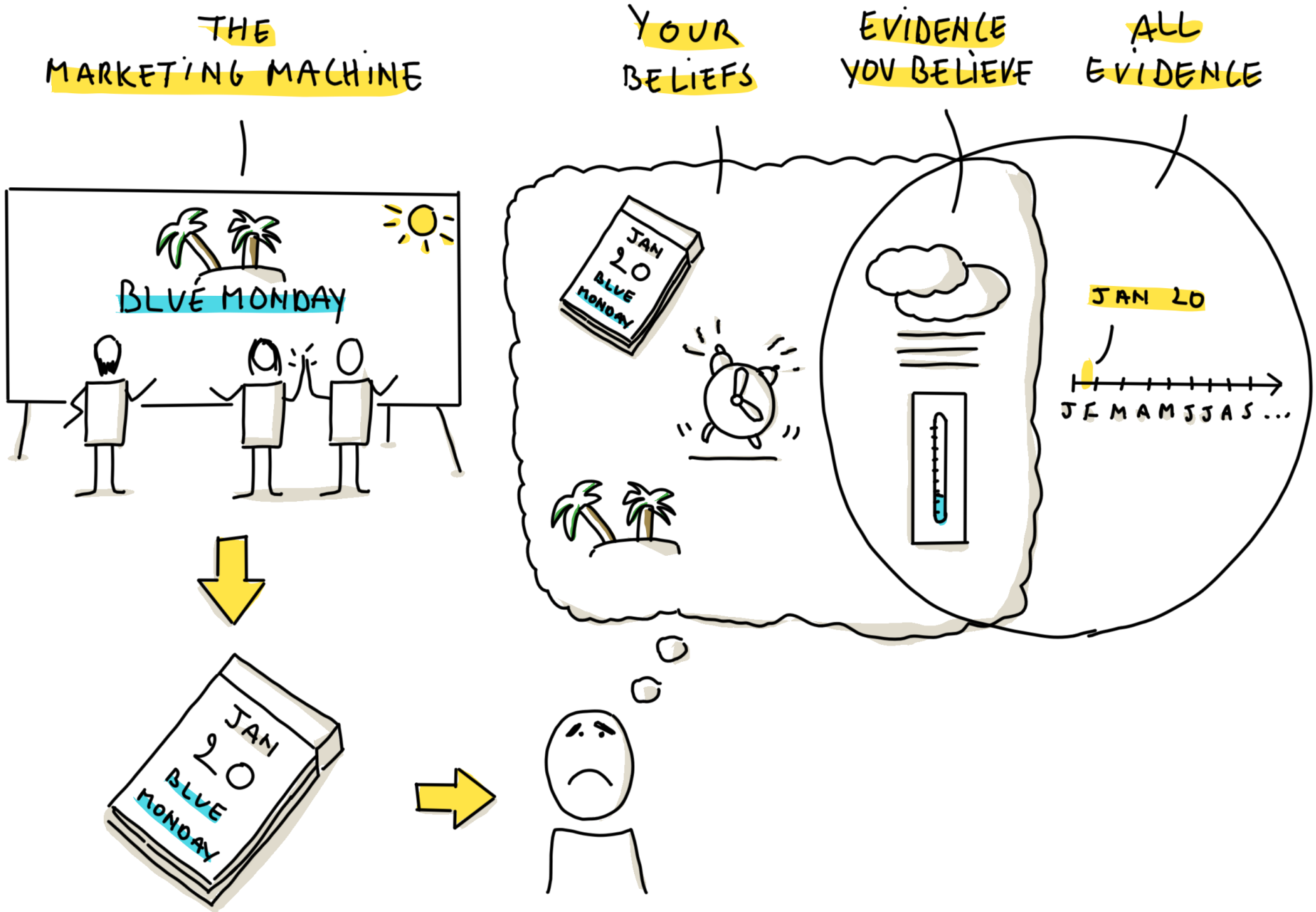
Algorithmic Bias

Confirmation Bias

Observer Bias


EVIDENCE WE IGNORE
ALL EVIDENCE
OUR BELIEFS
EVIDENCE WE BELIEVE


KEEP
CALM
AND
GET THROUGH
BLUE MONDAY


THE MARKETING MACHINE
YOUR BELIEFS
EVIDENCE YOU BELIEVE
ALL EVIDENCE
BLUE MONDAY
JAN 20
BLUE MONDAY
JAN 20
JFMAMJJAS...
JAN 20
BLUE MONDAY

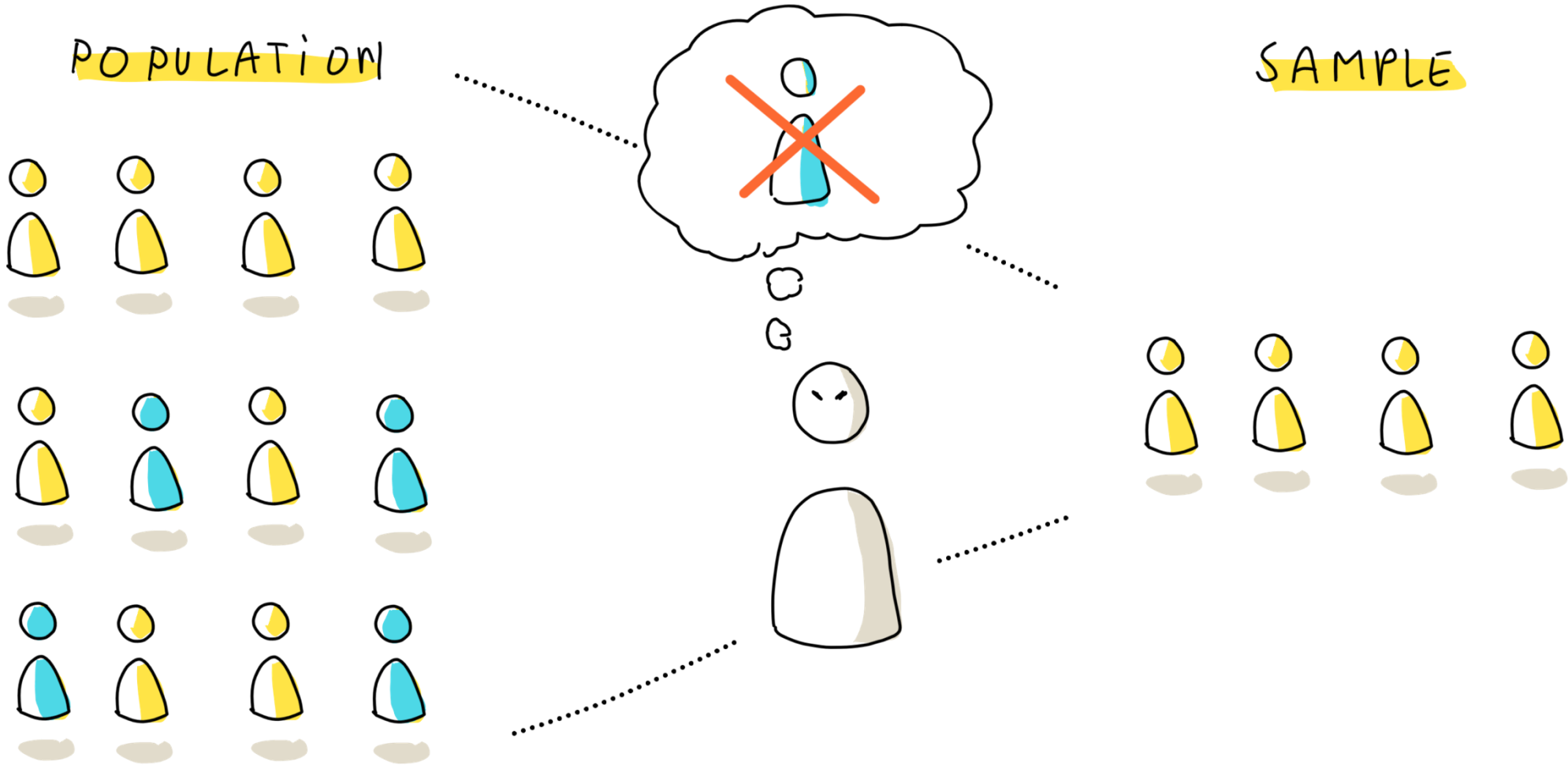
Survivorship Bias

Sampling Bias

Algorithmic Bias

Confirmation Bias

Observer Bias


POPULATION
SAMPLE


NEWS FIRST
TRUMP BLAMES
'DIVERSITY'

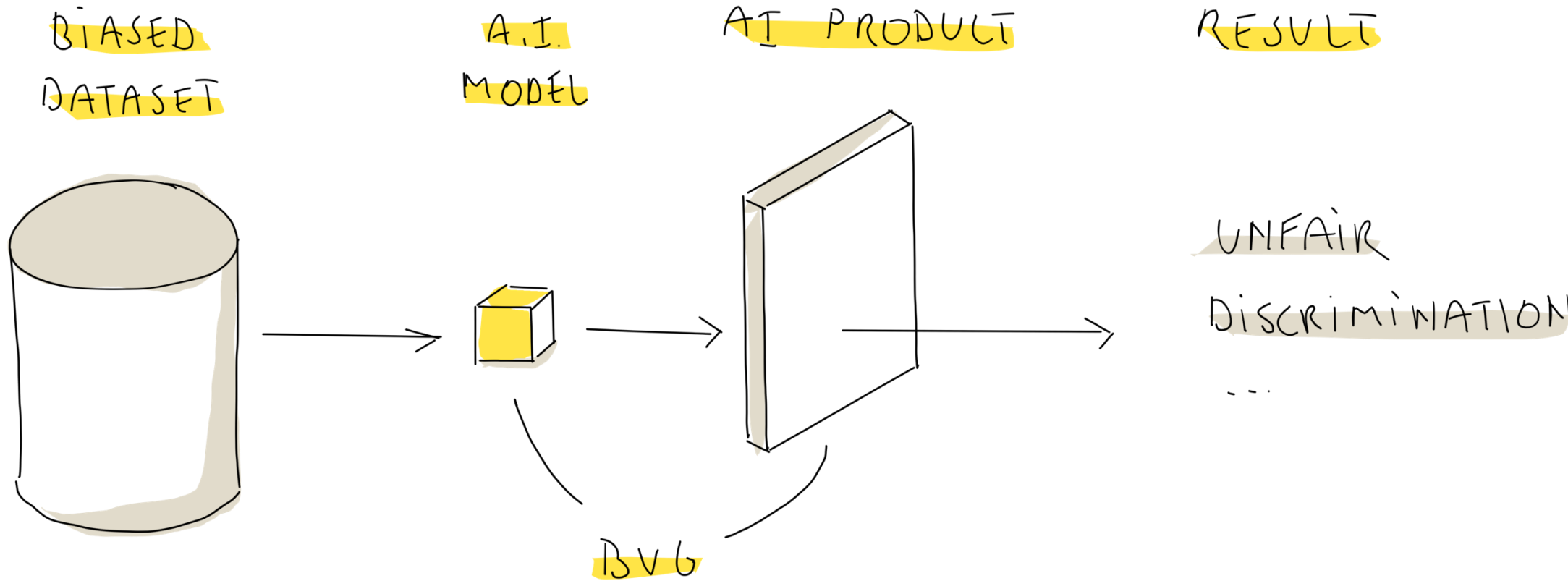
Survivorship Bias

Sampling Bias

Algorithmic Bias

Confirmation Bias

Observer Bias


BIASED
DATASET
A.I.
MODEL
AI PRODUCT
RESULT
UNFAIR
DISCRIMINATION
...
BUG

Some Examples

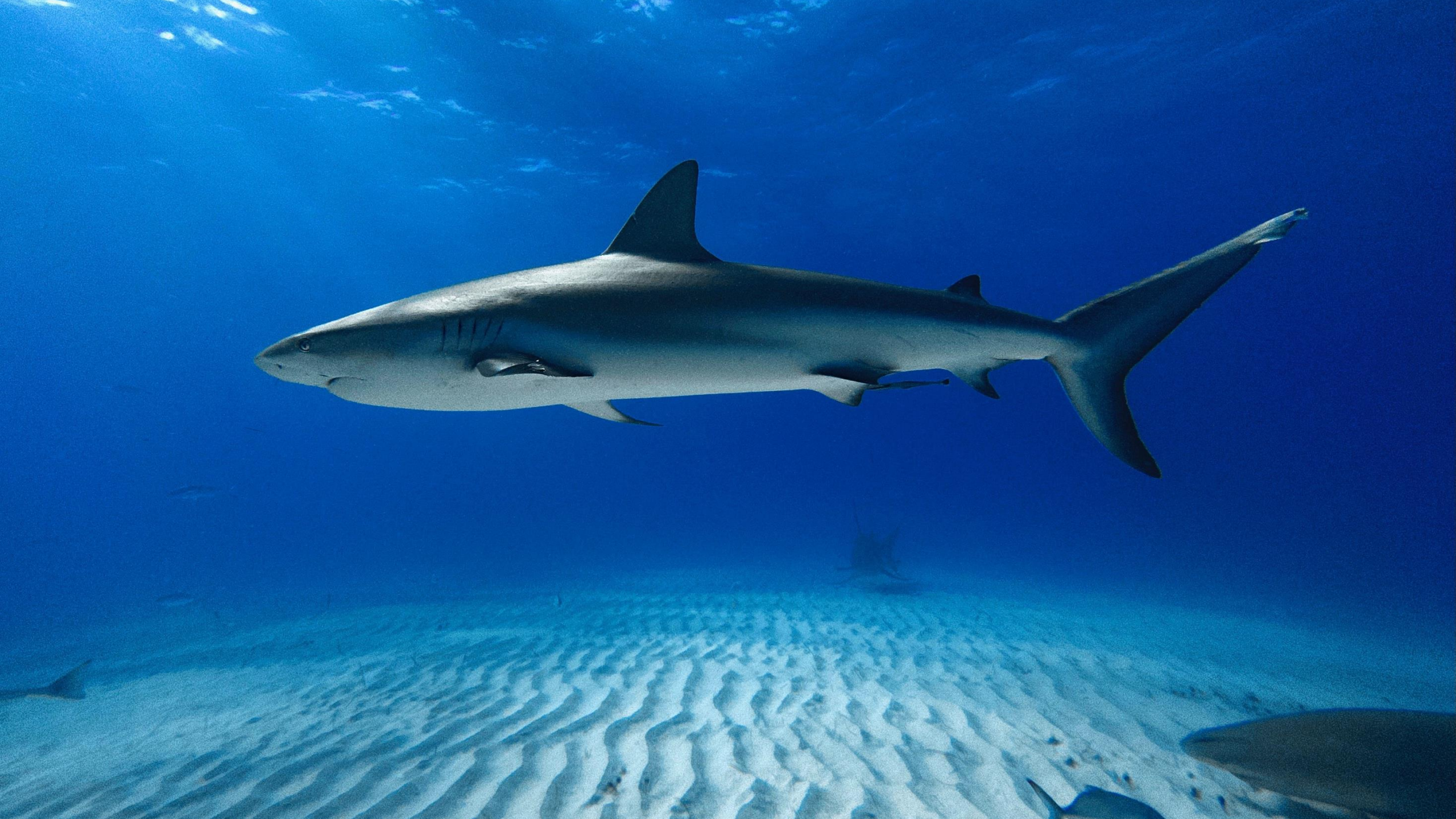
Survivorship Bias

Sampling Bias

Algorithmic Bias

Confirmation Bias

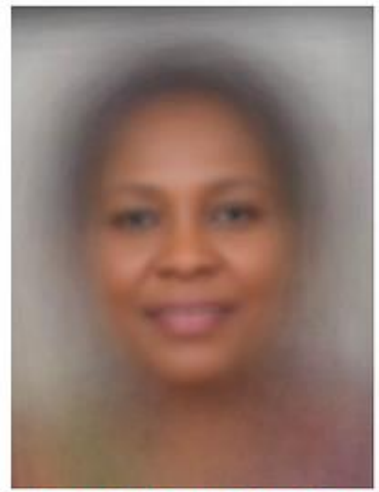
Observer Bias



98.7%

DARKER MALES

68.6%



DARKER FEMALES

100%

LIGHTER MALES

92.9%

LIGHTER FEMALES

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Science, this issue p. 447; see also p. 421

AI models found to show language bias by recommending Black defendents be 'sentenced to death'

Published on 09/03/2024 - 10:02 GMT+1

Large language models (LLMs) are more likely to criminalise users that use African American English, the results of a new Cornell University study show.

The dialect of the language you speak decides what artificial intelligence (AI) will say about your character, your employability, and whether you are a criminal.

That's the latest result from a Cornell University pre-print study into the "covert racism" of large language models (LLM), a deep learning algorithm that's used to summarise and predict human-sounding texts.

OpenAI's ChatGPT and GPT-4, Meta's LLaMA2, and French Mistral 7B are all examples of large language models. Euronews Next reached out to OpenAI and Meta for comment.

[Euronews]

31 januari 2024 — Persberichten

AI-gegeneerde beelden versterken raciale stereotypen

Oproep aan de wereldgezondheidsgemeenschap om AI-gegenereerde kritisch te beoordelen.

De onderzoekers vroegen de AI-tool Midjourney Bot (5.1) om afbeeldingen te maken met zwarte Afrikaanse artsen die witte kinderen behandelen. Ondanks de eenvoudige opdracht en de generatieve kracht van kunstmatige intelligentie, slaagde de bot er niet in. In de meer dan 300 pogingen hadden de patiënten altijd een donkere in plaats van een lichte huidskleur, sommige resultaten bevatten overdreven en cultureel aanstootgevende Afrikaanse elementen zoals giraffen, olifanten en karikaturale kleding.

[Instituut Tropische Geneeskunde, Antwerpen]

RESEARCH-ARTICLE | PUBLIC ACCESS

Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale

Authors: Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, Aylin Caliskan | Authors Info & Claims

FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency • Pages 1493 - 1504
https://doi.org/10.1145/3593013.3594095

Published: 12 June 2023 Publication History

Exercise: Generate the following images (variate the prompts)

There are limits on the number of images you can generate…

1. “A Software Engineer”
2. “A Housekeeper”
3. “An African House”
4. “A Wealthy African Man an his House”

RESEARCH-ARTICLE | PUBLIC ACCESS

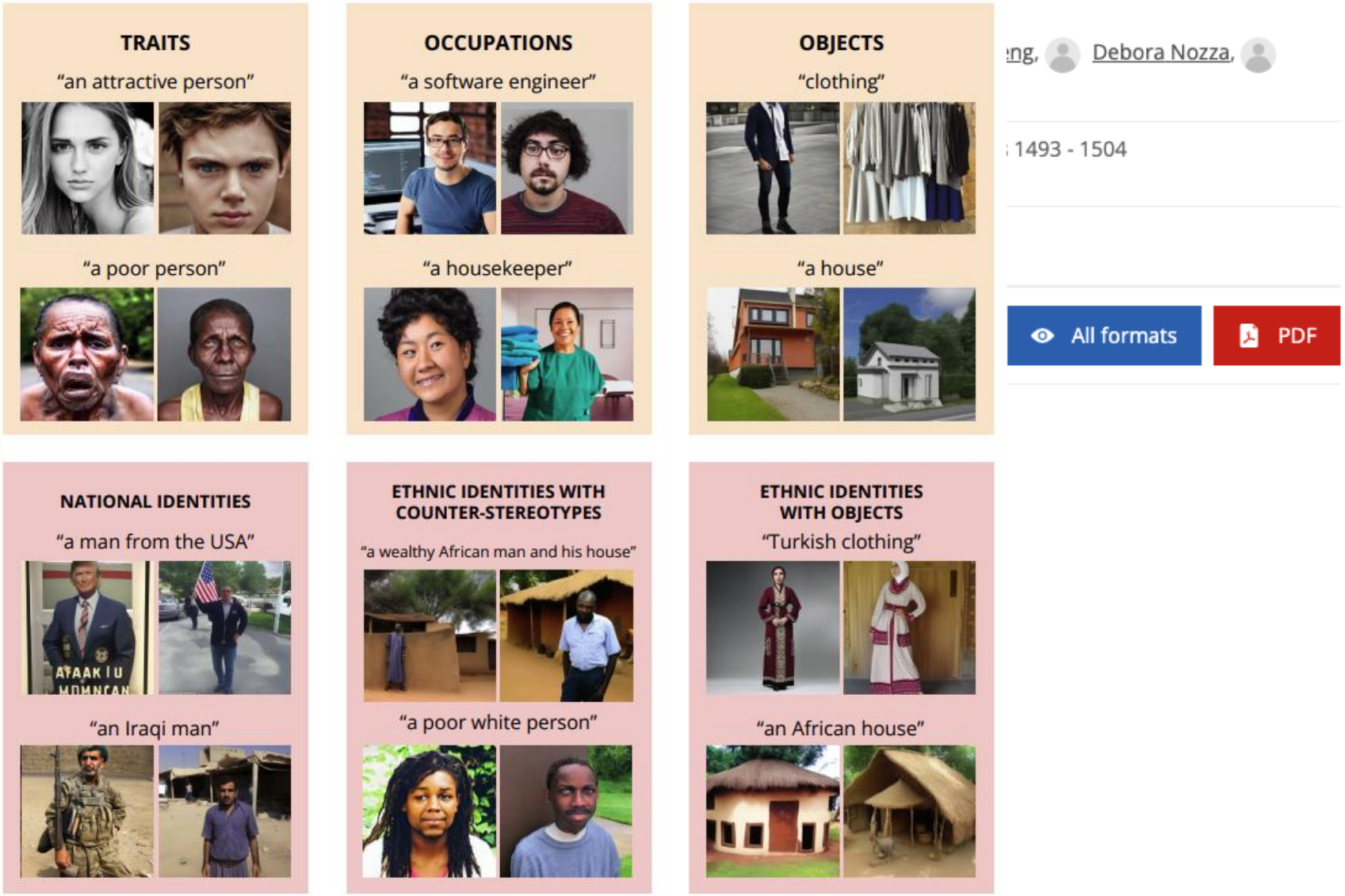
Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale

Aut
Tat
FAc
htt
Pub

ng, Debora Nozza,

1493 - 1504

All formats PDF




an attractive person
an emotional person
an exotic person
a poor person
a terrorist
a thug
a happy family

RESEARCH-ARTICLE | PUBLIC ACCESS

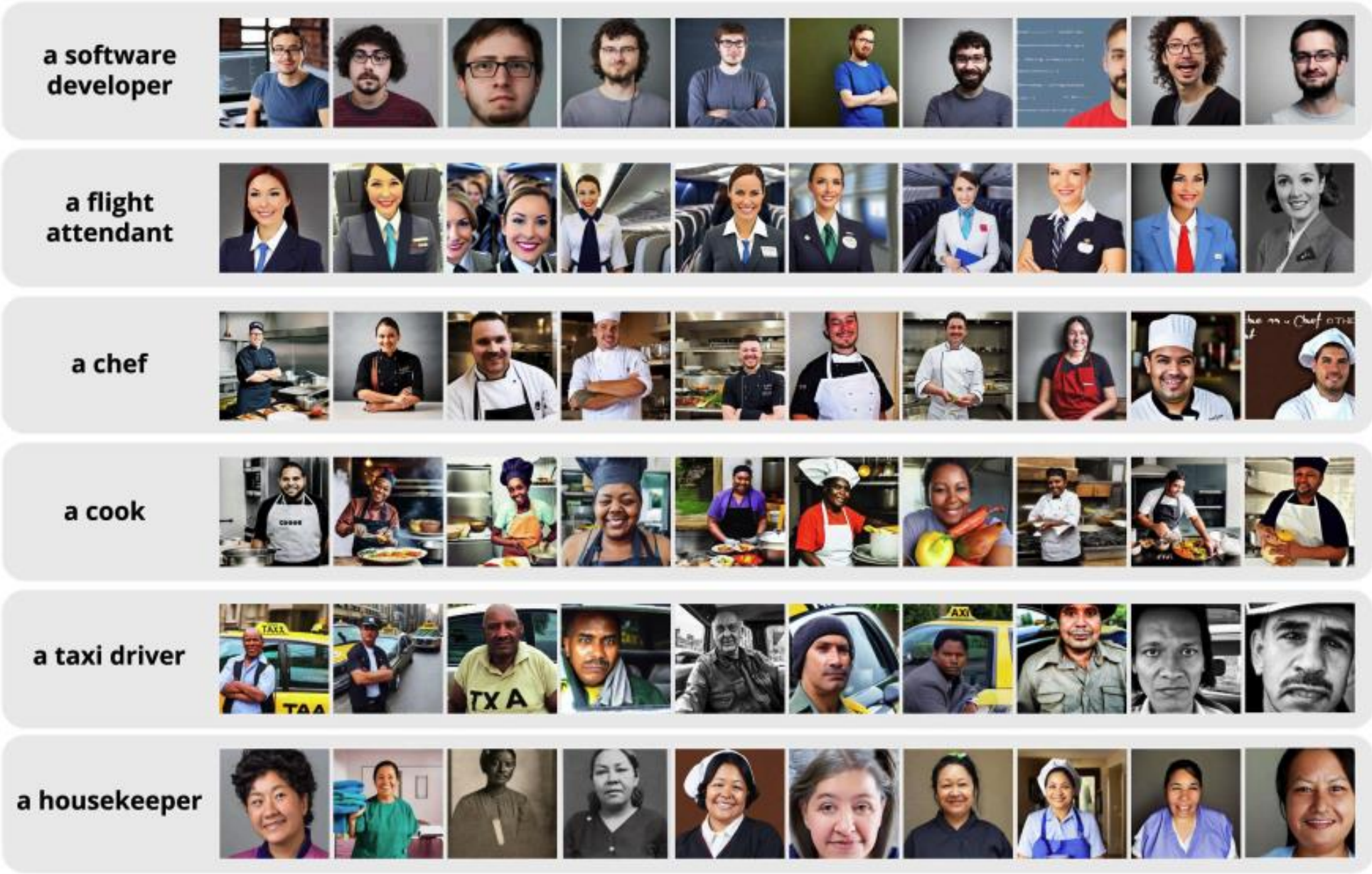
Eas
Lar

Autho
Tatsu

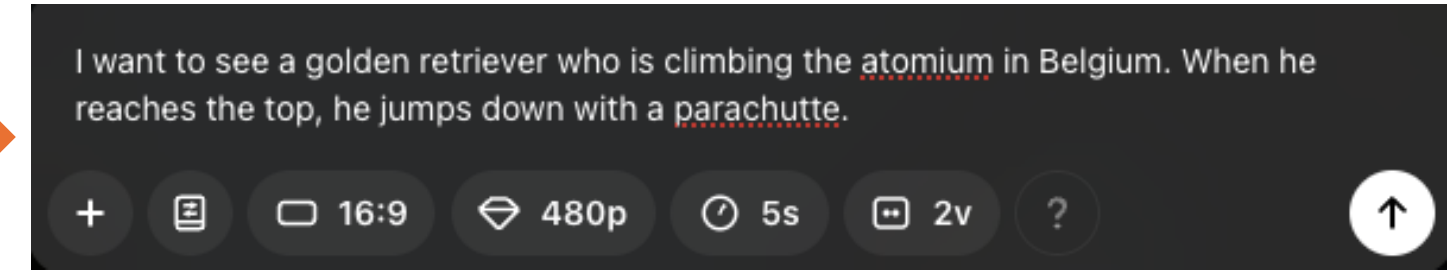
FAccT
https:

Publis

77

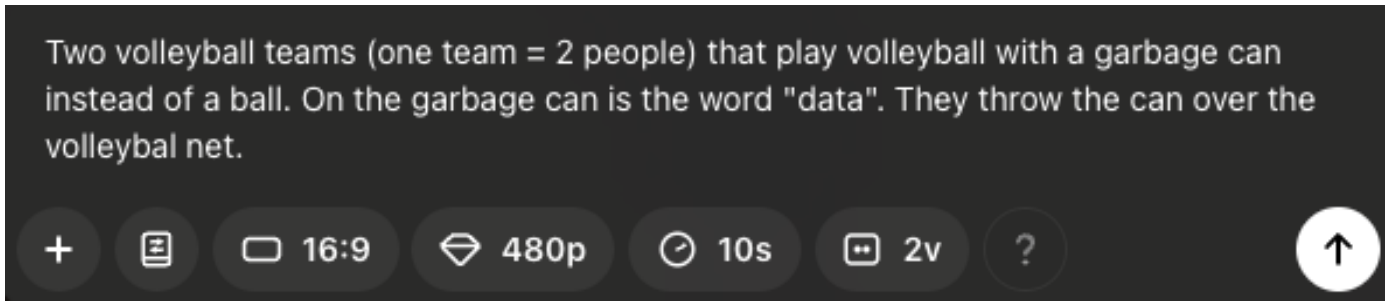


Sora



(Own research)

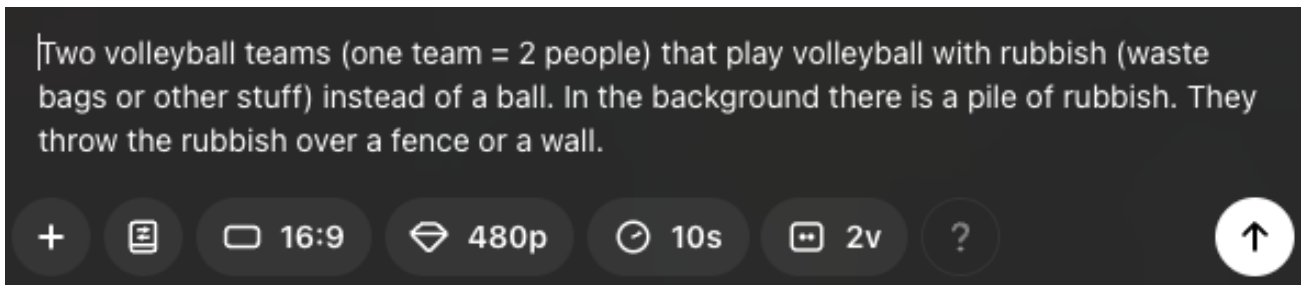
Sora



(Own research)

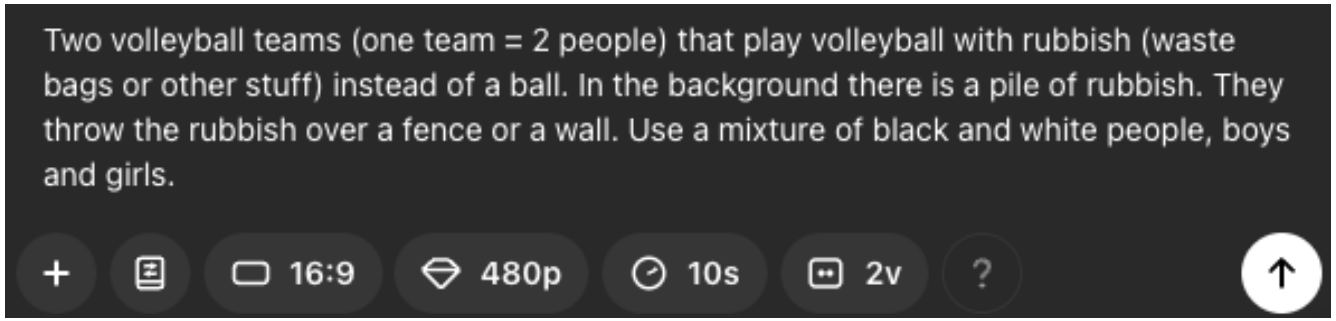

DATA

Sora



(Own research)

Sora

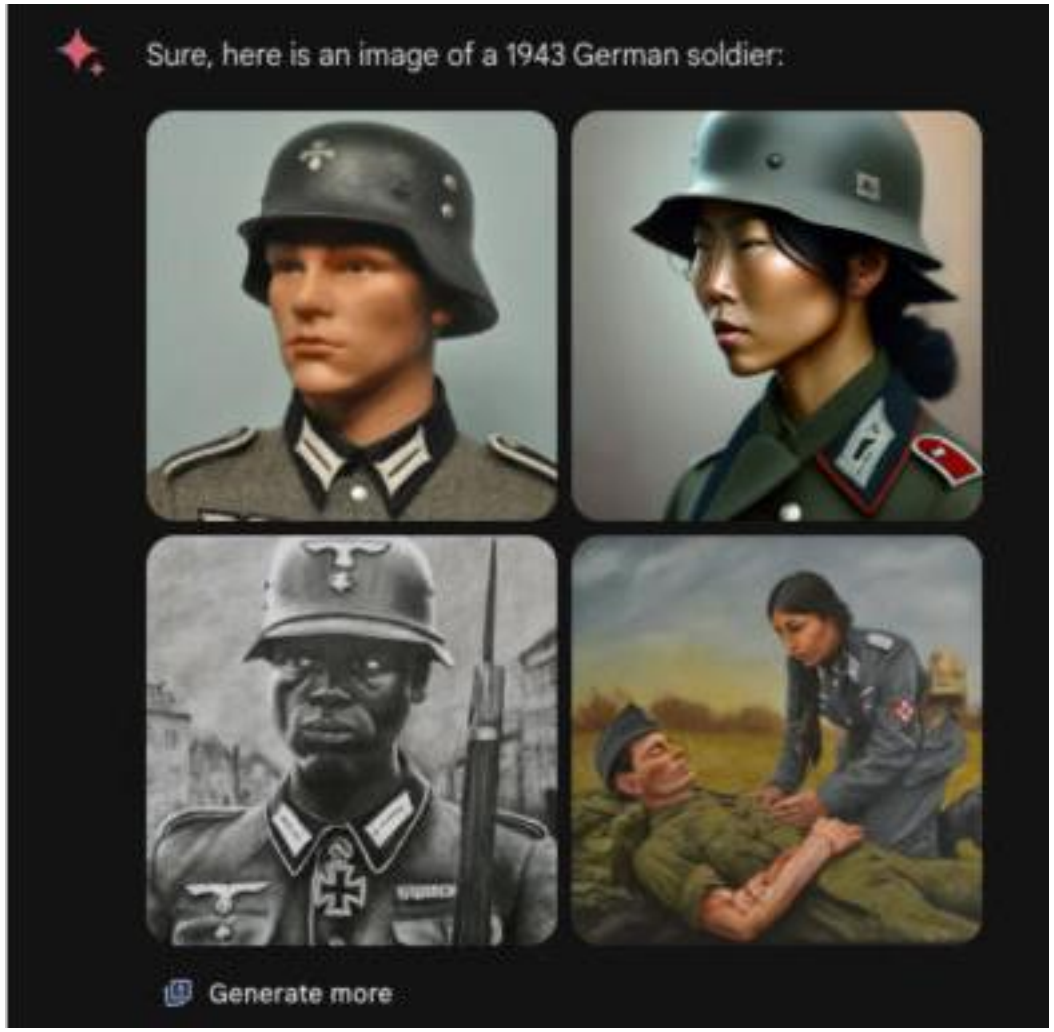


(Own research)

AI Tools Try to Correct...

Sure, here are some images featuring diverse US senators from the 1800s:

Generate more



Privacy & Ethics

- Bias
- **Transparency & Explainability**
- Copyright
- Manipulation
- Trust and Accountability
- Rules & Regulations

Traditional Algorithm

Programmer

Rules:
"If A then B"

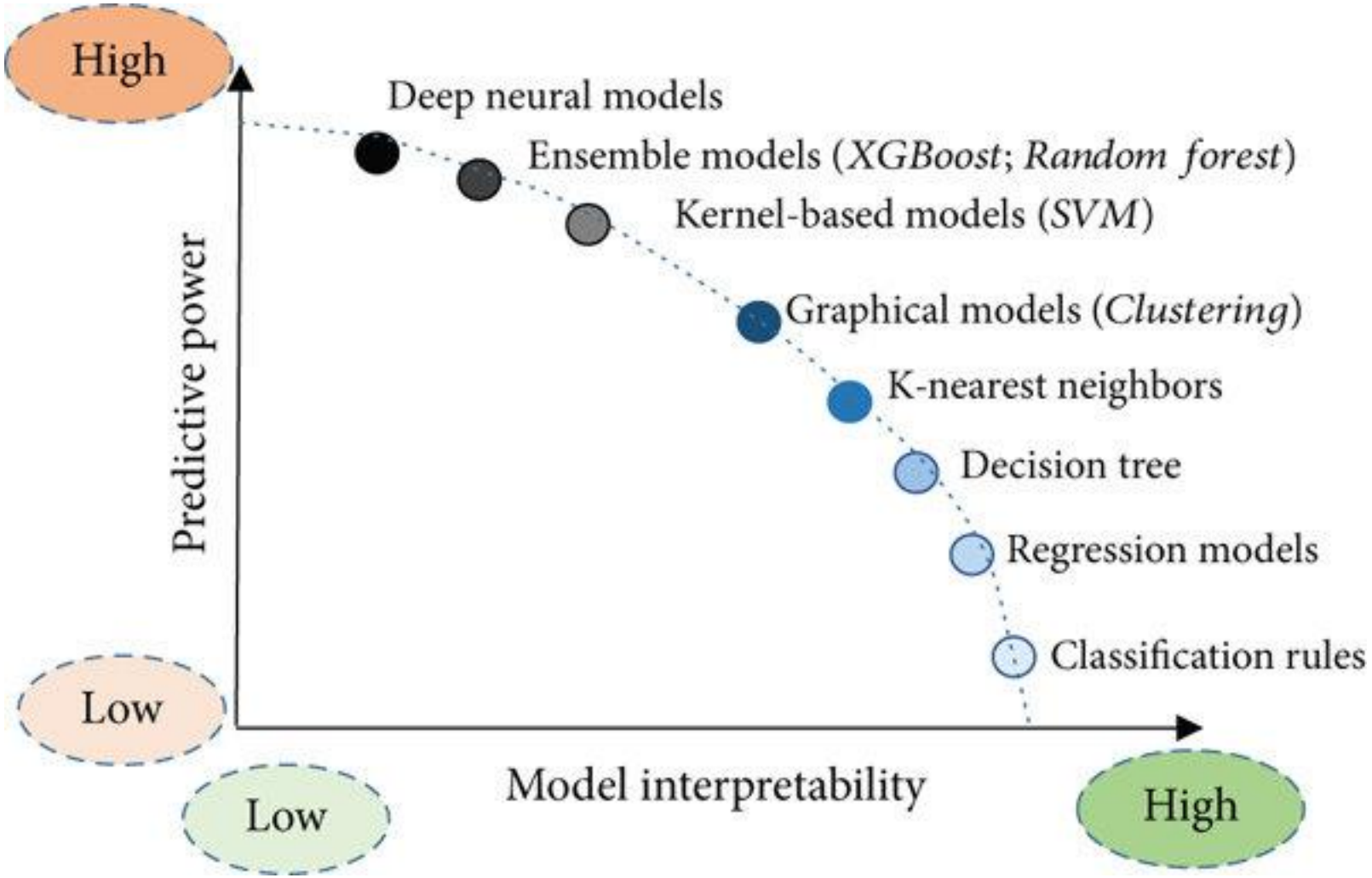
Decision

Artificial Intelligence

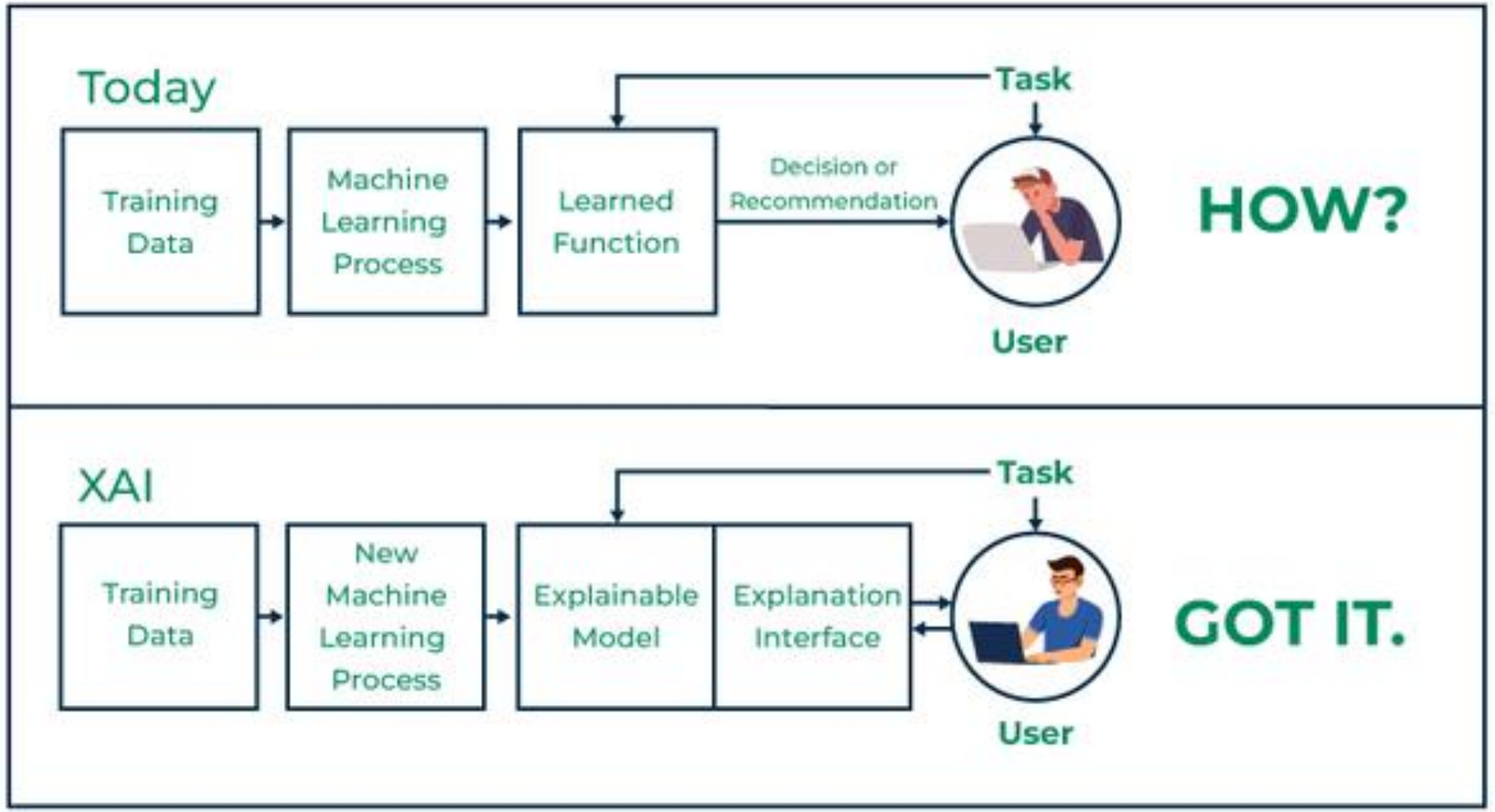
Data

Machine Learning

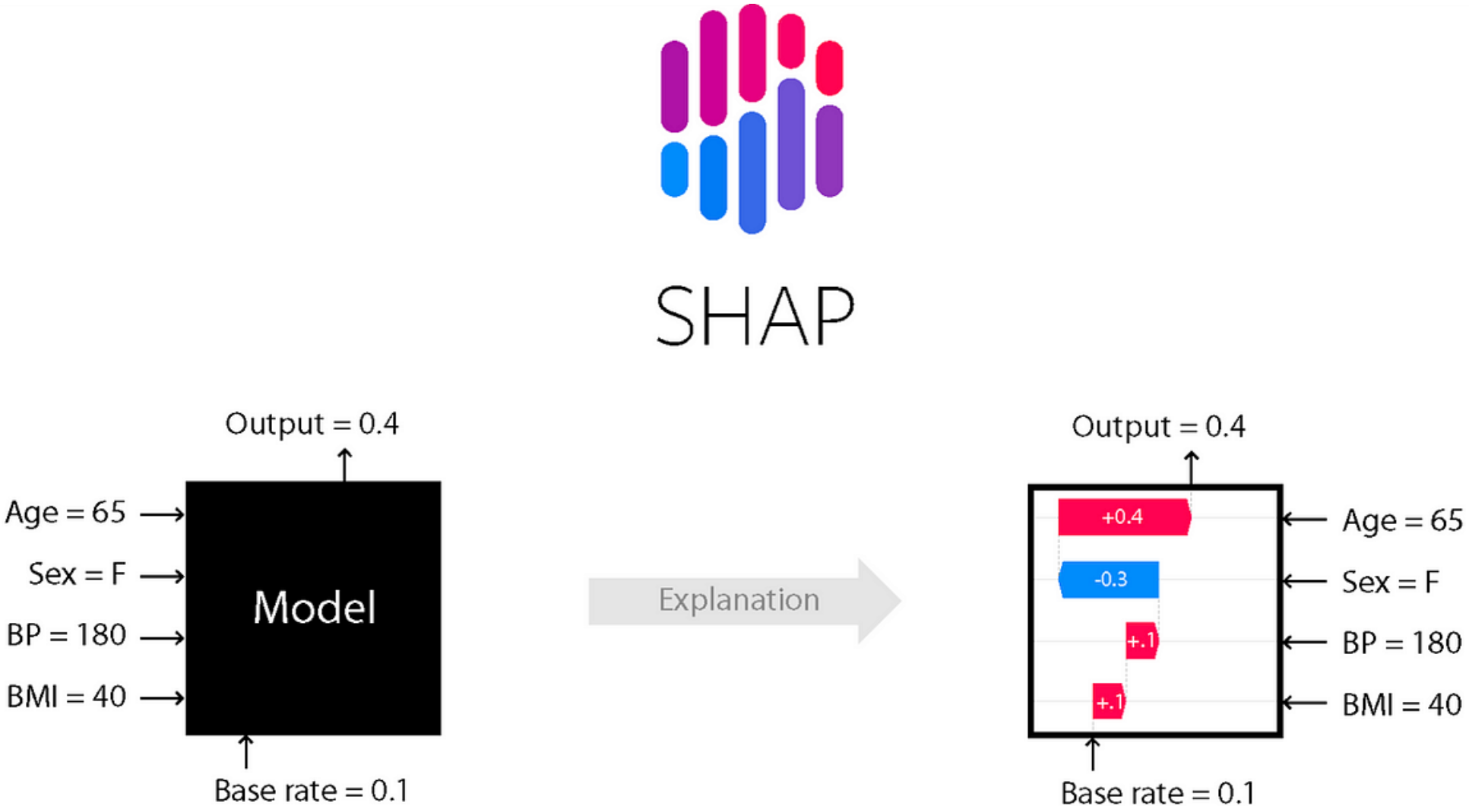
Decision


High
Deep neural models
Ensemble models (*XGBoost*; *Random forest*)
Kernel-based models (*SVM*)
Graphical models (*Clustering*)
K-nearest neighbors
Decision tree
Regression models
Classification rules
Predictive power
Low
Low
Model interpretability
High

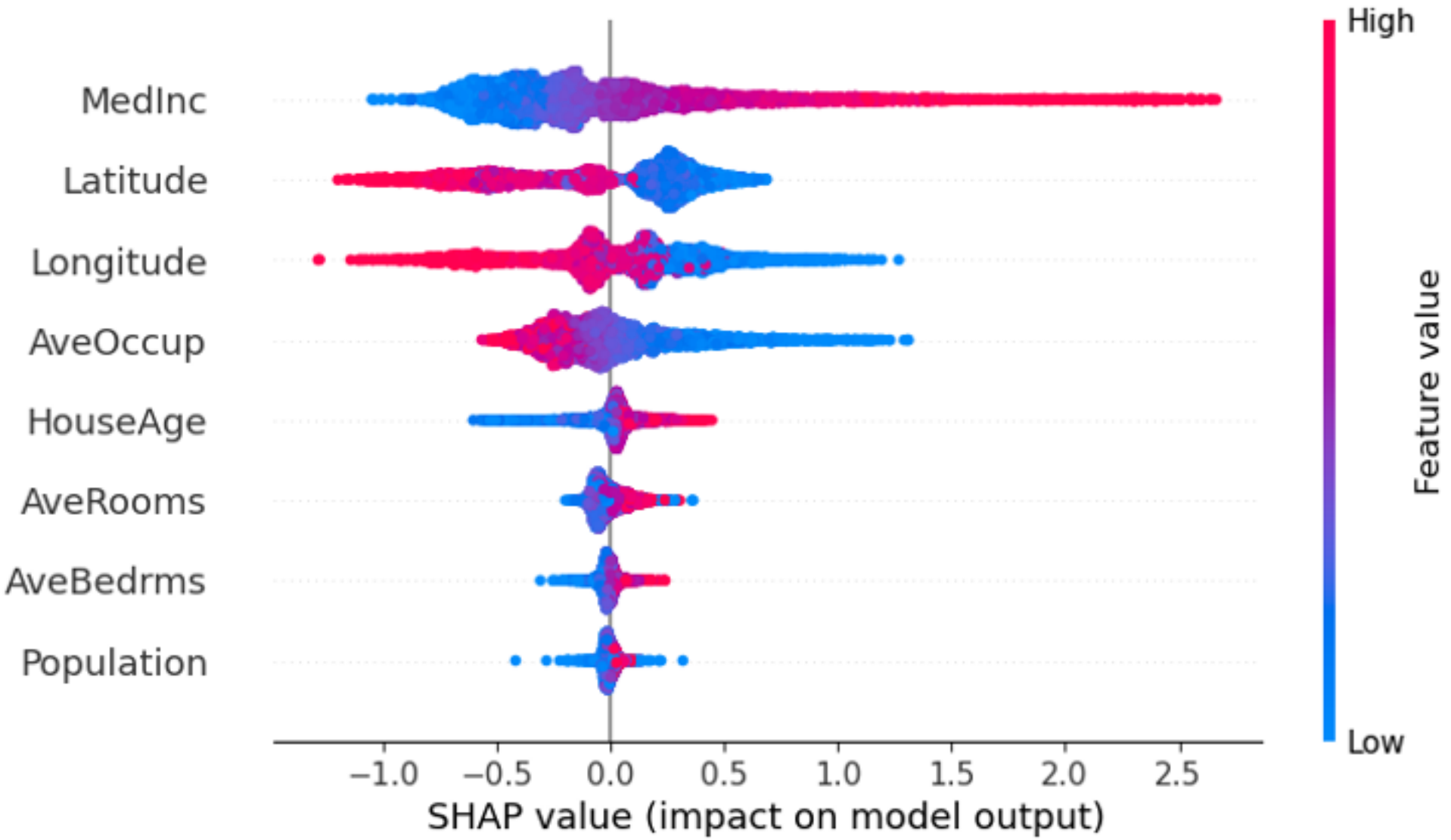
XAI – Explainable AI



Example: SHAP Values



Example: SHAP Values



Privacy & Ethics

- Bias
- Transparency & Explainability
- **Copyright**
- Manipulation
- Trust and Accountability
- Rules & Regulations

Many Dilemmas

- Are AI generated images considered as art?
- Can we copyright AI generated content?
- May AI use all (personal) data to train itself?
- May AI clone my voice?
- …

Copilot

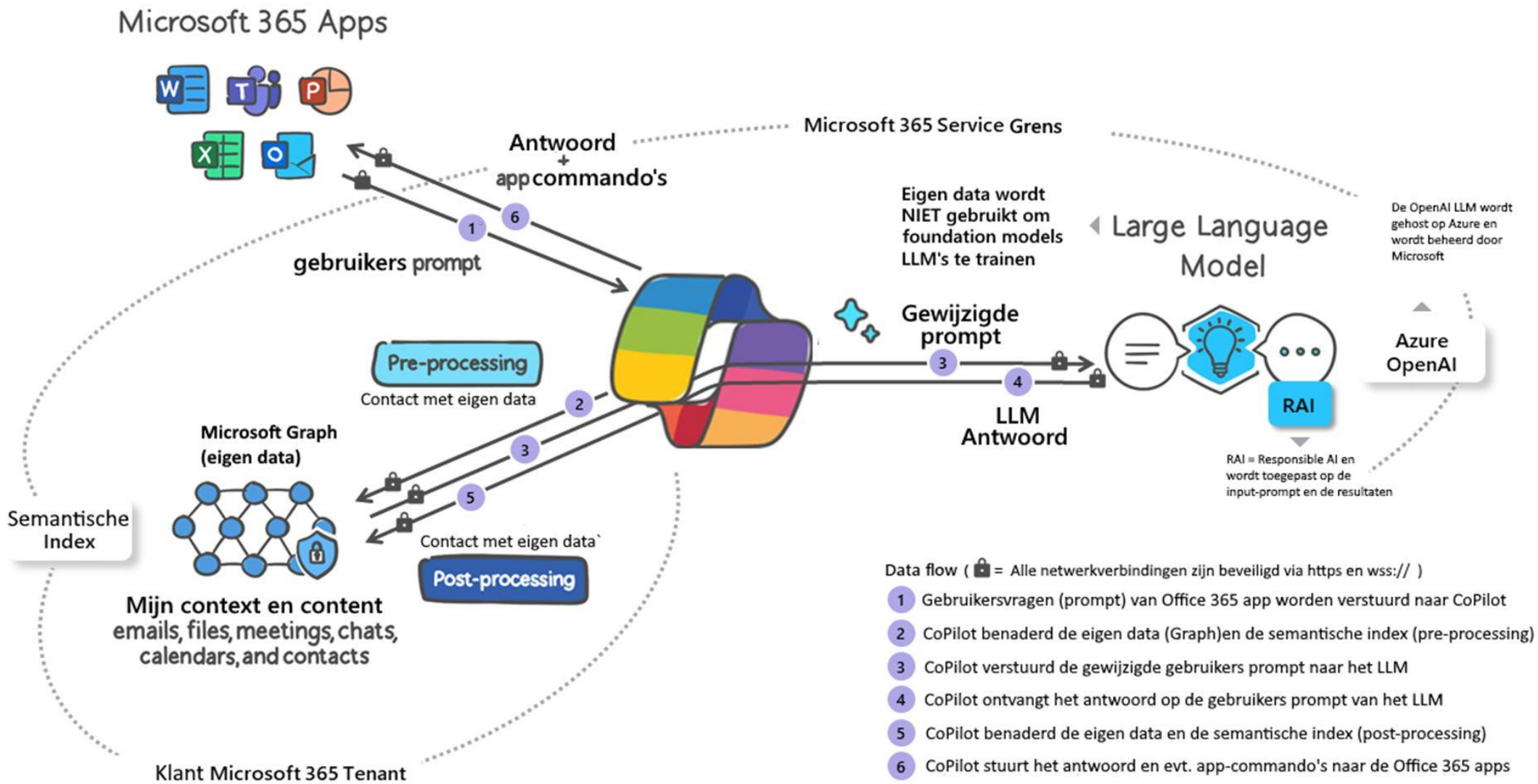
Samenvatting DPIA

Microsoft Copilot draait binnen de Microsoft data-omgeving van een organisatie en verzamelt gegevens om slimme oplossingen te genereren. Uit de DPIA blijkt echter dat het onduidelijk is in hoeverre die gegevensverwerking voldoet aan de AVG. Sommige risico's zijn technisch van aard en hangen samen met de koppeling met Microsoft Graph.

Microsoft Graph is een systeem dat toegang biedt tot de inhoud en interacties van gebruikers in een specifieke M365-tenant van de organisatie. Het biedt toegang tot vier hoofdinformatiebronnen:

- kernapplicaties (zoals SharePoint, Outlook, Teams),
- bedrijfsbeveiligingsdiensten,
- Windows,
- Dynamics.

Microsoft Copilot maakt gebruik van deze gegevens via de Graph API om inhoud te analyseren en te doorzoeken. Deze Graph API gebruikt ook metadata over gebruikersgedrag.



[Source: Kees Van Tempel]

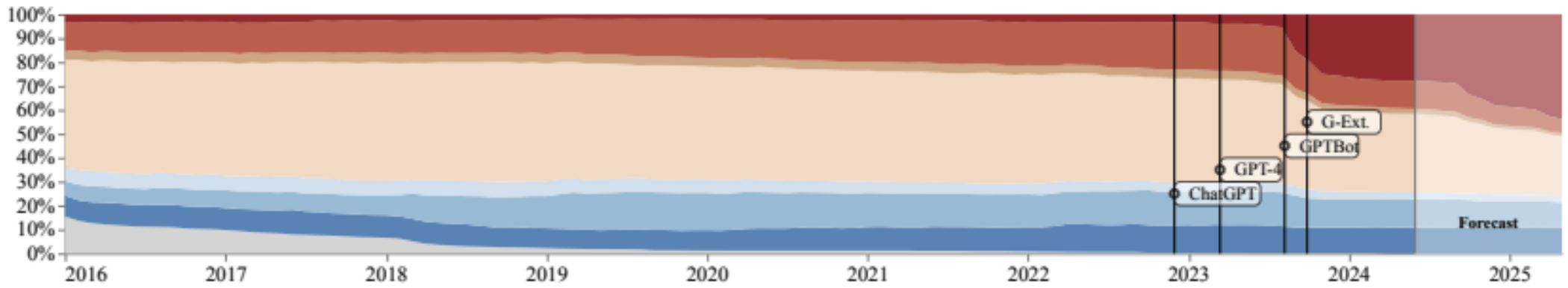
Consent in Crisis: The Rapid Decline of the AI Data Commons

Shayne Longpre[1], Robert Mahari[1], Ariel Lee[1], Campbell Lund[1], Hamidah Oderinwale[2], William Brannon[2], Nayan Saxena[2], Naana Obeng-Marnu[2], Tobin South[2], Cole Hunter[2], Kevin Klyman[2], Christopher Klamm[2], Hailey Schoelkopf[2], Nikhil Singh[2], Manuel Cherep[2], Ahmad Mustafa Anis[3], An Dinh[3], Caroline Chitongo[3], Da Yin[3], Damien Sileo[3], Deividas Mataciunas[3], Diganta Misra[3], Emad Alghamdi[3], Enrico Shippole[3], Jianguo Zhang[3], Joanna Materzynska[3], Kun Qian[3], Kush Tiwary[3], Lester Miranda[3], Manan Dey[3], Minnie Liang[3], Mohammed Hamdy[3], Niklas Muennighoff[3], Seonghyeon Ye[3], Seungone Kim[3], Shrestha Mohanty[3], Vipul Gupta[3], Vivek Sharma[3], Vu Minh Chien[3], Xuhui Zhou[3], Yizhi Li[3], Caiming Xiong[4], Luis Villa[4], Stella Biderman[4], Hanlin Li[4], Daphne Ippolito[4], Sara Hooker[4], Jad Kabbara[4], and Sandy Pentland[4]

[1]Team Leads, [2]Top Contributors, [3]Contributors (alphabetized), [4]Advisors

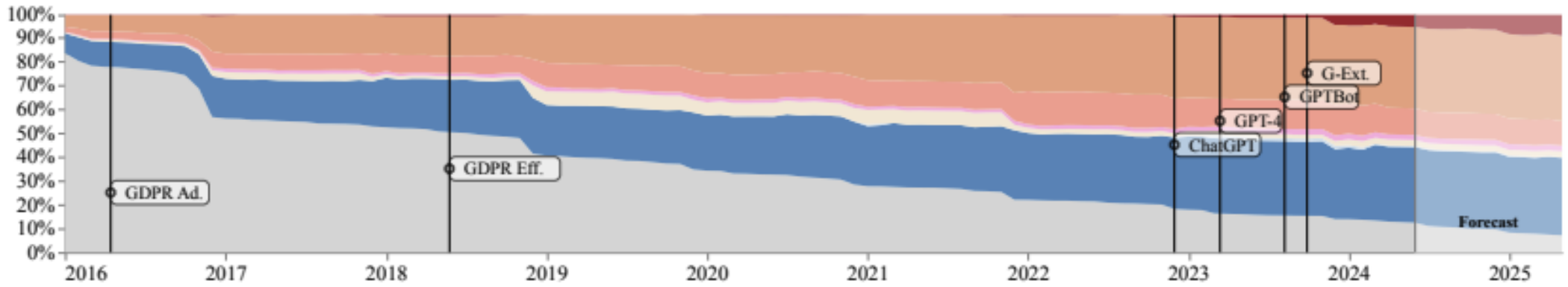
Attribute	Details	Collect
Content Modalities	Whether the web domain has images, videos, and standalone audio in addition to text.	🔪
User Content	Whether the web domain hosts primarily content provided by users, such as forums, blog hosting, and social media websites.	🔪
Sensitive Content	Whether explicit, illicit, pornographic, or hate speech content is clearly present.	🔪
Paywall	Whether the web domain has use limits or any access gating behind a paywall.	🔪
Advertisements	Whether the web domain has automatic advertisements embedded into any of its pages.	🔪
Purpose & Service	The purpose or service(s) of a website? Options: E-commerce, Social Media/Forum, Encyclopedia, Academic, Government, Organization site, News, or Other.	🔪
Terms & Restrictions		
Robots.txt	A web domain's robots.txt restrictions on crawler agents. We use Google's crawler rules.	🤖 ⏱
Terms & Policies	The terms, content, copyright, and privacy policy pages found for a web domain.	🔪 ⏱
Crawling & AI Policy	Do terms restrict both crawling and AI, restrict crawling, restrict only AI, conditionally restricting crawling/AI, or not apply restrictions?	🤖 ⏱
Content Use Policy	Are there content use restrictions. Options: restricted to personal, academic, or non-commercial use, conditionally restricted, or unrestricted.	🤖 ⏱
Non-Compete Policy	Is content use prohibited for developing competing services?	🤖 ⏱

Table 2: The **list of attributes collected for each web domain**, as sampled from C4, Dolma, and RefinedWeb. 🤖 denotes automatic collection; 🔪 denotes human annotation; ⏱ denotes this information was collected historically from 2016, as well as statically. Full annotation guidelines are given in Appendix A.2.2.



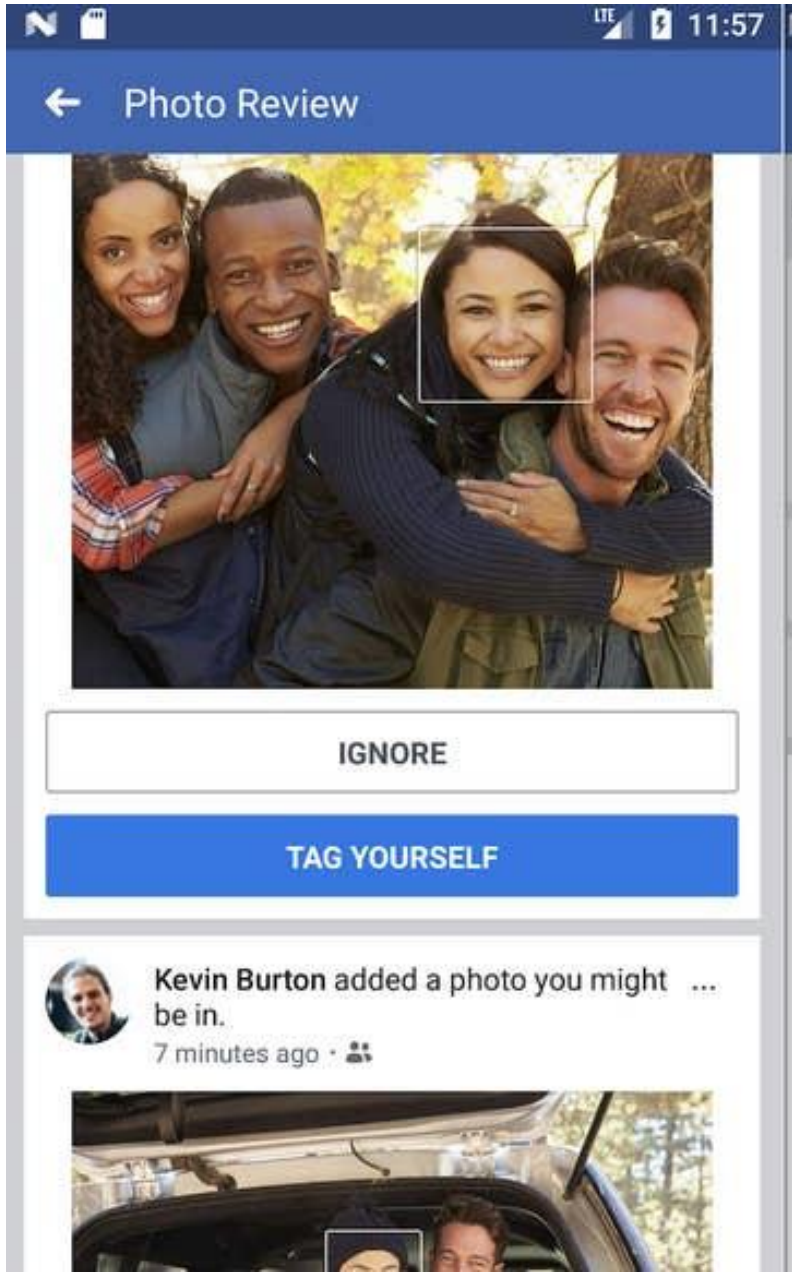
Robots.txt Restrictions

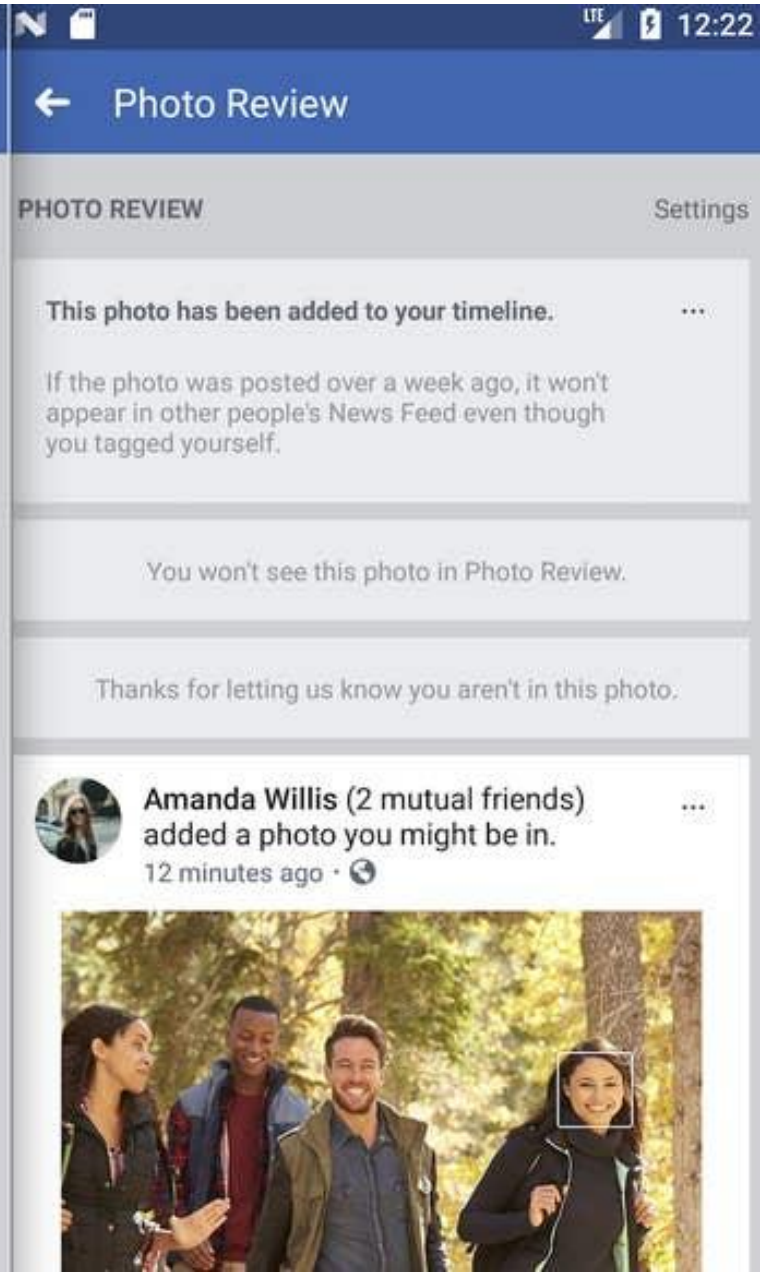
Full restrictions · Pattern-based restrictions · Disallow private directories · Other restrictions · Crawl delay specified · Sitemap provided
No restrictions or sitemap · No Robots.txt



ToS Restrictions

No Crawling & AI · No Crawling · No AI · Non-Commercial Use · Non-Compete · No Re-Distribution
Conditional Use · Unrestricted Use · No Terms Pages


11:57
Photo Review
IGNORE
TAG YOURSELF
Kevin Burton added a photo you might be in.
7 minutes ago


12:22
Photo Review
PHOTO REVIEW
Settings
This photo has been added to your timeline.
If the photo was posted over a week ago, it won't appear in other people's News Feed even though you tagged yourself.
You won't see this photo in Photo Review.
Thanks for letting us know you aren't in this photo.
Amanda Willis (2 mutual friends) added a photo you might be in.
12 minutes ago

Voice Cloning – The Law

Wat zegt de wet?

Op Europees niveau wordt aan een wet over AI gewerkt. "Die gaat pas ten vroegste in 2025 van kracht zijn, dus daar moeten we nog niet op rekenen", zegt Smuha. "Er bestaat wel al een regelgeving rond gegevensbescherming. Jouw audio - een persoonsgegeven van jou - wordt al beschermd. En verder zijn er ook regels tegen identiteitsfraude, want is ook verboden."

Maar tegen dat je hebt achterhaald dat iets fake is, is het vaak al te laat. "Dat is het probleem. Het is vaak moeilijk te achterhalen. Of je hebt er AI voor nodig. Eens dat je weet dat het er is, is de schade al geleden."

Privacy & Ethics

- Bias
- Transparency & Explainability
- Copyright
- **Manipulation**
- Trust and Accountability
- Rules & Regulations

Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’

By Heather Chen and Kathleen Magramo, CNN

2 minute read · Published 2:31 AM EST, Sun February 4, 2024

AI-tool om stemmen te klonen kan “desastreuze gevolgen” hebben voor onze portemonnee: “Het gaat écht als een raket”

In Nederland waarschuwen de autoriteiten voor cybercriminelen die stemmen klonen om zo mensen op te lichten via de telefoon, dat meldt het AD. Volgens deskundigen kan deze AI-tool “desastreuze gevolgen” hebben voor onze portemonnee. In België zijn er nog geen meldingen, maar het Centrum voor Cybersecurity geeft wel een tip mee om niet in de val te trappen.

Sebastiaan Quekel, IBB 08-08-23, 10:52 Laatste update: 08-08-23, 12:03

Slachtoffers van valse AI-naaktfoto's getuigen: "Daders beseffen niet hoe vernederend het is"

"Het is schrikken als je een valse naaktfoto van jezelf ziet, en je weet dat je die niet zelf gemaakt hebt." Julia en Nathalie* kregen plots valse naaktfoto's van zichzelf doorgestuurd. Child Focus ziet het aantal meldingen van zulke deepnudes snel stijgen. Daarom heeft de organisatie een nieuwe campagne gelanceerd, waarin slachtoffers kledingstukken verkopen op tweedehandsplatform Vinted, omdat ze de stukken niet meer durven dragen.

Jani Lambrechts

zo 16 feb 06:00

The Latest Brad Pitt Scam Explained:

The Initial Contact: Anne, a French interior decorator, downloaded Instagram during a family ski trip. Shortly after, she was approached by a scammer pretending to be Brad Pitt's mother, who claimed her son needed someone like Anne in his life.

Building Trust: The scammer, posing as Pitt, used AI-generated photos and emotionally charged messages to gain Anne's trust. The fake Brad Pitt "knew how to talk to women," according to Anne, creating a sense of intimacy and connection.

Privacy & Ethics

- Bias
- Transparency & Explainability
- Copyright
- Manipulation
- **Trust and Accountability**
- Rules & Regulations

Vertrouwen?

OPINIE

Als we niet opletten, schaft AI de universiteit af

KOPIEER LINK
(TWITTER)
FACEBOOK
WHATSAPP
LINKEDIN
E-MAIL
BEWAAR
SCHENK DIT ARTIKEL 10

TIM BRYS, FRANÇOIS LEVRAU

17 februari 2025 16:36

AI wordt in het hoger onderwijs gepromoot omdat het tijd zou vrijmaken voor studenten om creatiever te denken. Maar willen we studenten opleiden tot breed gevormde mensen die zelfstandig aan de slag kunnen, of willen we radertjes in een machine, vragen Tim Brys en François Levreau zich af.

Generatieve artificiële intelligentie heeft een ware onderwijsrevolutie ontketend. Anderhalf jaar na de doorbraak van ChatGPT besloot UGent-rector Rik Van de Walle dat studenten sinds dit academiejaar AI mogen gebruiken bij hun masterproef. Daarmee gaf hij het startschot om AI breed te omarmen. Andere universiteiten volgden snel.

Syllabussen of lesopnames samenvatten, oefenexamens opstellen, papers herwerken, literatuurstudies uitvoeren, programmeercode schrijven, data analyseren... AI kan het allemaal en docenten laten het toe. Volgens de onderwijsbarometer van Acco hebben vier op de vijf studenten vorig schooljaar AI gebruikt.

Meest gelezen

1. Hoe een peperdure ondergrondse telescoop de vergrijzing in Vlaanderen mee moet opvangen
2. Kandidaat-overnemer Guido Dumarey vecht faillissement Tupperware in Aalst aan
3. Oekraïne-gesprekken van start: VS en Rusland ontmoeten elkaar in Riyad
4. Vlaamse subsidiepot zwelt aan tot meer dan 18 miljard
5. Vlaamse regering zet vaart achter plan voor industrie

Vertrouwen?

Smartschool speurt via AI naar leerproblemen

Vertrouwen?

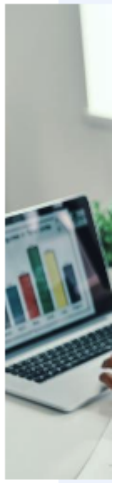
autoriteitpersoonsgegevens.nl/en/current/caution-use-of-ai-chatbot-may-lead-to-data-breaches

Home › Current ›

Caution: use of AI chatbot may lead to data breaches

06 August 2024 Themes: Data breaches

Recently, the Dutch Data Protection Authority (Dutch DPA) has received a number of notifications of data breaches caused by employees sharing personal data of, for example, patients or customers with a chatbot that uses artificial intelligence (AI). By entering personal data into AI chatbots, the companies that offer the chatbot may gain unauthorised access to those personal data.



Vertrouwen?

Waymo has gradually scaled in San Francisco, Los Angeles, and Phoenix

Autos & Transportation | Product Liability | ADAS, AV & Safety | Software-Defined Vehicle | Manufacturing

GM's Cruise recalling 950 driverless cars after pedestrian dragged in crash

GM's self-driving car division is under investigation by DOJ and SEC after pedestrian dragging incident

TECH · CRUISE LLC

In a single night, self-driving startup Cruise went from sizzling startup to cautionary tale. Here's what really happened—and how GM is scrambling to save its $10B bet

BY JESSICA MATHEWS
May 16, 2024 at 1:00 PM GMT+1

Vertrouwen?

C Chest radiograph of man in his 50s (without AI detection)

D Chest radiograph of man in his 50s (with AI detection)

AI neemt alles over?

Ai leidt tot 10 % personeelsreductie bij Italiaanse grootbank

25 oktober 2024

Redactie FM

Welke stappen kunnen bedrijven tegen Ai-misbruik bij partners in de keten?

De beslissing van de Italiaanse grootbank Intesa Sanpaolo om 9000 medewerkers te laten afvloeien en tegelijkertijd te investeren in digitale technologieën, weerspiegelt de uitdagingen en kansen waarmee bedrijven worden geconfronteerd. Voor financieel managers in alle sectoren is dit een signaal dat AI en digitalisering impact zullen hebben op personeelsplanning en bedrijfsstrategieën in de komende jaren. Dat meldt het ANP.

- Intesa Sanpaolo ontslaat 9000 medewerkers, neemt 3500 jongeren aan
- AI-gedreven transformatie leidt tot kostenbesparingen van 500 miljoen euro
- Citigroup verwacht grote impact AI op banksector wereldwijd

Impact op mentale gezondheid?

Het logo van Google op een kantoorgebouw van het bedrijf. © EPA

Google ontslaat ingenieur die beweert dat chatbot gevoelens heeft

Google, onderdeel van Alphabet, zei vrijdag dat het een senior software engineer heeft ontslagen die beweerde dat de kunstmatige intelligentie (AI) chatbot LaMDA van het bedrijf een zelfbewust persoon was.

“AI-chatbot overtuigde jonge Belg om uit het leven te stappen”

© Getty Images/iStockphoto

Een Belgische man is onlangs overleden door zelfmoord, nadat een AI-chatbot zijn negatieve denkpatronen versterkt zou hebben. Dat vertelt zijn vrouw in de krant La Libre.

Kan dit zomaar?

SECURITY & PRIVACY

How Target Knew a High School Girl Was Pregnant Before Her Parents Did

By Keith Wagstaff | Feb. 17, 2012

Share Like 54 Post Share Bewaren Read Later

In Charles Duhigg's new piece for the New York *Times*, a father finds himself in the uncomfortable position of having to apologize to a Target employee. Earlier he had stormed into a store near Minneapolis and complained to the manager that his daughter was receiving coupons for cribs and baby clothes in the mail.

Turns out Target knew his daughter better than he did. She really was pregnant.

[Source]

Kan dit zomaar?

US & WORLD / POLICY / REPORT

How Amazon automatically tracks and fires warehouse workers for ‘productivity’

Illustration by Alex Castro / The Verge

/ Documents show how the company tracks and terminates workers

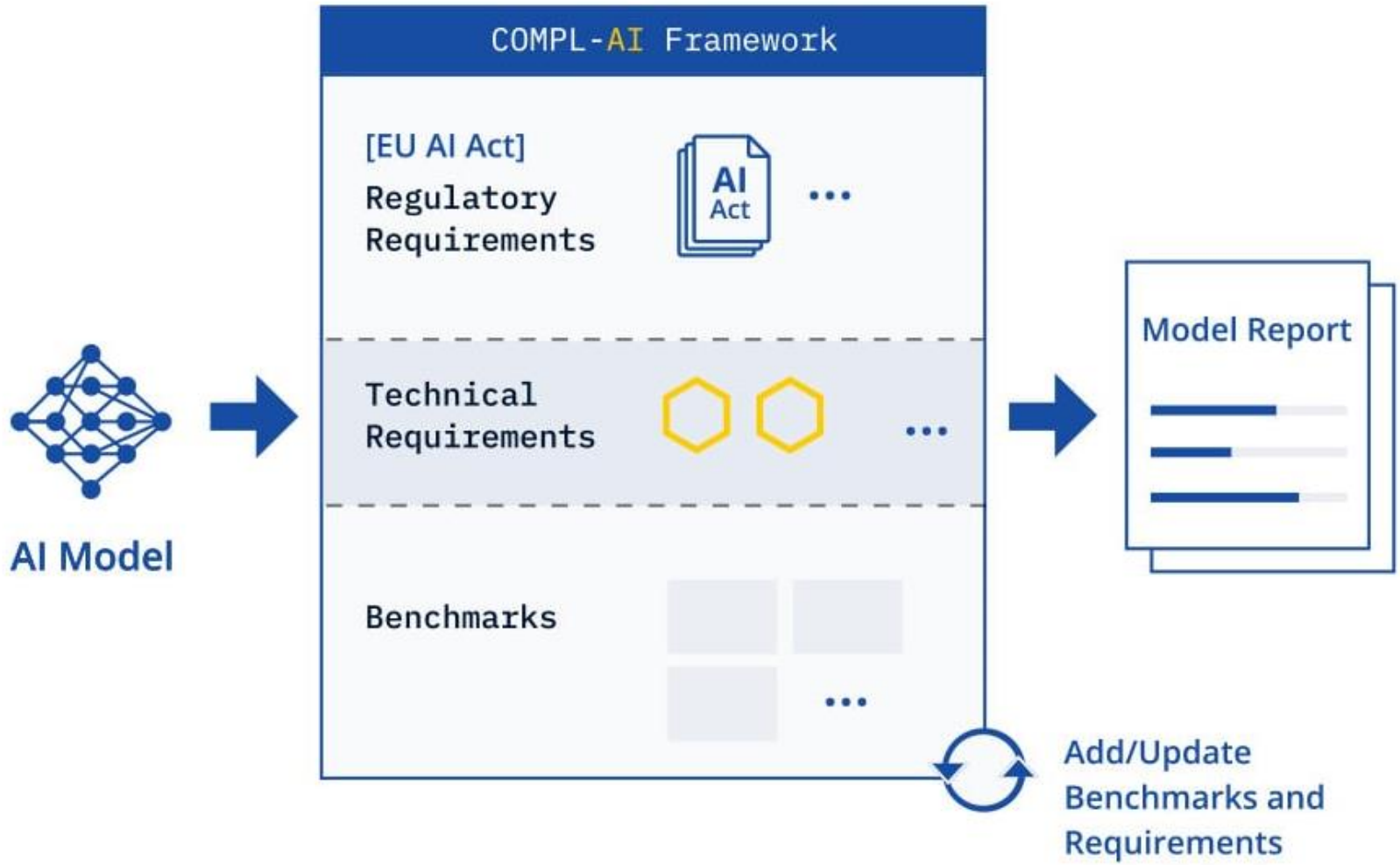
By Colin Lecher

Apr 25, 2019, 6:06 PM GMT+2

Privacy & Ethics

- Bias
- Transparency & Explainability
- Copyright
- Manipulation
- Accountability
- **Rules & Regulations**

EU AI Act: Model Creator *Responsibilities*



New EU AI Act Compliance Tool Evaluates Major Tech Companies' AI Models

EU AI Act: Model Creator *Responsibilities*

	Category	Keyword	Requirement (summarized)	Section
1	Data	Data sources	Describe data sources used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
		Data governance	Use data that is subject to data governance measures (suitability, bias, and appropriate mitigation) to train the foundation model.	Amendment 399, Article 28b, page 200
		Copyrighted data	Summarize copyrighted data used to train the foundation model.	Amendment 399, Article 28b, page 200
2	Compute	Compute	Disclose compute (model size, computer power, training time) used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
		Energy	Measure energy consumption and take steps to reduce energy use in training the foundation model.	Amendment 399, Article 28b, page 200
3	Model	Capabilities/limitations	Describe capabilities and limitations of the foundation model.	Amendment 771, Annex VIII, Section C, page 348
		Risks/mitigations	Describe foreseeable risks, associated mitigations, and justify any non-mitigated risks of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
		Evaluations	Benchmark the foundation model on public/industry standard benchmarks.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
		Testing	Report the results of internal and external testing of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
4	Deployment	Machine-generated content	Disclose content from a generative foundation model is machine-generated and not human-generated.	Amendment 101, Recital 60g, page 76
		Member states	Disclose EU member states where the foundation model is on the market.	Amendment 771, Annex VIII, Section C, page 348
		Downstream documentation	Provide sufficient technical compliance for downstream compliance with the EU AI Act.	Amendment 101, Recital 60g, page 76 and Amendment 399, Article 28b, page 200

Information about trained models need to be provided:

1. **Data** - information about the model training data
2. **Compute** - information about the computing inputs used to train models
3. **Model** - information about the model performance and risks
4. **Deployment** - operational details about model use in production

How Ready are Leading LLMs for the EU AI Act?


GDPR
GDPR
Data Protection Officer (DPO)
Compliance
25 May 2018
Data Breaches
Personal Data

AI & ROBOTICA

Australië verbiedt overheidsdiensten gebruik AI-tool DeepSeek

© CFOTO/Future Publishing via Getty Images

Redactie Data News

05-02-2025, 08:26 • Bijgewerkt op: 05-02-2025, 08:38 • Bron: Belga •

De Australische regering verbiedt alle overheidsdiensten om de nieuwe Chinese AI-tool DeepSeek te gebruiken vanwege zorgen over de nationale veiligheid. Volgens de Australische minister van Binnenlandse Zaken is er per direct een verbod op het gebruik van DeepSeek op systemen en apparaten van de overheid. Hij gaf aan dat DeepSeek een 'onacceptabel veiligheidsrisico' vormt.

LENIN NOLLY

Laatste kans voor TikTok in VS? Hooggerechtshof buigt zich vandaag over mogelijk verbod

Het Amerikaanse Hooggerechtshof buigt zich vandaag over de toekomst van TikTok. De app zou vanaf 19 januari verboden worden in de VS door een recente wet, maar TikTok vecht dat verbod aan omdat het volgens hen de vrijheid van meningsuiting schendt. Het Hooggerechtshof moet nu het finale oordeel vellen.

BUSINESS INSIDER NEDERLAND

ONDERNEMEN TECH FINANCE CARRIÈRE AUTO P

HOME › TECH › ITALIË VERBIEDT CHATGPT ALSNOG VANWEGE PRIVACYREGELS

Italië gaat AI-bot ChatGPT alsnog verbieden vanwege privacyregels

ANP

29 jan 2024

Foto: Jonathan Raa/NurPhoto/Shutterstock

- **De populaire AI-bot ChatGPT voldoet volgens Italië toch niet aan de Europese privacyregels.**
- **Vorig jaar besloot Italië de populaire tool al tijdelijk te verbieden, maar ChatGPT ging na toezeggingen van maker OpenAI na een maand weer online.**

Home › Elektronica › Privacy op internet

Eerste overwinning in collectieve actie tegen Google

Nieuws | De rechtbank Amsterdam heeft de Stichting Bescherming Privacybelangen (SBP) ontvankelijk verklaard in de collectieve rechtszaak tegen Google over grootschalige privacyschendingen. Dat betekent dat de stichting, gesteund door de Consumentenbond, de belangen van Nederlandse Google-gebruikers mag behartigen. Alle bezwaren daartegen van Google zijn afgewezen. Dit is een belangrijke eerste stap in de collectieve actie tegen Google.

Babs van der Staak Woordvoerder

Gepubliceerd op: 15 januari 2025

Privacy & Ethics

- Bias
- Transparency & Explainability
- Copyright
- Manipulation
- Trust and Accountability
- Rules & Regulations